Universidad de Cantabria

Facultad
de
Ciencias

Separación gráfica e independencia condicional en redes bayesianas

(Graphical separation and conditional independence in bayesian networks)

Trabajo Fin de Grado
para acceder al

GRADO EN MATEMÁTICAS

Autor: Merino Iglesias, Daniel

Director: Palazuelos Calderón, Camilo

Julio – 2023

Resumen

Las redes bayesianas y, en general, los modelos gráficos probabilísticos constan de un grafo que recoge relaciones de independencia condicional que se cumplen en la distribución de probabilidad que representan. Esta equivalencia entre el conocimiento que ofrece el grafo y el disponible en la distribución de probabilidad subyacente es lo que abre la puerta a enfoques algorítmicos para el cálculo de consultas probabilísticas con la red bayesiana.

En este Trabajo Fin de Grado, se estudiará la relación entre los conceptos de separación gráfica e independencia condicional en redes bayesianas. Para ello, se demostrará la equivalencia entre la factorización de la distribución de probabilidad y el cumplimiento de las propiedades global y local de Márkov por parte de la red bayesiana con respecto a su grafo. La demostración será autocontenida y no asumirá de partida, como suele ser habitual, que la red bayesiana satisface la propiedad local de Márkov.

Asimismo, se mostrará la aplicación de los resultados anteriores para la construcción de sistemas inteligentes con un conjunto de datos real. Este conjunto de datos contiene información temporal sobre las interacciones de grupos de estudiantes con una aplicación de escritorio para la consecución colaborativa de un objetivo común, así como las puntuaciones asignadas por su profesor a las soluciones presentadas por los grupos. Se construirá una red bayesiana que permita predecir, en distintos instantes de tiempo, si cada grupo de estudiantes conseguirá puntuaciones altas por su desempeño hasta el momento. La implementación se realizará en R con el paquete *bnlearn*.

Abstract

Bayesian networks, and probabilistic graphical models in general, consist of a graph that captures conditional independence relationships present in the probability distribution they represent. This equivalence between the knowledge provided by the graph and the one available in the underlying probability distribution opens the door to algorithmic approaches for computing probabilistic queries with the Bayesian network.

In this TFG (acronym of Bachelor's thesis in Spanish), we will study the relationship between the concepts of graphical separation and conditional independence in Bayesian networks. To do so, we will demonstrate the equivalence between the factorization of the probability distribution and the fulfillment of the global and local Markov properties by the Bayesian network with respect to its graph. The proof will be self-contained and will not assume, as is commonly done, that the Bayesian network satisfies the local Markov property.

Furthermore, we will demonstrate the application of the aforementioned results in building intelligent systems using a real dataset. This dataset contains temporal information about the interactions of student groups with a desktop application for the collaborative achievement of a common goal, as well as the scores assigned by their instructor to the solutions presented by the groups. We will construct a Bayesian network that allows us to predict, at different time points, whether each student group will achieve high scores based on their performance up to that moment. The implementation will be carried out in R using the *bnlearn* package.

Índice general

Índice de figuras

Capítulo 1

Introducción

En este capítulo, explicaremos las motivaciones y objetivos que tenemos en este Trabajo Fin de Grado. También proporcionaremos una breve explicación sobre la organización de la memoria.

1.1. Motivación personal

La educación es uno de los principales pilares del desarrollo personal y profesional del ser humano. En este contexto, los docentes juegan un papel importante al guiar y apoyar a los estudiantes en el camino hacia el éxito académico. Sin embargo, identificar y tratar con estudiantes en riesgo de suspender un trabajo es a menudo un gran desafío debido a la gran cantidad de estudiantes y al tiempo y recursos limitados. Este hecho me llevó a investigar y desarrollar un modelo predictivo basado en redes bayesianas para proporcionar a los docentes una herramienta eficiente y rápida para identificar a los estudiantes con dificultades y brindarles la ayuda necesaria antes de que sea demasiado tarde.

La capacidad de predecir el rendimiento de los estudiantes utilizando modelos predictivos como las redes bayesianas abre nuevas oportunidades para mejorar los resultados educativos. Mediante el uso de métodos probabilísticos y datos recopilados regularmente sobre el rendimiento de los estudiantes, este modelo proporcionará a los docentes una visión imparcial de los estudiantes que corren el riesgo de suspender un trabajo. Esta información permitirá a los docentes ayudar a cada estudiante para que puedan mejorar la situación lo antes posible y brindarles el apoyo que necesitan en un momento crítico.

Este trabajo tiene como objetivo estudiar la relación entre los conceptos de separación gráfica e independencia condicional en redes bayesiana, así como su aplicación para crear un modelo predictivo.

1.2. Objetivos del TFG

El objetivo principal de este Trabajo Fin de Grado tal y como hemos visto anteriormente es estudiar los conceptos de separación gráfica e independencia condicional, para ello deberemos realizar ciertos objetivos secundarios que nos irán ayudando a entender el objeto de estudio.

