Universidad Nacional de San Luis

Facultad de Ciencias Físico, Matemáticas y Naturales

Tesis para optar al grado de

Doctor en Ciencias de la Computación

Comparación eficiente de estructuras de independencia de modelos loglineales

---

Por:

Jan Strappa Figueroa

Director de tesis:

Dr. Facundo Bromberg

Mayo de 2020

«*You value your ignorance of what is to come?*»

«*That may be the most important thing to understand about humans. It is the unknown that defines our existence. We are constantly searching, not just for answers to our questions, but for new questions. We are explorers. We explore our lives, day by day, and we explore the galaxy, trying to expand the boundaries of our knowledge.*»

—"The Emissary", Star Trek: Deep Space Nine

«*Well, most people when they don't understand something, they frown. You… smile.*»

—"The Pilot", Doctor Who

Resumen

Las distribuciones de probabilidad discretas capturan diversas relaciones entre variables aleatorias, tales como independencias marginales o independencias condicionales. Estos patrones conforman en conjunto lo que se denomina una estructura de independencias. Estas estructuras cumplen dos roles: por un lado, facilitar múltiples reducciones en complejidad, y por otro, capturar conocimiento acerca del dominio de interés. Lo primero consiste en la capacidad de las estructuras para reducir la complejidad muestral de algoritmos que aprenden las distribuciones a partir de observaciones, la complejidad espacial para almacenar las distribuciones en memoria, y la complejidad temporal para realizar inferencia con ellas. Lo segundo ocurre debido a que las estructuras permiten extraer afirmaciones probabilísticas acerca de los eventos posibles de las variables. A pesar de la utilidad de este conocimiento estructural, no fue posible comenzar a obtenerlo, manipularlo e interpretarlo eficientemente hasta la introducción de los modelos probabilísticos gráficos en la década de 1980. Debido a su versatilidad, los modelos probabilísticos gráficos han tenido un gran auge hasta la actualidad. Sin embargo, su eficiencia tiene como contraparte una limitación de expresividad: su representación mediante grafos —dirigidos o no dirigidos— les impide capturar relaciones de grano más fino conocidas como *independencias específicas del contexto*. Estas definen una *estructura local* o *contextual* dentro del modelo. El interés en este tipo de estructuras contextuales ha resultado en muchos avances en una familia de distribuciones de probabilidad más general, que incluye a los modelos gráficos, conocida como *modelos loglineales*. En ellos, la estructura de independencias puede representar no solo independencias marginales y condicionales, sino

también independencias específicas del contexto. En los últimos años, ha habido un resurgimiento de interés en el desarrollo de enfoques para el aprendizaje automático de estructuras de independencias de modelos loglineales a partir de datos. Como resultado, existen actualmente diversos algoritmos que permiten el aprendizaje, la manipulación y la inferencia eficientes de estas estructuras. Sin embargo, no existía aún un método eficiente para evaluar estos enfoques de manera directa en términos de las estructuras de los modelos. Los únicos métodos conocidos evalúan estos enfoques de forma indirecta a través del modelo completo producido, el cual incluye no solo la estructura sino también los parámetros del modelo, introduciendo distorsiones potenciales en la comparación. Esta tesis presenta un método que permite una medición directa para comparar las estructuras de independencia de modelos loglineales, inspirada en la distancia de Hamming utilizada en modelos gráficos no dirigidos, los cuales constituyen solo un subconjunto de los modelos loglineales. La medida presentada provee las ventajas de poder ser computada eficientemente en términos del número de variables del dominio, y de permitir definir una distancia con las propiedades formales de una métrica.

Agradecimientos

Agradezco a CONICET la beca doctoral que ha financiado mi formación durante este recorrido. A la Universidad Tecnológica Nacional, Facultad Regional Mendoza, y a la Universidad Nacional de San Luis las oportunidades para crecer. A la educación pública.

Merece el más grande reconocimiento mi amigo Federico Schlüter, quien me acompañó durante gran parte de mi doctorado, enseñándome a investigar y a vivir con más alegría.

A mis colegas: Carlos Díaz, Ana Laura Diedrichs, Sebastián Pérez y Àlex Ribas les agradezco su compañerismo y apoyo. Son personas que me inspiran constantemente con su esfuerzo y dedicación.

A les amigues que me ayudaron a seguir durante los momentos más tormentosos: Emily Landry, Andrés Croba, Aaron Ruiz, Paula Simón, Tino Lucero, Daniela Soria, Majo Trigueiro, Ismael Ortiz, Emi Brunner, Fran Viladrich, Cristella Copparoni. A Ale Celi, quien me recordó que la vida es más importante que el doctorado.

A Jess, mi compañere, la persona más hermosa que he conocido, agradezco cada momento y espacio compartido, y sus incontables maneras de hacerme feliz.

A mi madre y a mi padre, por su amor y dedicación incondicional. Por trabajar y velar incansablemente por mi bienestar. Por acompañarme tantos años, alentándome a pensar y actuar libremente.

Índice general

Índice de figuras

Índice de tablas

A todas las personas que creen en la lucha colectiva
para construir un mundo más justo.

Capítulo 1

Introducción

Esta tesis presenta un método para la comparación eficiente de estructuras de dos modelos loglineales, los cuales son una conocida representación de distribuciones de probabilidad de las asignaciones de variables aleatorias discretas [Christensen, 2006, Agresti, 2002, Haberman, 1973]. Las distribuciones tienen implícita una estructura de dependencias e independencias probabilísticas, las cuales corresponden, respectivamente, a interacciones o falta de ellas entre las variables. La contribución de este trabajo está enmarcada en el problema de aprendizaje de estructuras de modelos loglineales, que tiene como fin obtener estructuras precisas para representar interacciones probabilísticas extraídas a partir de datos [Della Pietra et al., 1997, McCallum, 2003, Lee et al., 2006, Davis and Domingos, 2010, Lowd and Davis, 2014, Van Haaren and Davis, 2012]. Específicamente, provee un mecanismo para estimar la calidad de las estructuras obtenidas por medio de un algoritmo eficiente, permitiendo así comparar estructuras aprendidas con estructuras de buena calidad conocidas.

Así, el método presentado en este trabajo puede contribuir a la evaluación de técnicas de aprendizaje de estructuras, teniendo en cuenta que es, a mi entender,

el primer aporte que realiza una comparación estructural de manera directa. Es además eficiente en el número de variables del dominio, y constituye una métrica.

1.1. Preliminares

1.1.1. Modelos probabilísticos

Los modelos probabilísticos son herramientas para razonar bajo incertidumbre mediante un proceso llamado *inferencia* probabilística. La inferencia permite calcular la probabilidad de una proposición, la cual está compuesta por una asignación de valores a variables. Una variable es un atributo o propiedad del mundo y puede tomar un valor de entre un número de valores posibles. Esta tesis se enmarca en modelos para dominios finitos de variables discretas. Algunos ejemplos de este tipo de variables pueden ser *Edad* de una persona, *Fuma* (si fuma o no), y *Cardiopatía*. Estas variables podrían tener como valores, respectivamente: números enteros en el rango [18,90], las categorías {nunca, ocasionalmente, frecuentemente}, y {no, sí}. Una *distribución* consiste en una función p que asigna a valores de una o más variables una probabilidad, es decir, un número real entre 0 y 1. Las distribuciones de probabilidad deben cumplir con los conocidos axiomas de Kolmogorov. Para una variable, la distribución de probabilidades sobre cada valor es su *distribución marginal*. Por ejemplo, la distribución marginal para *cardiopatía* podría ser $p(\text{Cardiopatía} = \text{sí}) = 0{,}3$ y $p(\text{Cardiopatía} = \text{no}) = 0{,}7$[1]. Cuando el interés reside en las probabilidades de combinaciones de variables, se utilizan las *distribuciones conjuntas*. Por ejemplo, la Tabla 1.1.1 es una representación exhaustiva de una distribución de probabilidad conjunta. También podría representarse

[1] Todos los valores en este ejemplo son ficticios.

		Cardiopatía		
		no	sí	
Fuma	frecuentemente	0.07	0.18	0.25
	ocasionalmente	0.28	0.09	0.37
	nunca	0.35	0.03	0.38
		0.7	0.3	1

TABLA 1.1: Ejemplo de una distribución conjunta, $p(\text{Fuma}, \text{Cardiopatía})$. Las filas muestran probabilidades sobre los valores de *Fuma* y las columnas sobre la existencia de una *Cardiopatía*. La última columna muestra la probabilidad marginal sobre $p(\text{Fuma})$ y la última fila contiene la probabilidad marginal para $p(\text{Cardiopatía})$.

como una tabla de solo dos columnas, una con las posibles asignaciones conjuntas de las dos variables (un ejemplo de asignación es: <Fuma = ocasionalmente, Cardiopatía = no>) y otra con la probabilidad conjunta de cada asignación. Así tendríamos tantas filas como combinaciones posibles de los valores de las variables: en este caso, serían $|\text{Fuma}| \times |\text{Cardiopatía}| = 2 \times 3 = 6$ filas. Esta forma de representar una distribución requiere especificar tantos valores de probabilidad como asignaciones. Cada uno de estos valores es un *parámetro* de la distribución.

Posiblemente el ejemplo más común de inferencia es la predicción de una variable en base a otras observadas previamente. En general, en un dominio de variables $\{X_1, \ldots, X_n, Y\}$, predecir Y corresponde al cálculo de

$$p(Y|X_1, \ldots, X_n) = \frac{p(Y, X_1, \ldots, X_n)}{p(X_1, \ldots, X_n)}.$$

Esto constituye un caso particular de una *distribución condicional*, es decir, una distribución de las probabilidades de una variable (en este caso, Y) cuando se observan valores de otras variables, enumeradas en lo que se denomina su *conjunto condicionante* (en este caso, $\{X_1, \ldots, X_n\}$). El numerador se encuentra en las probabilidades conjuntas de la representación tabular. El denominador se obtiene

mediante la ley de probabilidad total:

$$p(X_1,\ldots,X_n) = \sum_{y \in val(Y)} p(Y = y, X_1,\ldots,X_n).$$

Siguiendo el ejemplo previo, el interés está en predecir la ocurrencia de una cardiopatía en personas de las cuales se conoce su condición como fumadoras. El numerador $p(\text{Cardiopatía}, \text{Fuma})$ se obtiene de forma directa de cada celda en la parte principal de la Tabla 1.1.1. La sumatoria para el denominador ya está resuelta en el margen de la columna derecha. Así, si observamos que una persona fuma ocasionalmente, obtenemos

$$p(\text{Cardiopatía} = \text{sí} | \text{Fuma} = \text{ocasionalmente}) = \frac{0{,}09}{0{,}37} = 0{,}2432.$$

Dado que la distribución condicional está normalizada, tenemos que

$$p(\text{Cardiopatía} = \text{no} | \text{Fuma} = \text{ocasionalmente}) = 1 - 0{,}2432 = 0{,}7568.$$

La complejidad de las representaciones tabulares para distribuciones conjuntas se vuelve rápidamente intratable a medida que aumenta el número de variables a considerar. En este ejemplo, la probabilidad conjunta sobre las dos variables mostradas solo requiere la especificación de seis valores. No obstante, si consideráramos información adicional, como los años que la persona lleva fumando, cuántos días a la semana realiza ejercicio, peso, antecedentes familiares de cardiopatía, o niveles de colesterol, la cantidad de asignaciones completas cuyas probabilidades debemos modelar aumentaría exponencialmente con el número de variables. En conjuntos de datos realistas no es inusual que se presenten cientos o miles de variables con cientos de miles o millones de observaciones. Con una

representación tabular, estos números resultan inmanejables tanto para su representación en memoria como en el tiempo de cómputo para realizar inferencia.

Para solucionar estos problemas de complejidad, existen en la actualidad numerosos modelos probabilísticos que permiten descomponer la expresión funcional de las distribuciones conjuntas representadas. Utilizando ciertas suposiciones sobre las propiedades de la distribución, un modelo introduce restricciones que permiten reducir el número de parámetros necesario tanto para representar la distribución como para el cómputo de cualquier probabilidad.

En el Capítulo 2 explicaré en detalle algunas familias de modelos y cómo permiten esta reducción exponencial. A continuación, proveo una intuición en base al concepto que emplean estos modelos para lograr dicha simplificación: el concepto de *independencia* probabilística. En general, los modelos probabilísticos — entre los cuales se encuentran los modelos loglineales— son capaces de capturar *independencias y dependencias condicionales*, las cuales son relaciones entre las variables acerca de cómo el conocimiento sobre el valor de unas afecta las probabilidades de los valores de otras, teniendo en cuenta otras variables observadas. Las variables observadas conforman el condicionante de una dependencia o independencia. La notación utilizada en este trabajo es $(X_A \perp\!\!\!\perp X_B \mid X_C)$ para independencia entre conjuntos de variables X_A y X_B condicionadas en un conjunto de variables X_C, y $(X_A \not\perp\!\!\!\perp X_B \mid X_C)$ para indicar dependencia condicional. Para facilitar la comprensión de este concepto presento un nuevo ejemplo.

Ejemplo 1. *Para el diagnóstico de celiaquía, existe una prueba sencilla en base a un análisis de sangre. Esto es porque las personas celíacas que consumen gluten, en general, tienen niveles más altos de lo normal de ciertos anticuerpos en su sangre. Entonces, podemos representar la presencia de enfermedad celíaca como C, una variable con los valores posibles $\{sí, no\}$, y el resultado de la prueba diagnóstica como*

$P \in \{positivo, negativo\}$. Con esta representación podemos afirmar, sin lugar a dudas, que existe una dependencia entre la celiaquía y el resultado positivo de la prueba para una población determinada:

$$(P \not\perp\!\!\!\perp C).$$

Sin embargo, estas pruebas, en algunos casos, pueden resultar en falsos positivos (personas sin celiaquía pero con resultado positivo en la prueba) o falsos negativos (personas con celiaquía pero con resultado negativo en la prueba). Esto sucede porque no todas las personas celíacas presentan los mismos anticuerpos, y algunas personas presentan los mismos anticuerpos que se asocian a la celiaquía, pero debido a otras condiciones. A veces es necesario recurrir a una biopsia intestinal, la cual ofrece un resultado sumamente confiable sobre la presencia o ausencia de enfermedad celíaca. Teniendo el resultado de la biopsia, saber el resultado del análisis de sangre es superfluo, es decir, es irrelevante, y podríamos representar ese conocimiento de la siguiente manera:

$$(P \perp\!\!\!\perp C \mid B),$$

donde B es el resultado de la biopsia intestinal, $B = \{positivo, negativo\}$.

Una particularidad de los modelos loglineales es que permiten representar cualitativamente la presencia de *independencias específicas del contexto*, una clase de estructura de grano más fino que no puede representarse con las familias de modelos probabilísticos más conocidos, los modelos probabilísticos *gráficos*. Estas son independencias que se cumplen solo para un subconjunto de los valores de las variables en el condicionante.

Ejemplo 1 (continuación)**.** *Siguiendo con el ejemplo anterior, es importante tener en cuenta que los análisis de sangre para detectar celiaquía solo son confiables cuando la persona ha tenido una alimentación que incluye gluten (por cierto tiempo) al realizar la prueba. Contemplando este conocimiento, podríamos incluir una variable que represente si la persona consume gluten,* $G = \{0 : \textit{no consume}; 1 : \textit{sí consume}\}$*, y nos encontraríamos con que la prueba es relevante para un solo valor de G:*

$$(C \not\perp\!\!\!\perp P \mid G = 1),$$

mientras que para el caso negativo, no podemos afirmar que el resultado de la prueba P nos de información sobre la celiaquía C:

$$(C \perp\!\!\!\perp P \mid G = 0).$$

Esta independencia es contextual, dado que solo se cumple para uno o más de los valores del condicionante (en este caso, para $G = 0$*), pero no para todos.*

Una subclase de los modelos loglineales son los modelos gráficos no dirigidos, también conocidos como *redes de Markov* o *Markov Random Fields*. Estos son una representación robusta para distribuciones de probabilidad conjunta, y permiten codificar de manera compacta un número exponencial de independencias condicionales [Pearl, 1988, Lauritzen, 1996, Koller and Friedman, 2009]. Para comprender la relación de modelos loglineales con el aprendizaje automático y las bases de esta tesis, es relevante conocer en qué consisten las redes de Markov y cómo se utilizan. Las redes de Markov usan dos componentes interdependientes: una estructura de independencias y un conjunto de pesos (o parámetros numéricos) sobre la estructura. La estructura consiste en un grafo no dirigido donde los nodos corresponden a las variables y las aristas codifican dependencias directas

entre ellas (es decir, no mediadas por otras variables). En la actualidad, el aprendizaje de estructuras, el aprendizaje de parámetros y la inferencia con esta clase de modelos probabilísticos son problemas importantes de aprendizaje automático, con una amplia variedad de aplicaciones en la literatura, tales como visión computacional y análisis de imágenes [Hwang and Kim, 2015, Peng et al., 2016], procesamiento del lenguaje [Tellex et al., 2011], biología computacional [Li et al., 2015], y biomedicina [Schmidt et al., 2008, Wan et al., 2015], entre muchas otras.

Los modelos loglineales [Christensen, 2006, Agresti, 2002, Haberman, 1973] son más flexibles que los grafos, permitiendo codificar no solo independencias condicionales sino también independencias específicas del contexto. La representación loglineal de la estructura se define como un *conjunto de funciones características*, cada una consistente en una asignación a algún subconjunto de variables del dominio. Las características representan dependencias entre subconjuntos de variables aleatorias. Dado un conjunto de características, la distribución de probabilidad conjunta queda completamente especificada por los pesos de las características, un número real por cada característica. Estos son los parámetros numéricos del modelo loglineal. Esto se explica en detalle en la Sección 2.3.

1.1.2. Aprendizaje de estructuras

Generalmente, se caracteriza el problema de aprendizaje de estructuras de modelos loglineales como un problema de selección de características. Este enfoque consiste en algoritmos que construyen un modelo loglineal seleccionando características desde un conjunto de datos, es decir, eligiendo qué funciones características se utilizarán para representar la distribución a partir de asignaciones de

subconjuntos de variables que han sido observadas (asignaciones que están presentes en los datos). Las estrategias de este tipo generalmente realizan una búsqueda que añade o elimina características incrementalmente [Della Pietra et al., 1997, Lee et al., 2006, Davis and Domingos, 2010, Van Haaren and Davis, 2012, Lowd and Rooshenas, 2013, Lowd and Davis, 2014].

Los trabajos relacionados demuestran que la ventaja en expresividad de los modelos loglineales ha generado considerable interés en su aprendizaje automático a partir de datos. Sin embargo, la evaluación de este aprendizaje se encuentra actualmente limitada por la imposibilidad de evaluar las estructuras aprendidas de forma directa. El aporte de este trabajo es proveer una técnica para comparar la similitud entre dos modelos loglineales de manera única en términos de las independencias y dependencias que estos codifican. Esta técnica puede mejorar la comparación de calidad entre diferentes técnicas de aprendizaje de estructuras, permitiendo a la vez incrementar las oportunidades de validar las propiedades de tales técnicas, al obtener información sobre las estructuras aprendidas. Por ejemplo, en la actualidad es posible hacer una estimación de qué métodos obtienen estructuras más densas, midiendo la densidad como cantidad de dependencias aprendidas en relación al total de dependencias posibles en un modelo saturado. Sin embargo, esta estimación es indirecta y provee información limitada. En particular, no es posible verificar si las estructuras más densas lo son porque contienen dependencias correctas más un número de dependencias espurias, o si omiten dependencias y además contienen un número aún mayor de dependencias espurias. El método propuesto en esta tesis permite realizar este tipo de análisis.

En el área de aprendizaje de estructuras de modelos probabilísticos en general, los problemas suelen ser formulados con alguno de dos objetivos diferentes: *estimación de densidad* o *descubrimiento de conocimiento* ([Koller and Friedman, 2009],

Capítulo 19). El primero hace referencia a las funciones de densidad de probabilidad, y consiste en utilizar las estructuras para construir modelos precisos para tareas de inferencia, tales como la estimación de probabilidades condicionales y marginales [Lowd and Davis, 2014, Van Haaren and Davis, 2012, Davis and Domingos, 2010]. Por ejemplo, muchos problemas importantes en el área de visión computacional consisten en el reconocimiento de caracteres [Lee et al., 2006] u objetos [Hwang and Kim, 2015] a partir de imágenes. Desde los modelos se obtienen las probabilidades de que una imagen contenga cada clase de carácter u objeto, y la clase con mayor probabilidad se usa como predicción. Cuando se busca el segundo objetivo, la estructura aprendida por un algoritmo puede emplearse como un modelo interpretable que muestra las interacciones más significativas de un dominio [Lee et al., 2006, Van Haaren et al., 2013, Claeskens et al., 2015, Nyman et al., 2014a, Pensar et al., 2017]. Un ejemplo, publicado en [Van Haaren et al., 2013], investiga relaciones entre tres o más enfermedades crónicas (multimorbilidad) a partir de modelos probabilísticos aprendidos por un número de algoritmos desde los datos clínicos de pacientes. En ambos escenarios, la estructura juega un rol importante. Esto ocurre de forma directa en el descubrimiento de conocimiento, puesto que la extracción de información se realiza de forma principalmente cualitativa analizando qué interacciones están presentes y cuáles ausentes, pero también de forma indirecta en la estimación de densidad, ya que los errores en la estructura derivan en interacciones faltantes o sobrantes. Las interacciones faltantes llevan a la imposibilidad de obtener estimaciones correctas por falta de representación en la forma funcional de la distribución, mientras que las sobrantes conducen a imprecisiones a causa de un aumento en la complejidad del cómputo de las estimaciones. Por estos motivos es de gran importancia la evaluación adecuada del aprendizaje de estructuras, para garantizar que los

algoritmos con una mejor evaluación son efectivamente los que tienden a aprender estructuras correctas. Esto no puede garantizarse con los métodos empleados actualmente, como se explica en la siguiente sección.

1.1.3. Aportes previos relacionados

Durante la revisión de la literatura, no he encontrado medidas de distancia estructural para evaluar la similitud de dos modelos loglineales en términos de las independencias específicas del contexto que codifican. La literatura de algoritmos de aprendizaje de estructuras de modelos loglineales solo evalúa algoritmos en términos de la *capacidad de inferencia* de las estructuras aprendidas, mientras que los análisis estructurales se reducen a comparaciones por inspección en casos de juguete o mediciones indirectas relacionadas a propiedades de las estructuras.

Una forma limitada de analizar estructuras por inspección visual deriva de los diversos aportes para la interpretabilidad de modelos loglineales. La extracción de información desde una representación analítica convencional es difícil y tediosa, puesto que leer independencias desde ella no es trivial. Debido a que las independencias específicas del contexto solo se cumplen en un subespacio de las configuraciones del condicionante, su representación conlleva una gran complejidad que no puede resolverse completamente con el uso de grafos. Por esta razón, recientemente, se han desarrollado una variedad de representaciones gráficas para modelos loglineales que generalizan los modelos gráficos no dirigidos. Entre ellas se encuentran los modelos de interacción específicos del contexto [Eriksen, 1999, Højsgaard, 2004], modelos divididos [Højsgaard, 2003] y modelos gráficos estratificados [Nyman et al., 2014b]. En estas representaciones, la semántica para leer independencias condicionales sirve de inspiración para la representación

gráfica de independencias específicas del contexto, con el fin de mantener la interpretabilidad del modelo. A pesar de estos importantes esfuerzos por mejorar la visualización y extracción de conocimiento, en ninguno de estos trabajos se han propuesto medidas para la comparación sistemática de las estructuras de modelos loglineales.

En la actualidad, al usar la representación loglineal, el desempeño de los algoritmos de aprendizaje usualmente se mide usando la *divergencia Kullback-Leibler* o *divergencia KL* [Kullback and Leibler, 1951, Cover and Thomas, 2012] para evaluar la similitud de las distribuciones completas codificadas por cada estructura [Della Pietra et al., 1997, Lee et al., 2006, Davis and Domingos, 2010, Van Haaren and Davis, 2012, Lowd and Rooshenas, 2013, Lowd and Davis, 2014]. Este es un procedimiento indirecto que tiene varias desventajas.

En primer lugar, la divergencia KL requiere aprender los parámetros del modelo además de la estructura. Por esto, la calidad de las estructuras es analizada haciendo una evaluación de la calidad de las distribuciones completas resultantes, sin tener en cuenta si la estructura a comparar tiene falsos positivos (dependencias sobrantes) o falsos negativos (dependencias faltantes) respecto a la estructura correcta. Precisamente, una desventaja de comparar modelos usando la divergencia KL es que los falsos positivos y los falsos negativos tienen diferente impacto en la calidad de la distribución, y esto no se refleja en la medida. Los falsos positivos pueden ser mitigados al aprender los parámetros, estableciendo ciertos parámetros en cero para codificar las independencias omitidas. Sin embargo, los falsos negativos no pueden ser mitigados por los parámetros numéricos, porque añaden suposiciones de independencia incorrectas en la distribución que pueden invalidar la inferencia estadística, llevando a conclusiones erróneas. En cambio, el método presentado en esta tesis no depende de los parámetros, por lo que permite un análisis individual de falsos positivos y falsos negativos.

Como segunda limitación, es importante señalar que el proceso de aprendizaje de parámetros es sensible a escasez de datos, y por lo tanto, la divergencia KL, dependiente de los parámetros, podría no ser precisa cuando los datos son insuficientes. El nuevo método presentado se calcula sobre las estructuras y por lo tanto no es influenciado por la escasez de datos. Al aprender estructuras para dominios de muchas variables, debido a que el cómputo de KL no es factible, se puede utilizar la Verosimilitud Condicional (Marginal) (*C(M)LL* por sus siglas en inglés) [Lee et al., 2006, Davis and Domingos, 2010, Van Haaren and Davis, 2012, Lowd and Rooshenas, 2013, Lowd and Davis, 2014]. Sin embargo, la C(M)LL es un método aproximado, y también presenta las dos limitaciones mencionadas, al requerir la tarea de aprender los parámetros numéricos de la estructura.

Para comprender cualidades estructurales sin tomar en consideración los parámetros, algunos trabajos han utilizado el número de características y la longitud promedio de características [Van Haaren and Davis, 2012, Edera et al., 2014a,b], los cuales son indicadores indirectos y resumidos y, como tales, son poco informativos; además, no permiten una comparación confiable entre diferentes estructuras.

1.1.4. Descripción del aporte de esta tesis

El método propuesto funciona contando el número de diferencias estructurales que aparecen entre dos modelos loglineales. Definiremos estas diferencias estructurales en términos de dependencias contextuales entre dos variables, dadas asignaciones completas de todas las restantes variables. Comparando qué dependencias e independencias se cumplen en ambos modelos y en cuáles no coinciden, es posible producir una matriz de confusión que cuenta los verdaderos positivos, falsos positivos, verdaderos negativos y falsos negativos que están presentes en

el segundo modelo en relación al primero. Así, los verdaderos positivos corresponden a dependencias que están presentes en ambos modelos; los verdaderos negativos corresponden a dependencias ausentes (independencias) en ambos modelos; los falsos positivos corresponden a dependencias presentes en la segunda estructura que están ausentes en la primera; y los falsos negativos corresponden a dependencias ausentes (independencias) en la segunda estructura que están presentes en la primera.

En grafos no dirigidos, la comparación se realiza de forma similar en base a aristas, que tienen una correspondencia con dependencias condicionadas en todas las restantes variables, con la diferencia de que son condicionantes no asignados, es decir, no contextuales. La suma de los valores de la contradiagonal de la matriz de confusión (falsos positivos más falsos negativos) es la conocida distancia de Hamming, usada ampliamente para la comparación estructural de redes de Markov [Bromberg et al., 2009, Schlüter et al., 2014, Pensar et al., 2017, Schlüter et al., 2018]. Sin embargo, en contraste con las redes de Markov, un cómputo exhaustivo para producir la matriz de confusión de un modelo loglineal presenta un costo computacional exponencial respecto al número de variables y al producto del número de características de cada modelo.

El método propuesto en esta tesis computa los valores de la matriz de confusión para estructuras de modelos loglineales de manera eficiente con respecto al número de variables. La eficiencia con respecto al número de características no está garantizada para un gran número de características en los modelos y está sujeta a trabajo futuro.

La propuesta puede contribuir a los dos objetivos del aprendizaje mencionados previamente: descubrimiento de conocimiento y estimación de densidad. Dado que una de las principales ventajas de los modelos es su interpretabilidad, el

reconocimiento de diferencias estructurales entre dos modelos puede ser utilizado para analizar las diferencias entre el modelo loglineal subyacente de un problema sintético y el modelo aprendido por un algoritmo en términos de interacciones faltantes o sobrantes. Además, puede ser utilizado para examinar diferencias estructurales entre modelos obtenidos por diferentes métodos de aprendizaje. Por otro lado, una mejor estructura puede mejorar tanto la calidad como la eficiencia del aprendizaje de parámetros, lo que conduce a mejores modelos para tareas de inferencia. La utilidad del método se enfatiza con la ventaja de que cumple con las propiedades de una métrica, con lo cual se puede contar con garantías teóricas sobre los resultados obtenidos mediante su uso.

1.2. Organización de los restantes capítulos

Los capítulos siguientes de este trabajo están estructurados como se describe a continuación.

El Capítulo 2 introduce conceptos relevantes sobre teoría de probabilidad y modelos probabilísticos gráficos. Se presentan las definiciones de distintos tipos de relaciones de independencia y el concepto de modelo de dependencias. Se ofrece una introducción al concepto de redes de Markov, presentando definiciones y propiedades requeridos para la comprensión de ciertos elementos de esta tesis que están basados en ellas. Se definen modelos loglineales y se provee una breve reseña de algunas propuestas para la representación gráfica de sus estructuras, con sus respectivas problemáticas concernientes a la comparación de las mismas. Por último, se resume el problema del aprendizaje automático de modelos loglineales a partir de datos, detallando metodologías experimentales y algoritmos existentes en la literatura.

El Capítulo 3 aborda tres temas principales. En primer lugar, se sirve del concepto matemático de medida de distancia (o métrica) para delinear las características deseables para la comparación de estructuras. En segundo lugar, presenta un método de comparación de estructuras de modelos probabilísticos gráficos no dirigidos. Por último, describe métodos indirectos para la comparación de estructuras de modelos loglineales.

El aporte de esta tesis es desarrollado en los Capítulos 4 y 5. El Capítulo 4 se enfoca, por un lado, en el análisis y justificación del problema, y por otro, en la solución propuesta. Esta segunda parte incluye un número de teoremas y demostraciones necesarios para sustentar el método. Por su parte, el Capítulo 5 ofrece la demostración de que este método es una métrica, de acuerdo a las propiedades especificadas en el Capítulo 3.

En el Capítulo 6 se enuncian las contribuciones y conclusiones de la tesis, así como las posibles líneas futuras de investigación relacionadas con dichas contribuciones.

Adicionalmente, este trabajo proporciona cuatro apéndices. Los tres primeros proveen definiciones y demostraciones auxiliares para los conceptos utilizados en este trabajo. En el último apéndice se comparan de manera ilustrativa el método directo de comparación propuesto y la divergencia KL, el ya mencionado método indirecto utilizado comúnmente en la literatura relacionada.

Capítulo 2

Marco teórico

En este capítulo se resumen conceptos básicos preliminares sobre teoría de probabilidad, modelos probabilísticos gráficos y algunas de sus representaciones. Para comenzar, se desarrollan conceptos y características de distribuciones de probabilidad y modelos probabilísticos, y se definen los conceptos de independencias marginales, condicionales y específicas del contexto. Además, se formaliza la noción de modelo de dependencias. Estos conceptos son la base de la semántica de los modelos presentados en las secciones subsiguientes. En ellas, se ofrece una breve introducción a dos clases de distribuciones y a algunas de sus distintas representaciones, cuya semántica está asociada a independencias probabilísticas. Una de las representaciones más utilizadas son las *redes de Markov*, basadas en grafos no dirigidos. Una clase más general de distribuciones puede ser representada a través de *modelos loglineales*, capaces de capturar estructura a nivel local o más específico, y en los cuales se enfoca el aporte de esta tesis. Posteriormente puede encontrarse una reseña sobre el aprendizaje automático de modelos loglineales a partir de datos, que abarca metodologías experimentales que han sido

utilizadas en la literatura así como también la presentación de algoritmos publicados. Al final de este capítulo se incluye un apartado que sirve como referencia de la notación empleada y que resume los principales conceptos presentados.