El primero de estos objetivos sera demostrar la equivalencia entre la factorización de la distribución de probabilidad y el cumplimiento de las propiedades tanto global como local de Markov por parte de la red bayesiana con respecto a su grafo.

El otro objetivo secundario sera la puesta en práctica del objeto de estudio. Se llevará a cabo la construcción de una red bayesiana que permita predecir, en distintos instantes de tiempo, si cada grupo de estudiantes conseguirá puntuaciones altas por su desempeño hasta el momento.

1.3. Organización de la memoria

El resto de la memoria está organizado como sigue. En el capítulo 2 se introducirá al lector a las redes bayesianas, dando distintas definiciones de elementos que las forman y dando algunas de sus características. En el capítulo 3 demostraremos distintos teoremas, lemas y corolarios que nos serán de gran utilidad a la hora de realizar el modelo predictivo. En el capítulo 4 definiremos este modelo y veremos cuál es la pregunta que nos interesa responder. Por último, en el 5, mostraremos los resultados, dando las curvas ROC para cada variable que predecimos en cada instante de tiempo y sus respectivas áreas bajo la curva.

Capítulo 2

Redes bayesianas

En este capítulo vamos a introducir al lector a las redes bayesianas. Primero veremos una serie de definiciones básicas para que queden claros todos los conceptos que vamos a utilizar y se pueda seguir con facilidad. Después veremos la representación de las redes bayesianas y algunas características como la de camino activo, la independencia de variables o caminos bloqueados. Por último, veremos la inferencia de las redes bayesianas sus factores y operaciones. Para las primeras definiciones hemos utilizado [Fio20], mientras que para los apartados de representación e inferencia de redes bayesianas hemos empleado [Cal22].

2.1. Definiciones

Vamos a definir ciertos conceptos antes de comenzar con las redes bayesianas.

Definición 1. Un **grafo finito** es una terna $G = (V, A, \varphi)$ donde $V = V(G)$ es un conjunto finito ,cuyos elementos se denominan nodos, no vació, $A = A(G)$ es un conjunto finito cuyos elementos se llaman aristas y $\varphi : A \to P(V)$ es una aplicación que asocia a cada arista un par de vértices.

Definición 2. Un **grafo dirigido** es una terna $G = (V, A, \varphi)$ donde V, A son como en la definición de grafo y $\varphi : A \to V \times V$ es una aplicación que asociada a cada arista un par de vértices ordenados. Es decir $\varphi(a) = (x, y)$ con $a \in A$ y $x, y \in V$ o, lo que es lo mismo, una arista que empieza en x y acaba en y.

Definición 3. Un **recorrido** en G es una sucesión $x_0a_1x_1a_2x_2 \cdots x_{k-1}a_kx_k$ donde $x_i \in V$ y a_i es una arista incidente a x_{i-1} y a x_i.

Definición 4. Un **camino** es un recorrido en el que todas las aristas son

distintas.

Definición 5. Un **ciclo** es un camino simple (no se repiten los vértices salvo el primero y el último), cerrado (empieza y acaba en el mismo vértice) y no trivial (no es un lazo, es decir, un camino de una única arista que empieza y acaba en el mismo vértice como podría ser $x_0 a_1 x_0$).

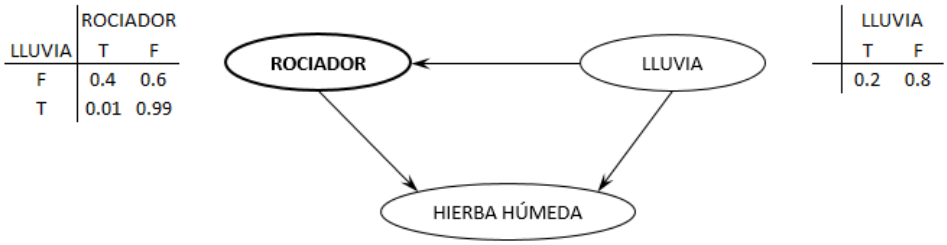
Definición 6. Un **grafo dirigido acíclico** es una terna $G = (V, A, \varphi)$ donde se cumple lo mismo que en la definición de grafo dirigido, pero, además, tiene la característica de no tener ciclos.

2.2. Representación de redes bayesianas

Una red bayesiana es un grafo dirigido acíclico $G = (V, E)$ en que cada $v_i \in V$ se asocia con una variable aleatoria $X_i, i \in \{1, \ldots, n\}$, y cada X_i sigue una distribución condicional $P(x_i \mid pa_G(X_i))$, donde $pa_G(X_i)$ son los padres de X_i en G.
Una red bayesiana describe una distribución factorizándola de acuerdo con

$$P(x_1, \ldots, x_n) = \prod_{i=1}^{n} P(x_i \mid pa_G(X_i))$$

Ahora veamos una red bayesiana para tener así un ejemplo práctico:



LLUVIA	ROCIADOR T	ROCIADOR F
F	0.4	0.6
T	0.01	0.99

LLUVIA T	LLUVIA F
0.2	0.8

ROCIADOR	LLUVIA	HIERBA HÚMEDA T	HIERBA HÚMEDA F
F	F	0	1
F	T	0.8	0.2
T	F	0.9	0.1
T	T	0.99	0.01

Figura 2.1: Red bayesiana ([Wik15])

Como podemos observar aquí, se establece una relación entre la lluvia, un rociador y hierba húmeda y podemos observar cómo el que se produzcan unas u otras afecta al resto.