2.1. Distribuciones de probabilidad y modelos probabilísticos

Esta tesis aborda modelos asociados a distribuciones de probabilidad sobre variables aleatorias discretas. Una distribución de probabilidad es una función $p(X_V)$ sobre un dominio finito de variables X_V, donde V es el conjunto de todos los índices de variables de dicho dominio. En un dominio, cada variable individual X_k puede tomar un valor $x_k \in val(X_k)$. Una observación de los valores de todas las variables se denomina *suceso*, observación o configuración completa, y lo denotaremos simplemente como x, mientras que el conjunto de todos los sucesos será denotado como $\mathcal{X}$. La distribución $p(X_V)$ asocia a cada posible suceso $x \in \mathcal{X}$ un valor real entre 0 y 1. Especificar una *distribución conjunta* sobre X_V nos permite describir probabilidades de ocurrencia de eventos descritos por subconjuntos de sus variables. Sin embargo, una representación directa de dicha distribución requiere la obtención de un número de parámetros (en este caso, las probabilidades de cada suceso completo x) exponencial en la cardinalidad de X_V. Esto es problemático, tanto para cómputo y almacenamiento, como para la obtención de estos valores y su posterior interpretación y comunicación por seres humanos. Hoy en día, estas dificultades son mitigadas enormemente mediante el uso de modelos probabilísticos como representación de distribuciones [Koller and Friedman, 2009]. Cada modelo asume ciertas características funcionales respecto a la distribución representada, permitiendo su *factorización* (descomposición) y

una subsecuente reducción en su parametrización. En los modelos probabilísticos, la factorización se determina en base a distintos tipos de *independencias* entre las variables del dominio, subsumidas en lo que se conoce como la *estructura* del modelo. La estructura ofrece una manera natural de conceptualizar las interacciones entre variables, por lo que se describe como el aspecto cualitativo del modelo, mientras que los parámetros definidos sobre la estructura, en cambio, constituyen el aspecto cuantitativo que caracteriza dichas interacciones.

2.1.1. Independencias

Los conceptos de independencia e independencia condicional son un principio fundamental de la teoría de la probabilidad.

Definición 1. ***Independencia.*** *Se dice que dos variables X_i y X_j son independientes entre sí, expresado como $(X_i \perp\!\!\!\perp X_j)_p$ — alternativamente, $(X_i \perp\!\!\!\perp X_j \mid \emptyset)_p$— si la distribución de la variable X_i es idéntica dado cualquier valor $x_j \in val(X_j)$, es decir,*

$$p(x_i|x_j) = p(x_i),$$

para todo $x_i \in val(X_i)$ y $x_j \in val(X_j)$.

Este tipo de independencia también suele denominarse *independencia marginal*.

Definición 2. ***Independencia condicional.*** *Una independencia condicional $(X_i \perp\!\!\!\perp X_j \mid X_Z)_p$ indica que la distribución de X_i dados los valores de la variable X_j y la conjunción de valores del conjunto de variables X_Z solo depende de los valores en X_Z:*

$$p(x_i|x_j, x_Z) = p(x_i|x_Z),$$

para todo $x_i \in val(X_i)$, $x_j \in val(X_j)$ y $x_Z \in val(X_Z)$.

Una interpretación usual del concepto de independencias es la de *irrelevancia* [Dawid, 1998, Pearl, 1988, Lauritzen, 1996]. Respectivamente, podemos reinterpretar las anteriores definiciones diciendo que la información provista por el valor de X_j es irrelevante para determinar los posibles valores de X_i, y que conocer X_j es irrelevante para X_i cuando conocemos el valor de X_Z. Sobre la independencia condicional también podemos decir que la interacción entre X_i y X_j está mediada por X_Z. Esta visión cualitativa de las relaciones de independencia condicional es axiomatizada siguiendo las propiedades presentadas en el Apéndice A.1.

Existe otro tipo de independencias, no tan ampliamente conocidas, pero que están sin embargo presentes en numerosos escenarios: estas son las llamadas *independencias específicas del contexto* [Boutilier et al., 1996, Højsgaard, 2004, Koller and Friedman, 2009]. Estos patrones aparecen cuando se cumple una independencia entre variables *solo para un subconjunto de estados u observaciones de un condicionante*, mientras que las mismas variables son dependientes para otros valores de ese condicionante. Formalizaremos esta noción a través de la siguiente definición, tomada de [Boutilier et al., 1996]:

Definición 3. ***Independencia específica del contexto.*** *Sean i, j índices, U y W conjuntos disjuntos de índices en V, con $x_U \in val(X_U)$. X_i y X_j son contextualmente independientes dados X_W y el contexto x_U, y se denota como $(X_i \perp\!\!\!\perp X_j \mid x_U, X_W)_p$, si y solo si*

$$p(x_i|x_j, x_U, x_W) = p(x_i|x_U, x_W),$$

para todo $x_i \in val(X_i)$, $x_j \in val(X_j)$ y $x_W \in val(X_W)$, siempre que $p(X_j, x_U, X_W) > 0$.

Podemos relacionar esta última definición con la anterior (Definición 2) para obtener la siguiente equivalencia (véanse también [Højsgaard, 2004], [Edera et al.,

2014a] y [Pensar et al., 2016]):

$$(X_i \perp\!\!\!\perp X_j \mid X_{U \cup W})_p \equiv \forall x_U \in val(X_U), (X_i \perp\!\!\!\perp X_j \mid x_U, X_W)_p. \tag{2.1}$$

2.1.2. Modelo de dependencias

Podemos caracterizar la estructura de una distribución p mediante un *modelo de dependencias*: una regla que determina el valor de verdad para todos los tripletes de conjuntos disjuntos de variables $I(X_i, X_j \mid X_Z)_p$ (Pearl [1988], sección 3.1.3). Este triplete se interpreta como una pregunta de independencia condicional, esto es, la consulta booleana de si la independencia $(X_A \perp\!\!\!\perp X_B \mid X_C)_p$ se cumple o no en una distribución o modelo; simbólicamente:

$$I(X_A, X_B \mid X_C)_p \text{ es verdadero} \iff (X_A \perp\!\!\!\perp X_B \mid X_C)_p$$

En otras palabras, un modelo de dependencias captura las independencias y dependencias codificadas en una distribución de probabilidad. Al modelar la estructura de una distribución mediante independencias marginales y condicionales, podemos expresar su modelo de dependencias como $\mathcal{D}^{NC}(p)$ (NC indica *no contextual*, para explicitar el hecho de que no enuncia dependencias específicas del contexto):

$$\mathcal{D}^{NC}(p) = \left\{ I(X_i, X_j \mid X_Z)_p \,\middle|\, i \neq j \in V; Z \subseteq V \setminus \{i, j\} \right\}.$$

Sin embargo, para modelar la estructura de grano más fino de modelos loglineales generales, es conveniente redefinir el modelo de dependencias incluyendo todas las posibles aserciones de dependencia específicas del contexto, de la siguiente

manera:

$$\mathcal{D}(p) = \left\{ I(X_i, X_j \mid x_U, X_W)_p \;\middle|\; \begin{array}{l} i \neq j \in V; \\ U, W \subseteq V \setminus \{i, j\}; \\ U \cap W = \emptyset; \\ x_U \in val(X_U) \end{array} \right\}. \tag{2.2}$$

Esta expresión es el punto de partida para la propuesta de comparación de estructuras de modelos loglineales de este trabajo. Es importante notar que a lo largo de esta tesis usaré indistintamente la notación relacionada con aserciones de dependencia o independencia para hablar de evaluación de estas expresiones no solo sobre distribuciones de probabilidad (por ejemplo, p) sino también sobre representaciones de su estructura (por ejemplo, un grafo o un conjunto de características, como se verá más adelante).

2.2. Redes de Markov

Una importante clase de modelos probabilísticos gráficos son las redes de Markov, basadas en grafos no dirigidos y empleadas para modelar correlaciones; en otras palabras, son modelos aplicables a fenómenos en los cuales se asume una influencia simétrica entre variables [Pearl, 1988, Lauritzen, 1996, Koller and Friedman, 2009, Spirtes et al., 2000, Gauraha, 2017]. En las redes de Markov, los grafos representan la estructura de distribuciones de probabilidad, codificando un modelo de dependencias (es decir, un conjunto de independencias y dependencias condicionales). Si bien la contribución de este trabajo no se basa exclusivamente en este tipo de estructuras, estas constituyen la base de algunas representaciones gráficas de modelos más generales —los modelos loglineales— y, lo que es más importante, el aporte está inspirado en un método para la comparación de estas estructuras presentado en la Sección 3.2.1.

Este apartado comienza por una introducción a las nociones de la teoría de grafos necesarias para definir una red de Markov y determinar sus propiedades. Luego, presenta la formalización de dichos modelos en cuanto a estructura y parámetros.

2.2.1. Teoría de grafos

Un grafo es un par ordenado $G = (V, E)$, donde V es un conjunto de nodos o vértices[1], y E es un conjunto de aristas. Los elementos de E son pares de nodos (a, b) donde $a, b \in V$. En la Figura 2.1, se ejemplifica un grafo que corresponde a $V = \{a, b, c, d\}$ y $E = \{(a, c), (b, c), (b, d), (c, d)\}$. Este trabajo se limita a grafos no dirigidos, donde las aristas no tienen direccionalidad y por lo tanto los pares en E no son ordenados. Si $(a, b) \in E$ se dice que a y b son *adyacentes*. Un grafo es *completo* si todos sus nodos son adyacentes, es decir, $\forall a, b \in V, (a, b) \in E$.

Un *camino* es una secuencia de nodos adyacentes en el grafo. Por ejemplo, la secuencia (a, c, b, d) es un camino en el grafo de la Figura 2.1 porque existe una arista en el grafo para cada par de elementos sucesivos en la secuencia. Se dice que un par de vértices a y b en un grafo G están *conectados* si existe algún camino desde a a b en G; de otra forma están *desconectados*.

Un conjunto de nodos S *separa* a los conjuntos A y B en G, denotado como $sep_G(A; B \mid S)$, si todo par de nodos (a, b) tal que $a \in A$ y $b \in B$ queda desconectado luego de eliminar los nodos S de G. En la Figura 2.1, c separa a $\{a\}$ y $\{b, d\}$; es decir, $sep_G(a; b, d \mid c)$.

Un subgrafo de $G = (V, E)$ es cualquier grafo $G' = (V', E')$ tal que $V' \subseteq V$, y $E' = \{(a, b) \in E \mid a, b \in V'\}$. Un *clique* en G es cualquier subgrafo completo de G. Por ejemplo, en la Figura 2.1, existen varios cliques de dos variables, por ejemplo

[1] Utilizo la misma notación V tanto para los nodos de un grafo como para los índices de las variables X_V de un dominio de manera intencional: en los modelos probabilísticos gráficos hay una correspondencia entre las variables y los nodos del grafo.

$G' = \{\{b,d\}, \{(b,d)\}\}$, y existe también un único clique de tres variables, $G'' = \{\{b,c,d\}, \{(b,c),(b,d),(c,d)\}\}$. Un clique es *máximo* cuando no es un subconjunto propio de otro clique. Siguiendo el mismo ejemplo, G'' es un clique máximo pero G' no lo es.

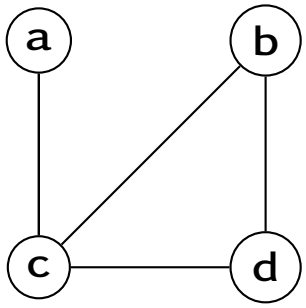


Figura 2.1: Ejemplo de un grafo no dirigido.

2.2.2. Parametrización y estructura de independencias en redes de Markov

2.2.2.1. Definiciones

Las redes de Markov son una clase de modelos probabilísticos gráficos cuya estructura es un grafo no dirigido y cuya parametrización consiste en un conjunto de *funciones potenciales* o *factores*. Los factores $\phi(X_{D_i})$, $i = 1 \dots k$ están definidos desde subconjuntos de las variables aleatorias $X_{D_i} \subseteq X_V$ hacia números no negativos. La forma de la distribución conjunta asociada a una red de Markov se expresa como

$$p(X_V) = \frac{1}{Z} \prod_{i=1}^{k} \phi_i(X_{D_i}).$$

Z es una constante de normalización que garantiza que p sea efectivamente una distribución de probabilidad, y se la conoce como la *función de partición*. Se define

como

$$Z = \sum_{x} \prod_{i=1}^{k} \phi_i(X_{D_i}).$$

En estos modelos, la parametrización define interacciones locales entre variables que se encuentran relacionadas directamente entre sí a través de los factores. Los modelos locales se vinculan multiplicando dichas interacciones; finalmente, la distribución se define tomando el producto de los factores locales y normalizándolos.

La estructura de estas interacciones es representable a través de un grafo no dirigido: los nodos corresponden a las variables en X_V, mientras que cada arista codifica una interacción directa entre un par de variables, es decir, una conexión que no está mediada por otras variables. Más concretamente, si la parametrización de p contiene un factor con un par de variables (X_i, X_j), entonces su grafo correspondiente contiene la arista (i, j). Se dice que una distribución p factoriza de acuerdo a un grafo G si para todos los cliques $X_C \subseteq X_V$ existen funciones no negativas ϕ_C que dependen de x solo a través de X_C. Es común asumir, sin pérdida de generalidad, que los conjuntos X_C son cliques máximos.

Dado un grafo G que representa a p, es posible leer independencias condicionales utilizando el concepto de separación de vértices en G. Este criterio nos permite obtener el siguiente conjunto de *independencias globales*:

$$\mathcal{I}(G) = \left\{ (X_A \perp\!\!\!\perp X_B \mid X_Z)_p \,\middle|\, sep_G(A; B \mid Z) \right\},$$

para todos los subconjuntos disjuntos A, B, Z en V.

En otras palabras, si G representa a p, podemos afirmar que $(X_A \perp\!\!\!\perp X_B \mid X_Z)_P$ si y solo si todo camino de A a B en G contiene un miembro de Z.

Ejemplo 2. *Por ejemplo, si suponemos que el grafo de la Figura 2.1 representa la estructura de una red de Markov G, podemos extraer las siguientes independencias:*

$$(X_a \perp\!\!\!\perp X_b \mid X_c)_G$$

$$(X_a \perp\!\!\!\perp X_b \mid X_c, X_d)_G$$

$$(X_a \perp\!\!\!\perp X_d \mid X_c)_G$$

$$(X_a \perp\!\!\!\perp X_d \mid X_b, X_c)_G,$$

además de múltiples dependencias, tales como $(X_a \not\perp\!\!\!\perp X_c \mid \emptyset)_G$, *o* $(X_a \not\perp\!\!\!\perp X_d \mid X_b)_G$.

2.2.2.2. Corrección y completitud

No siempre es posible usar la separación de vértices para extraer correctamente todas las dependencias e independencias contenidas en una distribución. En este contexto, se dice que un criterio es "correcto" cuando todas las independencias representadas en un grafo se satisfacen en cualquier distribución que factorice sobre ese grafo. Por otra parte, un criterio es "completo" cuando todas las independencias presentes en la distribución están representadas en su grafo asociado. Se ha demostrado que las únicas distribuciones para las cuales la separación es un criterio correcto y completo para determinar independencia condicional son aquellas para las cuales se cumplen ciertas propiedades, reproducidas a continuación desde [Pearl, 1988].

Definición 4. *Un grafo no dirigido G es un* ***mapa de dependencias*** *de un modelo de dependencias de p si existe una correspondencia uno a uno entre los elementos* X_V *del modelo y los vértices V de G, tal que para todos los subconjuntos disjuntos A, B y Z en V tenemos que*

$$(X_A \perp\!\!\!\perp X_B \mid X_Z)_p \implies sep_G(A; B \mid Z).$$

Similarmente, G es un ***mapa de independencias*** *de p si*

$$sep_G(A; B \mid Z) \implies (X_A \perp\!\!\!\perp X_B \mid X_Z)_p.$$

Si G es tanto un mapa de independencias como de dependencias de p, decimos que G es un ***mapa perfecto*** *de p.*

Decir que G es un mapa de independencias para p es equivalente a afirmar que p satisface las independencias en $\mathcal{I}(G)$. Esto quiere decir que, si un grafo es un mapa de independencias de un modelo de dependencias, los nodos que pueden ser separados en el grafo corresponden a variables condicionalmente independientes, pero no garantiza que los nodos conectados sean dependientes. Cuando esta última condición se cumple decimos que G es un *mapa de dependencias* de p. A la inversa, si se tiene que G es un mapa de dependencias de p, esto garantiza que los nodos conectados son dependientes en p, pero es posible que existan variables dependientes cuyos nodos están desconectados en G. Estas dificultades implican que deben existir restricciones sobre el conjunto de modelos de dependencia que pueden ser representados gráficamente por una red de Markov. Un modelo de este tipo se describe como *isomorfo a grafos*.

Definición 5. *Si existe un grafo no dirigido que constituye un mapa perfecto de p, es decir, si para todos los subconjuntos disjuntos A, B y Z en V se cumple que*

$$sep_G(A; B \mid Z) \iff (X_A \perp\!\!\!\perp X_B \mid X_Z)_p,$$

entonces se dice que p es ***isomorfa a grafos****.*

Las propiedades que deben cumplir las aserciones de independencia de un modelo para garantizar isomorfismo a grafos se encuentran caracterizadas a través de un conjunto de axiomas que pueden ser consultados en el Apéndice A.2.

2.2.2.3. Propiedades de Markov

El criterio de separación permite obtener un conjunto de independencias globales $\mathcal{I}(G)$. Adicionalmente, las independencias contenidas en una distribución se pueden definir de otras maneras, a través de lo que se conoce como las suposiciones locales de Markov. Las siguientes son las definiciones de *independencias por pares* e *independencias locales*, tomadas de [Koller and Friedman, 2009]:

Definición 6. *El conjunto de* ***independencias por pares*** *asociado con un grafo no dirigido* $G = (V, E)$ *se define como*

$$\mathcal{I}_p(G) = \left\{ (X_i \perp\!\!\!\perp X_j \mid X_V \setminus \{X_i, X_j\}) \,\middle|\, (i,j) \notin E \right\},$$

para todos los pares X_i, X_j *en* X_V.

Esta definición formaliza la idea de que dos variables que no están directamente enlazadas en un grafo solo pueden tener una correlación mediada por otra variable o conjunto de variables. Indirectamente, también establece que, si existe una correlación que no es mediada por ninguna otra variable o conjunto de variables, esta se representa a través de una arista.

Definición 7. *La* ***manta de Markov*** *de una variable* $X_i \in X_V$, *denotada como* $MB_G(X_i)$, *son los vecinos del nodo i en G. Las* ***independencias locales*** *asociadas con G se definen como*

$$\mathcal{I}_\ell(G) = \{ (X_i \perp\!\!\!\perp X_V \setminus \{X_i\} \setminus MB_G(X_i) \mid MB_G(X_i)) \,|\, X_i \in X_V \}.$$

Esta última definición implica que cualquier variable es independiente del resto dadas sus variables vecinas en el grafo. Visto de otra manera, dos variables que no están directamente enlazadas son independientes condicionadas en la manta de Markov de cualquiera de ellas.

Los tres conjuntos de independencias están relacionados entre sí. Para cualquier G, $\mathcal{I}(G) \implies \mathcal{I}_\ell(G)$, mientras que $\mathcal{I}_\ell(G) \implies \mathcal{I}_p(G)$. La recíproca solo se cumple para distribuciones positivas[2]; por lo tanto, si p es positiva, entonces $\mathcal{I}_p(G) = \mathcal{I}_\ell(G) = \mathcal{I}(G)$; esta equivalencia deriva de un importante resultado conocido como el *teorema de Hammersley-Clifford* [Hammersley and Clifford, 1971].

2.3. Modelos loglineales

El formalismo de modelos loglineales provee un marco general para el análisis de correlaciones. En particular, su parametrización permite visualizar patrones que involucran valores particulares de variables, es decir, permite codificar independencias específicas del contexto [Spirtes et al., 2000, Koller and Friedman, 2009]. En un modelo loglineal, la distribución conjunta sobre las variables es una ecuación del logaritmo de la suma de un número de funciones. Al igual que en el caso de redes de Markov, esta descomposición está dada por una estructura que consiste en un conjunto de funciones de las variables del dominio. De hecho, cualquier red de Markov parametrizada con factores positivos puede ser convertida a una representación logarítmica, reescribiendo los factores como $\phi(X_{D_i}) = \exp\left\{\ln \phi(X_{D_i})\right\}$ y haciendo

$$p(X_V) = \frac{1}{Z} \exp\left\{ \sum_{i=1}^{k} \ln \phi(X_{D_i}) \right\}.$$

La estructura de un modelo loglineal se expresa de forma general como un conjunto de *funciones características* (*features* en inglés) $F = \{f_i(X_{D_i})\}$, cada una de las cuales determina un valor numérico para cada asignación x_{D_i} a algún subconjunto $X_{D_i} \subseteq X_V$. Comúnmente, las características se definen como funciones

[2] Una distribución p es positiva si $\forall x \in \mathcal{X}$, $p(x) > 0$.

indicatrices, las cuales toman el valor 1 para un valor específico $x_{D_i} \in val(X_{D_i})$ y 0 en caso contrario. Estas permiten codificar una red de Markov estándar teniendo simplemente una característica por entrada en el potencial. Dado el conjunto F, los parámetros del modelo loglineal son los pesos $\theta = \{\theta_i : f_i \in F\}$. Luego, la distribución completa se define como

$$p(X_V) = \frac{1}{Z} \exp\left\{ \sum_{f_i \in F} \theta_i f_i(X_{D_i}) \right\}, \tag{2.3}$$

donde la función partición Z se redefine como

$$Z = \sum_x \exp\left\{ \sum_{f_i \in F} \theta_i f_i(x_{D_i}) \right\}.$$

La representación logarítmica asegura que la distribución es positiva; de hecho, los parámetros pueden tomar cualquier valor real. En particular, cuando $\theta_i = 0$ para algún i, se ignora la característica, y no tiene efecto en la distribución. Así, tenemos que en esta parametrización loglineal, la noción de independencia corresponde a la especificación de que uno o más términos de la suma en la Ec. (2.3) desaparecen. Los términos no nulos indican la presencia de interacciones probabilísticas entre las variables relacionadas por una misma característica. Al ser una representación muy genérica, la estructura puede ser global y local; es decir, dado un conjunto de características F, es posible verificar tanto independencias condicionales como independencias específicas del contexto. Para un contexto x_U, una independencia específica del contexto $(X_i \perp\!\!\!\perp X_j \mid x_U)_F$ se verifica con la ayuda de dos definiciones. Primero, el *alcance* S_f de una característica f como el conjunto de variables que están asignadas en ella. Segundo, $X_u(\cdot)$ es una función que devuelve el valor asignado a la variable X_u en una característica. Entonces, $(X_i \perp\!\!\!\perp X_j \mid x_U)_F$ se verifica, primero, tomando las características de F "compatibles" con x_U, es

decir, tomando $F' = \left\{ f \in F \middle| \forall u \in U, X_u \in S_f \implies X_u(x_U) = X_u(f) \right\}$. Luego, se verifica si para ninguna característica de F' las variables X_i y X_j están asignadas a la vez. Entonces, simbólicamente podemos decir que $(X_i \perp\!\!\!\perp X_j \mid x_U)_F$ se cumple si $\forall f \in F', \left\{ X_i, X_j \right\} \notin S_f$. Las independencias condicionales son deducibles aplicando la Ec. (2.1) a este procedimiento. Con este razonamiento, podemos afirmar que existe una interacción directa (una dependencia) entre cualquier par de variables $X_i, X_j \in X_V$ siempre que estas aparezcan juntas en alguna característica $f \in F$. De esta forma podemos asociar un modelo loglineal con un modelo de dependencias según la Ec. (2.2).

El siguiente es un ejemplo de la estructura de un modelo loglineal y su modelo de dependencias asociado.

Ejemplo 3. *Sea F un conjunto de características definidas sobre el conjunto X_V de variables binarias con $V = \{0,1,2\}$, compuesto de la siguiente manera:*

$$\begin{aligned} F = \{ \quad & < X_0 = 0, X_1 = 0 >, < X_0 = 0, X_1 = 1 >, \\ & < X_0 = 0, X_2 = 0 >, < X_0 = 0, X_2 = 1 >, \\ & < X_0 = 1, X_1 = 0, X_2 = 0 >, < X_0 = 1, X_1 = 0, X_2 = 1 >, \\ & < X_0 = 1, X_1 = 1, X_2 = 0 >, < X_0 = 1, X_1 = 1, X_2 = 1 > \}. \end{aligned}$$

Inspeccionando este conjunto es posible descubrir que X_1 y X_2 solo aparecen juntas en características en las que $X_0 = 1$. Esto implica que $(X_1 \perp\!\!\!\perp X_2 \mid X_0 = 0)_F$. En cualquier otro caso, todos los pares de variables son dependientes. Así, el modelo de dependencias de F podría ser expresado de forma más simple explicitando solo las independencias, y asumiendo que todo elemento no incluido corresponde a una aserción de dependencia[3]*. Tal conjunto es:*

$$\mathcal{I}(F) = \{ (X_1 \perp\!\!\!\perp X_2 \mid X_0 = 0)_F \}.$$

[3]Más adelante, en el Capítulo 4, emplearé la simplificación inversa.

2.3.1. Representación de la estructura de independencias de modelos loglineales

Si intentamos representar la estructura del Ejemplo 3 a través de un grafo no dirigido, podríamos hacerlo agregando una arista entre nodos correspondientes a cada par de variables que aparecen asignadas juntas en alguna característica. Esto produce el grafo completo de tres nodos, correspondiente a un modelo saturado, es decir, un modelo que contiene todas las posibles interacciones, y por lo tanto no codifica independencia alguna. Sin embargo, hemos visto que nuestra estructura original F contiene una independencia específica del contexto. Si, por el contrario, quisiéramos reflejar esta independencia en el grafo eliminando la arista entre los nodos 1 y 2, estaríamos introduciendo de forma espuria la independencia condicional $(X_1 \perp\!\!\!\perp X_2 \mid X_0)$, siendo que X_1 y X_2 son dependientes cuando $X_0 = 1$; es decir, estaríamos codificando una ausencia de interacción entre esas variables cuando sí existe una interacción (contextual). Esto nos demuestra que un grafo no dirigido, si bien es una metáfora intuitiva que cuenta con una semántica sólida y bien definida, no es lo suficientemente expresivo para representar la estructura local contenida en un modelo loglineal arbitrario[4]. No obstante, existen en la literatura algunas propuestas que permiten la visualización e interpretación gráfica de estos modelos. Concretamente, presentaré a continuación los *modelos de interacción específicos del contexto*, que admiten la representación con grafos para contextos arbitrarios; los *modelos divididos*, basados en una estructura jerárquica de grafos progresivamente más simples; y, finalmente, los *modelos estratificados*, los cuales codifican estructura local en un único grafo para un subconjunto de modelos loglineales.

[4]Lo mismo se puede afirmar de grafos dirigidos, aunque no estén cubiertos en el alcance de este trabajo. Véase [Boutilier et al., 1996] para un análisis en esta clase de modelos.

Este apartado sólo pretende ser una breve descripción de los modelos existentes para posibles representaciones gráficas de modelos loglineales, y exponer las limitaciones que hacen inviable el desarrollo de un método de comparación de modelos expresados con ellas. Sin embargo, en el Capítulo 5 y en el Apéndice D, aprovecharé la semántica asociada a los modelos CSI como elemento auxiliar en la exposición. Adicionalmente, en el Capítulo 4 aparece un caso ilustrativo del aporte de la tesis, para el cual la representación mediante grafos estratificados es sumamente práctica, y por lo tanto utilizo esta herramienta para facilitar la interpretación intuitiva del ejemplo desarrollado.

2.3.1.1. Modelos CSI

[Højsgaard, 2004] introduce una clase de modelos loglineales llamados *modelos de interacción específicos del contexto* (*modelos CSI* por sus siglas en inglés). En términos generales, la estructura de estos modelos es una descomposición que combina asignaciones y cliques, y su parametrización se define en base a estos elementos. A continuación introduciré los conceptos formales para definir un modelo CSI.

Sean A, B, C, D, E y F subconjuntos disjuntos de V. Un *generador* es un par $[x_B; X_A]$ donde $x_B \in val(X_B)$. Una *clase generadora* es un conjunto de generadores

$$\mathcal{C} = \{[x_B; X_A], [x_D; X_C], \ldots, [x_F; X_E]\};$$

también puede escribirse como $[x_B; X_A], [x_D; X_C], \ldots, [x_F; X_E]$. Una clase generadora especifica una familia de distribuciones de probabilidad con la siguiente factorización:

$$p(x) = \prod_{[x_B; X_A] \in \mathcal{C}} \psi_A^{x_B}(x_A | x_B).$$

La función $\psi_A^{x_B}(x_A|x_B)$ es un potencial que depende de $x_{A\cup B}$ y es igual a 1 para toda asignación de X_B distinta a x_B. Si $B = \emptyset$ el potencial se escribe ψ_A, mientras que si $A = \emptyset$ se escribe ψ^{x_B}.

La representación a través de una clase generadora admite la visualización de un grafo asociado a interacciones dado un contexto x_E^*. El razonamiento es el siguiente: si observáramos que las variables X_E toman los valores x_E^*, tendríamos que todos los factores para los cuales $x_{B\cap E} \neq x^*_{B\cap E}$ serían constantes (iguales a 1), y por lo tanto podrían ser ignorados. Así, en el contexto x_E^* solo existirían interacciones entre las variables X_C cuando $X_C \subset X_A \cup X_B$ para algún $[x_B; X_A] \in \mathcal{C}$ que satisficiera que $x_{B\cap E} \neq x^*_{B\cap E}$. Formalmente,

$$\mathcal{C}^*(x_E^*) = \left\{ X_B \cup X_A \,\middle|\, [x_B; X_A] \in \mathcal{C} \wedge x_{B\cap E} \neq x^*_{B\cap E} \right\}.$$

$\mathcal{C}^*(x_E^*)$ genera un grafo en el cual los nodos E están *instanciados* (asociados a una asignación) y los nodos $B \cup A \setminus E$ determinan los cliques del grafo para cada generador presente en $\mathcal{C}^*(x_E^*)$. Los nodos E pueden ser eliminados si así se desea. El resultado es llamado el *grafo instanciado* por x_E^* y se denota $\mathcal{G}_{x_E^*}(\mathcal{C})$. El resultado es equivalente a la estructura de una red de Markov tradicional sobre un dominio reducido —sin las variables en E. Así, este grafo permite leer independencias y dependencias usando el criterio de separación presentado en la Sección 2.2. Este hecho está formalizado en el siguiente teorema:

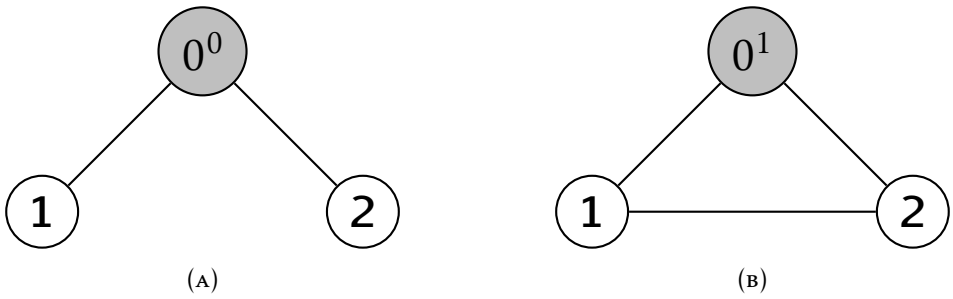
Teorema 1. *[Højsgaard [2004]] Sea $\mathcal{G}(x_U)$ el grafo instanciado por x_U. Entonces $(X_i \perp\!\!\!\perp X_j \mid x_U, X_W)$ si y solo si W separa a i y j en $\mathcal{G}(x_U)$.*

Ejemplo 4. *Este ejemplo reproduce la estructura del modelo loglineal del Ejemplo 3 como una clase generadora de un modelo CSI. Para facilitar la lectura estableceremos*

$x_{10} \equiv (X_1 = 0)$ *y* $x_{11} \equiv (X_1 = 1)$. *Sea*

$$\mathcal{C}_1 = \{[x_{00}; X_1], [x_{00}; X_2], [x_{01}; X_1, X_2]\}.$$

Los grafos instanciados por los contextos x_{00} *y* x_{01} *están representados en la Figura 2.2. A partir del Teorema 1 y de estos grafos instanciados, podemos extraer las siguientes afirmaciones:* $(X_1 \perp\!\!\!\perp X_2 \mid X_0 = 0)$ *por* $\mathcal{G}_{x_{00}}(\mathcal{C}_1)$, *y* $(X_1 \not\perp\!\!\!\perp X_2 \mid X_0 = 1)$ *por* $\mathcal{G}_{x_{01}}(\mathcal{C}_1)$.



(A) (B)

FIGURA 2.2: Grafos instanciados por $\mathcal{G}_{x_{00}}(\mathcal{C}_1)$ (2.2(a)), y por $\mathcal{G}_{x_{01}}(\mathcal{C}_1)$ (2.2(b)). Los nodos etiquetados como 0^0 y 0^1 (en gris) representan los contextos x_{00} y x_{01}, respectivamente.Del primero podemos leer $(X_1 \perp\!\!\!\perp X_2 \mid X_0 = 0)$ y del segundo $(X_1 \not\perp\!\!\!\perp X_2 \mid X_0 = 1)$.

Es muy importante señalar que un modelo loglineal puede ser expresado mediante distintas clases generadoras equivalentes entre sí. Respecto a la representación gráfica, son necesarios múltiples grafos instanciados para mostrar todas las independencias específicas del contexto presentes en el modelo. Además, no hay un método para estandarizar esta representación, lo que obstaculizaría su comparación. Una manera de sistematizar la visualización de modelos CSI podría ser un conjunto de grafos instanciados, uno por cada x_B tal que $[x_B; X_A] \in \mathcal{C}$, pero esto presenta dificultades, como muestra el siguiente ejemplo.

Ejemplo 5. *Siguiendo el Ejemplo 4, sea*

$$\mathcal{C}_2 = \{[x_{00}; X_1], [x_{00}; X_2], [x_{01}, x_{10}; X_2], [x_{01}, x_{11}; X_2]\}.$$

Podríamos querer comparar estas estructuras creando grafos instanciados de acuerdo a cada contexto en los generadores de $\mathcal{C}_1$ y $\mathcal{C}_2$. Con este método obtendríamos tres grafos instanciados: $\mathcal{G}_{x_{00}}(\mathcal{C}_2)$, $\mathcal{G}_{x_{01},x_{10}}(\mathcal{C}_2)$ y $\mathcal{G}_{x_{01},x_{11}}(\mathcal{C}_2)$. No está claro cómo efectuar una comparación entre ellos y los dos grafos instanciados obtenidos en el Ejemplo 4.

Sin embargo, si quisiéramos condicionar en x_{00} y x_{01},tendríamos que los grafos instanciados $\mathcal{G}_{x_{00}}(\mathcal{C}_2)$ y $\mathcal{G}_{x_{01}}(\mathcal{C}_2)$ codifican las mismas independencias o dependencias que $\mathcal{G}_{x_{00}}(\mathcal{C}_1)$ y $\mathcal{G}_{x_{01}}(\mathcal{C}_1)$. Podemos concluir entonces que $\mathcal{C}_1 \equiv \mathcal{C}_2$.

2.3.1.2. Modelos divididos

Los *modelos divididos* (*split models* en inglés) son un subconjunto de los modelos CSI y fueron presentados en [Højsgaard, 2003]. Un modelo dividido es un modelo CSI cuya clase generadora está dada por un *grafo dividido* (*split graph*). Debido a su complejidad, he decidido omitir las definiciones formales, limitándome a incluir una intuición suficiente para comprender su funcionamiento a rasgos generales.

Un grafo dividido es una colección de grafos cada vez más simples, organizados en una estructura jerárquica. Su definición es recursiva y se basa en dos operaciones realizadas sobre grafos: eliminación de aristas y creación de nuevos grafos mediante divisiones. Para definir un grafo dividido, se parte de un grafo no dirigido tradicional G. Se selecciona una variable X_a sobre la cual realizar la división en sus posibles valores $val(X_a)$. Cada valor se asocia a un subgrafo de G quitando a X_a, llamado *grafo contextual*. Sobre estos grafos se realizan reducciones, es decir, eliminación de aristas. Estas corresponden a independencias específicas del contexto presentes en la estructura local para cada $x_a \in val(X_a)$.

Ejemplo 6. *Para la clase generadora $\mathcal{C}_1$ del Ejemplo 4, es posible generar un grafo dividido como el de la Figura 2.3. Primero, realizando una división en los valores de X_0, es decir, en x_{00} y x_{01}. Luego, en cada nivel, tenemos un subgrafo con nodos*

$\{1,2\}$: *para* x_{00}, *su grafo no tiene aristas, codificando de esta forma la independencia* $(X_1 \perp\!\!\!\perp X_2 \mid X_0 = 0)$; *para* x_{01}, *la arista* $(1,2)$ *está presente, conservando la dependencia representada por la interacción de* X_1 *y* X_2 *en el generador* $[x_{01}; X_1, X_2]$ *de* $\mathcal{C}_1$.

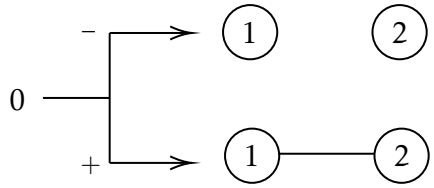


FIGURA 2.3: Ejemplo de un grafo dividido asociado a la clase generadora $\mathcal{C}_1$. La raíz del árbol corresponde a la variable X_0. Los valores 0 y 1 se han reemplazado por $-$ y $+$, respectivamente, para evitar ambigüedad.

Un grafo dividido determina una clase generadora; por lo tanto, es una forma de representación de la estructura de un modelo CSI. Por otra parte, una clase generadora, en general, no identifica unívocamente a un grafo dividido. En consecuencia, existen grafos divididos distintos pero equivalentes (cuando determinan clases generadoras equivalentes). Además, existen ciertas restricciones sobre cuáles divisiones son aceptables y cuáles no, dado que una división puede dar lugar a reducciones no válidas en el modelo, no producir reducción alguna, o incluso añadir interacciones (véase [Højsgaard, 2003], Sección 4.2). Por estas razones es claro que una comparación de estructuras de modelos CSI representadas como grafos divididos es incluso más compleja que la comparación de clases generadoras.

2.3.1.3. Modelos gráficos estratificados

Una extensión al formalismo de modelos probabilísticos gráficos tradicional fue presentada por primera vez en [Corander, 2003] y formalizada luego en [Nyman et al., 2014b]. Esta clase de modelos, llamados *modelos gráficos estratificados* (*stratified graphical models*), permite la representación de múltiples independencias específicas del contexto sobre un grafo.

Informalmente, estos modelos representan una distribución con un grafo no dirigido modificado por un conjunto de etiquetas llamadas *estratos*. Los estratos están asociados a aristas y codifican independencias específicas del contexto entre las variables conectadas por ellas. Cada estrato es un conjunto de configuraciones de las variables adyacentes al par de una arista determinada, indicando que se cumple una independencia condicionando en cada una de esas configuraciones. Formalmente:

Definición 8 ([Nyman et al., 2014b]). ***Estrato.*** *Sea el par (G,p) un modelo gráfico para X_V. Para toda $\{\delta,\gamma\} \in E$, sea $L_{\{\delta,\gamma\}}$ el conjunto de nodos adyacentes tanto a δ como a γ. Para $L_{\{\delta,\gamma\}}$, se define el estrato de la arista $\{\delta,\gamma\}$ como el conjunto $\mathcal{L}_{\{\delta,\gamma\}}$ que contiene a todas las configuraciones $x_{L_{\{\delta,\gamma\}}} \in \mathcal{X}_{L_{\{\delta,\gamma\}}}$ para las cuales X_δ y X_γ son independientes dado $x_{L_{\{\delta,\gamma\}}}$, es decir,*

$$\mathcal{L}_{\{\delta,\gamma\}} = \left\{ x_{L_{\{\delta,\gamma\}}} \in \mathcal{X}_{L_{\{\delta,\gamma\}}} \,\middle|\, (X_\delta \perp\!\!\!\perp X_\gamma \mid x_{L_{\{\delta,\gamma\}}})_p \right\}.$$

Usando un conjunto de estratos es posible representar las independencias específicas del contexto de un modelo loglineal. Esta representación define una clase de modelos, formalizada como sigue:

Definición 9 ([Nyman et al., 2014b]). ***Modelo gráfico estratificado.*** *Un modelo gráfico estratificado se define por la tripla (G,L,p), donde G es el grafo subyacente, L es la colección de todos los estratos $\mathcal{L}_{\{\delta,\gamma\}}$ para las aristas de G, y p es una distribución conjunta sobre X_V que factoriza de acuerdo a las restricciones impuestas por G y L.*

Ejemplo 7. *La Figura 2.4 muestra el grafo estratificado del modelo CSI del Ejemplo 4. Esta estructura consiste en un grafo subyacente completo y un conjunto de estratos L con un único elemento: $\mathcal{L}_{\{1,2\}} = \{< X_0 = 0 >\}$.*

A pesar de su interpretabilidad y eficiencia en la representación, los modelos gráficos estratificados requieren de fuertes restricciones para aprendizaje e

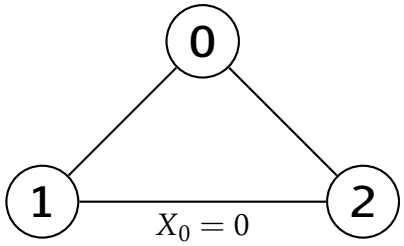


FIGURA 2.4: Grafo estratificado correspondiente al modelo del Ejemplo 4. La etiqueta en la arista $(1, 2)$ representa el estrato $X_0 = 0$.

inferencia. En cuanto a comparación, si bien es posible imaginar una medición directa (aproximada) entre dos modelos estratificados con el mismo grafo subyacente en base a un conteo de estratos, intentar comparar dos modelos con distintos grafos subyacentes presenta los mismos inconvenientes que las representaciones tratadas previamente. Esta complejidad deriva de las distintas formas de definir estratos al partir de adyacencias de variables, las cuales pueden variar considerablemente de un grafo a otro.

2.4. Aprendizaje automático de modelos loglineales

La comparación de estructuras de modelos loglineales, y de modelos probabilísticos en general, puede ser pensada como un problema independiente del área de aprendizaje automático; sin embargo, la idea para la contribución de este trabajo se originó en esta área y es aplicable a la misma: la medida propuesta puede emplearse en la evaluación de estructuras aprendidas por algoritmos. Esta evaluación consiste en comparar estructuras sintéticas, a partir de las cuales se muestrearon datos para el aprendizaje automático, con las estructuras aprendidas por cada algoritmo desde esos datos. En consecuencia, presento a continuación una reseña del problema de aprendizaje de estructuras y descripciones de algoritmos para el aprendizaje de modelos loglineales.

2.4.1. El problema del aprendizaje de estructuras

Un problema de aprendizaje o descubrimiento de estructuras consiste en encontrar información sobre una estructura que es la fuente de un conjunto de datos, teniendo un conjunto de posibles estructuras alternativas a considerar [Spirtes et al., 2000].

Parte de la literatura de aprendizaje de modelos probabilísticos separa los usos posibles de las estructuras aprendidas (o distribuciones completas en el caso de un posterior aprendizaje de parámetros) en dos categorías. Por un lado, se pueden utilizar las estructuras para construir modelos precisos para tareas de inferencia, es decir, para la toma de decisiones sobre una o más hipótesis, que podrían ser verdaderas en algunas estructuras candidatas y falsas en otras. En los modelos probabilísticos la inferencia se traduce en la estimación de probabilidades condicionales y marginales [Lowd and Davis, 2014, Van Haaren and Davis, 2012, Davis and Domingos, 2010].

Por otro lado, la información a obtener desde el aprendizaje puede ser la estructura que modele por completo los datos para comprender un fenómeno. Así, es posible descubrir dependencias significativas entre variables que permitan interpretar interacciones de los efectos medidos en el dominio de interés [Lee et al., 2006, Van Haaren et al., 2013, Claeskens et al., 2015, Nyman et al., 2014a, Pensar et al., 2017]. En la literatura, el primer objetivo se conoce como *estimación de densidad* mientras que el segundo es llamado *descubrimiento de conocimiento* (véase [Koller and Friedman, 2009], Capítulo 16).

Otra forma de clasificar enfoques consiste en el planteo del problema de aprendizaje, directamente relacionado con el diseño del algoritmo. En modelos probabilísticos gráficos, el problema suele plantearse como una búsqueda con una función de puntaje de estructuras, dando como resultado algoritmos basados en

puntaje, o bien como la construcción de una estructura a partir de pruebas de independencia, para proponer entonces algoritmos basados en restricciones o basados en independencias. Muchos de los algoritmos existentes de búsqueda de estructuras (para modelos probabilísticos en general) parten desde una estructura arbitraria, una estructura completa (es decir, sin restricciones de parámetros) o una estructura vacía (es decir, completamente restringida en la que todas las variables son independientes). El procedimiento aplica una medida de evaluación (llamada frecuentemente puntaje o *fitness*) para determinar qué parámetro fijar o liberar, según el caso, definiendo así independencias y dependencias en la estructura. Los algoritmos basados en independencias, en cambio, descubren una estructura aplicando una sucesión de pruebas de independencia estadística entre variables, tales como Información Mutua [Cover and Thomas, 2012], χ^2 de Pearson o G^2 [Agresti, 2002] , entre otros. Con los resultados de estas pruebas, los algoritmos deciden qué restricciones aplicar a la estructura, por ejemplo, a través de aristas o ausencia de ellas en grafos.

Para modelos loglineales, se utiliza preponderamente el enfoque de búsqueda, con dos variantes: la búsqueda global tradicional que optimiza una función de puntaje de modelos, y una búsqueda local a nivel de características, que evalúa características candidatas a ser agregadas o eliminadas del modelo en cada paso. En este último caso, la construcción del modelo se efectúa evaluando características candidatas generadas siguiendo algún criterio, o convirtiendo las mismas observaciones del conjunto de datos en características [Della Pietra et al., 1997, Lee et al., 2006, Davis and Domingos, 2010, Van Haaren and Davis, 2012, Lowd and Rooshenas, 2013, Lowd and Davis, 2014]. Esta caracterización suele denominarse problema de selección de características. A mi entender, solo existe hasta el momento una instancia de aprendizaje de estructuras de modelos loglineales basada en independencias, en [Edera et al., 2014a].

En lo concerniente a modelos gráficos no dirigidos, el aprendizaje de su estructura ha sido ampliamente investigado, con numerosos algoritmos propuestos, tanto desde la perspectiva basada en puntaje como basada en restricciones. Una reseña amplia y concisa se encuentra en [Schlüter, 2014]. Además, otros algoritmos han sido publicados posteriormente en [Pensar et al., 2014, Bresler, 2015, Gao and Ji, 2017, Kuronen et al., 2019].

Más recientemente, el aprendizaje automático de estructuras de modelos loglineales ha recibido considerable interés [Della Pietra et al., 1997, McCallum, 2003, Lee et al., 2006, Davis and Domingos, 2010, Van Haaren and Davis, 2012, Lowd and Rooshenas, 2013, Lowd and Davis, 2014]. El mayor poder expresivo de estos modelos tiene como desventajas una complejidad computacional de aprendizaje mucho mayor y dificultades en la interpretabilidad del modelo. La primera se traduce en espacios de búsqueda extremadamente amplios, o bien en un gran número de características o pruebas de independencia candidatas, según la perspectiva del algoritmo en concreto. La segunda ha llevado al desarrollo de representaciones gráficas que auxilian en la interpretación mas no resuelven el problema de comparabilidad entre estructuras, e introducen limitaciones en el conjunto de modelos representables, como señala la Sección 2.3.1.

2.4.2. Metodologías de evaluación del aprendizaje de estructuras

En general, en el área de aprendizaje de estructuras de modelos probabilísticos gráficos se utilizan dos clases de metodologías experimentales con el objetivo de evaluar y comparar la calidad de las estructuras aprendidas por distintos algoritmos —y, en algunos casos, la eficiencia del aprendizaje. El primer caso se observa en trabajos relacionados con el aprendizaje de modelos probabilísticos

gráficos tradicionales, mientras que el segundo es ampliamente utilizado para la evaluación de algoritmos que aprenden estructuras de modelos loglineales.

El primer caso, esquematizado en la Figura 2.5, consiste en la experimentación con modelos sintéticos. Estos modelos son diseñados a mano y, por lo tanto, tanto su estructura como sus parámetros son conocidos. Estos modelos sintéticos se consideran los "reales" y su estructura se utiliza al final de la experimentación para ser comparadas con aquellas estructuras aprendidas por distintos algoritmos competidores. La metodología consiste, en primer lugar, en diseñar uno o más modelos sintéticos adecuados al problema. Por ejemplo, si se desea demostrar las ventajas de un algoritmo muy eficiente pueden diseñarse redes de gran cantidad de variables y muchos datos, o si el algoritmo detecta independencias específicas del contexto puede diseñarse un modelo que contenga muchas independencias de este tipo. Luego, se realizan múltiples muestreos de cada modelo sintético para obtener conjuntos de datos de entrenamiento. Posteriormente se ejecutan los algoritmos con sus respectivas configuraciones que se han seleccionado para la experimentación sobre cada conjunto de datos. Por último, las estructuras obtenidas en cada ejecución son comparadas con la estructura sintética correspondiente utilizando una medida de distancia; por ejemplo, la distancia de Hamming (véase la Sección 3.2). Los resultados se interpretan como mejor calidad de aprendizaje cuanto menor es la distancia de la estructura aprendida respecto a la sintética. Algunos ejemplos de trabajos del área en que se utiliza esta metodología son: [Tsamardinos et al., 2006, Bromberg et al., 2009, Pensar et al., 2017, Schlüter et al., 2014, Schlüter et al., 2018].

El segundo caso es un procedimiento ampliamente utilizado tanto para modelos contextuales como no contextuales, pero resulta la única alternativa en el aprendizaje de estructuras de modelos loglineales, debido a la carencia de una medida de distancia entre tales estructuras. Existen dos variantes: en la primera y

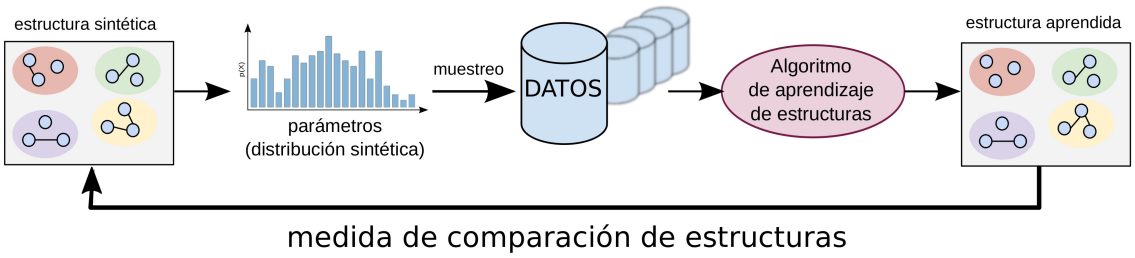


FIGURA 2.5: Metodología experimental basada en similitud entre estructuras aprendidas y estructuras subyacentes conocidas.

más utilizada, se recolecta una colección de conjuntos de datos comúnmente llamados "del mundo real", porque los datos son capturados desde dominios reales y no generados por algoritmos en base a modelos sintéticos. El aprendizaje se realiza de la misma manera que en el caso sintético, pero requiere un segundo paso de aprendizaje de parámetros de todos los modelos aprendidos (utilizando los mismos conjuntos de datos). De esta forma se obtienen modelos completos que pueden ser evaluados en base a su capacidad predictiva, respondiendo preguntas de inferencia. Algunos ejemplos de uso de esta metodología son: [Lowd and Rooshenas, 2013, Lowd and Davis, 2014, Davis and Domingos, 2010, Van Haaren and Davis, 2012]. La otra variante, resumida en la Figura 2.7, se usa en el caso particular de modelos contextuales, e involucra el diseño de modelos sintéticos y muestreo a partir de ellos tal como en el primer caso. Sin embargo, al no existir medida de comparación de estructuras, se requiere también el aprendizaje de parámetros. Posteriormente las distribuciones aprendidas se comparan con el modelo sintético con una medida de comparación de distribuciones. Esta metodología aparece en [Edera et al., 2014a] y [Edera et al., 2014b].

Las medidas de evaluación de estructuras y distribuciones más comunes se encuentran descritas en detalle en el Capítulo 3.

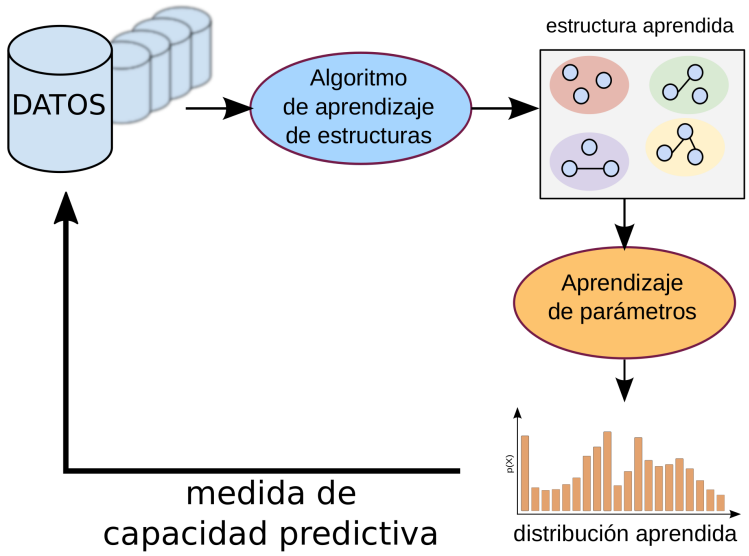


FIGURA 2.6: Metodología experimental basada en capacidad predictiva (modelo subyacente desconocido).

2.4.3. Algoritmos de aprendizaje de modelos loglineales

En lo que resta de esta sección expongo algunos enfoques específicos de aprendizaje de estructuras de modelos loglineales.

En [Della Pietra et al., 1997] se presenta un algoritmo que construye un modelo loglineal de manera incremental. El modelo inicial está compuesto por un conjunto de características atómicas. En cada iteración el algoritmo genera características candidatas añadiéndoles asignaciones a las existentes (asignaciones de una variable a la vez). Se agrega al modelo la característica que maximice una función de ganancia. Luego, se estiman los parámetros del nuevo modelo y se repite el proceso. Consiste, por lo tanto, en una búsqueda voraz (*greedy* en inglés) en el espacio de conjuntos de características. El algoritmo es evaluado en la aplicación a un problema de extracción de características morfológicas de palabras. Los autores muestran cómo mejora la caracterización de palabras en inglés a medida que el algoritmo induce más características.

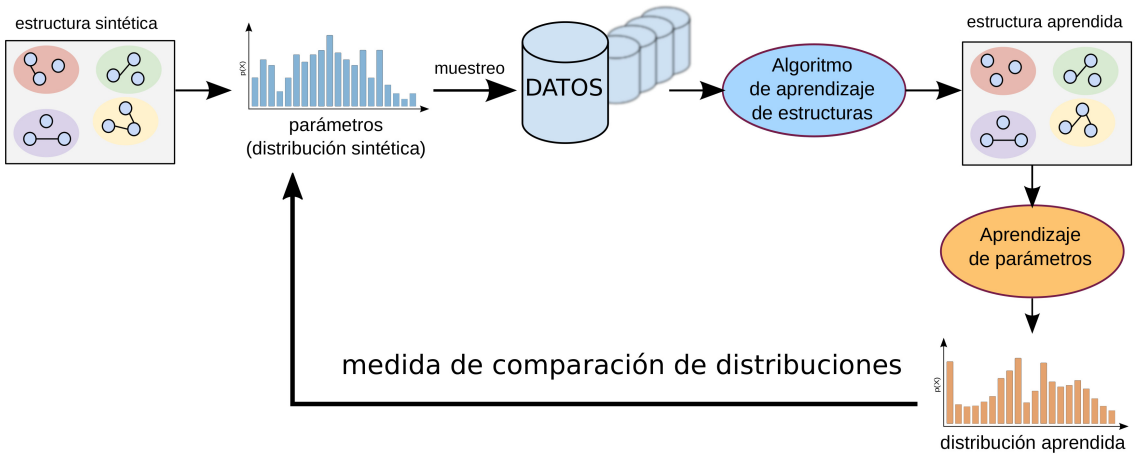


FIGURA 2.7: Metodología experimental basada en modelo subyacente conocido y evaluación mediante comparación de distribuciones.

Otro algoritmo, en [McCallum, 2003], también sigue un método incremental basado en una función de ganancia, que añade múltiples características en cada iteración. El enfoque se ejemplifica en dos tareas de procesamiento del lenguaje natural: extracción de entidades nombradas y segmentación de frases nominales.

El algortimo ACMN [Lowd and Rooshenas, 2013] se basa en la representación de circuitos aritméticos y sigue el enfoque de [Della Pietra et al., 1997] y [McCallum, 2003]. Comienza con características atómicas y realiza progresivamente *splits*, que consisten en la generación de características con cada asignación de alguna variable no asignada en la característica original. Este algoritmo y los siguientes fueron evaluados siguiendo la metodología experimental de *datos del mundo real*, como se explicó en la sección anterior.

El algoritmo BLM, de *Bottom-up Learning of Markov Networks* ("aprendizaje de abajo hacia arriba") [Davis and Domingos, 2010], utiliza el enfoque opuesto a los anteriores: toma como conjunto de características iniciales el conjunto de datos,

para luego generalizar las características iterativamente, es decir, quitar asignaciones a cada una de forma que representen un número mayor de observaciones.

Otro algoritmo "de abajo hacia arriba" es GSSL (*Generate-Select Structure Learning*) [Van Haaren and Davis, 2012]. Su funcionamiento consiste en la generación de un gran número de características a partir del conjunto de datos, eliminando aleatoriamente asignaciones de un cierto número de variables. En un paso posterior, el conjunto resultante se reduce tomando solo las características que aparecen un determinado número de veces. Lo que destaca de este enfoque es que tanto la selección de la característica a generalizar como el número de variables a eliminar y selección de estas variables se realizan al azar de manera uniforme. Esto lo convierte en un algoritmo muy eficiente, si bien más costoso en memoria que sus competidores.

Un enfoque alternativo a los anteriores consiste en aprender modelos locales y luego combinarlos. Dos ejemplos de este enfoque son descritos a continuación.

[Ravikumar et al., 2010] presenta un algoritmo de regresión logística con regularización L_1 para el aprendizaje de modelos locales, que luego construye un modelo global con una característica por cada par de variables que tuvo un peso de interacción no nulo en la regresión logística.

Por su parte, DTSL (*Decision Tree Structure Learning*) [Lowd and Davis, 2014] opera sobre cada variable del dominio aprendiendo un árbol de decisión para predecir dicha variable condicionada en las demás, y convierte cada árbol en un conjunto de características; todas las características juntas conforman luego la estructura del modelo aprendido.

Existen dos algoritmos basados en independencias que aprenden modelos loglineales: CSPC (*Context-Specific PC*) [Edera et al., 2014a] y CSGS (*Context-Specific*

GSMN) [Edera et al., 2014b]. Ambos utilizan estrategias existentes para el aprendizaje de redes de Markov no contextuales, respectivamente, PC [Spirtes and Glymour, 1991] y GSMN (*Grow-Shrink Markov Network Structure Learning*) [Bromberg et al., 2009], pero utilizando una representación exhaustiva y redundante para codificar independencias específicas del contexto, y realizando pruebas de independencia contextualizadas. Las pruebas de independencia utilizadas son básicamente pruebas de independencia condicional tradicionales, pero realizadas sobre un subconjunto del conjunto de datos para el cual se satisface un contexto dado. Cada iteración ejecuta el algoritmo para un contexto diferente. Estos algoritmos fueron comparados con competidores con medidas predictivas; pero también cuentan con experimentación en base a modelos sintéticos, en la cual se compararon distribuciones aprendidas con las originantes de los datos.

2.5. Notación y definiciones

Sea V un conjunto finito de índices para un conjunto de variables aleatorias X_V. Los subíndices en minúsculas denotan índices individuales (por ejemplo, $X_i, X_j \in X_V$ donde $i, j \in V$), mientras que los subíndices en mayúsculas denotan subconjuntos de índices (por ejemplo, $X_A \subseteq X_V$ donde $A \subseteq V$).

La notación $(X_A \perp\!\!\!\perp X_B \mid X_C)_\gamma$ denota que las variables en el conjunto X_A son (conjuntamente) independientes de aquellas en X_B condicionadas en los valores de las variables en X_C, para conjuntos disjuntos de índices A, B, y C en el modelo γ, mientras que $(X_A \not\perp\!\!\!\perp X_B \mid X_C)_\gamma$ denota dependencia condicional. $I(X_A, X_B \mid X_C)_\gamma$ denota una pregunta de independencia condicional en un modelo; simbólicamente:

$$I(X_A, X_B \mid X_C)_\gamma \text{ es verdadero} \iff (X_A \perp\!\!\!\perp X_B \mid X_C)_\gamma$$

El subíndice en $(X_A \perp\!\!\!\perp X_B \mid X_C)_\gamma$ se utiliza equivalentemente para distintas expresiones de un modelo o estructura, sea este una distribución de probabilidad, o bien una representación de la misma o de uno de sus componentes (tal como un modelo de dependencias, un conjunto de características o una característica individual); lo mismo para preguntas de independencia condicional.

Una variable X_k puede tomar un valor de un conjunto finito de configuraciones o asignaciones, denotadas por $val(X_k)$ e indexadas por V_k. Por ejemplo, para una variable binaria X_0, $val(X_0) = \{0,1\}$. Una configuración arbitraria será expresada en minúsculas, por ejemplo, x_K ($K \subseteq V$); la configuración de una sola variable será x_k.

Para hacer referencia a un valor específico se indexan los valores en el dominio de la variable, y se usa un segundo índice, genéricamente denotado por la letra m, para referirse al m-ésimo valor de la variable. Por ejemplo, x_{km} es el m-ésimo valor de la k-ésima variable.

Una independencia específica del contexto entre las variables X_A y X_B dadas las variables X_C y una configuración (contexto) x_D, con $D \not\subseteq A \cup B \cup C$, se denota como $(X_A \perp\!\!\!\perp X_B \mid X_C, x_D)_\gamma$ mientras que $(X_A \not\perp\!\!\!\perp X_B \mid X_C, x_D)_\gamma$ denota una dependencia específica del contexto; $I(X_A, X_B \mid X_C, x_D)_\gamma$ es una aserción de independencia específica del contexto.

Un modelo loglineal está compuesto por una estructura y un conjunto de parámetros. La estructura es un conjunto de características, usualmente denotado por F o G. Las características son funciones denotadas por letras minúsculas, tales como f, g o h. El valor que la variable $X_k \in X_V$ toma en la característica f es $X_k(f)$. Por ejemplo, si $f =< X_0 = 1, X_2 = 0, X_3 = 1 >$, entonces $X_2(f) = 0$. El conjunto de variables que se encuentran asignadas en una característica f es llamado el *alcance* de f y se designa S_f. Para el último ejemplo, $S_f = \{X_0, X_2, X_3\}$. La notación f^{ij} designa una característica de la cual se han eliminado las asignaciones de las

variables X_i y X_j. Por ejemplo, $f^{03} =< X_2 = 0 >$.

El espacio de todas las configuraciones para X_V se denota como $\mathcal{X}$. Un *contexto canónico* x es una asignación completa en un dominio, es decir, $x \in \mathcal{X}$, y $S_x = X_V$.

Capítulo 3

Análisis del estado del arte

En este capítulo presento medidas de comparación de modelos probabilísticos utilizadas en trabajos de aprendizaje automático de sus estructuras. No me es posible contrastar métodos alternativos para la comparación directa de estructuras de modelos loglineales, puesto que a mi entender no existen tales medidas en la literatura actual. Es relevante, sin embargo, tener en cuenta otros métodos, aplicables a otro tipo de modelos o bien a distribuciones completas (en lugar de estar enfocados solo en las estructuras). Por un lado, la distancia de Hamming es una medida ampliamente empleada para evaluar estructuras de modelos gráficos no dirigidos. Por otro lado, es posible comparar estructuras de modelos loglineales a través de algunas medidas indirectas. En las siguientes secciones comenzaremos por mencionar las propiedades deseables en medidas para comparar estructuras o distribuciones, y luego analizaremos los métodos existentes —para los casos mencionados— y sus limitaciones.

3.1. Definición y propiedades de una métrica

El concepto de medida de distancia, o simplemente métrica, es útil para analizar las ventajas y desventajas de algunos métodos discutidos en este trabajo. Esta definición y sus propiedades tomarán especial relevancia en el Capítulo 5, donde presento una demostración de que el aporte de esta tesis es, efectivamente, una métrica.