2.2.1. Camino activo

Un camino activo es una secuencia de nodos y enlaces que conectan dos nodos específicos en la red. Es activo porque se utiliza para inferir la probabilidad de un evento dado, utilizando la información disponible de la red.

Un camino es activo si, para cada trío de variables consecutivas A, B, C se está en una de las siguientes situaciones:

1. $A \leftarrow B \leftarrow C$ y no hemos observado B
2. $A \rightarrow B \rightarrow C$ y no hemos observado B
3. $A \leftarrow B \rightarrow C$ y no hemos observado B
4. $A \rightarrow B \leftarrow C$ y $\exists D \in De_G(B) \cup \{B\}$ tal que D esta observado

Donde $De_G(B)$ son los descendientes de B en G.
Veamos una serie de ejemplos para que sea todo más visual y quede más claro. Para ello, necesitaremos dos redes bayesianas donde los nodos rellenos son aquellos que han sido observados.

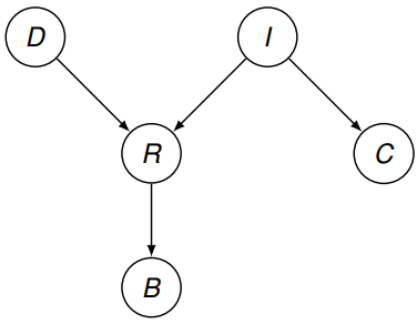


Figura 2.2: Primera red bayesiana

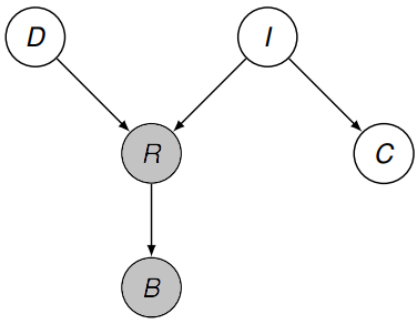


Figura 2.3: Segunda red bayesiana

Ahora veamos un ejemplo para cada uno de los 4 tipos de caminos activos que hemos visto:

1. Cogemos el camino activo $B \leftarrow R \leftarrow D$ en la figura 2.2
2. El camino activo $I \rightarrow R \rightarrow B$ en la figura 2.2
3. Para este caso cogemos $R \leftarrow I \rightarrow C$ en la figura 2.2
4. En este usaremos el camino $D \rightarrow R \leftarrow I$ en la figura 2.3

2.2.2. Visión intuitiva de independencia

Dadas una red bayesiana y dos variables X e Y, vamos a conocer el significado intuitivo de independencia.

- X e Y son dependientes bajo una condición C si solo si el conocimiento sobre una variable influye sobre la otra bajo la condición C.
- X e Y son independientes bajo una condición C si solo si el conocimiento sobre una variable no influye sobre la otra bajo la condición C.

Veamos los distintos casos de conexiones para observar dependencias e independencias.

Caso 1: conexión directa

Si X e Y están conectados por un arco, entonces X e Y son dependientes.

Caso 2: conexión en serie

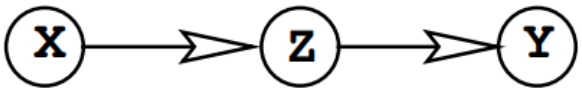


Figura 2.4: Ejemplo de conexión en serie

- Si Z no está observado, X e Y son dependientes. La información puede pasar entre X e Y a través de Z si este último no está observado.
- Si Z está observado, X e Y son independientes. La información no puede ser transmitida entre X e Y a través de Z si este último está observado.

Caso 3: conexión divergente

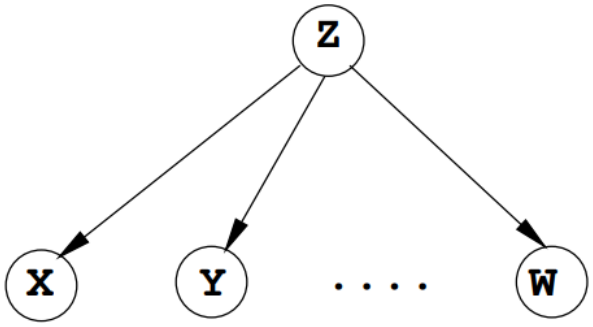


Figura 2.5: Ejemplo de conexión divergente

- Si Z no está observado, X e Y son dependientes. La información puede transmitirse a través de Z entre sus hijos si Z no está observado.
- Si Z está observado, X e Y son independientes. La información no puede transmitirse a través de Z entre sus hijos si Z está observado. Al observar Z se bloquean los caminos de información.