Una métrica es una función definida sobre duplas de un conjunto $d : (x, y) \rightarrow \mathbb{R}_{\geq 0}$, donde $\mathbb{R}_{\geq 0}$ es el conjunto de los números reales no negativos. Para todos los elementos x, y, z en el dominio, una métrica debe satisfacer las siguientes cuatro propiedades:

- Axioma de separación (no negatividad): $d(x, y) \geq 0$.
- Axioma de coincidencia: $d(x, y) = 0 \Leftrightarrow x = y$.
- Simetría: $d(x, y) = d(y, x)$.
- Desigualdad triangular (o subaditividad): $d(x, z) \leq d(x, y) + d(y, z)$.

Se dice que una función que cumple las primeras dos condiciones es *positiva*.

3.2. Comparación de modelos gráficos no dirigidos

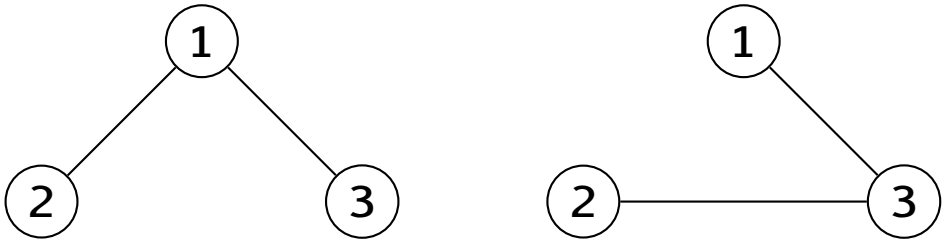
3.2.1. Distancia de Hamming

La distancia de Hamming es una *métrica* que tiene su origen en la teoría de la información, siendo su uso más difundido la detección y corrección de errores en teoría de códigos. Está definida en general sobre dos secuencias de símbolos de igual longitud, y es igual al número de posiciones en que el símbolo de una

secuencia difiere al de la otra secuencia en la misma posición. Dicho de otra manera, es la cantidad de sustituciones necesarias para convertir una secuencia en otra. Esta distancia es un caso particular de la llamada *distancia de Levenshtein* o *distancia de edición*.

En el caso de grafos no dirigidos, la distancia de Hamming es equivalente a la suma de falsas aristas (falsos postivios) y falsas ausencias de aristas (falsos negativos) de un grafo respecto a otro. Podemos comprender esto fácilmente con la ayuda del siguiente ejemplo. Si consideramos cada grafo con una representación alternativa, concretamente, con su *matriz de adyacencias*, obtenemos una secuencia de ceros y unos asociados a cada uno. En cada matriz, un uno indica una arista (correspondiente a una dependencia directa entre variables) entre los nodos asociados a la fila y la columna en que se encuentra dicho valor, y un cero corresponde a la ausencia de arista (independencia condicional) entre esos nodos. Por lo tanto, cuando en una matriz del grafo a comparar aparece un uno en una posición en la que la matriz del grafo original contiene un cero, esto se interpreta como un falso positivo, mientras que en el caso inverso se trata de un falso negativo. En la Figura 3.1 tenemos un ejemplo de grafos donde el segundo presenta un falso positivo y un falso negativo respecto al primero, y en la Figura 3.2 podemos observar sus respectivas matrices de adyacencias. Debido a que se trata de grafos no dirigidos, sus matrices son simétricas, por lo que he omitido los valores de la submatriz triangular inferior. Además, esta clase de estructuras no contiene bucles, por lo que los elementos de las diagonales están vacíos.

Probablemente gracias a sus propiedades y facilidad de interpretación e implementación, esta métrica es extensamente empleada para la evaluación de algoritmos de aprendizaje de estructuras [Pensar et al., 2014, Schlüter et al., 2014, Schlüter et al., 2018]. Desafortunadamente, no tiene la precisión necesaria para ser aplicable de igual manera a modelos loglineales en los que la estructura puede



(A) (B)

FIGURA 3.1: Dos grafos no dirigidos de tres nodos cada uno. El grafo en 3.1(b) contiene un falso positivo respecto al grafo en 3.1(a) (la arista entre los nodos 2 y 3), y también contiene un falso negativo (ausencia de arista entre los nodos 1 y 2). Nótese que la arista entre 1 y 3 constituye un verdadero positivo.

$$\begin{bmatrix} \cdot & 1 & 1 \\ & \cdot & 0 \\ & & \cdot \end{bmatrix} \qquad \begin{bmatrix} \cdot & \mathbf{0} & 1 \\ & \cdot & \mathbf{1} \\ & & \cdot \end{bmatrix}$$

(A) (B)

FIGURA 3.2: Matrices de adyacencia asociadas a los grafos de la Figura 3.1. Los valores señalados en negrita en 3.2(b) son los que difieren respecto a 3.2(a); concretamente, el valor 1 corresponde al falso positivo y el 0 al falso negativo. En este caso, la distancia de Hamming entre estos grafos es igual a 2.

contener independencias más refinadas entre variables. Sin embargo, el aporte de esta tesis está inspirado en este método, como veremos en la Sección 4.1.

3.3. Comparación de modelos loglineales

3.3.1. Divergencia KL

La *divergencia de Kullback-Leibler* (divergencia KL) [Kullback and Leibler, 1951, Cover and Thomas, 2012], también llamada *entropía relativa*, es una medida de la similitud de dos distribuciones de probabilidad $p(X)$ y $q(X)$, definidas sobre el mismo dominio. Se define como

$$D(p\|q) = \sum_{x \in \mathcal{X}} p(x) log \frac{p(x)}{q(x)},$$

siguiendo las convenciones de que $0\ log \frac{0}{q(x)} = 0$ y $p(x)\ log \frac{p(x)}{0} = \infty$.

La divergencia KL se interpreta como la pérdida de información asociada a utilizar $q(x)$ como aproximación de $p(x)$. Alternativamente, puede pensarse como una aproximación de la distancia entre las dos distribuciones, pues satisface la intuición de que el costo de aproximar $p(x)$ con $q(x)$ es menor mientras más parecida sea esta a aquella, siendo nulo solo cuando estas son idénticas y positivo en cualquier otro caso. Por lo tanto, la divergencia KL satisface los axiomas de separación y de coincidencia (es decir, es positiva). Sin embargo, no es una métrica, debido a que no satisface simetría ni la desigualdad triangular.

A pesar de estas características, la divergencia KL es útil como medida de la calidad de una distribución aprendida a partir de un conjunto de datos muestreados desde una distribución conocida. La verdadera desventaja de este método consiste en que no es un indicador directo de la similitud de las estructuras de los modelos, dado que involucra también los parámetros del modelo aprendido. Esto tiene dos

desventajas. En primer lugar, debido a que el aprendizaje de parámetros depende de la complejidad muestral del método empleado, los parámetros obtenidos son propensos a errores de estimación por escasez de datos. En segundo lugar, es posible que los parámetros anulen interacciones espurias presentes en la distribución aproximada, ocultando la existencia de las mismas. Esto se traduce en una dificultad para evaluar posibles tendencias de los algoritmos de aprendizaje de estructuras a introducir falsos positivos (falsas interacciones entre variables).

Para ilustrar las limitaciones mencionadas y para brindar una intuición de la utilidad del método propuesto en este trabajo, he incluido en el Apéndice D un ejemplo que contrasta los valores de divergencia KL obtenidos desde estructuras con distintas cantidades de falsos positivos y falsos negativos, medidos según la métrica propuesta.

3.3.2. Logverosimilitud condicional y logverosimilitud condicional marginal

Los métodos mencionados anteriormente tienen sentido cuando la comparación se realiza en base a una distribución conocida. Sin embargo, para una evaluación más cercana al comportamiento de los algoritmos en aplicaciones reales, es común conducir en la práctica experimentos de aprendizaje de estructuras sobre conjuntos de datos donde su distribución de probabilidad subyacente es desconocida. En estos escenarios, el análisis de resultados se realiza en base a la capacidad de predicción de los modelos aprendidos: se utilizan conjuntos de datos de prueba para una tarea de inferencia determinada, y se cuantifica y compara qué tan precisas son las predicciones de cada modelo. Si bien estas no son técnicas de comparación de estructuras o modelos, resulta de interés mencionarlas por su aplicación como formas de comparar la calidad de modelos aprendidos.

Una función de evaluación de este tipo es la *logverosimilitud condicional* (CLL por sus siglas en inglés) [Kok and Domingos, 2010, Lowd and Rooshenas, 2013], mientras que una versión aproximada, la *logverosimilitud condicional marginal* tiene un uso más amplio [Lee et al., 2007, Lowd and Rooshenas, 2013, Lowd and Davis, 2014, Van Haaren and Davis, 2012, Davis and Domingos, 2010, Edera et al., 2014a,b].

La CLL de una distribución aprendida p se calcula para un conjunto de *variables evidencia* X_E (es decir, que se consideran observadas) y un conjunto de *variables de consulta* X_Q, sobre un conjunto de datos de prueba $\mathcal{D}$:

$$CLL(\mathcal{D}) = \sum_{x \in \mathcal{D}} log\ p(x_Q | x_E).$$

Usualmente, la probabilidad conjunta es difícil de estimar, por lo que se emplea la suma de las probabilidades marginales de las variables de consulta:

$$CMLL(\mathcal{D}) = \sum_{x \in \mathcal{D}} \sum_{x_i \in X_Q} log\ p(x_i | x_E).$$

Para eliminar cualquier sesgo introducido por la selección del conjunto X_Q, es posible repetir el procedimiento particionando el conjunto completo de variables X_V y empleando en cada iteración una partición como X_Q, y el resto, $X_E = X_V \backslash X_Q$, como variables evidencia.

Al igual que con la divergencia KL, el cómputo de estas funciones de evaluación requiere el aprendizaje de parámetros del modelo, pero tienen la ventaja de poder aplicarse a situaciones realistas de distribuciones subyacentes desconocidas.

3.3.3. Indicadores basados en características

Otros indicadores relacionados con cualidades de modelos loglineales que merecen mención son algunos valores computados sobre el conjunto de características de las estructuras [Lowd and Davis, 2014, Van Haaren and Davis, 2012, Edera et al., 2014a,b]. En particular, podemos nombrar el *número de características* y la *longitud promedio de características*. El primero no requiere explicación, mientras que el segundo consiste en el promedio de la cantidad de variables en el alcance de cada característica de la estructura del modelo. La utilidad del número de características reside en poder observar una dimensión de la complejidad de los modelos, lo que permite, por un lado, comparar modelos aprendidos con modelos sintéticos, y por otro lado, evaluar la complejidad de las estructuras resultantes de utilizar distintos algoritmos de aprendizaje. Con respecto a la longitud promedio de características, esta constituye un método de análisis indirecto de la presencia o ausencia de independencias específicas del contexto, y posibilita discriminar entre algoritmos de aprendizaje que producen estructuras con mayor o menor complejidad en cuanto a la longitud de características. Conocer estas cualidades es importante cuando se desea obtener estructuras que tiendan a ser más simples en su representación o que permitan realizar inferencia de manera más eficiente.

Capítulo 4

Medida de comparación de estructuras de independencia de modelos loglineales

Se presenta el principal aporte de la tesis, comenzando, por un lado, con un análisis del problema a nivel computacional y una intuición sobre la justificación de la propuesta. Por otro lado, se expone en detalle en qué consiste el enfoque, proveyendo definiciones formales de los valores de la matriz de confusión propuesta, su descomposición y el desarrollo de ecuaciones y operaciones para el cómputo eficiente de sus valores. Se incluyen luego todas las demostraciones asociadas a estos procedimientos, con algunos ejemplos. Al final del capítulo se presenta un resumen del método y un ejemplo de aplicación para facilitar su comprensión.

4.1. Enfoque de fuerza bruta

4.1.1. Complejidad del enfoque de fuerza bruta

Comparar las estructuras de dos modelos loglineales F y G implica comparar todas las independencias y dependencias codificadas en cada uno de ellos. Este conjunto exhaustivo puede definirse formalmente como

$$\mathcal{M}(F) = \left\{ I(X_i, X_j \mid x_U, X_W)_F \;\middle|\; \begin{array}{l} i \neq j \in V; \\ U \cup W \subseteq V \setminus \{i, j\}; \\ U \cap W = \emptyset; \\ x_U \in val(X_U) \end{array} \right\},$$

y una definición similar para G. Un enfoque simple de *fuerza bruta* consistiría en evaluar cada elemento de $\mathcal{M}(F)$ y $\mathcal{M}(G)$ y comparar sus valores de verdad. Por cada elemento, cuando ambos valores indican dependencia se cuenta un verdadero positivo; cuando F indica dependencia y G independencia se cuenta un falso negativo; cuando F indica independencia y G dependencia se cuenta un falso positivo; y cuando ambas independencias son verdaderas se cuenta un verdadero negativo. Finalmente se deberían reportar los valores totales para cada uno de los cuatro casos de verdaderos positivos, verdaderos negativos, falsos positivos y falsos negativos en una matriz de confusión.

En lo que sigue se utilizará una expresión más simple que contiene solo las dependencias:

$$\mathcal{D}(F) = \left\{ (X_i \not\perp\!\!\!\perp X_j \mid x_U, X_W)_F \;\middle|\; \begin{array}{l} i \neq j \in V; \\ U \cup W \subseteq V \setminus \{i,j\}; \\ U \cap W = \emptyset; \\ x_U \in val(X_U) \end{array} \right\}, \tag{4.1}$$

donde se interpreta que las independencias son simplemente su complemento, y que la dependencia o independencia de alguna aserción $t = I(X_i, X_j \mid x_Z)$ en el modelo F es capturada por su inclusión o exclusión de $\mathcal{D}(F)$, es decir, $t \in \mathcal{D}(F)$ indica que t es dependiente en F, mientras que $t \notin \mathcal{D}(F)$ indica que t es independiente en F.

Desafortunadamente, la complejidad de esta evaluación es exponencial en tres maneras:

1. existe un número exponencial de subconjuntos de $V \setminus \{i,j\}$;
2. para cada subconjunto de $V \setminus \{i,j\}$ existe un número exponencial de formas de descomponerlo en conjuntos disjuntos U y W (todas las particiones en dos conjuntos junto con $U = \emptyset, W = V \setminus \{i,j\}$ y $U = V \setminus \{i,j\}, W = \emptyset$);
3. por cada combinación de U y W, existe un número de contextos que es exponencial en la cardinalidad de U.

Estas dificultades computacionales son probablemente la razón por la que no se han ofrecido hasta la fecha medidas de comparación de estructuras contextuales. El desarrollo de esta tesis aborda estos tres puntos.

Con respecto a la primer complejidad exponencial, la propuesta que aporta esta tesis sigue un enfoque utilizado por los modelos no dirigidos (no contextuales).

Sin embargo, dicho enfoque no es directamente aplicable a estructuras contextuales, como se explicará en la siguiente sección (Sección 4.1.2). De la misma forma que en los modelos contextualizados, la estructura de un modelo gráfico (no contextualizado) H puede ser representada por el siguiente modelo de dependencias:

$$\mathcal{D}^{NC}(H) = \{(X_i \not\perp\!\!\!\perp X_j \mid X_U)_H \mid i \neq j \in V,\ U \subseteq V \setminus \{i,j\}\}, \tag{4.2}$$

donde es posible observar lo mencionado en el punto 1, dado que existe un número exponencial de subconjuntos $U \in V \setminus \{i,j\}$. En la práctica, e implícitamente, los grafos no dirigidos son comparados empleando los *modelos de dependencias por pares* $\mathcal{D}^{NC}_{\mathcal{P}}(H)$, un subconjunto de la Ec. (4.2) con cardinalidad polinomial, definido como

$$\mathcal{D}^{NC}_{\mathcal{P}}(H) \equiv \{(X_i \not\perp\!\!\!\perp X_j \mid X_{V\setminus\{i,j\}})_H \mid i \neq j \in V\},$$

es decir, dependencias donde el condicionante es de tamaño fijo y contiene a todas las variables excepto el par (X_i, X_j) (por esto las llamamos "por pares"). Esto es simplemente la distancia de Hamming de las aristas en su representación de grafo no dirigido, debido a que por la *propiedad de Markov por pares* (Pearl [1988]) tenemos que

$$(X_i, X_j) \in E \equiv (X_i \not\perp\!\!\!\perp X_j \mid X_{V\setminus\{i,j\}})_H,$$

donde E es el conjunto de aristas del grafo asociado a un modelo H.

En modelos no dirigidos, la comparación a través de modelos de dependencias por pares está justificada al demostrar que por cada par de modelos no dirigidos H_1 y H_2, la igualdad de sus modelos de dependencias por pares $\mathcal{D}^{NC}_{\mathcal{P}}(H_1)$ y $\mathcal{D}^{NC}_{\mathcal{P}}(H_2)$ implica la igualdad de sus modelos de dependencias generales $\mathcal{D}^{NC}(H_1)$ y $\mathcal{D}^{NC}(H_2)$ (la implicación en sentido opuesto es trivial ya que los modelos generales incluyen a todas las aserciones incluidas en los modelos por pares).

4.1.2. Una primera intuición a través de modelos CSI

Las estructuras de modelos loglineales no pueden compararse de forma precisa mediante la distancia de Hamming que se usa para grafos, porque, si se usara un grafo para representar una estructura arbitraria, se perderían independencias específicas del contexto. Solo quedarían representadas las que forman parte de independencias condicionales, debido a la equivalencia en la Ec. (2.1). Sería razonable pensar, entonces, que una forma de trasladar esta distancia estandarizada a la clase de estructuras de modelos loglineales sin perder precisión consistiría en emplear los modelos CSI (explicados en la Sección 2.3.1.1), puesto que permiten el uso de grafos pero también logran codificar estructuras locales.

Desde esta perspectiva, se debe tomar un conjunto de contextos y obtener sus grafos instanciados, para luego comparar estos grafos con la distancia de Hamming convencional. Sin embargo, se debe resolver de forma sistemática qué clase generadora usar, ya que, en general, F y G tendrán clases generadoras diferentes. Además, un mismo modelo loglineal puede tener distintas clases generadoras equivalentes entre sí. La dificultad es aún mayor teniendo en cuenta que, dependiendo de qué contextos se utilicen para instanciar grafos y qué nodos queden sin instanciar en cada uno, puede haber independencias específicas del contexto que queden ocultas y terminen sin formar parte de la comparación.

La solución de este trabajo puede verse desde esta perspectiva de modelos CSI como una solución a los problemas mencionados, de la siguiente manera: se debe tomar cada par de variables (X_i, X_j) del dominio, e instanciar las restantes, obteniendo un grafo instanciado de solo dos variables, X_i y X_j, por cada asignación posible de las variables en $X_{V\setminus\{i,j\}}$. Introduciremos para este fin la notación $\mathcal{X}^{ij} \equiv X_{V\setminus\{i,j\}}$, y llamaremos a cada elemento $x_Z \in \mathcal{X}^{ij}$ un *contexto por pares*, ya que pueden ser considerados como contextos canónicos en el espacio reducido de

configuraciones $\mathcal{X}^{ij}$. Entonces, en cada grafo instanciado, existirá o no una arista dependiendo de si se cumple o no la dependencia $(X_i \not\perp\!\!\!\perp X_j \mid x_Z)$. En la Sección 4.2, veremos que estas instanciaciones completas nos permiten identificar todas las independencias y dependencias específicas del contexto que pueden existir en un modelo loglineal arbitrario, a pesar de traducirse en un subconjunto de su modelo de dependencias. La formalización detallada de estas dependencias y su justificación se presentan a continuación.

4.1.3. Tratamiento de las complejidades para una comparación eficiente

El enfoque propuesto está inspirado en la reducción exponencial en complejidad que provee la distancia de Hamming para el caso de modelos no dirigidos al utilizar un subconjunto de su modelo de dependencias. Con este nuevo enfoque, dadas las estructuras de dos modelos loglineales F y G, estas se comparan a través de una versión reducida de $\mathcal{D}(F)$ y $\mathcal{D}(G)$. Esta versión solo contiene aserciones de independencias por pares con condicionante completo, es decir, donde el condicionante también involucra las variables $X_{V\setminus\{i,j\}}$, pero, en este caso, dicho condicionante será contextual: una conjunción de todas las variables restantes con alguna asignación en particular. Entonces, la comparación incluirá una aserción $I(X_i, X_j \mid x_Z)$, para cada $i \neq j \in V$ y cada $x_Z \in \mathcal{X}^{ij}$; formalmente,

$$\mathcal{D}_{\mathcal{P}}(F) \equiv \left\{ (X_i \not\perp\!\!\!\perp X_j \mid x_Z)_F \,\middle|\, i \neq j \in V,\ x_Z \in \mathcal{X}^{ij} \right\}. \tag{4.3}$$

Al igual que en el caso no dirigido, para justificar el uso de la versión por pares de los modelos para comparar cualquier par de modelos contextualizados, se debe demostrar que para cualquier par de modelos loglineales F y G, la igualdad de sus modelos de dependencias por pares $\mathcal{D}_{\mathcal{P}}(F)$ y $\mathcal{D}_{\mathcal{P}}(G)$ implica la igualdad

de sus modelos de dependencias generales $\mathcal{D}(F)$ y $\mathcal{D}(G)$. El Capítulo 5 está dedicado a esto, donde también son demostradas otras propiedades, que en conjunto demuestran que la suma de falsos positivos y falsos negativos sobre $\mathcal{D}_\mathcal{P}$ es una métrica.

Para enfrentar la segunda complejidad exponencial, el enfoque utiliza un conjunto reducido de aserciones de dependencia específicas del contexto mutuamente excluyentes, fijando los conjuntos U y W de la Ec. (4.1) de manera que solo incluya aserciones donde $W = \emptyset$ y $U = V \setminus \{i, j\}$; es decir, solo se utilizan condicionantes contextualizados. Esto mantiene la especificidad necesaria para codificar la estructura de independencias de un modelo loglineal arbitrario.

La tercera complejidad es significativamente más difícil de resolver, además de ser el aporte central de este manuscrito, por lo que es desarrollada en su propio apartado a continuación.

4.2. Enfoque para una comparación eficiente

Esta sección desarrolla el enfoque para comparar de manera eficiente los modelos de dependencias por pares de dos modelos loglineales F y G, de acuerdo a la Ec. (4.3).

Como una forma de simplificar esta presentación, comenzaremos por señalar que $\mathcal{D}_\mathcal{P}(F)$ (y, similarmente, $\mathcal{D}_\mathcal{P}(G)$), puede ser particionado en conjuntos de dependencias mutuamente excluyentes $\mathcal{D}^{ij}_\mathcal{P}(F)$, uno por cada par de variables $(X_i, X_j), con i \neq j \in V$, definido formalmente como

$$\mathcal{D}^{ij}_\mathcal{P}(F) \equiv \left\{ (X_i \not\perp\!\!\!\perp X_j \mid x_Z)_F \,\middle|\, x_Z \in \mathcal{X}^{ij} \right\},$$

y por tanto $\mathcal{D}_\mathcal{P}(F) = \bigcup_{i \neq j \in V} \mathcal{D}^{ij}_\mathcal{P}(F)$. Además, observando que, dados i y j, existe

exactamente una aserción por cada configuración (cada instancia de x_Z, o cada elemento de $\mathcal{X}^{ij}$), podemos reexpresar el modelo de dependencias de F para las variables X_i y X_j en términos del conjunto $\mathcal{X}^{ij}(F)$ definido como las configuraciones para las cuales las variables son dependientes de acuerdo a F, concretamente,

$$\mathcal{X}^{ij}(F) \equiv \left\{ x_Z \in \mathcal{X}^{ij} \,\middle|\, (X_i \not\perp\!\!\!\perp X_j \mid x_Z)_F \right\}. \tag{4.4}$$

Esta definición puede utilizarse para leer $I(X_i, X_j \mid x_Z)$ en F, dado que es fácil demostrar que $x_Z \in \mathcal{X}^{ij}(F)$ si y solo si $I(X_i, X_j \mid x_Z) \in \mathcal{D}_{\mathcal{P}}(F)$.

Para comparar F y G, comparamos el valor de dependencia o independencia para cada aserción $I(X_i, X_j \mid x_Z)$ en F y G, contando las coincidencias y diferencias dentro de una matriz de confusión, compuesta del número total de *verdaderos positivos*, *verdaderos negativos*, *falsos negativos* y *falsos positivos*, denotados como VP, VN, FN y FP, respectivamente. Para una aserción dada $I(X_i, X_j \mid x_Z)$, estas cantidades son medidas contando, para cada par $i \neq j \in V$, en cuántos conjuntos condicionantes por pares $x_Z \in \mathcal{X}^{ij}$ esas variables son dependientes tanto en F como en G, es decir, $x_Z \in \mathcal{X}^{ij}(F)$ y $x_Z \in \mathcal{X}^{ij}(G)$; en cuántos son independientes tanto en F como en G ($x_Z \notin \mathcal{X}^{ij}(F)$ y $x_Z \notin \mathcal{X}^{ij}(G)$); cuándo son dependientes en F e independientes en G ($x_Z \in \mathcal{X}^{ij}(F)$ pero $x_Z \notin \mathcal{X}^{ij}(G)$), y el caso opuesto (es decir, $x_Z \notin \mathcal{X}^{ij}(F)$ pero $x_Z \in \mathcal{X}^{ij}(G)$). Formalmente, esto es equivalente a

$$VP = \sum_{i \neq j \in V} VP_{ij}; \quad VP_{ij} = \left| \left\{ x_Z \in \mathcal{X}^{ij} \,\middle|\, x_Z \in \mathcal{X}^{ij}(F) \wedge x_Z \in \mathcal{X}^{ij}(G) \right\} \right|; \tag{4.5}$$

$$FN = \sum_{i \neq j \in V} FN_{ij}; \quad VP_{ij} = \left| \left\{ x_Z \in \mathcal{X}^{ij} \,\middle|\, x_Z \in \mathcal{X}^{ij}(F) \wedge x_Z \notin \mathcal{X}^{ij}(G) \right\} \right|; \tag{4.6}$$

$$FP = \sum_{i \neq j \in V} FP_{ij}; \quad VP_{ij} = \left| \left\{ x_Z \in \mathcal{X}^{ij} \,\middle|\, x_Z \notin \mathcal{X}^{ij}(F) \wedge x_Z \in \mathcal{X}^{ij}(G) \right\} \right|; \tag{4.7}$$

$$VN = \sum_{i \neq j \in V} VN_{ij}; \quad VP_{ij} = \left| \left\{ x_Z \in \mathcal{X}^{ij} \,\middle|\, x_Z \notin \mathcal{X}^{ij}(F) \wedge x_Z \notin \mathcal{X}^{ij}(G) \right\} \right|. \tag{4.8}$$

En base a equivalencias básicas de conjuntos, la conjunción en la definición de VP_{ij} lo hace equivalente a la intersección de dos conjuntos, uno por cada término en la conjunción, a saber,

$$VP_{ij} = \left| \left\{ x_Z \in \mathcal{X}^{ij} \,\middle|\, (X_i \not\perp\!\!\!\perp X_j \mid x_Z)_F \right\} \bigcap \left\{ x_Z \in \mathcal{X}^{ij} \,\middle|\, (X_i \not\perp\!\!\!\perp X_j \mid x_Z)_G \right\} \right|,$$

lo cual, por la Ec. (4.4) resulta ser

$$VP_{ij} = \left| \mathcal{X}^{ij}(F) \cap \mathcal{X}^{ij}(G) \right|. \tag{4.9}$$

Similarmente, por equivalencias de conjuntos, las conjunciones de la inclusión y exclusión de FN_{ij} y FP_{ij} pueden reexpresarse como la diferencia de dos conjuntos, para obtener

$$FN_{ij} = \left| \mathcal{X}^{ij}(F) \setminus \mathcal{X}^{ij}(G) \right|, \tag{4.10}$$

$$FP_{ij} = \left| \mathcal{X}^{ij}(G) \setminus \mathcal{X}^{ij}(F) \right|. \tag{4.11}$$

Finalmente, para simplificar la expresión para VN_{ij}, primero extraemos la negación para obtener $\neg(x_Z \in \mathcal{X}^{ij}(F) \vee x_Z \in \mathcal{X}^{ij}(G))$, y reescribimos la negación como un complemento de conjunto y la disyunción como una unión de conjuntos, lo que resulta en

$$VN_{ij} = \left| \overline{\mathcal{X}^{ij}(F) \cup \mathcal{X}^{ij}(G)} \right|. \tag{4.12}$$

Asumiendo que contamos con los tres valores anteriores, este último número se obtiene con una operación simple que involucra el producto de las cardinalidades de todas las variables del dominio exceptuando al par (X_i, X_j). Este producto

es equivalente al número de posibles asignaciones $\mathcal{X}^{ij}$. Sumando este valor para todo par de variables obtenemos el número total de comparaciones, y solo queda restar las otras tres cantidades:

$$VN = \left(\sum_{i \neq j \in V} \prod_{k \in V \setminus \{i,j\}} \left| val(X_k) \right| \right) - VP - FN - FP. \tag{4.13}$$

En el caso particular en que $|val(X_k)| = m$ para todo $k \in V$, el cálculo total de verdaderos negativos se simplifica a

$$VN = m^{|V|-2} \binom{|V|}{2} - VP - FN - FP. \tag{4.14}$$

Por ejemplo, en un dominio de 6 variables binarias, el total de comparaciones es $m^{|V|-2} \binom{|V|}{2} = 2^4 \times \frac{6 \times 5}{2} = 16 \times 15 = 240$.

Por otro lado, y a pesar de su simplicidad, las definiciones en las ecuaciones (4.9), (4.10) y (4.11) aún conllevan un cómputo exponencial para la comparación de $\mathcal{X}^{ij}(F)$ y $\mathcal{X}^{ij}(G)$, lo cual requiere recorrer cada uno de los elementos del conjunto exponencial $x_Z \in \mathcal{X}^{ij}$, para cada par (X_i, X_j).

Abordaremos este problema en dos pasos. En primer lugar consideraremos el escenario simple en el cual tanto F como G están compuestos por una única característica, denotadas estas como f y g respectivamente. Más adelante nos extenderemos al caso de múltiples características.

4.2.1. Caso de una única característica

Las Ecuaciones (4.9), (4.10), (4.11) y (4.12) muestran que para computar eficientemente VP_{ij}, FN_{ij}, FP_{ij} y VN_{ij} para el caso de una única característica, es suficiente crear un procedimiento eficiente para computar las cardinalidades de

$\mathcal{X}^{ij}(f) \cap \mathcal{X}^{ij}(g)$ y de $\mathcal{X}^{ij}(f) \setminus \mathcal{X}^{ij}(g)$. Esto se debe a que el procedimiento para computar la diferencia funciona para características arbitrarias, por lo que también puede obtener $\mathcal{X}^{ij}(g) \setminus \mathcal{X}^{ij}(f)$; mientras que VN_{ij} se calcula eficientemente sustrayendo las otras tres cantidades al número total de comparaciones. Estos procedimientos eficientes se encuentran descritos y justificados a continuación, en los Teoremas 2 y 3.

4.2.1.1. Cómputo de $|\mathcal{X}^{ij}(f) \cap \mathcal{X}^{ij}(g)|$

El Teorema 2 provee un cómputo eficiente de la intersección, efectuando el conteo sin la necesidad de pasar por el número exponencial de contextos por pares. En cambio, puede llegar a los mismos valores computando —en el peor caso— el producto de las cardinalidades de un subconjunto de variables en X_V. Esto se formaliza a continuación.

Teorema 2. *Sean f y g dos características arbitrarias en F y en G, respectivamente. Entonces, la cardinalidad de la intersección de sus dependencias por pares $\mathcal{X}^{ij}(f)$ y $\mathcal{X}^{ij}(g)$ puede computarse como*

$$\left|\mathcal{X}^{ij}(f) \cap \mathcal{X}^{ij}(g)\right| = \begin{cases} 0 & \text{si } C_1 \vee C_2 \\ \left|f^{ij} \cup g^{ij}\right| & \text{en otro caso,} \end{cases} \tag{4.15}$$

donde f^{ij} y g^{ij} son f y g eliminando las asignaciones a las variables X_i y X_j, y

$$\begin{aligned} C_1 &\equiv (X_i \notin S_f) \vee (X_i \notin S_g) \vee (X_j \notin S_f) \vee (X_j \notin S_g), \\ C_2 &\equiv \exists X_h \in S_f \cap S_g \setminus \{X_i, X_j\} \mid X_h(f) \neq X_h(g), \end{aligned}$$

mientras que la unión $f^{ij} \cup g^{ij}$ se basa en la Definición 10 en el Apéndice B. Intuitivamente, la unión de dos características f y g es otra característica cuyas asignaciones son

la unión de las asignaciones de f y de g, siempre y cuando no contengan una asignación incompatible (es decir, que para ninguna variable X_k ocurra que $X_k(f) \neq X_k(g)$).