Caso 4: conexión convergente

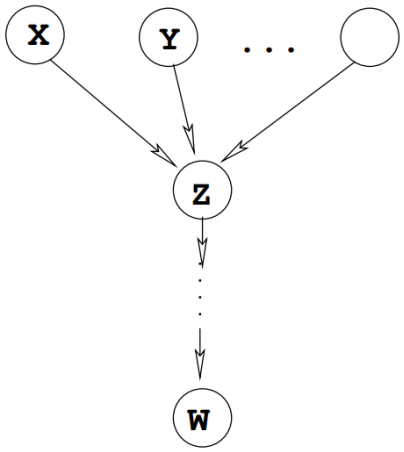


Figura 2.6: Ejemplo de conexión convergente

- Si Z ni ninguno de sus descendientes está observado, X e Y son independientes. La información no puede transmitirse a través de Z entre los padres de este último. La información se filtra a través de Z y sus descendientes.

- Si Z o alguno de sus descendientes está observado, X e Y son dependientes. La información puede transmitirse a través de Z entre los padres de este último si Z o alguno de sus descendientes es observado. Observar Z o sus descendientes abre los caminos de información.

2.2.3. Caminos bloqueados

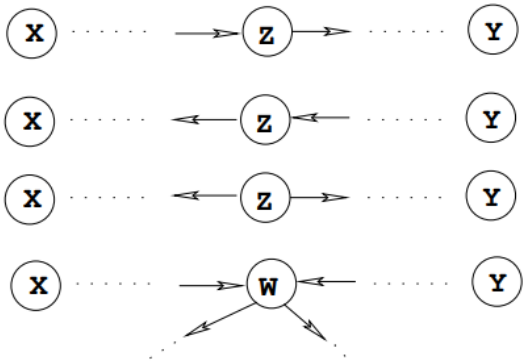


Figura 2.7: Ejemplo de camino bloqueado

Un camino entre X e Y está bloqueado por un conjunto Z de nodos si:

- El camino contiene un nodo z que está en Z y la conexión en ese nodo es en serie o divergente.
- El camino contiene un nodo w tal que él y sus descendientes no están en Z y la conexión con w es convergente.

Supongamos que todos los nodos en Z están observados, entonces un camino X e Y bloqueado por Z implica:

- La información no puede transmitirse a través de z porque observar a z bloquea este camino.
- La información no puede transmitirse a través de w ya que se filtra a través del mismo w.

En ambos casos, la información no puede transmitirse entre X e Y a lo largo del camino. En cambio, si el camino no está bloqueado, la información puede fluir entre X e Y.

2.3. Inferencia en redes bayesianas

2.3.1. Factores y sus operaciones

Sea A un conjunto de variables discretas. Un factor ϕ es una función entre el conjunto de valores de A y $\mathbb{R}$. Al conjunto de variables A se lo denomina ámbito del factor para todo $i, P(x_i \mid pa_G(X_i))$ es un factor con ámbito $Pa_G(X_i) \cup \{X_i\}$.

A continuación, vamos a ver dos operaciones que nos serán de gran utilidad para manipular la distribución de probabilidad. Veremos cada una de ellas junto con su expresión y un ejemplo para entenderlas con mayor facilidad. Una vez que hayamos visto las dos operaciones, veremos un caso de cómo aplicar ambas en una red bayesiana para calcular una distribución de probabilidad.

Producto de dos factores

Veamos su expresión para saber su funcionamiento y después un ejemplo visual. Sean ϕ_1, ϕ_2 y ψ tres funciones, las cuales dependen de los conjuntos de variables $\{a, b\}$, $\{b, c\}$ y $\{a, b, c\}$ respectivamente.

$$\psi(a, b, c) = \phi_1(a, b) \times \phi_2(b, c)$$

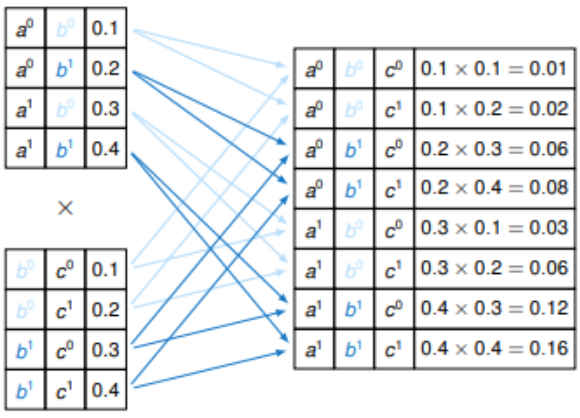


Figura 2.8: Ejemplo del producto de dos factores

Para entenderlo mejor, tomemos uno de los resultados vistos en la figura 2.8, por ejemplo $\psi(a^0, b^1, c^0) = 0{,}06$, el cual se forma con $\phi_1(a^0, b^1) = 0{,}2$ y $\phi_2(b^1, c^0) = 0{,}3$. Por lo tanto, observamos:

$$\psi(a^0, b^1, c^0) = \phi_1(a^0, b^1) \times \phi_2(b^1, c^0) = 0{,}2 \times 0{,}3 = 0{,}06$$

Porque la intersección de los ámbitos ϕ_1 y ϕ_2 es $\{b\}$. Entonces, elegimos aquellas filas en que b es igual para ϕ_1 y ϕ_2.

Marginalización

Al igual que en la operación anterior veremos su expresión y un ejemplo para facilitar su comprensión. Sean τ y ψ dos funciones, las cuales dependen de los conjuntos de variables $\{a, c\}$ y $\{a, b, c\}$ respectivamente.

$$\tau(a, c) = \sum_{b} \psi(a, b, c)$$

Ejemplo:

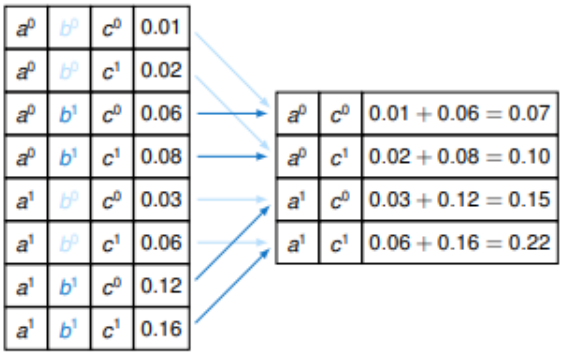


Figura 2.9: Ejemplo marginalización

Veamos uno de los resultados vistos en la imagen como por ejemplo $\tau(a^1, c^0) = 0{,}15$, donde para calcularlo necesitaremos $\psi(a^1, b^0, c^0) = 0{,}03$ y $\psi(a^1, b^1, c^0) = 0{,}12$ con esto ya podemos calcularlo:

$$\tau(a^1, c^0) = \psi(a^1, b^0, c^0) + \psi(a^1, b^1, c^0) = 0{,}03 + 0{,}12 = 0{,}15$$

Porque la intersección de los ámbitos de los ψ del sumatorio es $\{a, c\}$. Entonces elegimos aquellas filas en las que a, c sean iguales para ψ.

Ahora veamos un ejemplo de cómo utilizar ambas operaciones para calcular una probabilidad. En este caso calcularemos $p(b)$ en la figura 2.1.
La red bayesiana se factoriza de la siguiente manera:

$$p(b, c, d, i, r) = \phi_1(b, r) \times \phi_2(c, i) \times \phi_3(d) \times \phi_4(i) \times \phi_5(d, i, r)$$

Una vez factorizada comenzaremos eliminando el nodo d utilizando tanto la definición de producto de dos factores ($\phi_3(d) \times \phi_5(d, i, r)$) como la de

marginalización ($\sum_d \psi_1(d,i,r)$).

$$p(b,c,i,r) = \phi_1(b,r) \times \phi_2(c,i) \times \phi_4(i) \times \overbrace{\sum_d \underbrace{\phi_3(d) \times \phi_5(d,i,r)}_{\psi_1(d,i,r)}}^{\tau_1(i,r)}$$

A continuación eliminaremos el nodo i tal y como hicimos anteriormente.

$$p(b,c,r) = \phi_1(b,r) \times \overbrace{\sum_i \underbrace{\phi_2(c,i) \times \phi_4(i) \times \tau_1(i,r)}_{\psi_2(c,i,r)}}^{\tau_2(c,r)}$$

Repetiríamos este proceso tanto para eliminar el nodo c como el nodo r para así llegar a lo que buscábamos $p(b)$

$$p(b,r) = \phi_1(b,r) \times \overbrace{\sum_c \tau_2(c,r)}^{\tau_3(r)}$$

$$p(b) = \overbrace{\sum_r \underbrace{\phi_1(b,r) \times \tau_3(r)}_{\psi_4(b,r)}}^{\tau_4(b)}$$

2.3.2. Red bayesiana con variables continuas

En el caso de que las variables sigan una distribución normal, su función de densidad de probabilidad no puede resumirse en una tabla sino que se calcula como sigue:

Normal multivariante

$P(x_i \mid pac_G(X_i)) = f(x_i \mid \mu_i, \sigma_i^2) = \frac{e^{-\frac{(x_i-\mu_i)^2}{2\sigma_i^2}}}{\sigma_i\sqrt{2\pi}}$ para todo i

- $\mu_i = \alpha_i + \sum_{x_k \in pac_G(X_i)} \beta_{ik} x_k$
- σ_i^2 no depende de $pac_G(X_i)$