La cardinalidad de la unión en la Ec. (4.15), $\left|f^{ij} \cup g^{ij}\right|$, se calcula eficientemente como

$$\left|f^{ij} \cup g^{ij}\right| = \begin{cases} 1 & si \ \left|S_{f\cup g}\right| = |X_V| \\ \prod_{X_k \in X_V \setminus (S_f \cup S_g)} \left|val(X_k)\right| & en \ otro \ caso, \end{cases} \tag{4.16}$$

La Ec. (4.15) indica que tanto X_i como X_j deben estar presentes en ambas características para que exista una intersección no vacía (C_1); además, si estas características tienen una asignación incompatible, la intersección también es vacía (C_2). En cualquier otro caso, la cardinalidad se calcula como en la Ec. (4.16). Si existen variables no asignadas, el resultado es el producto de las cardinalidades de las variables que no están asignadas en f ni en g. En caso contrario, la cardinalidad es 1 —si todas las variables resultan asignadas, la intersección es igual a un único contexto por pares.

Demostración. Nos basamos en la siguiente definición:

$$\mathcal{X}^{ij}(f) \cap \mathcal{X}^{ij}(g) = \begin{cases} \emptyset & si \ C_1 \vee C_2 \\ \mathcal{X}^{ij}(f^{ij} \cup g^{ij}) & en \ otro \ caso. \end{cases} \tag{4.17}$$

Este hecho es demostrado en el Lema 1, presentado inmediatamente después de este Teorema.

Según la igualdad de la Ec. (4.17), para $C_1 \vee C_2$ tenemos que $\mathcal{X}^{ij}(f) \cap \mathcal{X}^{ij}(g)$ es el conjunto vacío, cuya cardinalidad es 0. Por lo tanto, solo resta demostrar el caso de negación $\neg(C_1 \vee C_2)$, es decir, que la cardinalidad de la segunda condición es igual al lado derecho de la Ec. (4.16).

En primer lugar observemos que, de acuerdo a la Ec. (4.17), la intersección es $\mathcal{X}^{ij}(f^{ij} \cup g^{ij})$. Para demostrar cuál es su cardinalidad, primero encontremos una expresión equivalente para la cardinalidad del caso más sencillo $\mathcal{X}^{ij}(h)$, y luego apliquémosla al caso de $h \equiv f^{ij} \cup g^{ij}$.

Del Lema Auxiliar 1 (expuesto en el Apéndice C.1), un conjunto condicionante por pares $x_Z \in \mathcal{X}^{ij}$ también está en $\mathcal{X}^{ij}(h)$ siempre que $h \subseteq x_Z$ y tanto X_i como X_j estén en S_h. Es importante señalar que esto es así sin importar los valores que tomen las restantes variables, es decir, las variables en $X_V \setminus S_h$. En el caso particular en que $X_V \setminus \{S_{f^{ij} \cup g^{ij}} \cup \{X_i, X_j\}\} = \emptyset$, h tiene todas sus variables asignadas, y por lo tanto es equivalente a un contexto canónico por pares, lo que hace que su cardinalidad sea igual a 1. Este es el primer caso de la Ec. (4.16). En caso contrario —cuando existen una o más variables sin asignar—, por cada asignación de las variables restantes tendríamos un conjunto por pares en $\mathcal{X}^{ij}(h)$, por lo tanto

$$\left|\mathcal{X}^{ij}(h)\right| = \prod_{X_k \in X_V \setminus \{S_h \cup \{X_i, X_j\}\}} \left|val(X_k)\right|. \tag{4.18}$$

Ahora lo único que falta para obtener la cardinalidad de $\mathcal{X}^{ij}(f^{ij} \cup g^{ij})$ es aplicar la Ec. (4.18) a la unión de las características f^{ij} y g^{ij} para obtener

$$\begin{aligned}\left|\mathcal{X}^{ij}(f^{ij} \cup g^{ij})\right| &= \prod_{X_k \in X_V \setminus \left\{S_{f^{ij} \cup g^{ij}} \cup \{X_i, X_j\}\right\}} \left|val(X_k)\right|, \\ &= \prod_{X_k \in X_V \setminus (S_f \cup S_g)} \left|val(X_k)\right|,\end{aligned}$$

donde esta última expresión deriva de la definición de la unión de características, ya que $S_{f^{ij} \cup g^{ij}} = S_{f^{ij}} \cup S_{g^{ij}}$, y entonces $S_{f^{ij}} = S_f \setminus \{X_i, X_j\}$ (y lo mismo para g). Esto concluye la demostración para el caso de negación en la Ec. (4.15).

□

Lema 1. *Sean F y G dos modelos loglineales sobre X_V, $f \in F$ y $g \in G$ dos características arbitrarias en F y G, respectivamente, y sean $X_i \neq X_j$ dos variables cualesquiera en X_V. La intersección de sus conjuntos de dependencias por pares $\mathcal{X}^{ij}(f)$ y $\mathcal{X}^{ij}(g)$ es*

$$\mathcal{X}^{ij}(f) \cap \mathcal{X}^{ij}(g) = \begin{cases} \emptyset & \text{si } C_1 \vee C_2 \\ \mathcal{X}^{ij}(f^{ij} \cup g^{ij}) & \text{en otro caso,} \end{cases} \tag{4.19}$$

donde C_1 y C_2 son las mismas condiciones que en el Teorema 2 (Ec. (4.15)).

Demostración. Consideremos cada caso por separado.

Caso C_1: Si $X_i \notin S_f$ o $X_j \notin S_f$, entonces X_i y X_j no aparecen juntas en f, y por lo tanto la segunda condición en el lado derecho de la Ec. (C.1) en el Lema Auxiliar 1 (Apéndice C.1) es falsa. Esto implica que para cada x_Z, el lado izquierdo es falso, es decir, $x_Z \notin \mathcal{X}^{ij}(f)$, o lo que es equivalente, $\mathcal{X}^{ij}(f) = \emptyset$. En cambio, si $X_i \notin S_g$ o $X_j \notin S_g$, siguiendo el mismo razonamiento, $\mathcal{X}^{ij}(g) = \emptyset$. En ambos casos llegamos a la conclusión de que $\mathcal{X}^{ij}(f) \cap \mathcal{X}^{ij}(g) = \emptyset$.

Caso C_2: En este caso, no están vacíos $\mathcal{X}^{ij}(f)$ ni $\mathcal{X}^{ij}(g)$, pero sí su intersección. Para demostrarlo, argumentamos que dados dos contextos por pares cualesquiera $x_Z \in \mathcal{X}^{ij}(f)$ y $x'_Z \in \mathcal{X}^{ij}(g)$, estos difieren en la asignación de al menos una de sus variables. Dado que se cumple C_2, debe haber al menos una variable X_h distinta de X_i y de X_j que está tanto en f como en g, es decir, $X_h \in f^{ij}$ y $X_h \in g^{ij}$, tal que $X_h(f) \neq X_h(g)$. De acuerdo a la Ec. (C.1) del Lema Auxiliar 1 (Apéndice C.1), para cualquier $x_Z \in \mathcal{X}^{ij}(f)$ se debe cumplir que $X_h(x_Z) = X_h(f^{ij})$ y para cualquier $x'_Z \in \mathcal{X}^{ij}(g)$ se debe cumplir que $X_h(x'_Z) = X_h(g^{ij})$. Por lo tanto, no puede ocurrir que $x_Z = x'_Z$; en otras palabras, la intersección de los contextos por pares de f y g también es vacía para este caso.

Caso $\neg C_1 \wedge \neg C2$: Debemos demostrar que $\mathcal{X}^{ij}(f) \cap \mathcal{X}^{ij}(g) \equiv \mathcal{X}^{ij}(f^{ij} \cup g^{ij})$. Según la Ec. (4.4),

$$\mathcal{X}^{ij}(f) \cap \mathcal{X}^{ij}(g) \equiv \left\{ x_Z \in \mathcal{X}^{ij} \mid (X_i \not\perp\!\!\!\perp X_j \mid x_Z)_f \wedge (X_i \not\perp\!\!\!\perp X_j \mid x_Z)_g \right\}.$$

Aplicando el Lema Auxiliar 1 al lado derecho obtenemos

$$\mathcal{X}^{ij}(f) \cap \mathcal{X}^{ij}(g) \equiv \left\{ x_Z \in \mathcal{X}^{ij} \mid f^{ij} \subseteq x_Z \wedge g^{ij} \subseteq x_Z \right\},$$

donde el operador de subconjunto ($\subseteq$) para características sigue la Definición 11 del Apéndice B. Luego, por teoría de conjuntos sabemos que para conjuntos arbitrarios A, B y C, $A \subseteq C \wedge B \subseteq C$ es equivalente a $A \cup B \subseteq C$, y por lo tanto

$$\mathcal{X}^{ij}(f) \cap \mathcal{X}^{ij}(g) \equiv \left\{ x_Z \in \mathcal{X}^{ij} \mid f^{ij} \cup g^{ij} \subseteq x_Z \right\},$$

lo que demuestra el caso $\neg C_1 \wedge \neg C2$, concluyendo la demostración del lema. □

Por último, ilustraremos el teorema con un ejemplo sencillo.

Ejemplo 8. *Consideramos un dominio $X_V, V = \{0,1,2,3,4,5\}$, donde todas las variables son binarias. Sean f y g dos características definidas como*

$$f = < X_0 = 1, X_1 = 1, X_2 = 0, X_5 = 0 >,$$
$$g = < X_0 = 0, X_1 = 1, X_2 = 0, X_4 = 1 > .$$

Tomando como ejemplo el par (X_0, X_1), tenemos que no se cumplen C_1 ni C_2, y podemos definir las características por pares f^{01} y g^{01}:

$$f^{01} = < X_2 = 0, X_5 = 0 >,$$

$$g^{01} =< X_2 = 0, X_4 = 1 > .$$

La unión de f^{01} *y* g^{01} *es:*

$$f^{01} \cup g^{01} =< X_2 = 0, X_4 = 1, X_5 = 0 > .$$

Este resultado corresponde al siguiente conjunto de contextos por pares:

$$\begin{aligned} \mathcal{X}^{01}(f^{01} \cup g^{01}) = \quad & \{< X_2 = 0, X_3 = 0, X_4 = 1, X_5 = 0 >, \\ & < X_2 = 0, X_3 = 1, X_4 = 1, X_5 = 0 >\}. \end{aligned} \tag{4.20}$$

De hecho, podemos observar que $X_V \setminus \{S_f \cup S_g\} = X_V \setminus \{X_0, X_1, X_2, X_4, X_5\} = \{X_3\}$, *y* $\left|val(X_3)\right| = 2$, *lo que satisface la Ec.* (4.15).

Ahora lo compararemos con el resultado de hacer una intersección de manera exhaustiva entre todos los contextos representados por las dos características. Para visualizar mejor este procedimiento usaremos una representación tabular. El resultado se presenta en la Tabla 4.1. Puede observarse que los dos contextos mostrados en la última división de la tabla ($\mathcal{X}^{01}(f) \cap \mathcal{X}^{01}(g)$) *son los mismos que en* (4.20).

4.2.1.2. Cómputo de $|\, \mathcal{X}^{ij}(f) \setminus \mathcal{X}^{ij}(g) \,|$

Para obtener la diferencia entre dos características, se presenta a continuación el Teorema 3, que provee un procedimiento para el cómputo eficiente de las diferencias de las Ec. (4.10) y (4.11), sin la necesidad de contar a través del número exponencial de contextos por pares. En lugar de esto, obtiene los mismos resultados sumando las cardinalidades de los conjuntos de contextos por pares de características individuales, lo cual por la Ec. (4.18) requiere de un cómputo simple y eficiente.

	X_2	X_3	X_4	X_5
f^{01}	0			0
g^{01}	0		1	
$\mathcal{X}^{01}(f)$	0	0	0	0
	0	**0**	**1**	**0**
	0	1	0	0
	0	**1**	**1**	**0**
$\mathcal{X}^{01}(g)$	**0**	**0**	**1**	**0**
	0	**1**	**1**	**0**
	0	0	1	1
	0	1	1	1
$\mathcal{X}^{01}(f) \cap \mathcal{X}^{01}(g)$	0	0	1	0
	0	1	1	0

TABLA 4.1: Ejemplo de intersección exhaustiva de f^{01} y g^{01}. En cada columna se muestran los valores que toma cada variable para las características (dos primeras filas) y los conjuntos de contextos (filas restantes).

Teorema 3. *La cardinalidad de la diferencia de conjuntos por pares de dos características arbitrarias f y g puede computarse eficientemente como*

$$\left|\mathcal{X}^{ij}(f) \setminus \mathcal{X}^{ij}(g)\right| \equiv \begin{cases} 0 & \text{si } g^{ij} \subseteq f^{ij} \ \wedge \neg C_1(g), \text{ o si } \ C_1(f) \\ \left|\mathcal{X}^{ij}(f)\right| & C_2 \vee C_1(g) \\ \sum_{d \in D^E_{fg}} \left|\mathcal{X}^{ij}(d)\right| & \text{en otro caso,} \end{cases} \tag{4.21}$$

recordando que $g^{ij} \subseteq f^{ij}$ se obtiene a partir de la Definición 11 (Apéndice B), así como también que tanto $\left|\mathcal{X}^{ij}(f)\right|$ como $\left|\mathcal{X}^{ij}(d)\right|$ pueden ser calculadas eficientemente utilizando la Ec. (4.18) (producto de cardinalidades), y donde $C_1(f)$ y $C_1(g)$ son condiciones definidas por

$$\begin{aligned} C_1(f) &\equiv X_i \notin S_f \vee X_j \notin S_f \\ C_1(g) &\equiv X_i \notin S_g \vee X_j \notin S_g, \end{aligned}$$

C_2 es la misma condición definida para la Ec. (4.15),

$$D^E_{fg} = \bigcup_{X_k \in S_g \setminus S_f} D^E_{fg_{(k)}}, \tag{4.22}$$

con

$$D^E_{fg_{(k)}} = \left\{ caract.\ d \left| \begin{array}{l} S_d = S_f \cup S_g^{\leq k}; \\ \forall X_m \in S_{f^{ij}}, X_m(d) = X_m(f^{ij}); \\ \forall m < k, S_g^{\leq k} \setminus S_f, X_m(d) = X_m(g); \\ X_k(d) \neq X_k(g) \end{array} \right. \right\}, \tag{4.23}$$

y $S_g^{\leq k} = \left\{ X_m \in S_g \middle| m \leq k \right\}$.

En lo que sigue se presenta la demostración de este teorema, más simple para los dos primeros casos de la Ec. (4.21), y más compleja para el tercero, el cual requiere de los Lemas 2, 3 y 4. Al final de la sección, proveo un ejemplo por cada caso (Ejemplos 9, 10 y 11).

Demostración. Para el primer caso de la Ec. (4.21), si se cumple $C_1(f)$ entonces $\mathcal{X}^{ij}(f) = \emptyset$, y por lo tanto la diferencia es claramente vacía. En cambio, si $g^{ij} \subseteq f^{ij}$ y $\neg C_1(g)$, entonces la Ec. (C.1) nos dice que $\mathcal{X}^{ij}(f)$ contiene todos los x_Z tal que $f^{ij} \subseteq x_Z$ y que $\mathcal{X}^{ij}(g)$ contiene todos los x'_Z tal que $g^{ij} \subseteq x'_Z$. Luego, por $g^{ij} \subseteq f^{ij}$, todo x_Z en $\mathcal{X}^{ij}(g)$ también está en $\mathcal{X}^{ij}(f)$. Esto significa que todos los elementos en $\mathcal{X}^{ij}(f)$ serán sustraídos, resultando en un conjunto vacío. Para el segundo caso de la Ec. (4.21), si se cumple C_2, algún valor en g no concuerda con el valor correspondiente en f, y por lo tanto ningún elemento de $\mathcal{X}^{ij}(g)$ está en $\mathcal{X}^{ij}(f)$. En cambio, si se cumple $C_1(g)$, entonces $\mathcal{X}^{ij}(g) = \emptyset$. En ambos casos, ningún elemento puede ser sustraído de $\mathcal{X}^{ij}(f)$.

La prueba para el tercer caso de la Ec.(4.21) consiste en dos partes. Primero, demostraremos la igualdad de los conjuntos dentro de las cardinalidades, es decir,

$$\mathcal{X}^{ij}(f) \setminus \mathcal{X}^{ij}(g) = \bigcup_{d \in D_{fg}^{E}} \mathcal{X}^{ij}(d).$$

y luego demostraremos la igualdad de sus cardinalidades.

En primer lugar mostraremos que, por el Lema 2, esta igualdad se cumple para D_{fg}, una versión más simple de D_{fg}^{E}. Este conjunto tiene una definición más directa que D_{fg}^{E}, pero contiene más características, conformando un conjunto de contextos que no son mutuamente excluyentes. Luego, veremos que de acuerdo al Lema 3, D_{fg} y D_{fg}^{E} representan exactamente los mismos contextos por pares, es decir,

$$\bigcup_{d \in D_{fg}} \mathcal{X}^{ij}(d) = \bigcup_{d \in D_{fg}^{E}} \mathcal{X}^{ij}(d).$$

demostrando así la igualdad para D_{fg}^{E}. Por último, el Lema 4 muestra que los contextos por pares de cada característica en D_{fg}^{E} son mutuamente excluyentes, lo que garantiza que la cardinalidad de la unión sea igual a la suma de las cardinalidades individuales, es decir,

$$\left| \bigcup_{d \in D_{fg}^{E}} \mathcal{X}^{ij}(d) \right| = \sum_{d \in D_{fg}^{E}} \left| \mathcal{X}^{ij}(d) \right|,$$

con lo cual queda completo el tercer caso de la Ec. (4.21) y el teorema. □

Lema 2.

$$\mathcal{X}^{ij}(f) \setminus \mathcal{X}^{ij}(g) = \bigcup_{d \in D_{fg}} \mathcal{X}^{ij}(d). \tag{4.24}$$

donde

$$D_{fg} = \left\{ caract.\ d \;\middle|\; \begin{array}{l} S_d = S_f \cup S_g; \forall X_m \in S_{f^{ij}}, X_m(d) = X_m(f^{ij}); \\ \exists X_k \in S_g \setminus S_f, X_k(d) \neq X_k(g) \end{array} \right\}. \tag{4.25}$$

Demostración. Por equivalencia de conjuntos, la Ec. (4.24) puede reformularse como

$$\forall x_Z \in \mathcal{X}^{ij}, x_Z \in \mathcal{X}^{ij}(f) \wedge x_Z \notin \mathcal{X}^{ij}(g) \iff \exists d \in D_{fg}, x_Z \in \mathcal{X}^{ij}(d).$$

Como tenemos $\neg C_1(f)$ y $\neg C_1(g)$, tanto X_i como X_j están en f y g, por lo tanto, por la Ec. (C.1) la expresión de arriba es equivalente a

$$f^{ij} \subseteq x_Z \wedge g^{ij} \not\subseteq x_Z \iff \exists d \in D_{fg}, d^{ij} \subseteq x_Z. \tag{4.26}$$

para un $x_Z \in \mathcal{X}^{ij}$ arbitrario.

Comenzamos con la implicación de derecha a izquierda en la Ec. (4.26), y consideramos cada término en el lado izquierdo por separado, es decir,

$$\begin{aligned} I_1: &\quad \exists d \in D_{fg}, d^{ij} \subseteq x_Z \implies f^{ij} \subseteq x_Z \\ I_2: &\quad \exists d \in D_{fg}, d^{ij} \subseteq x_Z \implies g^{ij} \not\subseteq x_Z \end{aligned}$$

Caso I_1: Debemos demostrar $f^{ij} \subseteq x_Z$. Para esto aplicamos el Lema Auxiliar 3 (Apéndice C.3) para $a = f^{ij}$ y $b = d^{ij}$, para lo cual es suficiente demostrar que $\exists d \in D_{fg}$ tal que $f^{ij} \subseteq d^{ij}$ (esto es, $a \subseteq b$). Por definición, $S_d = S_f \cup S_g$, así que $S_f \subseteq S_d$. Además, por definición de cada $d \in D_{fg}$, $\forall X_m \in S_{f^{ij}}, X_m(d) = X_m(f^{ij})$, lo que en combinación con $S_f \subseteq S_d$ implica $f^{ij} \subseteq d^{ij}$. □

Caso I_2: Primero observamos que

$$d^{ij} \subseteq x_Z \implies \forall X_m \in S_{d^{ij}}, X_m(d^{ij}) = X_m(x_Z). \tag{4.27}$$

Luego, a partir de la definición de cada $d \in D_{fg}$, $\exists X_m \in S_g \setminus S_f$ tal que $X_m(d) \neq X_m(g)$. Además, dado que X_i y X_j están en S_f, entonces $S_g \setminus S_f = S_{g^{ij}} \setminus S_{f^{ij}}$, por lo que $\exists X_m \in S_{g^{ij}} \setminus S_{f^{ij}}$ tal que $X_m(d^{ij}) \neq X_m(g^{ij})$. En combinación con la Ec. (4.27), esto resulta en que $\exists X_m \in S_{g^{ij}} \setminus S_{f^{ij}}$ tal que $X_m(g^{ij}) \neq X_m(x_Z)$, desde lo que podemos concluir que $g^{ij} \not\subseteq x_Z$.

□

A continuación, procedemos con la implicación de izquierda a derecha de la Ec. (4.26). Demostraremos que dada la condición en el lado izquierdo para un x_Z arbitrario, existe algún $d \in D_{fg}$ para el cual se satisface el lado derecho, es decir, $d^{ij} \subseteq x_Z$.

Comenzamos por reinterpretar el lado izquierdo:

$(f^{ij} \subseteq x_Z)$:

$$(f^{ij} \subseteq x_Z) : \forall X_m \in S_{f^{ij}}, X_m(f^{ij}) = X_m(x_Z). \tag{4.28}$$

$(g^{ij} \not\subseteq x_Z)$:

Dado que tenemos $S_{x_Z} = X_{V \setminus \{i,j\}}$ y $S_g^{ij} \subseteq S_{x_Z}$, entonces $g^{ij} \not\subseteq x_Z$ solo puede ocurrir debido a que alguna asignación en g^{ij} toma un valor que es diferente en x_Z, es decir,

$$\exists X_m \in S_{g^{ij}} \text{ tal que } X_m(g^{ij}) \neq X_m(x_Z). \tag{4.29}$$

Sin embargo, por la Ec. (4.28) y C_2, este no puede ser el caso para $X_m \in S_{g^{ij}} \cap S_{f^{ij}}$ y por lo tanto la Ec. (4.29) puede ser reexpresada como

$$\exists X_m \in S_{g^{ij}} \setminus S_{f^{ij}} \text{ tal que } X_m(g^{ij}) \neq X_m(x_Z). \tag{4.30}$$

Finalmente, debido a que no se cumple $C_1(f)$, X_i y X_j están en S_f, por lo que $S_{g^{ij}} \setminus S_{f^{ij}} = Sg \setminus S_f$, es decir, la Ec. (4.30) se convierte en

$$\exists X_m \in S_g \setminus S_f, X_m(g^{ij}) \neq X_m(x_Z). \tag{4.31}$$

Sea M el conjunto de índices que satisfacen la Ec. (4.31), es decir,

$$\forall m \in M, X_m(g^{ij}) \neq X_m(x_Z). \tag{4.32}$$

Procedemos proponiendo una característica d definida sobre $S_f \cup S_g$ que satisface $d^{ij} \subseteq x_Z$, es decir,

$$\forall X_m \in S_{d^{ij}} = S_f \cup S_g, X_m(d) = X_m(x_Z), \tag{4.33}$$

y demostramos que $d \in D_{fg}$. Para esto, demostramos que d satisface las tres condiciones en la definición de $d \in D_{fg}$:

1. La primera condición, $S_d = S_f \cup S_g$ se satisface por la definición de d.

2. Por las Ecuaciones (4.28) y (4.33), y por el hecho de que $S_{f^{ij}} \subseteq S_d$ (por la definición de d), tenemos que $\forall X_m \in S_{f^{ij}}, X_m(f^{ij}) = X_m(x_Z) = X_m(d)$, satisfaciendo la segunda condición de D_{fg} para $X_m(f^{ij})$ y $X_m(d)$.

3. Por las Ecuaciones (4.32) y (4.33), y por el hecho de que $X_M \subseteq S_g \setminus S_f$ (por la definición de M), y $S_g \setminus S_f \subseteq S_d$ (por la definición de d), y en consecuencia $M \subseteq S_d$, tenemos que

$$\forall m \in M, X_m(g^{ij}) \neq X_m(x_Z) \wedge X_m(x_Z) = X_m(d),$$

entonces $\forall m \in M, X_m(d) \neq X_m(g^{ij})$, satisfaciendo la tercera condición de D_{fg}.

□

Lema 3.

$$\bigcup_{d \in D_{fg}} \mathcal{X}^{ij}(d) = \bigcup_{d \in D^E_{fg}} \mathcal{X}^{ij}(d). \tag{4.34}$$

para D^E_{fg} definido por la Ec.((4.22)) y D_{fg} definido por la Ec. (4.25).

Demostración. Para un x_Z arbitrario, la Ec. (4.34) es equivalente a

$$\exists d \in D_{fg}, x_Z \in \mathcal{X}^{ij}(d) \iff \exists d' \in D^E_{fg}, x_Z \in \mathcal{X}^{ij}(d').$$

Dado que $\neg C_1(f)$, $S_f \subseteq S_d$ y $S_f \subseteq S_{d'}$, tenemos que X_i y X_j están tanto en S_d como en $S_{d'}$, y por lo tanto se cumple que $\neg C_1(d)$ y $\neg C_1(d')$. Podemos entonces aplicar la Ec. (C.1) arriba para obtener

$$\exists d \in D_{fg}, d \subseteq x_Z \iff d \in D^E_{fg}, d \subseteq x_Z. \tag{4.35}$$

Para la implicación de izquierda a derecha, por el Lema Auxiliar 1 en el Apéndice C.1, es suficiente demostrar que

$$\exists d \in D_{fg} \wedge \exists d' \in D^E_{fg} \text{ tal que } d' \subseteq d, \tag{4.36}$$

para $a = d', b = d$. Cualquier $d' \in D^E_{fg}$ y $d \in D_{fg}$ satisfacen $S_f \subseteq S_{d'}$ y $S_f \subseteq S_d$, respectivamente, y concuerdan con los valores de f^{ij}; por lo tanto, para demostrar $d' \subseteq d$, podemos enfocarnos en los valores para g. Para esto, comenzamos por señalar que cada característica $d' \in D^E_{fg}$ se define sobre algún subconjunto de $S_g \setminus S_f$

(dependiente de X_k), sobre el cual se garantiza que tendrá una asignación diferente a la de g en X_k, es decir, $X_k(d') \neq X_k(g)$. Luego, cualquier d que satisfaga estas asignaciones para estas variables en $S_{d'} \setminus S_f = S_d^{\leq k} \setminus S_f$, y cualquier valor para las asignaciones restantes en $S_g \setminus S_f$, satisfarían que $\exists X_m \in S_g \setminus S_f, X_m(d) \neq X_m(g)$, y por lo tanto D_{fg}.

Para la implicación de derecha a izquierda de la Ec. (4.35), tenemos que, por la Ec. (4.36), para cada $d' \in D_{fg}^E$ existe un $d \in D_{fg}$ tal que $d' \subseteq d$. Es suficiente entonces con completar d' con asignaciones para las restantes variables con valores iguales a los de x_Z:

$$\forall X_k \in S_g^{>k} \setminus S_f, X_k(d) = X_k(x_Z). \tag{4.37}$$

Entonces, por la Ec. (4.37) y por el hecho de que $d' \subseteq x_Z$, concluimos que $d \subseteq x_Z$.

□

Lema 4. *Los contextos por pares de cada característica en D_{fg}^E son mutuamente excluyentes:*

$$\forall d, d' \in D_{fg}^E, \mathcal{X}^{ij}(d) \cap \mathcal{X}^{ij}(d') = \emptyset$$

Demostración. Dada la definición de D_{fg}^E en la Ec. (4.23), todas las características de orden k en algún $D_{fg(k)}^E$ tienen valores diferentes en X_k, por lo cual no pueden tener contextos por pares en común.

Adicionalmente, para otro k' tal que $k' > k$, no solo son sus contextos por pares mutuamente excluyentes entre sí, si no que además son diferentes de todas las características para k, debido a que estas toman valores diferentes a $X_k(g)$ en X_k, mientras que las características en k' toman el valor $X_k(g)$ en X_k.

□

	X_2	X_3
f^{01}	0	0
g^{01}	0	
$\mathcal{X}^{01}(f)$	**0**	**0**
$\mathcal{X}^{01}(g)$	**0**	**0**
	0	1

TABLA 4.2: Ejemplo de diferencia para el primer caso de la Ec. (4.21). Al quitar $< X_2 = 0, X_3 = 0 >$ de $\mathcal{X}^{01}(f)$ obtenemos el conjunto vacío.

Ejemplo 9. *En el caso en que se cumple $C_1(f)$, $\mathcal{X}^{ij}(f)$ no tiene ningún elemento por definición; por lo tanto, solo resulta de interés ejemplificar el caso alternativo, $g^{ij} \subseteq f^{ij} \wedge \neg C_1(g)$.*

Sea X_V un dominio de variables binarias con $V = \{0,1,2,3\}$, y sean f y g dos características definidas como

$$f =< X_0 = 1, X_1 = 1, X_2 = 0, X_3 = 0 >,$$
$$g =< X_0 = 0, X_1 = 1, X_2 = 0 > .$$

Tomando nuevamente el par (X_0, X_1), vemos que se cumple $\neg C_1(g)$ porque $X_0 \in S_g \wedge X_1 \in S_g$, y tenemos que $g^{01} \subset f^{01}$ (porque $X_2(g) = X_2(f) = 0$ y X_3 no está asignada en g). Las características por pares son:

$$f^{01} =< X_2 = 0, X_3 = 0 >,$$
$$g^{01} =< X_2 = 0 > .$$

La Tabla 4.2 muestra los contextos representados por las dos características. En este caso observamos que $\mathcal{X}^{01}(f) \subset \mathcal{X}^{01}(g)$, por lo tanto $\mathcal{X}^{01}(f) \setminus \mathcal{X}^{01}(g) = \emptyset$ y su cardinalidad es 0.

Ejemplo 10. *Cuando es cierto $C_1(g)$, no hay elementos en $\mathcal{X}^{ij}(g)$ y por lo tanto el resultado de la Ec. (4.21) es trivialmente $|\mathcal{X}^{ij}(f)|$. Ilustraremos, entonces, el otro caso*

	X_2	X_3
f^{01}	0	0
g^{01}	1	
$\mathcal{X}^{01}(f)$	**0**	**0**
$\mathcal{X}^{01}(g)$	1	0
	1	1

TABLA 4.3: Ejemplo de diferencia para el segundo caso de la Ec. (4.21). $\mathcal{X}^{01}(g)$ no tiene elementos que puedan sustraerse de $\mathcal{X}^{01}(f)$; el resultado es $\mathcal{X}^{01}(f)$, que contiene un único contexto por pares.

(es decir, cuando se cumple C_2).

Para el mismo dominio que en el Ejemplo (9), supongamos que

$$f =< X_0 = 1, X_1 = 1, X_2 = 0, X_3 = 0 >,$$
$$g =< X_0 = 0, X_1 = 1, X_2 = 1 >,$$

y consideremos

$$f^{01} =< X_2 = 0, X_3 = 0 >,$$
$$g^{01} =< X_2 = 1 > .$$

En la Tabla 4.3 aparecen sus contextos por pares. Allí veremos que no hay elementos en $\mathcal{X}^{01}(g)$ que también estén en $\mathcal{X}^{01}(f)$, por lo tanto la resta nos devuelve $\mathcal{X}^{01}(f) = \{< X_2 = 0, X_3 = 0 >\}$.