Normal multivariante condicional (RB híbrida)

$P(x_i \mid pac_G(X_i), pad_G(X_i)) = f(x_i \mid \mu_{ij}, \sigma_{ij}^2) = \frac{e^{-\frac{(x_i-\mu_{ij})^2}{2\sigma_{ij}^2}}}{\sigma_{ij}\sqrt{2\pi}}$ para todo i, j tal que

- X_i es continua y $pad_G(X_i)$ es el j-ésimo valor de $Pad_G(X_i)$
- $\mu_{ij} = \alpha_{ij} + \sum_{x_k \in pac_G(X_i)} \beta_{ijk} x_k$

Capítulo 3

Separación gráfica e independencia condicional

Durante este capítulo realizaremos una serie de demostraciones extraídas de [Zha08] con las cuales queremos ver que las siguientes tres afirmaciones son equivalentes en un grafo dirigido acíclico $G = (V, E)$ y una distribución de probabilidad P.

1. $P(V)$ se factoriza de acuerdo con G.
2. $P(V)$ cumple la propiedad global de Markov según G.
3. $P(V)$ cumple la propiedad local de Markov según G.

Para ello será necesario demostrar resultados como la propiedad global de Markov o la propiedad local de Markov. Todos estos lemas, teoremas, corolarios y proposiciones están extraídos de [Zha08].

Primero explicaremos la regla de la cadena de probabilidad que nos será de gran utilidad en los lemas y proposiciones siguientes.

3.0.1. Regla de la cadena de probabilidad

Si tenemos un conjunto de variables $X_1, X_2, \ldots, X_n$ podemos calcular la probabilidad conjunta de estas variables como:

$$P(X_1, X_2, \ldots, X_n) = P(X_1)P(X_2|X_1)\cdots P(X_n|X_1, X_2, \cdots, X_{n-1})$$

Donde $P(X_i|X_1, \cdots, X_{i-1})$ es la probabilidad condicional de la variable X_i dadas todas sus variables padre.

3.1. Equivalencia de redes bayesianas cuando no se observan nodos hoja

Este lema es bastante útil ya que en la proposición 1 lo utilizaremos para hacer una estructura repetitiva con la cual ir eliminando los nodos hojas. Para realizar esta demostración haremos uso de definiciones vistas anteriormente y de la regla de la cadena de probabilidad.

Lema 1. Sea $\mathcal{N}$ una red bayesiana e Y un nodo hoja. $\mathcal{N}'$ es una red bayesiana obtenida al quitarle el nodo Y a $\mathcal{N}$. Si X son todos los nodos de $\mathcal{N}'$.

$$P_{\mathcal{N}}(X) = P_{\mathcal{N}'}(X)$$

Demostración:

$$\begin{aligned} P_{\mathcal{N}}(X) &= \sum_{Y} P_{\mathcal{N}}(X, Y) \\ &= \sum_{Y} [\prod_{W \in X} P(W|Pa(W))] P(Y|Pa(Y)) \\ &= \prod_{W \in X} P(W|Pa(W)) \sum_{Y} P(Y|Pa(Y)) \\ &= \prod_{W \in X} P(W|Pa(W)) \\ &= P_{\mathcal{N}'}(X) \end{aligned}$$

A continuación se procederá a explicar cada uno de los pasos de la demostración:

1. Hacemos uso de la definición de inferencia marginal.
2. La siguiente igualdad se consigue mediante la regla de la cadena de probabilidad.
3. Podemos extraer el productorio del sumatorio ya que Y no pertenece a X y obviamente tampoco es ningún padre de alguno de sus nodos ya que Y es un nodo hoja (lo cual implica que no tiene hijos).
4. $\sum_Y P(Y|Pa(Y)) = 1$.
5. Para llegar a esta igualdad se vuelve a usar la regla de la cadena de probabilidad.

3.2. Equivalencia de redes bayesianas cuando no observamos los nodos fuera de un conjunto ancestral

Que un conjunto sea ancestral implica que para cada nodo de este conjunto sus antecesores también están en él.
La demostración de esta proposición nos será de gran ayuda para demostrar el teorema 1, la propiedad global de Markov, donde podremos coger el conjunto $an(\{x, y\} \cup Z)$ como el conjunto de todos los nodos.
Para realizar esta demostración se hará un proceso iterativo junto con el lema visto anteriormente.

Proposición 1. X conjunto de nodos de una red bayesiana $\mathcal{N}$. Supongamos X ancestral. $\mathcal{N}'$ una red bayesiana que obtenemos de eliminar de $\mathcal{N}$ todos los nodos que no pertenecen a X. Entonces:

$$P_{\mathcal{N}}(X) = P_{\mathcal{N}'}(X)$$

Demostración:
Sea n la cantidad de nodos de $\mathcal{N}$ y sea k la cantidad de nodos de X, llamaremos $Y_1, Y_2, \ldots, Y_k$ a los nodos que no pertenecen a X pero sí a $\mathcal{N}$, los cuales seguirán un orden topológico (con lo que nos aseguramos que Y_k sea un nodo hoja). Sabiendo esto y utilizando el lema 1 visto anteriormente observamos:

$$P_{\mathcal{N}}(X) = P_{\mathcal{N}\setminus\{Y_k\}}(X)$$

Al haber eliminado Y_k de la red bayesiana, entonces Y_{k-1}, si no era un nodo hoja, pasara a serlo. Siguiendo utilizando el lema 1 tenemos:

$$P_{\mathcal{N}\setminus\{Y_k\}}(X) = P_{\mathcal{N}\setminus\{Y_k, Y_{k-1}\}}(X)$$

Siguiendo el mismo razonamiento que hasta ahora acabaríamos con lo siguiente:

$$P_{\mathcal{N}}(X) = P_{\mathcal{N}\setminus\{Y_k\}}(X) = P_{\mathcal{N}\setminus\{Y_k, Y_{k-1}\}}(X) = P_{\mathcal{N}\setminus\{Y_k, Y_{k-1}, \ldots, Y_2, Y_1\}}(X)$$

De donde es obvio observar que:

$$P_{\mathcal{N}'}(X) = P_{\mathcal{N}\setminus\{Y_k, Y_{k-1}, \ldots, Y_2, Y_1\}}(X)$$

Por ende concluimos:

$$P_{\mathcal{N}}(X) = P_{\mathcal{N}'}(X)$$

3.3. Equivalencia entre independencia de dos variables y su distribución de probabilidad como dos funciones

Esta proposición será de gran utilidad para demostrar la propiedad global de Markov ya que se utiliza para cerrar la demostración. Para hacer la demostración la dividiremos en dos casos, primero el caso de izquierda a derecha, en el cual usaremos la regla de la cadena de probabilidad.El segundo caso es la probabilidad de derecha a izquierda, el cual habremos desarrollando la expresión que tenemos y usando sus características.

Proposición 2. $X \perp Y|Z \iff P(X,Y,Z) = f(X,Z)g(Y,Z)$ donde f, g son dos funciones y $\perp$ representa la independencia de las variables a sus lados (X, Y), sin embargo en este caso concreto significa independencia condicional, es decir las variables X, Y son independientes si observamos la variable Z.
Demostración:
Comenzaremos con la implicación de izquierda a derecha:
Supongamos que $X \perp Y|Z$ o lo que es lo mismo que $P(X,Y|Z) = P(X|Z)P(Y|Z)$ y queremos ver que $P(X,Y,Z) = f(X,Z)g(Y,Z)$

$$\begin{aligned} P(X,Y,Z) &= P(X,Y|Z)P(Z) \\ &= P(X|Z)P(Y|Z)P(Z) \\ &= f(X,Z)g(Y,Z) \end{aligned}$$

Y con esto quedaría demostrado, ahora vamos a explicar cada uno de los pasos que hemos hecho:

- En la primera igualdad hemos utilizado la regla de la cadena de probabilidad.
- La segunda igualdad es trivial teniendo en cuenta la hipótesis inicial de la demostración.
- Por ultimo la tercera igualdad sale de nombrar $f(X,Z) = P(X|Z)$ y $g(Y,Z) = P(Y|Z)P(Z)$

Ahora veamos la implicación de derecha a izquierda:
Supongamos que $P(X,Y,Z) = f(X,Z)g(Y,Z)$ y queremos demostrar que $X \perp Y|Z$ o lo que es lo mismo que $P(X,Y|Z) = P(X|Z)P(Y|Z)$. Comenzamos viendo que $P(X,Y|Z) = \frac{P(X,Y,Z)}{P(Z)}$. Esto se debe a que usando la regla de Bayes sabemos que $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$ por lo que $P(B|A) = \frac{P(A,B)}{P(A)} \implies P(A,B) = P(B|A)P(A)$ y podemos escribir la regla de Bayes como $P(A|B) = \frac{P(A,B)}{P(B)}$ donde $A = X \cup Y$ y $B = Z$.

Sabiendo esto y utilizando la hipótesis inicial podemos ver que:

$$\begin{aligned} P(X,Y|Z) &= \frac{f(X,Z)g(Y,Z)}{P(Z)} \\ &= \frac{f(X,Z)g(Y,Z)}{\sum_{X',Y'} P(X',Y',Z)} \\ &= \frac{f(X,Z)g(Y,Z)}{\sum_{X',Y'} f(X',Z)g(Y',Z)} \\ &= \frac{f(X,Z)g(Y,Z)}{\sum_{X'} f(X',Z) \sum_{Y'} g(Y',Z)} \end{aligned}$$