Ejemplo 11. *En el tercer caso de la Ec. (4.21) no se cumplen $g^{ij} \subseteq f^{ij}$, $C_1(f)$, $C_1(g)$ ni C_2. Para proporcionar un ejemplo que cumpla estas condiciones consideremos esta vez variables un dominio X_V con $V = \{0,1,2,3,4\}$ y $\forall X_k \in X_V, val(X_k) = \{0,1,2\}$. El par considerado será (X_0, X_1) como en los ejemplos anteriores. Sean*

$$f =< X_0 = 2, X_1 = 1, X_3 = 0 >, y$$

$$g =< X_0 = 0, X_1 = 1, X_2 = 0, X_3 = 0, X_4 = 1 >,$$

tenemos que $S_g \setminus S_f = \{X_2, X_4\}$*; por lo tanto, por la Ec.* (4.22),

$$D^E_{fg} = D^E_{fg_{(2)}} \cup D^E_{fg_{(4)}}.$$

Construyamos a continuación cada uno de estos dos conjuntos. A partir de la definición en la Ec. (4.23)*, para todo* $d \in D^E_{fg_{(2)}}$*:*

1. $S_d = S_{f^{01}} \cup S^{\leq 2}_{g^{01}} = \{X_2, X_3\}$ *(toda* d *tiene asignadas las variables* X_2 *y* X_3*);*
2. $S_{f^{01}} = \{X_3\} \implies X_3(d) = X_3(f^{01}) = 0$ *(en toda* d*,* X_3 *toma el valor* 0*);*
3. $S^{\leq 2}_{g^{01}} \setminus S_f = \emptyset$*;*
4. $X_2(d) \neq X_2(g) \implies X_2(d) = 1 \vee X_2(d) = 2.$

Esto produce dos características, debido a que la última condición (ítem 4) permite que X_2 *tome los valores 1 o 2 (porque* $val(X_2) = \{0, 1, 2\}$ *y* $X_2(g) = 0$*). Por lo tanto obtenemos:*

$$D^E_{fg_{(2)}} = \{< X_2 = 1, X_3 = 0 >, < X_2 = 2, X_3 = 0 >\}.$$

Ahora pasaremos a construir $D^E_{fg_{(4)}}$ *donde para cada característica* d*,*

1. $S_d = S_{f^{01}} \cup S^{\leq 4}_{g^{01}} = \{X_2, X_3, X_4\}$ *(toda* d *tiene asignadas estas tres variables);*
2. $S_{f^{01}} = \{X_3\} \implies X_3(d) = X_3(f^{01}) = 0$ *(toda* d *tiene la asignación* $X_3(d) = 0$*);*
3. $\forall m < k, X_k \in S^{\leq 4}_{g^{01}} \setminus S_f = \{X_2\}, X_2(d) = X_2(g) = 0$ *(toda* d *tiene la asignación* $X_2(d) = 0$*);*
4. $X_4(d) \neq X_4(g) \implies X_4(d) = 0 \vee X_4(d) = 2.$

Esto resulta en

$$D^E_{fg_{(4)}} = \{< X_2 = 0, X_3 = 0, X_4 = 0 >, < X_2 = 0, X_3 = 0, X_4 = 2 >\}.$$

Finalmente tenemos:

$$\begin{aligned} D^E_{fg} &= D^E_{fg_{(2)}} \cup D^E_{fg_{(4)}} \\ &= \{< X_2 = 1, X_3 = 0 >, \\ &\quad < X_2 = 2, X_3 = 0 >, \\ &\quad < X_2 = 0, X_3 = 0, X_4 = 0 >, \\ &\quad < X_2 = 0, X_3 = 0, X_4 = 2 >\}. \end{aligned}$$

Observando las diferencias en los valores de X_2 y X_4 podemos concluir rápidamente que las características en D^E_{fg} son mutuamente excluyentes.

Por un lado, si aplicamos la definición como en la Ec. (4.21), y calculamos las cardinalidades de cada característica en D^E_{fg} según la Ec. (4.18), obtendremos:

$$\left|\mathcal{X}^{01}(f) \setminus \mathcal{X}^{01}(g)\right| = \left|val(X_4)\right| + \left|val(X_4)\right| + 1 + 1 = 8$$

Por otro lado, podemos comparar el conjunto de contextos por pares $\mathcal{X}^{01}(D^E_{fg})$ con la enumeración exhaustiva en la Tabla 4.4, para verificar que la construcción es correcta, y que, efectivamente, el resultado es 8, dado que solo se sustrae uno de los 9 contextos por pares de $\mathcal{X}^{01}(f)$.

4.2.2. Caso de múltiples características

Procedemos a continuación con la discusión del caso más general de comparación de estructuras de modelos loglineales F y G con más de una característica,

	X_2	X_3	X_4
f^{01}		0	
g^{01}	0	0	1
$\mathcal{X}^{01}(f)$	0	0	0
	0	**0**	**1**
	0	0	2
	1	0	0
	1	0	1
	1	0	2
	2	0	0
	2	0	1
	2	0	2
$\mathcal{X}^{01}(g)$	**0**	**0**	**1**

TABLA 4.4: Ejemplo de diferencia para el tercer caso de la Ec. (4.21). $\mathcal{X}^{01}(f)$ tiene nueve elementos. $\mathcal{X}^{01}(g)$ tiene un solo elemento, el cual está en $\mathcal{X}^{01}(f)$. El resultado son todos los elementos de $\mathcal{X}^{01}(f)$ excepto el que está en negrita ($< X_2 = 0, X_3 = 0, X_4 = 1 >$).

para lo cual debemos elaborar un cómputo eficiente de

$$\left|\mathcal{X}^{ij}(F) \cap \mathcal{X}^{ij}(G)\right|, y$$
$$\left|\mathcal{X}^{ij}(F) \setminus \mathcal{X}^{ij}(G)\right|,$$

formalizados en los Teoremas 4 y 5 más abajo, para los casos de intersección y diferencia, respectivamente. Ambos teoremas utilizan el resultado del Lema Auxiliar 2, presentado en el Apéndice C.2, el cual afirma que para cualquier modelo loglineal F, $\mathcal{X}^{ij}(F) \equiv \cup_{f \in F} \mathcal{X}^{ij}(f)$. En ambos casos, además, se utiliza un procedimiento para convertir un conjunto de características en otro conjunto que cumpla las propiedades de ser una partición de contextos. Este resultado se presenta en el Algoritmo 1, en la Sección 4.2.2.1.

Teorema 4. *La cardinalidad de la intersección de conjuntos por pares de dos modelos loglineales arbitrarios F y G sobre X_V puede ser calculada eficientemente en términos*

de $|V|$ de la siguiente manera:

$$\left|\mathcal{X}^{ij}(F)\cap\mathcal{X}^{ij}(G)\right| \equiv \sum_{p\in P}\left|\mathcal{X}^{ij}(p)\right|, \tag{4.38}$$

donde P es el conjunto de características producido por el Algoritmo 1 provisto del conjunto de características $H = \left\{\mathcal{X}^{ij}(g)\cap\mathcal{X}^{ij}(f)\right\}_{f\in F, g\in G}$ como entrada, con las características equivalentes a $\mathcal{X}^{ij}(g)\cap\mathcal{X}^{ij}(f)$ como aquellas computadas por la Ec. (4.19).

Demostración. Comenzamos demostrando que

$$\mathcal{X}^{ij}(F)\cap\mathcal{X}^{ij}(G) \equiv \bigcup_{f\in F}\bigcup_{g\in G}\mathcal{X}^{ij}(g)\cap\mathcal{X}^{ij}(f). \tag{4.39}$$

La demostración usa el resultado del Lema Auxiliar 2, presentado y demostrado en el Apéndice C.2,

$$\mathcal{X}^{ij}(F)\cap\mathcal{X}^{ij}(G) = \mathcal{X}^{ij}(F)\cap\underbrace{\left(\bigcup_{g\in G}\mathcal{X}^{ij}(g)\right)}_{\text{por Lema Auxiliar 2 sobre } G}$$

$$= \underbrace{\bigcup_{g\in G}\mathcal{X}^{ij}(g)\cap\mathcal{X}^{ij}(F)}_{\text{por distribución de intersección sobre unión}}$$

$$= \bigcup_{g\in G}\left(\mathcal{X}^{ij}(g)\cap\underbrace{\bigcup_{f\in F}\mathcal{X}^{ij}(f)}_{\text{por Lema Auxiliar 2 sobre } F}\right)$$

$$= \underbrace{\bigcup_{g\in G}\bigcup_{f\in F}\mathcal{X}^{ij}(g)\cap\mathcal{X}^{ij}(f).}_{\text{por distribución de intersección sobre unión.}}$$

No obstante, esto es insuficiente, porque la cardinalidad no puede ser introducida

dentro de las uniones del lado derecho, puesto que podrían existir superposiciones entre las intersecciones de características individuales dando como resultado un conteo doble de algunos contextos por pares. Esto justifica emplear el Algoritmo 1 (introducido al final de esta sección) con H como entrada, conteniendo todos los conjuntos intersección dentro de la doble unión del lado derecho, ya que garantiza producir como salida un conjunto de características P cuyo conjunto de contextos por pares $\mathcal{X}^{ij}(P)$ es una partición del conjunto de contextos de las características de entrada H. Simbólicamente,

$$\bigcup_{h\in H}\mathcal{X}^{ij}(h)=\bigcup_{p\in P}\mathcal{X}^{ij}(p)\quad y\quad \forall p,p'\in P,\mathcal{X}^{ij}(p)\cap\mathcal{X}^{ij}(p')=\emptyset.$$

□

Teorema 5. *La cardinalidad de la diferencia de conjuntos por pares de dos modelos loglineales arbitrarios F y G puede ser calculada eficientemente en términos de $|V|$ de la siguiente manera:*

$$\left|\mathcal{X}^{ij}(F)\setminus\mathcal{X}^{ij}(G)\right|\equiv\sum_{p\in P}\left|\mathcal{X}^{ij}(p)\right| \tag{4.40}$$

donde P es el conjunto de características producido por el Algoritmo 1 provisto como entrada del conjunto de características $\cup_{d\in\mathbf{d}}d^{ij}$, una por cada $\mathbf{d}\in\bigtimes_{g\in G}D^E_{fg}$ [1], y para toda $f\in F$.

[1] Si $g_k,g_{k'}\in G$ son características distintas de G, el producto cartesiano $\bigtimes_{g\in G}D^E_{fg}$ es el resultado de generar todos los pares $(d^l_{g_k},d^n_{g_{k'}})$ para todo $d^l_{g_k}\in D^E_{fg_k}$ y $d^n_{g'_k}\in D^E_{fg_{k'}}$. Por ejemplo, si $G=\{g_1,g_2\}$, $D^E_{fg_1}=\left\{d^1_{g_1},d^2_{g_1}\right\}$ y $D^E_{fg_2}=\left\{d^1_{g_2},d^2_{g_2}\right\}$, entonces

$$\bigtimes_{g\in G}D^E_{fg}=\left\{(d^1_{g_1},d^1_{g_2}),(d^1_{g_1},d^2_{g_2}),(d^2_{g_1},d^1_{g_2}),(d^2_{g_1},d^2_{g_2})\right\}.$$

Demostración. Dado que el Algoritmo 1 produce una partición, es suficiente demostrar que

$$\mathcal{X}^{ij}(F) \setminus \mathcal{X}^{ij}(G) \equiv \bigcup_{f \in F} \bigcup_{\mathbf{d} \in \bigtimes_{g \in G} D^E_{fg}} \mathcal{X}^{ij}(\cup_{d \in \mathbf{d}} d^{ij}), \tag{4.41}$$

señalando que la cardinalidad puede ser introducida dentro de las uniones del lado derecho luego de particionar el conjunto de todas las características $\cup_{d \in \mathbf{d}} d^{ij}$.

$$\begin{aligned}
\mathcal{X}^{ij}(F) \setminus \mathcal{X}^{ij}(G) &= \underbrace{\mathcal{X}^{ij}(F) \cap \overline{\mathcal{X}^{ij}(G)}}_{\text{por equivalencia de conjuntos}} \\
&= \overline{\mathcal{X}^{ij}(G)} \cap \underbrace{\bigcup_{f \in F} \mathcal{X}^{ij}(f)}_{\text{por Lema Auxiliar 2 sobre } F} \\
&= \underbrace{\bigcup_{f \in F} \mathcal{X}^{ij}(f) \cap \overline{\mathcal{X}^{ij}(G)}}_{\text{por distribución de intersección sobre unión}} \\
&= \bigcup_{f \in F} \underbrace{\mathcal{X}^{ij}(f) \setminus \mathcal{X}^{ij}(G)}_{\text{por equivalencia de conjuntos}} \\
&= \bigcup_{f \in F} \mathcal{X}^{ij}(f) \setminus \underbrace{\bigcup_{g \in G} \mathcal{X}^{ij}(g)}_{\text{por Lema Auxiliar 2 sobre } G} \\
&= \bigcup_{f \in F} \underbrace{\bigcap_{g \in G} \mathcal{X}^{ij}(f) \setminus \mathcal{X}^{ij}(g)}_{\text{por propiedad de } \textit{complementos relativos}} \\
&= \bigcup_{f \in F} \bigcap_{g \in G} \underbrace{\bigcup_{d \in D^E_{fg}} \mathcal{X}^{ij}(d)}_{\text{por Ec. (4.21)}}
\end{aligned}$$

$$= \bigcup_{f \in F} \underbrace{\bigcup_{\mathbf{d} \in \bigtimes_{g \in G} D^E_{fg}} \bigcap_{d \in \mathbf{d}} \mathcal{X}^{ij}(d)}_{\text{por distribución de intersección sobre unión}}$$

$$= \bigcup_{f \in F} \bigcup_{\mathbf{d} \in \bigtimes_{g \in G} D^E_{fg}} \underbrace{\mathcal{X}^{ij}(\cup_{d \in \mathbf{d}} d^{ij})}_{\text{por Ec. (4.19)}},$$

donde aplicamos equivalencias generales de conjuntos: sustraer un conjunto es equivalente a intersectar su complemento (en otras palabras, $A \setminus B = A \cap \overline{B}$); la distribución de la intersección sobre la unión (por ejemplo, $(A \cup B) \cap (C \cup D) = (A \cap C) \cup (A \cap D) \cup (B \cap C) \cup (B \cap D)$); y en el último paso una propiedad de conjuntos conocida como *complementos relativos*, que afirma que para tres conjuntos A, B, y C, $C \setminus (A \cup B) = (C \setminus A) \cap (C \setminus B)$. □

4.2.2.1. Algoritmo de particionamiento de características

Para concluir, describamos el procedimiento para convertir un conjunto de características en una partición, desarrollado en el Algoritmo 1. En este trabajo, entendemos una partición de características como una partición de conjuntos de contextos por pares, ya que podemos interpretar cada característica como la expresión resumida de un conjunto de estos contextos. La idea general es que, a partir de los contextos por pares de cada característica, necesitamos sustraer aquellos que están duplicados en cualquier otra característica. La partición comienza con la primera característica (primer resultado de las líneas 1, 3 y 11) y a continuación la característica que le sigue es sustraída de ella (por medio de la operación de diferencia de características definida en la Sección 4.2.1.2 en la línea 7. Las características resultantes son almacenadas en D_{hp}. Estas corresponden al conjunto de diferencias entre el h actual y cada $p \in P$. Fuera del bucle, el resultado de sustraer de h todas las características que están actualmente en la partición P es

Algoritmo 1: *partición*(H).

Entrada: conjunto de características H.
Salida: una partición para los contextos representados por H, es decir, un conjunto de características tal que cada contexto representado en H esté representado exactamente por una característica en la partición.

```
1  P ⟵ ∅;
2  para h ∈ H hacer
3  |   D_h ⟵ {h};
4  |   para p ∈ P hacer
5  |   |   D_hp ⟵ ∅;
6  |   |   para h' ∈ D_h hacer
7  |   |   |   D_hp ⟵ D_hp ∪ D_h'\p ;
8  |   |   fin
9  |   |   D_h ⟵ D_hp;
10 |   fin
11 |   P ⟵ P ∪ D_h;
12 fin
13 devolver P;
```

asignado a D_h en la línea 9. Este bucle y asignación (líneas 6-8, 9) son necesarios ya que el resultado de una diferencia es un conjunto de características (no una única característica) y cada $p \in P$ debe ser sustraído de cada una de ellas. Luego, esta diferencia es añadida a la partición en la línea 11. Antes de este punto, D_h no tiene ningún contexto por pares en común con P; por lo tanto, su unión sigue siendo una partición. La operación de diferencia garantiza que cada característica h de la entrada es convertida en un conjunto de características h' que representa todos los contextos por pares que son representados por h pero no aquellos que son representados por algún $p \in P$. Esto garantiza que todas las características en la salida P sean mutuamente excluyentes y que $\bigcup_{p \in P} \mathcal{X}^{ij}(p) = \mathcal{X}^{ij}(H)$, obteniendo por tanto una partición.

4.3. Resumen general del método

La siguiente es una guía que resume la aplicación de la medida de comparación de estructuras paso a paso. Además, al final de esta sección hay un ejemplo de aplicación del cálculo de la medida a un caso concreto.

En los casos de verdaderos positivos (VP), falsos negativos (FN) y falsos positivos (FP), el procedimiento algorítmico en general tiene ciertos elementos en común. En primer lugar, el cómputo de cardinalidades se descompone por pares de variables y los resultados de cada par se suman para producir el valor total. En segundo lugar, para cada par de variables, las operaciones internas se realizan para cada posible par de características donde una pertenece a F y la otra a G. Por último, el resultado de esta operatoria interna es ingresado como entrada del Algoritmo 1, produciendo siempre una partición P que finalmente permite el cómputo de la cardinalidad como una suma de las cardinalidades de las características $p \in P$, las cuales se calculan de manera eficiente usando la Ec. (4.18).

Las siguientes ecuaciones integran las principales definiciones necesarias para el cómputo de las cuatro cantidades de la matriz de confusión. En cada una de estas cuatro cantidades están indicados teoremas (T), lemas (L) y ecuaciones relacionados.

(4.5) → (4.9) → (4.39)

$$VP = \sum_{i \neq j \in V} \left| \bigcup_{f \in F} \bigcup_{g \in G} \underbrace{\mathcal{X}^{ij}(g) \cap \mathcal{X}^{ij}(f)}_{\text{T2, L1: (4.19)}} \right| = \sum_{i \neq j \in V} \underbrace{\sum_{p \in P} \left| \mathcal{X}^{ij}(p) \right|}_{\text{T4: (4.38)}}, \tag{4.42}$$

(4.6) → (4.10) → T5: (4.41)

$$FN = \sum_{i \neq j \in V} \left| \bigcup_{f \in F} \underbrace{\bigcup_{\mathbf{d} \in \bigtimes_{g \in G} D^E_{fg}} \mathcal{X}^{ij}(\cup_{d \in \mathbf{d}} d^{ij})}_{\text{T3: (4.22) y (4.23)}} \right| = \sum_{i \neq j \in V} \underbrace{\sum_{p \in P} \left| \mathcal{X}^{ij}(p) \right|}_{\text{T5: (4.40)}}, \tag{4.43}$$

(4.7) → (4.11) → T5: (4.41)

$$FP = \sum_{i \neq j \in V} \left| \bigcup_{g \in G} \bigcup_{\mathbf{d} \in \bigtimes_{f \in F} D^E_{gf}} \mathcal{X}^{ij}(\cup_{d \in \mathbf{d}} d^{ij}) \right| = \sum_{i \neq j \in V} \sum_{p \in P} \left| \mathcal{X}^{ij}(p) \right|, \tag{4.44}$$

(4.8) → (4.12) → (4.13)

$$VN = \left(\sum_{i \neq j \in V} \prod_{k \in V \setminus \{i,j\}} \left| val(X_k) \right| \right) - VP - FN - FP. \tag{4.45}$$

Puede ser de interés señalar que, para utilizar el método como medida de la distancia entre las estructuras, simplemente se deben sumar FP y FN. Esta suma puede ser interpretada como el número de errores de la segunda estructura respecto a la primera.

A continuación expondré en detalle el procedimiento para computar cada valor de la matriz de confusión.

4.3.1. Verdaderos positivos

1. El procedimiento comienza por un par de índices distintos (i, j) (sumatoria en la Ec. (4.42)), empezando por una característica $f \in F$.

2. Por cada f, iterar por cada característica $g \in G$.

3. Con esta combinación de f y g, se realiza una intersección entre los conjuntos de contextos por pares $\mathcal{X}^{ij}(f)$ y $\mathcal{X}^{ij}(g)$. Para esto, siguiendo la Ec. (4.19), vemos que se deben evaluar las siguientes condiciones:

$$\begin{aligned} C_1 &\equiv (X_i \notin S_f) \vee (X_i \notin S_g) \vee (X_j \notin S_f) \vee (X_j \notin S_g), \\ C_2 &\equiv \exists X_h \in S_f \cap S_g \setminus \{X_i, X_j\} \mid X_h(f) \neq X_h(g), \end{aligned}$$

   a) Si cualquiera de estas condiciones se cumple, es decir, si alguna variable de este par no está en el alcance de f y/o de g, o bien si las variables están en los alcances de ambas pero alguna de ellas tiene distinto valor asignado en f y en g, entonces esta intersección es vacía y no suma verdaderos positivos. En este caso se procede a la siguiente característica en G y se repite esta evaluación.

   b) Si ninguna se cumple, se computa la intersección $\mathcal{X}^{ij}(f) \cap \mathcal{X}^{ij}(g)$, la cual se obtiene, primero, eliminando las asignaciones de X_i y X_j tanto en f como en g para obtener f^{ij} y g^{ij}, y luego, realizando la operación de unión de características $f^{ij} \cup g^{ij}$. De esta unión se obtiene una característica que se incluye en el conjunto resultante. Aquí se procede a la siguiente característica en G.

4. Luego de haber evaluado f con todas las características en G, se procede de la misma forma con la siguiente característica en F desde el paso 2.

5. Al finalizar con todas las características en F, ha sido generado un conjunto de uniones con a lo sumo $|F| \times |G|$ características.[2]

6. Este conjunto se utiliza como entrada en el Algoritmo 1.

[2]Estas cardinalidades deben entenderse como el número de características contenidas en cada conjunto, y no deben confundirse con la cardinalidad de un conjunto de contextos por pares asociados a características o conjuntos de ellas ($|F| \neq |\mathcal{X}^{ij}(F)|$).

7. La salida de este algoritmo es un conjunto de características P. Su cardinalidad se obtiene eficientemente con la Ec. (4.18):

$$\sum_{p\in P}\left|\mathcal{X}^{ij}(p)\right| = \sum_{p\in P}\prod_{X_k\in X_V\setminus\{S_p\cup\{X_i,X_j\}\}}\left|val(X_k)\right|.$$

   En el caso particular de variables binarias esto es simplemente

$$\sum_{p\in P}\left|\mathcal{X}^{ij}(p)\right| = \sum_{p\in P}2^{|X_V|-|S_p|-2}.$$

   Esta suma de todas las cardinalidades en P es el subtotal VP_{ij}, como aparece en la Ec. (4.42).

8. El procedimiento se repite para cada par de índices de variables distintos. Finalmente, las cardinalidades obtenidas para cada par se suman para obtener el total de verdaderos positivos de G respecto a F.

4.3.2. Falsos negativos y falsos positivos

Por otra parte, los falsos negativos se obtienen, en los casos más complejos, en base a diferencias de conjuntos de contextos por pares asociados a características. En términos generales, a cada conjunto para $f \in F$ se le deben sustraer los conjuntos asociados a cada $g \in G$. El cálculo de esta diferencia se define en la igualdad en la Ec. (4.21). El resultado que se muestra en las ecuaciones (4.43) y (4.44) fue derivado en la Sección 4.2.2.

1. Al igual que con verdaderos positivos, se procede por cada par de índices distintos (i,j), por cada característica $f \in F$ y, para cada f, se itera por cada característica $g \in G$, como lo indican la sumatoria y las uniones en las ecuaciones (4.43) y (4.44).

2. En base a la Ec. (4.21), se deben tener en cuenta las siguientes definiciones:

$$\begin{aligned} C_1(f) &\equiv X_i \notin S_f \vee X_j \notin S_f \\ C_1(g) &\equiv X_i \notin S_g \vee X_j \notin S_g, \end{aligned}$$

y C_2, que es la misma condición definida para verdaderos positivos.

Con estas definiciones se deben evaluar dos casos:

a) En primer lugar, si $(g^{ij} \subseteq f^{ij} \wedge \neg C_1(g)) \vee C_1(f)$: esto puede significar, si se cumple el lado izquierdo de la disyunción, que el par de variables está en el alcance de g y que las asignaciones en g^{ij} son un subconjunto de las asignaciones de f^{ij}, es decir, f es más "específica" que g, representando un subconjunto de sus contextos por pares; y si se cumple el lado derecho, que alguna de las variables (o ambas) no están en el alcance de f. En este caso, el resultado es vacío y no suma a la cardinalidad de FN_{ij}.

b) En segundo lugar, si $C_2 \vee C_1(g)$: ningún contexto por pares de $\mathcal{X}^{ij}(f)$ coincide con ninguno de $\mathcal{X}^{ij}(g)$, o bien alguna variable del par no está en el alcance de g. En ambos casos, ningún contexto por pares en $\mathcal{X}^{ij}(g)$ está incluido en $\mathcal{X}^{ij}(f)$. En estos dos casos, se agrega f al conjunto de características resultantes.

c) Si no se cumple ninguno de los casos anteriores, se construye un conjunto de características D^E_{fg} siguiendo la Ec. (4.22).

3. Luego de hacer la evaluación anterior para toda $g \in G$, se obtiene el producto cartesiano entre todos los D^E_{fg}, obteniendo un *conjunto de conjuntos* de características $\mathbf{d}$, donde cada $d \in \mathbf{d}$ contiene $|G|$ características (véase la Sección 4.2.2, Teorema 5).

4. Con los conjuntos $\mathbf{d} \in \bigtimes_{g \in G} D^E_{fg}$ resultantes, para cada $\mathbf{d}$ se eliminan las asignaciones a X_i y X_j de todos sus elementos para obtener las d^{ij}, y se realiza la unión de todas las características $\cup_{d \in \mathbf{d}} d^{ij}$. Se obtiene entonces un conjunto de características, a lo sumo una por cada $\mathbf{d}$.

5. Se repite el procedimiento para la siguiente característica en F.

6. Finalmente, el conjunto de todas las características obtenidas es proporcionado como entrada del Algoritmo 1.

7. Desde el resultado del particionamiento, la cardinalidad de FN_{ij} se obtiene de la misma forma que en el caso de verdaderos positivos (véase el paso 7 del cálculo de verdaderos positivos), y estos se suman para producir el total FN.

Los falsos positivos se calculan con el mismo procedimiento que los falsos negativos, pero intercambiando F y G.

4.3.3. Verdaderos negativos

1. Finalmente, los verdaderos negativos requieren, para cada par de variables, el producto de cardinalidades de las variables restantes, para computar el número total de comparaciones TC:

$$TC = \sum_{i \neq j \in V} \prod_{k \in V \setminus \{i,j\}} \left| val(X_k) \right|,$$

lo que produce el primer término en la Ec. (4.45).

2. Luego, simplemente se restan los restantes valores de la matriz de confusión, computados previamente:

$$TN = TC - VP - FN - FP.$$

4.3.4. Ejemplo de aplicación de la medida

El siguiente es un ejemplo no exhaustivo de aplicación de la medida a un par de estructuras. El propósito de este ejemplo es transmitir una noción intuitiva del cálculo de los valores de la matriz de confusión teniendo en cuenta cómo se combinan múltiples características en las operaciones del método. Las condiciones y operaciones no cubiertas por este ejemplo ya han sido ilustradas por los ejemplos provistos para pares de características a lo largo de este trabajo. Por simplicidad, el caso mostrado tampoco incluye una explicación del particionamiento, al tratarse de un caso particular en el que los resultados de intersección y diferencia producen conjuntos ya particionados.

Ejemplo 12. *La notación de características se encuentra abreviada por claridad. Sean*

$$\begin{aligned} F = \{ \quad & < X_0, X_2 >, \\ & < X_1, X_2 >, \\ & < X_1, X_3 >, \\ & < X_0, X_1, X_2 = 0, X_3 = 1 >, \\ & < X_0, X_1, X_2 = 1, X_3 = 1 > \}, y \end{aligned}$$

$$\begin{aligned} G = \{ \quad & < X_0, X_2 >, \\ & < X_1, X_2 >, \\ & < X_1, X_3 > \}, \end{aligned}$$

donde abreviamos características de la siguiente manera:

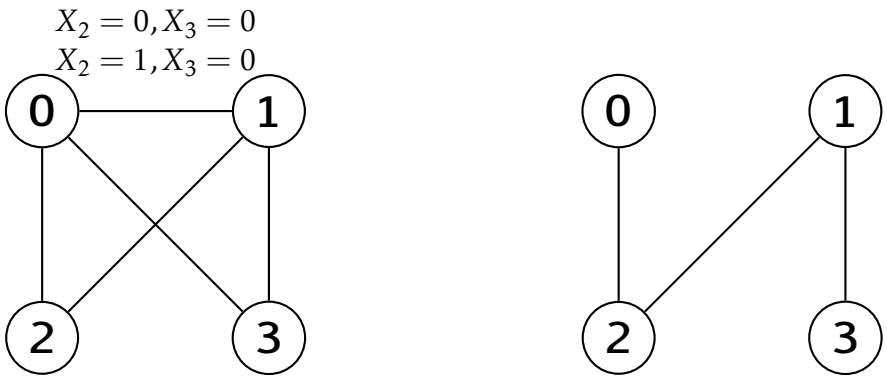
$$< X_0, X_2 >= \{ \quad < X_0 = 0, X_2 = 0 >, < X_0 = 0, X_2 = 1 >,$$
$$< X_0 = 1, X_2 = 0 >, < X_0 = 1, X_2 = 1 > \},$$

y así sucesivamente.

Nótese que en F tenemos las independencias específicas del contexto:

$$(X_0 \perp\!\!\!\perp X_1 \mid X_2 = 0, X_3 = 1) \wedge (X_0 \perp\!\!\!\perp X_1 \mid X_2 = 1, X_3 = 1) \implies (X_0 \perp\!\!\!\perp X_1 \mid X_2, X_3 = 1),$$

mientras que en G observamos la independencia condicional $(X_0 \perp\!\!\!\perp X_1 \mid X_2)$*. Es posible facilitar la interpretación de estas estructuras a través de sus grafos estratificados en la Figura 4.1.*



(A) F (B) G

FIGURA 4.1: Grafo estratificado correspondiente al Ejemplo 12. En el caso de F, existen interacciones específicas del contexto representadas por los estratos $\mathcal{L}_{\{0,1\}} = \{< X_2 = 0, X_3 = 0 >, < X_2 = 1, X_3 = 0 >\}$.

Supongamos que queremos obtener los valores de la matriz de confusión para $i = 0$ *y* $j = 1$*. Veamos a continuación cada caso:*

- VP_{01}*: En F, tenemos un total de 8 características con* X_0 *y* X_1 *en su alcance. Sin embargo, no hay ninguna con ambas variables juntas en G; por lo tanto, no hay ningún verdadero positivo para el par* $(0,1)$*. Esto es fácil de ver dado que no hay*

arista entre los nodos 0 y 1 en el grafo correspondiente a G, indicando que no existen interacciones entre ellas.

- FN_{01}*: Para las características* $< X_0, X_1, X_2 = 0, X_3 = 1 >$ *y* $< X_0, X_1, X_2 = 1, X_3 = 1 >$*, debemos realizar una serie de sustracciones entre conjuntos de contextos por pares, por cada característica en G (Ec. (4.21)). En este caso en particular todos los sustraendos nos llevan a la condición en que no hay contextos para sustraer, es decir, el sustraendo es nulo. Por lo tanto, los falsos negativos son la cardinalidad de los contextos por pares en*

$$\mathcal{X}^{01}(< X_0, X_1, X_2 = 0, X_3 = 1 >) \text{ y}$$
$$\mathcal{X}^{01}(< X_0, X_1, X_2 = 1, X_3 = 1 >).$$

En realidad, cada uno de estos dos conjuntos es el resultado de una unión entre cuatro conjuntos de contextos por pares, uno por cada característica representada en esta notación abreviada. Sin embargo, al eliminar X_0 *y* X_1*, los contextos* $\mathcal{X}^{01}$ *son iguales a* $< X_2 = 0, X_3 = 1 >$ *y* $< X_2 = 1, X_3 = 1 >$*, respectivamente, repetidos cuatro veces. En ambos casos, al obtener la unión de los cuatro nos queda, simplemente, un contexto por conjunto. Por lo tanto, tenemos en total* $FN_{01} = 2$*. (No hay particionamiento en este ejemplo: el resultado ya es una partición debido a que los contextos por pares son completos.)*

- FP_{01}*: Al no existir características con* X_0 *y* X_1 *juntas en G, no hay falsos positivos para este par. Al igual que en el caso de verdaderos positivos, podemos observar esto en el grafo, donde no tenemos interacciones entre los nodos correspondientes.*

- VN_{01}*: El total de comparaciones —solo para* $(0,1)$*— es* $2^2 = 4$*. Tenemos entonces* $VN_{01} = 4 - 2 - 0 - 0 = 2$*. En el grafo, podemos interpretar esto como la*

coincidencia entre la ausencia de arista entre los nodos 0 y 1 para dos contextos en F y en G.

Por completitud, veamos un caso de verdaderos positivos mayor a cero, por ejemplo, $i = 0$ *y* $j = 2$*. En este caso, todas las combinaciones de pares de características a intersectar tendrán un resultado nulo, excepto para la combinación de la característica* $< X_0, X_2 >$ *de F y de G. Al realizar su intersección de contextos por pares, obtendremos una unión igual a dichas características, y sus contextos por pares* $\mathcal{X}^{02}(< X_0, X_2 >)$ *son 4, por las asignaciones posibles de* X_1 *y* X_3*. Nuevamente, al ser una unión de contextos sin* X_0 *ni* X_2 *en su alcance, los conjuntos de contextos por pares para todas las características representadas son iguales, y el resultado final es* $VP_{02} = 4$*. (No hay particionamiento porque el resultado es una única característica.)*

Capítulo 5

Métrica de distancia

Este capítulo está dedicado a la propuesta de una métrica[1] basada en la matriz de confusión presentada en el Capítulo 4, y a proveer una demostración de que tal propuesta cumple con las propiedades que permiten definirla como tal, enunciadas en la Sección 3.1.

5.1. Demostración de las propiedades de métrica

Una métrica es una función que describe una idea de "distancia" entre elementos de un conjunto. Una noción esencial para evaluar si una función cumple las propiedades de una métrica es la de identidad entre dichos elementos. Debido a que la estructura de modelos loglineales es determinada en términos de independencias específicas del contexto, utilizaré como definición de identidad de dos estructuras de modelos loglineales la identidad de sus respectivos modelos de dependencias. Siguiendo la Ec. (2.2) (Sección 2.1.2), podemos describir las

[1] Los términos *métrica* y *medida de distancia* son equivalentes y se refieren a una función que cumple un conjunto de propiedades. A lo largo de esta tesis he utilizado el término (más genérico) *medida* para referirme a una función que puede utilizarse para comparaciones, satisfaga o no las propiedades de una métrica.

estructuras evaluando el conjunto de *todas* las preguntas de independencia específica del contexto para un dominio X_V, que definiremos como

$$\mathcal{D} = \left\{ I(X_i, X_j \mid x_U, X_W) \;\middle|\; \begin{array}{l} \forall i \neq j \in V; \\ U \cup W \subseteq V \setminus \{i, j\}; \\ U \cap W = \emptyset; \\ x_U \in val(X_U) \end{array} \right\}.$$

Así, podemos decir que dos modelos de dependencias D_1 y D_2 son idénticos cuando coinciden los valores evaluados en ambos modelos para toda $I \in \mathcal{D}$, es decir

$$D_1 = D_2 \iff \forall I \in \mathcal{D}, I_{D_1} = I_{D_2}, \tag{5.1}$$

donde I_{D_1} es el valor de verdad devuelto al consultar si la aserción $I \in \mathcal{D}(D_1)$ (y lo mismo para D_2).

Con el fin de proporcionar una demostración clara tendremos en cuenta una distancia definida de manera directa y exhaustiva, que contaría el número de aserciones cuyos valores de verdad difieren en D_1 y D_2. Tal medida se definiría como

$$d(D_1, D_2) = \left| \left\{ I \in \mathcal{D} \mid I_{D_1} \neq I_{D_2} \right\} \right|. \tag{5.2}$$

El conjunto incluye las aserciones que son independientes en D_1 y dependientes en D_2, y viceversa; por lo tanto, su cardinalidad es equivalente a la suma de falsos positivos y falsos negativos de la matriz de confusión.

Ahora recordemos el conjunto reducido utilizado en el enfoque de esta tesis (definido a partir de la Ec. (4.3)):

$$\mathcal{D}_{\mathcal{P}} \equiv \left\{ I(X_i, X_j \mid x_Z) \mid i \neq j \in V, x_Z \in \mathcal{X}^{ij} \right\}.$$

Con esta propuesta estamos definiendo la medida como el conteo de valores de verdad incompatibles sobre el conjunto por pares:

$$d_{\mathcal{P}}(D_1, D_2) = \left| \left\{ I \in \mathcal{D}_{\mathcal{P}} \,\middle|\, I_{D_1} \neq I_{D_2} \right\} \right|. \tag{5.3}$$

A pesar de esta aproximación, podemos demostrar que la medida resultante es una métrica.

Teorema 6. *La medida entre dos modelos de dependencia específicos del contexto D_1 y D_2 en la Ec. (5.3) es una métrica, es decir, satisface las propiedades de no negatividad, identidad de los indiscernibles, simetría y subaditividad (desigualdad del triángulo).*

Demostración. Consideremos cada propiedad por separado:

I *No negatividad:* se cumple trivialmente puesto que la medida está definida como la cardinalidad de un conjunto, por lo tanto no es negativa.

II *Identidad de los indiscernibles:* puede expresarse como

$$d_{\mathcal{P}}(D_1, D_2) = 0 \iff D_1 = D_2. \tag{5.4}$$

Para la implicación de derecha a izquierda, la demostración es trivial ya que $D_1 = D_2$ por definición (Ec. (5.1)) implica que *cualquier conjunto de dependencias* evaluadas en los modelos tendrá valores iguales. Expresado de otra forma, en el caso exhaustivo tenemos que $\{ I \in \mathcal{D} \mid I_{D_1} \neq I_{D_2} \} = \emptyset$ y la cardinalidad de este conjunto es cero; entonces, dado que $\mathcal{D}_{\mathcal{P}} \subset \mathcal{D}$, tenemos que lo mismo se cumple para $\mathcal{D}_{\mathcal{P}}$.

La implicación de izquierda a derecha requiere un análisis más detallado. La demostración aquí propuesta es por transitividad, al mostrar que

$$d_{\mathcal{P}}(D_1, D_2) = 0 \implies d(D_1, D_2) = 0, \tag{5.5}$$

y luego verificando que la implicación se cumple para la medida en la Ec. (5.2), es decir, que

$$d(D_1, D_2) = 0 \implies D_1 = D_2. \tag{5.6}$$

En palabras, esto significa que siempre que la medida es cero sobre el conjunto de independencias *por pares*, entonces la medida sobre *todas* las independencias debe ser también cero (Ec. (5.5)), y que siempre que dos modelos tengan una distancia de cero (medida al comparar *todas* las independencias) entonces son idénticos (Ec. (5.5)).

Para la implicación en la Ec. (5.5), procedemos por contradicción, asumiendo que es posible tener una discordancia en alguna aserción de independencia *que no sea por pares* en D_1 y D_2 (es decir, $d(D_1, D_2) \neq 0$), mientras que simultáneamente se cumpla que $d_{\mathcal{P}}(D_1, D_2) = 0$, es decir, que todas las aserciones de independencia *por pares* concuerden en ambos modelos.

Asumimos que una independencia arbitraria —que no es por pares— se cumple en uno de los modelos pero no en el otro, es decir, $(X_i \perp\!\!\!\perp X_j \mid x_U, X_W)_{D_1}$ y $(X_i \not\perp\!\!\!\perp X_j \mid x_U, X_W)_{D_2}$, *sin producir discrepancia alguna* en el conjunto de aserciones de independencia por pares. Formalmente,

$$(X_i \perp\!\!\!\perp X_j \mid x_U, X_W)_{D_1} \wedge (X_i \not\perp\!\!\!\perp X_j \mid x_U, X_W)_{D_2} \wedge \forall I \in \mathcal{D}_{\mathcal{P}}, I_{D_1} = I_{D_2}.$$

La demostración puede comprenderse con facilidad utilizando el concepto de

caminos en grafos instanciados[2] y el Teorema 1. Por un lado, por $(X_i \not\perp\!\!\!\perp X_j \mid x_U, X_W)_{D_2}$ y el teorema mencionado, tenemos que, en el grafo instanciado $\mathcal{G}_{D_2}(x_U)$ (Figura 5.1(b)), existe un camino desde i hasta j que satisface que ninguno de sus nodos está en W. Denotemos este camino como P^2_{ij}. En términos de independencias, tomaremos en consideración la siguiente equivalencia,

$$(X_i \perp\!\!\!\perp X_j \mid x_U, X_W) \equiv \forall x_W \in val(X_W), (X_i \perp\!\!\!\perp X_j \mid x_U, x_W), \tag{5.7}$$

o bien

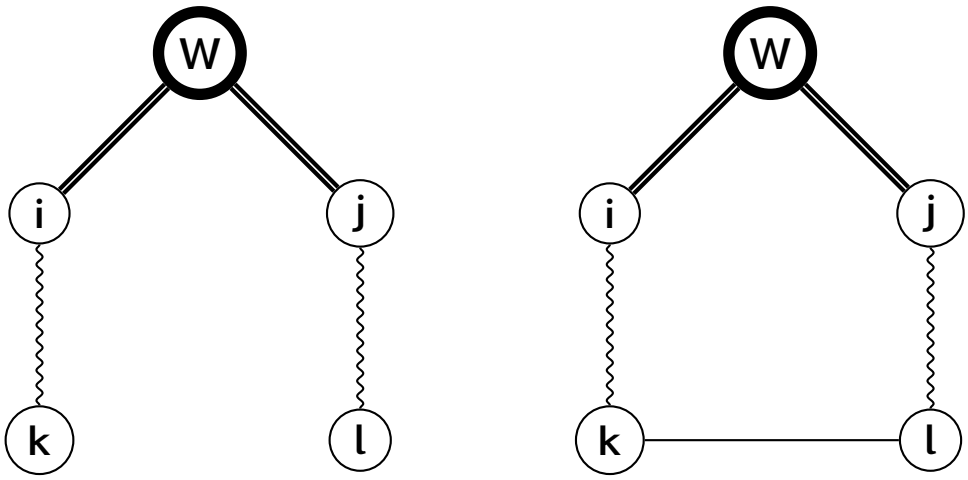
$$(X_i \not\perp\!\!\!\perp X_j \mid x_U, X_W) \equiv \exists x_W \in val(X_W), (X_i \not\perp\!\!\!\perp X_j \mid x_U, x_W),$$

para todos los $i \neq j \in V$, $U \cup W \subseteq V^{ij}$, $U \cap W = \emptyset$, $x_U \in val(X_U)$.

Esto implica que, para al menos un contexto x_Z, donde $x_U \subseteq x_Z$, $(X_i \not\perp\!\!\!\perp X_j \mid x_Z)_{D_2}$.

Por otro lado, por $(X_i \perp\!\!\!\perp X_j \mid x_U, X_W)_{D_1}$ y el Teorema 1, en el grafo instanciado $\mathcal{G}_{D_1}(x_U)$ (Figura 5.1(a)) no existe camino alguno desde i a j "por fuera" de W, es decir, un camino que satisfaga que ninguno de sus nodos está en W. Por lo tanto, la secuencia de nodos de P^2_{ij} no puede ser un camino en D_1. Esto implica que al menos un par de nodos adyacentes en P^2_{ij} no está conectado por una arista en $\mathcal{G}_{D_1}(x_U)$ (ejemplificados como k y l en la figura) mientras que sí ocurre que hay una arista entre ellos en $\mathcal{G}_{D_2}(x_U)$. En términos de separación, podemos decir que W separa a i y j en $\mathcal{G}_{D_1}(x_U)$ pero no en $\mathcal{G}_{D_2}(x_U)$. La implicación es, en primer lugar, que por el axioma de Unión Fuerte (véase el Apéndice A.2) tenemos $(X_i \perp\!\!\!\perp X_j \mid X_{V\setminus\{k,l\}\cup U})_{D_1}$, y por la Ec. (5.7), tenemos $(X_i \perp\!\!\!\perp X_j \mid x_Z)_{D_1}$, para $x_U \subseteq x_Z$ y para todo $x_{V\setminus(\{k,l\}\cup U)} \subseteq x_{Z\setminus U}$. En segundo lugar, tenemos que $(X_i \not\perp\!\!\!\perp X_j \mid X_{V\setminus(\{k,l\}\cup U)})_{D_2}$, y por la Ec. (5.7), debe existir al

[2] Un resumen del concepto de grafos instanciados se presentó en la Sección 2.3.1.1.



(A) $\mathcal{G}_{D_1}(x_U)$ (B) $\mathcal{G}_{D_2}(x_U)$

FIGURA 5.1: Dos posibles grafos instanciados en el contexto x_U para $(X_i \perp\!\!\!\perp X_j \mid X_W, x_U)_{D_1}$ y $(X_i \not\perp\!\!\!\perp X_j \mid X_W, x_U)_{D_2}$. W representa un conjunto de nodos y las líneas dobles gruesas representan aristas entre los nodos i y j hacia todas las variables en W. Las líneas onduladas entre nodos representan que hay un camino que une a esos nodos. En este ejemplo, existe un camino desde i a j en D_2 por fuera de W, pero para algún par de nodos (k,l), D_1 no tiene una arista entre ellos.

menos un contexto x_Z (con $x_U \subseteq x_Z$) tal que $(X_i \perp\!\!\!\perp X_j \mid x_Z)_{D_2}$. En consecuencia, D_1 y D_2 tendrán que diferir en una aserción por pares, contradiciendo así nuestra suposición de que todas ellas concuerdan.

Para la implicación en la Ec. (5.6), es suficiente observar que la medida solo puede ser cero si el conjunto está vacío, es decir, si *todas* las aserciones de independencia tienen el mismo valor de verdad en cada modelo. El único caso en que esto es posible es cuando $D_1 = D_2$, demostrando la implicación. Así, tenemos que por transitividad de las implicaciones en las Ecuaciones (5.5) y (5.6), la implicación de izquierda a derecha en la Ec. (5.4) se cumple.

III *Simetría:* está garantizada debido a que la medida es un conteo de igualdades (negadas), las cuales son conmutativas.

iv *Desigualdad triangular.* Debemos demostrar que, para tres modelos de dependencias específicos del contexto A, B y C,

$$d_{\mathcal{P}}(A,C) \leq d_{\mathcal{P}}(A,B) + d_{\mathcal{P}}(B,C). \tag{5.8}$$

Podemos reescribir la medida como una suma de términos,

$$d_{\mathcal{P}}(D_1,D_2) = \sum_{I \in \mathcal{D}_{\mathcal{P}}} \tilde{d}_i(D_1,D_2), \tag{5.9}$$

donde

$$\tilde{d}_i(D_1,D_2) = \begin{cases} 1 & \text{si } I_{D_1} \neq I_{D_2} \\ 0 & \text{si } I_{D_1} = I_{D_2}. \end{cases}$$

La interpretación es que cada uno de estos términos es un indicador que informa si la aserción de independencia I tiene el mismo valor en ambos modelos, es decir, si ambos modelos tienen la misma dependencia o independencia para las variables y conjunto condicionante especificados por I.

Podemos reexpresar la Ec. (5.8) utilizando la Ec. (5.9) como

$$\sum_{I \in \mathcal{D}_{\mathcal{P}}} \tilde{d}_i(A,C) \leq \sum_{I \in \mathcal{D}_{\mathcal{P}}} \tilde{d}_i(A,B) + \sum_{I \in \mathcal{D}_{\mathcal{P}}} \tilde{d}_i(B,C).$$

Considerando que la suma de dos o más desigualdades válidas lado a lado también es una desigualdad válida, es suficiente demostrar que para tres modelos cualesquiera A, B, C y todas las aserciones de independencia $I \in \mathcal{D}_{\mathcal{P}}$,

$$\tilde{d}_i(A,C) \leq \tilde{d}_i(A,B) + \tilde{d}_i(B,C). \tag{5.10}$$

Considerando los modelos A y B, tenemos que para el término del lado izquierdo en la Ec. (5.10), hay dos casos posibles: $\tilde{d}_i(A,C) = 0$ o $\tilde{d}_i(A,C) = 1$. Dado cada uno de estos, los posibles valores para $\tilde{d}_i(A,B)$ y $\tilde{d}_i(B,C)$ no pueden ser combinaciones arbitrarias sino que están restringidas como sigue:

$$\begin{aligned} \tilde{d}_i(A,C) = 0 \implies & \left(\tilde{d}_i(A,B) = 0 \wedge \tilde{d}_i(B,C) = 0\right) \\ & \vee \left(\tilde{d}_i(A,B) = 1 \wedge \tilde{d}_i(B,C) = 1\right), \end{aligned} \tag{5.11}$$

$$\begin{aligned} \tilde{d}_i(A,C) = 1 \implies & \left(\tilde{d}_i(A,B) = 0 \wedge \tilde{d}_i(B,C) = 1\right) \\ & \vee \left(\tilde{d}_i(A,B) = 1 \wedge \tilde{d}_i(B,C) = 0\right). \end{aligned} \tag{5.12}$$

Para el primer caso (Ec. (5.11)), tenemos que el valor de la aserción I es el mismo en A y C, lo que significa que los tres modelos contienen la misma dependencia o independencia, y entonces $\tilde{d}_i(A,B) = 0 \wedge \tilde{d}_i(B,C) = 0$; o bien que A y C concuerdan pero ambas difieren de B (ya que existen solo dos valores), lo cual es representado por $\tilde{d}_i(A,B) = 1 \wedge \tilde{d}_i(B,C) = 1$. Las otras dos combinaciones no pueden ocurrir porque negarían el lado izquierdo de la implicación.

Con respecto al segundo caso (Ec. (5.12)), siguiendo un razonamiento similar también existen solo dos posibilidades: $\tilde{d}_i(A,B) = 0 \wedge \tilde{d}_i(B,C) = 1$, o bien $\tilde{d}_i(A,B) = 1 \wedge \tilde{d}_i(B,C) = 0$; en consecuencia la igualdad se satisface en ambos casos. Debido a que todos los casos posibles satisfacen la Ec. (5.10) y a que la suma en ambos lados para todo $I \in \mathcal{D}_\mathcal{P}$ conserva la desigualdad, ahora podemos concluir que la propiedad en la Ec. (5.8) se satisface.

□

Capítulo 6

Resumen y conclusiones

En este trabajo fue presentada una métrica para la comparación directa y eficiente de las estructuras de dos modelos loglineales. Estos modelos son más expresivos que los que utilizan grafos no dirigidos como estructura, debido a su capacidad para representar independencias específicas del contexto; sin embargo, la interpretación de la estructura de independencias de estos modelos es compleja y a mi entender no se conocía un método confiable para realizar comparaciones de calidad directas de estas estructuras antes del diseño de esta métrica. La importancia de un método que compara estructuras de independencias reside en que puede ser utilizado no solo para el objetivo de aprendizaje de estimación de densidad, sino también para descubrimiento de conocimiento. Por un lado, es posible analizar diferencias en las estructuras de independencia obtenidas con algoritmos de aprendizaje de estructuras con respecto a estructuras sintéticas subyacentes, o comparar las estructuras aprendidas por distintos algoritmos. Por otro lado, se pueden sacar conclusiones cualitativas acerca de estructuras, ya sean aquellas aprendidas por algoritmos o las que fueren diseñadas por seres humanos expertos (o ambas), las cuales no sería posible obtener por mera observación excepto en

escenarios simples (de baja dimensionalidad).

Además, el método presentado provee más garantías que técnicas del estado del arte para evaluar estructuras de independencias de modelos loglineales en base a la distribución completa. Para esta representación, los algoritmos de aprendizaje usualmente son evaluados con la medida de divergencia KL para distribuciones completas, o el método aproximado CMLL para dominios de alta dimensionalidad, lo cual requiere aprender los parámetros numéricos de los modelos y son por lo tanto indirectos. Además, estos métodos tampoco poseen las propiedades de una métrica. Existen algunas mediciones directas de propiedades de las estructuras, que han sido utilizadas en varios trabajos, tales como la longitud promedio de características o el número de características, pero estos solo proveen información muy limitada sobre las estructuras, y no existe ninguna garantía de su validez. En este trabajo se ha demostrado que la técnica propuesta constituye una métrica, haciéndola así adecuada para obtener conclusiones confiables sobre las comparaciones realizadas con ella.

6.1. Trabajo futuro

Algunas posibles líneas futuras de investigación sobre este método podrían incluir la búsqueda de un procedimiento eficiente para cómputo sobre modelos loglineales de múltiples características. Si bien la métrica propuesta tiene la ventaja de omitir los procedimientos costosos, y a menudo aproximados, de aprendizaje de parámetros para la evaluación de estructuras, tiene la desventaja de que no se puede garantizar el cómputo eficiente respecto a la cantidad de características de los modelos loglineales a comparar.

También resultaría de interés conducir una exploración del comportamiento de la métrica en comparación con otras medidas tales como divergencia KL o

C(M)LL en aplicaciones reales. Por ejemplo, podrían reproducirse experimentos de aprendizaje automático de modelos loglineales sintéticos y aplicar la métrica sobre las estructuras aprendidas con distintos algoritmos del estado del arte para efectuar una evaluación de las distancias entre estructuras aprendidas y la(s) estructura(s) sintética(s), con el fin de comparar la calidad de los resultados de cada algoritmo. Esto también permitiría verificar si los resultados obtenidos de hacer la misma comparación con otra estimación, tal como la divergencia KL, produce diferencias en el ranking de algoritmos, y extraer conclusiones sobre el impacto de falsos positivos y falsos negativos en el aprendizaje. La métrica permitiría observar qué algoritmos son más propensos a cometer uno u otro tipo de error, y por lo tanto ayudaría a diseñar nuevas estrategias que mejoren el aprendizaje de estructuras.

Apéndice A

Axiomas de independencia

A.1. Caracterización axiomática de la relación de independencia condicional

Teorema 7 ([Pearl, 1988], Sección 3.1.2)**.** *Sean X_A, X_B, X_Z y X_W conjuntos disjuntos de variables en X_V. La relación $(X_A \perp\!\!\!\perp X_B \mid X_Z)$, leída como "$X_A$ es independiente de X_B dado X_Z" en algún modelo probabilístico, debe satisfacer las siguientes condiciones independientes:*

- *Simetría:*

  $$(X_A \perp\!\!\!\perp X_B \mid X_Z) \iff (X_B \perp\!\!\!\perp X_A \mid X_Z)$$

- *Descomposición:*

  $$(X_A \perp\!\!\!\perp X_B \mid X_{Z\cup W}) \implies (X_A \perp\!\!\!\perp X_B \mid X_Z) \wedge (X_A \perp\!\!\!\perp X_W \mid X_Z)$$

- *Unión débil:*

  $$(X_A \perp\!\!\!\perp X_{B\cup W} \mid X_Z) \implies (X_A \perp\!\!\!\perp X_B \mid X_{Z\cup W})$$

- *Contracción:*

  $(X_A \perp\!\!\!\perp X_B \mid X_Z) \wedge (X_A \perp\!\!\!\perp X_W \mid X_{Z \cup B}) \implies (X_A \perp\!\!\!\perp X_{B \cup W} \mid X_Z)$.

Si p es estrictamente positiva, se satisface una quinta condición:

- *Intersección:*

  $(X_A \perp\!\!\!\perp X_B \mid X_{Z \cup W}) \wedge (X_A \perp\!\!\!\perp X_W \mid X_{Z \cup B}) \implies (X_A \perp\!\!\!\perp X_{B \cup W} \mid X_Z)$.

A.2. Caracterización axiomática del isomorfismo a grafos de independencias

Teorema 8 ([Pearl, 1988], Sección 3.1.3)**.** *Una condición necesaria y suficiente para que un modelo de dependencias M de una distribución sea isomorfo a grafos es que* $(X_A \perp\!\!\!\perp X_B \mid X_Z)_M$ *satisfaga los siguientes axiomas (por claridad, se omite el subíndice M):*

- *Simetría*
- *Descomposición*
- *Intersección*
- *Unión fuerte:*

  $(X_A \perp\!\!\!\perp X_B \mid X_Z) \implies (X_A \perp\!\!\!\perp X_B \mid X_{Z \cup W})$

- *Transitividad:*

  $(X_A \perp\!\!\!\perp X_B \mid X_Z) \implies (X_A \perp\!\!\!\perp X_c \mid X_Z) \vee (X_c \perp\!\!\!\perp X_B \mid X_Z)$,

donde X_c *es una variable individual en* $X_{V \setminus (A \cup B \cup Z \cup W)}$.

Apéndice B

Definiciones auxiliares

B.1. Operaciones con características

Las siguientes definiciones han sido construidas durante el desarrollo de este trabajo con el fin de proveer una descripción clara de los procedimientos contenidos en el mismo:

Definición 10. ***Unión de características.*** *Una unión de dos características $f \cup g$ se define sobre cualquier par f y g que satisfaga que para todo $X_k \in S_f \cap S_g$, $X_k(f) = X_k(g)$, y produce como resultado una característica h tal que, para todo $X_k \in S_f \cup S_g$, $X_k \in S_h$ y*

$$X_k(h) \equiv \begin{cases} X_k(f) & si\ X_k \in S_f \\ X_k(g) & si\ X_k \in S_g. \end{cases}$$

Ejemplo 13. *Si $f =< X_0 = 1, X_2 = 0, X_3 = 1 >$ y $g =< X_0 = 1, X_1 = 0, X_4 = 0 >$, su unión es*

$$f \cup g =< X_0 = 1, X_1 = 0, X_2 = 0, X_3 = 1, X_4 = 0 > .$$

Definición 11. ***Subconjuntos de características.*** *Sean a y b dos características. Decimos que a es subconjunto de b, denotado como $a \subseteq b$, si se satisface lo siguiente:*

$$a \subseteq b \implies \forall k \in S_a, k \in S_b \wedge X_k(a) = X_k(b).$$

Ejemplo 14. *Dadas*

$$f =< X_0 = 1, X_2 = 0, X_3 = 1 >,$$
$$g =< X_0 = 0, X_3 = 1 >,$$
$$l =< X_0 = 1 > \ y$$
$$h =< X_0 = 1, X_3 = 1 >,$$

tenemos que $f \subset h$, $g \subset h$, y $l \not\subset h$.

Apéndice C

Lemas auxiliares y sus demostraciones

C.1. Relación entre contextos por pares y subconjuntos de una característica

Lema Auxiliar 1. *Sean F el modelo loglineal de alguna distribución X_V, $f \in F$ alguna característica en F, $X_i \neq X_j \in X_V$ dos variables distintas, $x_Z \in \mathcal{X}^{ij}$ algún conjunto condicionante por pares, y f^{ij} denote la característica compuesta por las mismas asignaciones en f pero sin asignación a las variables X_i y X_j. Entonces,*

$$x_Z \in \mathcal{X}^{ij}(f) \iff f^{ij} \subseteq x_Z \wedge X_i, X_j \in S_f \tag{C.1}$$

donde la operación de subconjunto $f^{ij} \subseteq x_Z$ se establece sobre asignaciones, es decir, se interpreta que cada variables en la característica f que no sea X_i ni X_j tiene asignado el mismo valor tanto en f^{ij} como en x_Z.

Demostración. El lema deriva de manera directa de hechos conocidos a partir de la teoría de modelos loglineales. Por la definición de $\mathcal{X}^{ij}(f)$ en la Eq. 4.4, $x_Z \in \mathcal{X}^{ij}(f)$ siempre que $(X_i \not\perp\!\!\!\perp X_j \mid x_Z)$. De acuerdo a [[Koller and Friedman, 2009], Capítulo 4], cualquier distribución estructurada por algún modelo loglineal F, contiene una interacción directa (una dependencia) entre cualquier par de variables X_i y X_j, siempre que aparecen juntas en al menos una característica $f \in F$. Además, la interacción se sigue cumpliendo cuando está condicionada en alguna asignación parcial de variables, cuando existe al menos una característica que satisfaga que todas sus asignaciones concuerden con el conjunto condicionante. En particular, para el caso de conjuntos condicionantes por pares, $I(X_i, X_j \mid x_Z)$ es una dependencia siempre que cada variable en x_Z (que también está en el alcance S_f de f) tiene un valor que concuerda en ambos. Finalmente, el lema queda demostrado al observar que x_Z no contiene a X_i ni a X_j.

□

C.2. Descomposición de contextos por pares de un conjunto de características en la unión de contextos por pares de características individuales

Lema Auxiliar 2. *El conjunto de contextos por pares $\mathcal{X}^{ij}(F)$ de un modelo loglineal F es equivalente a la unión de los conjuntos por pares $\mathcal{X}^{ij}(f)$ de cada una de sus características $f \in F$; formalmente,*

$$\mathcal{X}^{ij}(F) = \bigcup_{f \in F} \mathcal{X}^{ij}(f)$$

Demostración. De la Ec. (4.4), $\mathcal{X}^{ij}(F) \equiv \left\{ x_Z \in \mathcal{X}^{ij} \,\middle|\, (X_i \not\perp\!\!\!\perp X_j \mid x_Z)_F \right\}$; además, siguiendo el razonamiento presentado en el Lema Auxiliar 1, si una aserción $I(X_i, X_j \mid x_Z)$ es dependiente de acuerdo a alguna característica $f \in F$, entonces es dependiente de acuerdo al modelo loglineal completo:

$$(X_i \not\perp\!\!\!\perp X_j \mid x_u, X_W)_f \implies (X_i \not\perp\!\!\!\perp X_j \mid x_u, X_W)_F.$$

Por lo tanto,

$$\begin{aligned}
\mathcal{X}^{ij}(F) &\equiv \left\{ x_Z \in \mathcal{X}^{ij} \,\middle|\, (X_i \not\perp\!\!\!\perp X_j \mid x_Z)_F \right\} \\
&\equiv \left\{ x_Z \in \mathcal{X}^{ij} \,\middle|\, \bigvee_{f \in F} (X_i \not\perp\!\!\!\perp X_j \mid x_Z)_f \right\} \\
&\equiv \bigcup_{f \in F} \left\{ x_Z \in \mathcal{X}^{ij} \,\middle|\, (X_i \not\perp\!\!\!\perp X_j \mid x_Z)_f \right\} \\
&\equiv \bigcup_{f \in F} \mathcal{X}^{ij}(f)
\end{aligned}$$

□

Ilustraremos la definición de este lema auxiliar con un ejemplo:

Ejemplo 15. *Sea* $V = \{k\}, k = 1 \ldots 4$, $\forall X_k, val(X_i) = \{0, 1\}$, *y* $f, f', f'' \in F$ *donde*

$$f = < X_1 = 0, X_3 = 0, X_4 = 1 >,$$
$$f' = < X_2 = 1, X_3 = 0, X_4 = 0 >, \text{ y}$$
$$f'' = < X_1 = 0, X_2 = 0 > .$$

Sean $i = 3$ *y* $j = 4$. *Entonces,*

$$\mathcal{X}^{34}(f) = \{< X_1 = 0, X_2 = 0 >, < X_1 = 0, X_2 = 1 >\}$$

$$\mathcal{X}^{34}(f') = \{< X_1 = 0, X_2 = 1 >, < X_1 = 1, X_2 = 1 >\}$$
$$\mathcal{X}^{34}(f'') = \emptyset.$$

Así, $\mathcal{X}^{34}(F)$ es la unión de los contextos por pares $\mathcal{X}^{34}(f)$ y $\mathcal{X}^{34}(f')$, asumiendo que f y f' son las únicas características en F que contienen X_3 y X_4 en sus alcances, dando como resultado

$$\begin{aligned}\mathcal{X}^{34}(F) =& \mathcal{X}^{34}(f) \cup \mathcal{X}^{34}(f') \\ =& \{< X_1 = 0, X_2 = 0 >, < X_1 = 0, X_2 = 1 >, < X_1 = 1, X_2 = 1 >\}.\end{aligned}$$

C.3. Transitividad entre subconjuntos de características

Lema Auxiliar 3. *Sean a, b y c tres características cualesquiera; entonces*

$$a \subseteq b \wedge b \subseteq c \implies a \subseteq c.$$

Demostración. Por la definición de subconjuntos de características (Definición 11),

$$a \subseteq b \implies \forall k \in S_a, k \in S_b \wedge X_k(a) = X_k(b),$$

entonces,

$$
\begin{aligned}
b \subseteq c \quad &\Longrightarrow \forall k \in S_b, k \in S_c \wedge X_k(b) = X_k(c) \\
&\Longrightarrow \forall k \in S_a, k \in S_c \wedge X_k(b) = X_k(c), \text{ porque } S_a \subseteq S_b, \\
&\Longrightarrow \forall k \in S_a, k \in S_c \wedge X_k(a) = X_k(c), \text{ por la Ec. C.3} \\
&\Longrightarrow a \subseteq c.
\end{aligned}
$$

□

Apéndice D

Comparación entre la métrica propuesta y la divergencia de Kullback–Leibler

Las medidas existentes para la comparación de modelos loglineales, entre las cuales la más utilizada es la *divergencia de Kullback–Leibler* [Kullback and Leibler, 1951, Cover and Thomas, 2012] no permiten la comparación directa de estructuras; en cambio, comparan las distribuciones completas (estructura más parámetros de los modelos), lo cual introduce varias limitaciones, como se describió en el Capítulo 3.

El método propuesto en esta tesis, consistente en computar una matriz de confusión —abarcando las medidas de *verdaderos positivos* (VP), *verdaderos negativos* (VN), *falsos positivos* (FP), y *falsos negativos* (FN)—, no solo se computa de forma directa sobre estructuras de dependencias de los modelos a comparar, sino que además permite discriminar entre distintos tipos de errores.

Para facilitar una comprensión intuitiva del impacto de estas diferencias, presentaré a continuación un ejemplo de aplicación de la métrica de esta tesis a una comparación exhaustiva entre un modelo sintético original y modelos con distancias variables respecto a él.

D.1. Ejemplo: Discriminando falsos positivos de falsos negativos

En el siguiente ejemplo se utilizan modelos sintéticos con distintos números de falsos positivos y falsos negativos y parámetros aleatorios, con el fin de ilustrar limitaciones de una medida de comparación de distribuciones de modelos loglineales para la estimación de la disimilitud de sus estructuras, en contraste con la métrica propuesta. Específicamente, se evalúa la divergencia KL sobre todos los modelos sintéticos generados y los valores obtenidos se visualizan uno a uno contra los valores de errores (falsos positivos o falsos negativos) obtenidos con la métrica, dando una idea de su correlación.

El preprocesamiento de los resultados y su visualización fueron realizados en el lenguaje estadístico `R`, utilizando principalmente los paquetes `shiny` y `plotly`. El cómputo de la métrica fue realizado con una implementación en Java, accesible públicamente en el siguiente repositorio de GitHub: github.com/jstrappa/jllcm. La divergencia KL se computó con la implementación del siguiente repositorio: bitbucket.org/ystrappa/kl. El trabajo contenido en esta sección fue presentado oralmente en la conferencia *LatinR 2019*, llevada a cabo en el mes de septiembre de 2019 en Santiago, Chile.

D.1.1. Metodología

El procedimiento consiste, en primer lugar, en la utilización de un modelo sintético M_O —que también puede llamarse "original"—, definido sobre un dominio de 6 variables: $\{X_0,\ldots,X_5\}$. Fue seleccionado debido a su aparición en trabajos relacionados (véanse [Edera et al., 2014a] y [Edera et al., 2014b]), y a que posee una estructura que puede modificarse de manera sencilla para generar variantes con un número considerable de falsos positivos y de falsos negativos.

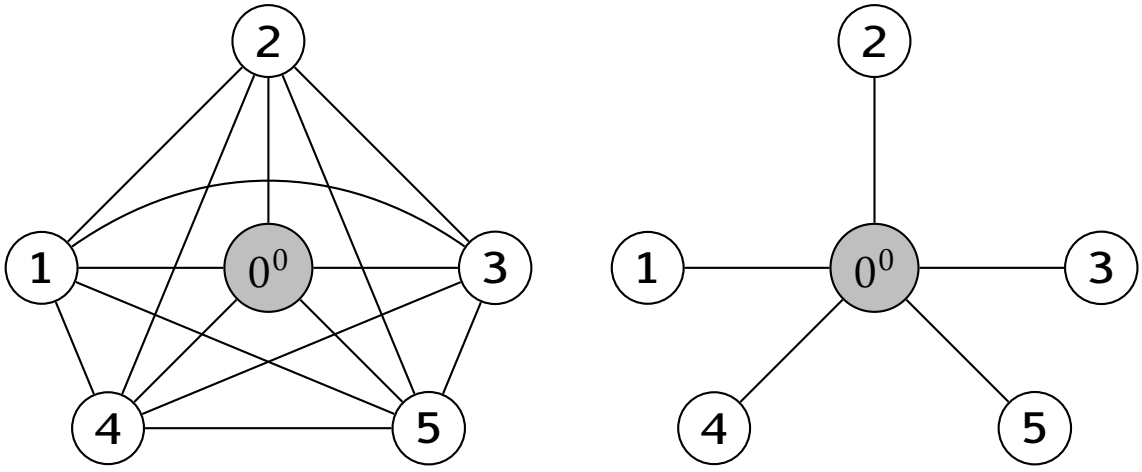
El modelo original se muestra en la Figura D.1, utilizando la representación de grafos instanciados descrita en la Sección 2.3.1.1. En este caso se incluye la variable instanciada en el grafo. El mismo consiste en dos estructuras locales: un modelo saturado (subgrafo completo) para un contexto dado por una variable, y un modelo independiente (subgrafo vacío). De esta forma la estructura global contiene un número de independencias específicas del contexto. El modelo de dependencias asociado es:

$$\mathcal{D}_M = \Big\{(X_i \not\perp\!\!\!\perp X_j \mid X_0 = 0)\Big\} \cup \Big\{(X_i \perp\!\!\!\perp X_j \mid X_0 = 1); \forall i \neq j \in \{1,\cdots,5\}\Big\}.$$

Los parámetros fueron generados variando los pesos de las características con el fin de mantener interacciones fuertes. El diseño de los mismos se explica en detalle en [[Edera et al., 2014a], Apéndice B]. He utilizado los modelos generados para la experimentación del citado trabajo, con el permiso de sus autores.

La comparación se divide en dos partes. Por un lado, veremos la evaluación sobre un conjunto de estructuras $\mathcal{M}_{FP}$ que solo tienen falsos positivos respecto a M_O y, por otro, sobre estructuras que solo tienen falsos negativos respecto a M_O, $\mathcal{M}_{FN}$.

Una vez obtenidas las estructuras el siguiente paso fue computar la medida



(A) $\mathcal{G}_{x_{60}}(\mathcal{M}_O)$ (B) $\mathcal{G}_{x_{61}}(\mathcal{M}_O)$

FIGURA D.1: Grafos instanciados asociados al modelo sintético $\mathcal{M}_O$.

de distancia de estructuras loglineales ($d_{\mathcal{P}}$) directamente entre la estructura sintética M_O y cada estructura generada aleatoriamente. Esto consiste en obtener, para cada estructura $M_{FN} \in \mathcal{M}_{FN}$, el valor $d_{\mathcal{P}}(M_O, M_{FN})$, y para cada estructura $M_{FP} \in \mathcal{M}_{FP}$, $d_{\mathcal{P}}(M_O, M_{FP})$.

El cálculo de la divergencia KL requirió pasos adicionales. En primer lugar, fue necesario generar conjuntos de datos de distintos tamaños desde el modelo sintético M_O. Específicamente, utilicé los tamaños $s \in \{50, 100, 1000, 10000\}$, y el método de muestreo fue un muestreo de Gibbs utilizando el paquete de software de código abierto *Libra toolkit*[1].

En segundo lugar, realicé aprendizaje de parámetros desde estos datos para todas las estructuras aleatorias $\mathcal{M}_{FN}$ y $\mathcal{M}_{FP}$, para obtener la distribución completa estimada con cada uno. Por último, calculé la divergencia KL entre M_O y cada modelo obtenido en el segundo paso, y promedié los valores sobre los 10 conjuntos de parámetros del modelo correspondientes a cada s.

[1] Disponible en http://libra.cs.uoregon.edu/

D.1.2. Resultados

La visualización se muestra en las Figuras D.2 a D.5. En cada una, los resultados para $\mathcal{M}_{FN}$ y $\mathcal{M}_{FP}$ están graficados por separado. Las gráficas muestran a su vez una comparación de los resultados para distintos tamaños de muestras utilizados por la divergencia KL para el cómputo de similitud.

En cada gráfica, el eje X representa el porcentaje de errores, mientras que el eje Y muestra el valor de la divergencia KL para la estructura correspondiente. Cada punto es una estructura diferente. La desviación estándar es la correspondiente al conjunto de modelos obtenidos por el aprendizaje de parámetros desde los datos generados por los 10 conjuntos de parámetros originales de M_O.

D.1.3. Conclusiones

Se observa una correlación positiva entre la divergencia KL y el cómputo de falsos negativos de la métrica. Esto es de esperar, porque este tipo de error implica que existen interacciones que no pueden ser cuantificadas con los parámetros, sin importar la cantidad de datos. En consecuencia, la divergencia KL muestra la disimilitud causada por la ausencia de interacciones en los modelos a comparar que sí están presentes en el modelo original. Sin embargo, no hay correlación con el porcentaje de falsos positivos. La divergencia KL, por otro lado, previsiblemente oscurece las diferencias estructurales entre modelos con este tipo de errores. Por último, es posible señalar que cuanto mayor es la cantidad de datos utilizada para el aprendizaje de parámetros de los modelos evaluados con divergencia KL, peor es la estimación de diferencias estructurales cuando hay falsos positivos presentes. Esto es a causa de la compensación de los parámetros aprendidos, ya que pueden ser estimados con mayor precisión cuantos más datos se utilicen.

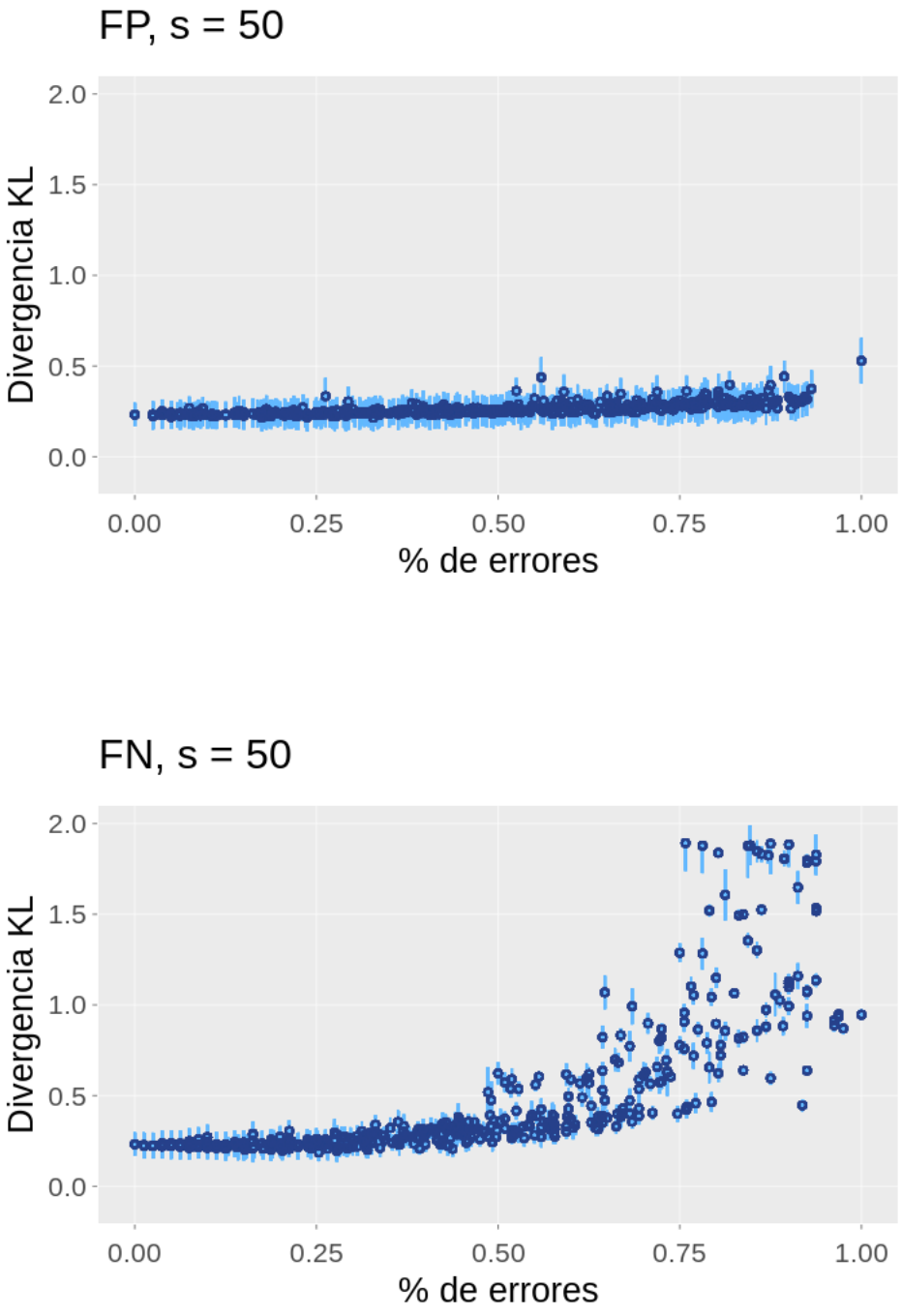


FIGURA D.2: Comparación de errores para la métrica propuesta (eje X) contra divergencia KL (eje Y) para conjuntos de datos sintéticos de 6 variables, para $s = 50$. Arriba: % de falsos positivos. Abajo: % de falsos negativos. Cada punto en las gráficas es una estructura diferente.

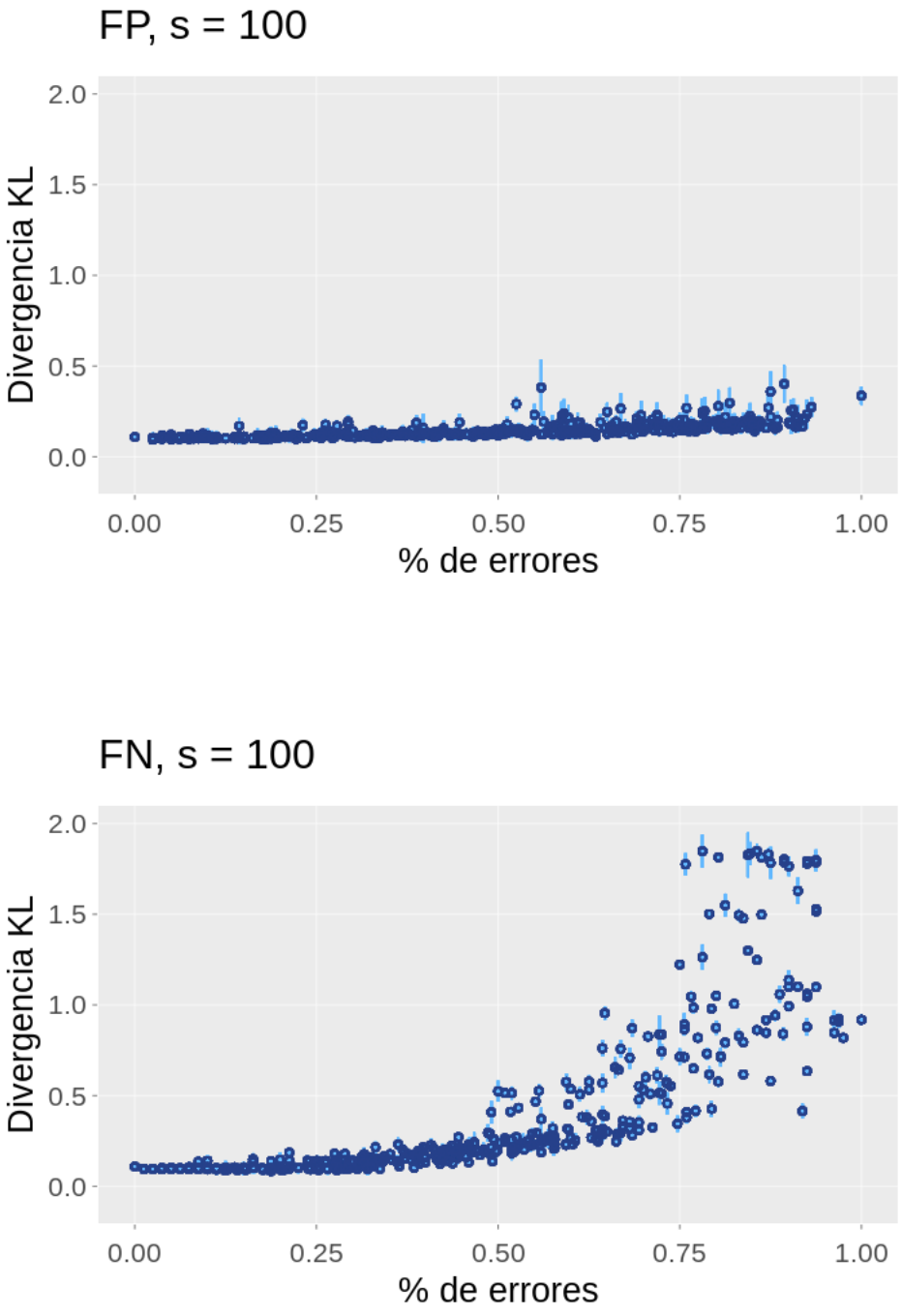


Figura D.3: Comparación de errores para la métrica propuesta (eje X) contra divergencia KL (eje Y) para conjuntos de datos sintéticos de 6 variables, para $s = 100$. Arriba: % de falsos positivos. Abajo: % de falsos negativos. Cada punto en las gráficas es una estructura diferente.

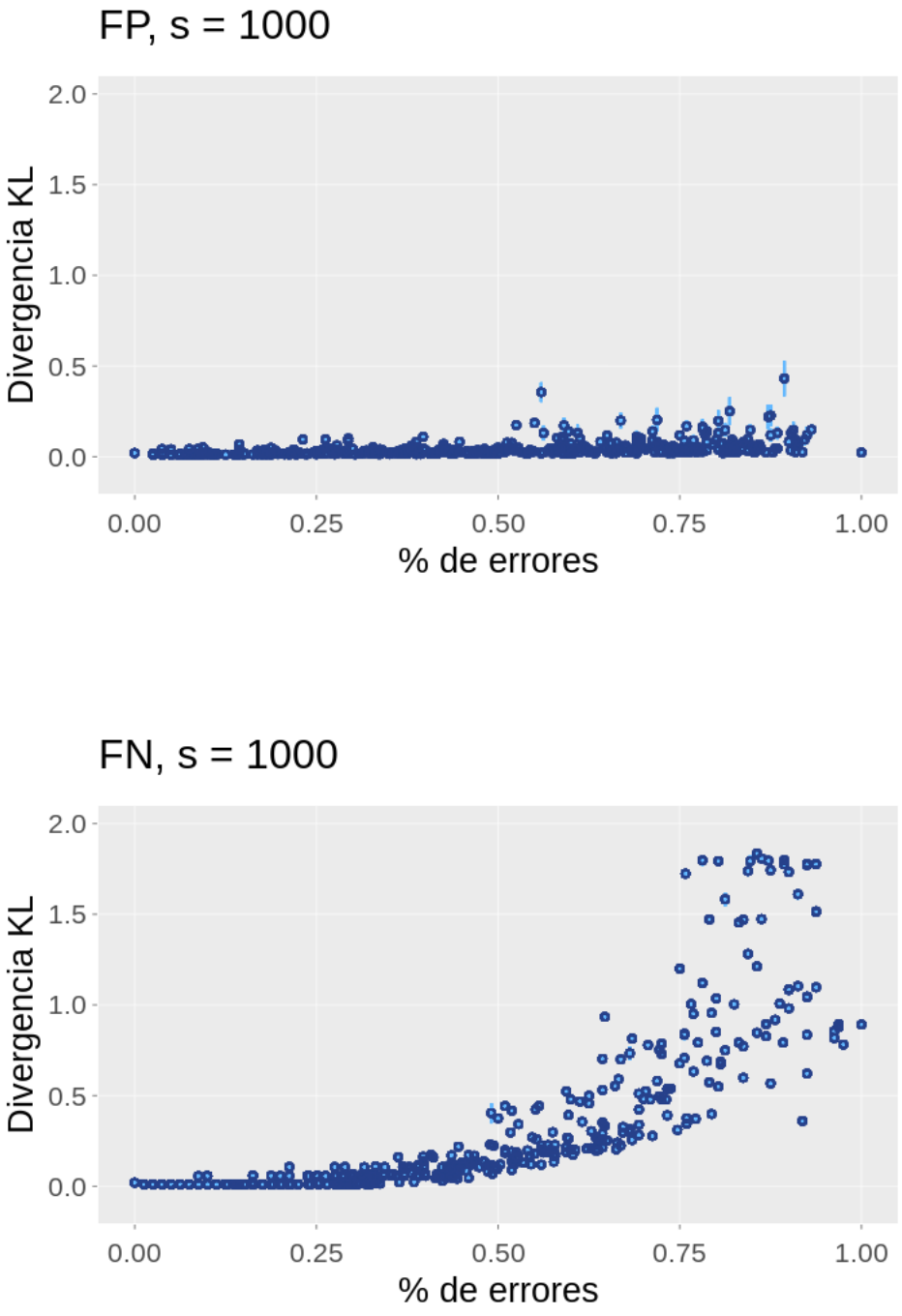


Figura D.4: Comparación de errores para la métrica propuesta (eje X) contra divergencia KL (eje Y) para conjuntos de datos sintéticos de 6 variables, para $s = 1000$. Arriba: % de falsos positivos. Abajo: % de falsos negativos. Cada punto en las gráficas es una estructura diferente.

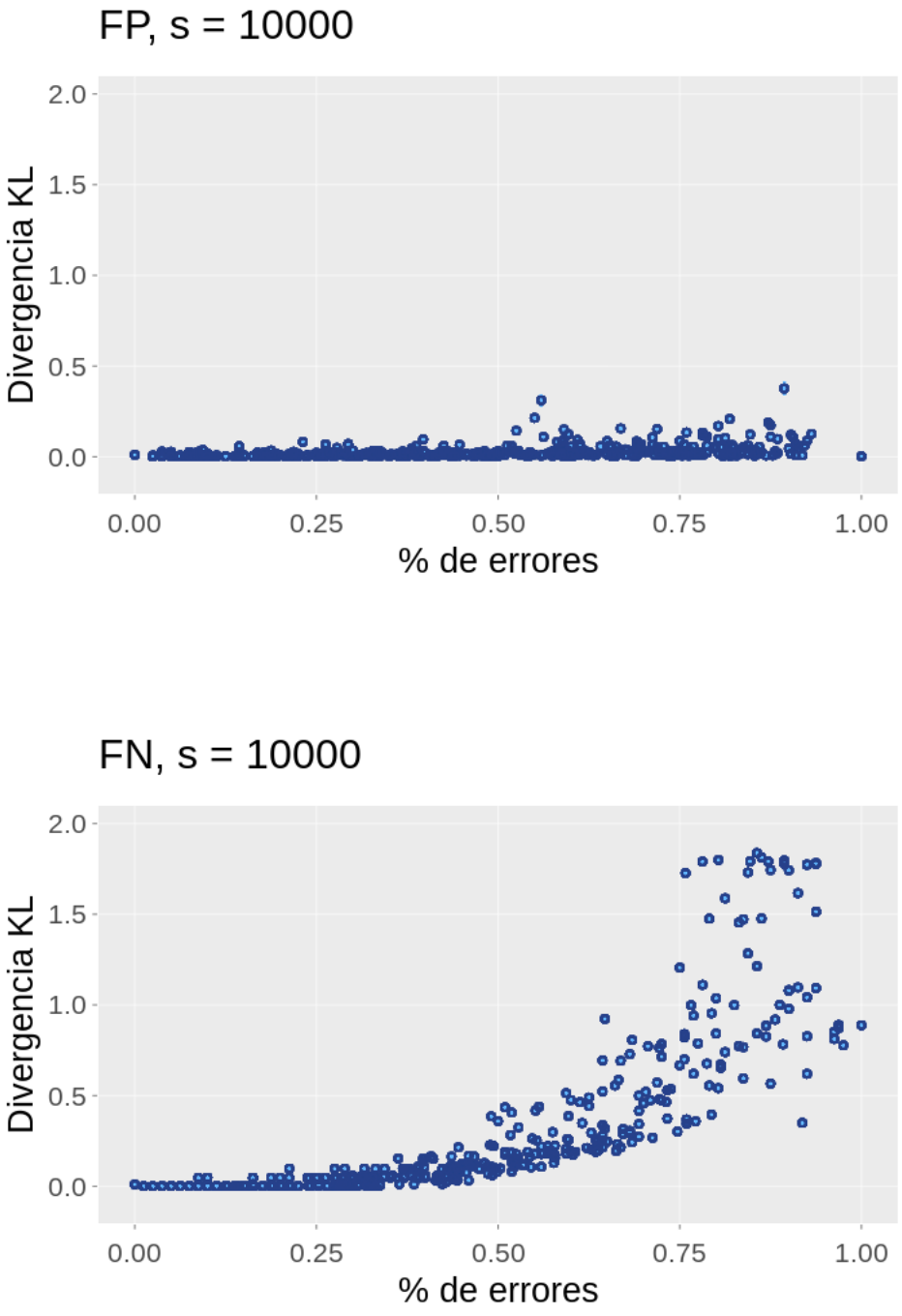


Figura D.5: Comparación de errores para la métrica propuesta (eje X) contra divergencia KL (eje Y) para conjuntos de datos sintéticos de 6 variables, para $s = 10000$. Arriba: % de falsos positivos. Abajo: % de falsos negativos. Cada punto en las gráficas es una estructura diferente.

Bibliografía

A. Agresti. *Categorical Data Analysis*. Wiley, 2nd edition, 2002.

Craig Boutilier, Nir Friedman, Moises Goldszmidt, and Daphne Koller. Context-specific independence in bayesian networks. In *Proceedings of the Twelfth international conference on Uncertainty in artificial intelligence*, pages 115–123. Morgan Kaufmann Publishers Inc., 1996.

Guy Bresler. Efficiently learning ising models on arbitrary graphs. In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, pages 771–782. ACM, 2015.

F. Bromberg, D. Margaritis, and V. Honavar. Efficient Markov network structure discovery using independence tests. *JAIR*, 35:449–485, July 2009.

Ronald Christensen. *Log-linear models and logistic regression*. Springer Science & Business Media, 2006.

Gerda Claeskens, Eugen Pircalabelu, and Lourens Waldorp. Constructing graphical models via the focused information criterion. In *Modeling and Stochastic Learning for Forecasting in High Dimensions*, pages 55–78. Springer, 2015.

Jukka Corander. Labelled graphical models. *Scandinavian Journal of Statistics*, 30 (3):493–508, 2003.

Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

J. Davis and P. Domingos. Bottom-up learning of Markov network structure. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 271–278, 2010.

A Philip Dawid. Conditional independence. *Encyclopedia of statistical sciences, update*, 2:146–153, 1998.

S. Della Pietra, V. Della Pietra, and Lafferty J. Inducing features of random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(4):390–393, 1997.

A. Edera, F. Schlüter, and F. Bromberg. Learning Markov network structures constrained by context-specific independences. *International Journal on Artificial Intelligence Tools (IJAIT)*, 2014a. ISSN 0218-2130.

Alejandro Edera, Yanela Strappa, and Facundo Bromberg. The grow-shrink strategy for learning markov network structures constrained by context-specific independences. In *Ibero-American Conference on Artificial Intelligence*, pages 283–294. Springer, 2014b.

P Svante Eriksen. *Context specific interaction models*. Department of Mathematical Sciences, Aalborg University, 1999.

Tian Gao and Qiang Ji. Efficient score-based markov blanket discovery. *International Journal of Approximate Reasoning*, 80:277–293, 2017.

Niharika Gauraha. Graphical log-linear models: Fundamental concepts and applications. *Journal of Modern Applied Statistical Methods*, 16(1):30, 2017.

Shelby J Haberman. Log-linear models for frequency data: Sufficient statistics and likelihood equations. *The Annals of Statistics*, pages 617–632, 1973.

J. M. Hammersley and P. E. Clifford. Markov random fields on finite graphs and lattices. Unpublished manuscript, 1971.

Søren Højsgaard. Split models for contingency tables. *Computational statistics & data analysis*, 42(4):621–645, 2003.

Søren Højsgaard. Statistical inference in context specific interaction models for contingency tables. *Scandinavian journal of statistics*, 31(1):143–158, 2004.

Wonjun Hwang and Junmo Kim. Markov network-based unified classifier for face recognition. *IEEE Transactions on Image Processing*, 24(11):4263–4275, 2015.

Stanley Kok and Pedro Domingos. Learning markov logic networks using structural motifs. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 551–558, 2010.

D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 2009.

Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

Juri Kuronen, Jukka Corander, and Johan Pensar. Learning pairwise markov network structures using correlation neighborhoods. *arXiv preprint arXiv:1910.13832*, 2019.

S. Lauritzen. *Graphical Models*. Oxford University Press, 1996.

S. Lee, V. Ganapathi, and D. Koller. Efficient structure learning of Markov networks using L1-regularization. In *NIPS*, 2006.

Su-In Lee, Varun Ganapathi, and Daphne Koller. Efficient structure learning of markov networks using l_1-regularization. In *Advances in neural Information processing systems*, pages 817–824, 2007.

Yupeng Li, Stephanie A Pearl, and Scott A Jackson. Gene networks in plant biology: approaches in reconstruction and analysis. *Trends in plant science*, 20(10): 664–675, 2015.

D. Lowd and J. Davis. Improving markov network structure learning using decision trees. *Journal of Machine Learning Research*, 15:501–532, 2014.

Daniel Lowd and Amirmohammad Rooshenas. Learning markov networks with arithmetic circuits. In *AISTATS*, pages 406–414, 2013.

A. McCallum. Efficiently inducing feeatures of conditional random fields. *Proceedings of Uncertainty in Artificial Intelligence (UAI)*, 2003.

Henrik Nyman, Johan Pensar, Timo Koski, and Jukka Corander. Context-specific independence in graphical log-linear models. *Computational Statistics*, pages 1–20, 2014a.

Henrik Nyman, Johan Pensar, Timo Koski, Jukka Corander, et al. Stratified graphical models-context-specific independence in graphical models. *Bayesian Analysis*, 9(4):883–908, 2014b.

J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, San Francisco, CA, 2nd edition, 1988.

Furong Peng, Jianfeng Lu, Yongli Wang, Richard Yi-Da Xu, Chao Ma, and Jingyu Yang. N-dimensional markov random field prior for cold-start recommendation. *Neurocomputing*, 191:187–199, 2016.

Johan Pensar, Henrik Nyman, Juha Niiranen, and Jukka Corander. Marginal pseudo-likelihood learning of markov network structures. *arXiv preprint arXiv:1401.4988*, 2014.

Johan Pensar, Henrik Nyman, Jarno Lintusaari, and Jukka Corander. The role of local partial independence in learning of Bayesian networks. *International Journal of Approximate Reasoning*, 69:91–105, 2016. ISSN 0888-613X. doi: https://doi.org/10.1016/j.ijar.2015.11.008. URL http://www.sciencedirect.com/science/article/pii/S0888613X15001759.

Johan Pensar, Henrik Nyman, Juha Niiranen, and Jukka Corander. Marginal pseudo-likelihood learning of discrete markov network structures. *Bayesian Analysis (2017)*, pages 1–21, 2017.

Pradeep Ravikumar, Martin J Wainwright, and John D Lafferty. High-dimensional Ising model selection using L1-regularized logistic regression. *Annals of Statistics*, 38:1287–1319, 2010. doi: 10.1214/09-AOS691.

F. Schlüter, F. Bromberg, and A. Edera. The IBMAP approach for Markov network structure learning. *Annals of Mathematics and Artificial Intelligence*, pages 1–27, 2014. ISSN 1012-2443.

Federico Schlüter. A survey on independence-based markov networks learning. *Artificial Intelligence Review*, 42(4):1069–1093, 2014.

Federico Schlüter, Yanela Strappa, Diego H Milone, and Facundo Bromberg. Blankets joint posterior score for learning markov network structures. *International Journal of Approximate Reasoning*, 92:295–320, 2018.

M. Schmidt, K. Murphy, G. Fung, and R. Rosales. Structure learning in random fields for heart motion abnormality detection. In *Computer Vision and Pattern*

Recognition, 2008. CVPR 2008. IEEE Conference on, pages 1 –8, june 2008. doi: 10.1109/CVPR.2008.4587367.

P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*. Adaptive Computation and Machine Learning Series. MIT Press, 2nd edition, January 2000.

Peter Spirtes and Clark Glymour. An algorithm for fast recovery of sparse causal graphs. *Social science computer review*, 9(1):62–72, 1991.

Stefanie A Tellex, Thomas Fleming Kollar, Steven R Dickerson, Matthew R Walter, Ashis Banerjee, Seth Teller, and Nicholas Roy. Understanding natural language commands for robotic navigation and mobile manipulation. 2011.

Ioannis Tsamardinos, Laura E Brown, and Constantin F Aliferis. The max-min hill-climbing bayesian network structure learning algorithm. *Machine learning*, 65(1):31–78, 2006.

J. Van Haaren and J. Davis. Markov network structure learning: A randomized feature generation approach. In *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence*, 2012.

J. Van Haaren, J. Davis, M. Lappenschaar, and A. Hommersom. Exploring disease interactions using markov networks. In *Workshops at the Twenty-Seventh AAAI Conference on Artificial Intelligence*, 2013.

Ying-Wooi Wan, Genevera I Allen, Yulia Baker, Eunho Yang, Pradeep Ravikumar, Zhandong Liu, and Maintainer Ying-Wooi Wan. Package 'xmrf'. 2015.