Viendo esto ahora vamos a demostrar que $P(X|Z)P(Y|Z) = P(X,Y|Z)$.Veamos paso a paso:

$$\begin{aligned} P(X|Z) &= \sum_{Y} P(X,Y|Z) \\ &= \sum_{Y} \frac{f(X,Z)g(Y,Z)}{\sum_{X'} f(X',Z) \sum_{Y'} g(Y',Z)} \\ &= \frac{f(X,Z)}{\sum_{X'} f(X',Z)} \frac{\sum_{Y} g(Y,Z)}{\sum_{Y'} g(Y',Z)} \\ &= \frac{f(X,Z)}{\sum_{X'} f(X',Z)} \frac{g'(Z)}{g'(Z)} = \frac{f(X,Z)}{\sum_{X'} f(X',Z)} \end{aligned}$$

Y también podemos ver:

$$\begin{aligned} P(Y|Z) &= \sum_{Y} P(X,Y|Z) \\ &= \sum_{X} \frac{f(X,Z)g(Y,Z)}{\sum_{X'} f(X',Z) \sum_{Y'} g(Y',Z)} \\ &= \frac{\sum_{X} f(X,Z)}{\sum_{X'} f(X',Z)} \frac{g(Y,Z)}{\sum_{Y'} g(Y',Z)} \\ &= \frac{f'(Z)}{f'(Z)} \frac{g(Y,Z)}{\sum_{Y'} g(Y',Z)} = \frac{g(Y,Z)}{\sum_{Y'} g(Y',Z)} \end{aligned}$$

Con lo cual con lo que acabamos de ver queda demostrado que $P(X|Z)P(Y|Z) = P(X,Y|Z)$ ya que simplemente sustituyendo tenemos:

$$P(X|Z)P(Y|Z) = \frac{f(X,Z)}{\sum_{X'} f(X',Z)} \frac{g(Y,Z)}{\sum_{Y'} g(Y',Z)} = P(X,Y|Z)$$

3.4. Separación gráfica implica independencia

Esta proposición también nos será de utilidad para demostrar la propiedad global de Markov, conectando el desarrollo de la demostración con la proposición 1.
Para realizar esta demostración necesitaremos tanto la proposición 1 como la regla de la cadena de probabilidad.

Proposición 3. Sea X, Y, Z conjuntos de nodos disjuntos en una red bayesiana tal que la unión de estos tres conjuntos son todos los nodos de la red bayesiana. Si Z separa X e Y entonces $X \perp Y|Z$.

Demostración:

- Cogemos Z_1 como los nodos en Z que tienen padres en X
- Cogemos $Z_2 = Z \backslash Z_1$
- Como Z separa X e Y
  - Para todo $w \in X \cup Z_1$, $Pa(w) \subseteq X \cup Z$
  - Para todo $w \in Y \cup Z_2$, $Pa(w) \subseteq Y \cup Z$

Como se puede observar en la proposición 2:

$$X \perp Y|Z \iff P(X, Y, Z) = f(X, Z)g(Y, Z)$$

para algunas funciones f, g. Por ende intentaremos llegar a este resultado con lo anteriormente visto.

$$P(X, Y, Z) = \prod_{w \in X \cup Y \cup Z} P(w|Pa(w)) \quad = \prod_{w \in X \cup Z_1} P(w|Pa(w)) \prod_{w \in Y \cup Z_2} P(w|Pa(w))$$

La primera igualdad se consigue mediante la regla de la cadena de probabilidad.
Siguiendo con la demostración podemos observar que $\prod_{w \in X \cup Z_1} P(w|Pa(w))$ es una función de X y Z la cual denotaremos como f y $\prod_{w \in Y \cup Z_2} P(w|Pa(w))$ es una función de Z e Y la cual denotaremos como g.
Por ende podemos observar que hemos llegado al resultado buscado:

$$P(X, Y, Z) = f(X, Z)g(Y, Z)$$

Y como vimos anteriormente si esto se produce entonces $X \perp Y|Z$.

3.5. Propiedad global de Markov

Este teorema es de los más importantes, ya que nos ayuda a demostrar la propiedad local de Markov (corolario 2) además de demostrar la equivalencia entre la afirmación (1) → (2) del teorema 2. Para realizar esta demostración haremos uso de las proposiciones vistas anteriormente.

Teorema 1. Sea una red bayesiana, x e y dos nodos y Z un conjunto de nodos que no contiene a x ni a y. Si Z separa a x e y, entonces $X \perp Y|Z$
Demostración:
Gracias a la proposición 1 podemos decir que $an(\{x, y\} \cup Z)$ es el conjunto de todos los nodos, ya que, como hemos visto en esa proposición, $P_{\mathcal{N}}(an(\{x, y\} \cup Z)) = P_{\mathcal{N}'}(an(\{x, y\} \cup Z))$ siendo $\mathcal{N}'$ tan solo los nodos que forman $an(\{x, y\} \cup Z)$.
Como dice la proposición tan solo necesitamos que el conjunto de nodos al cual le hacemos la probabilidad sea ancestral.

- $x \perp y|Z$ en la red bayesiana $\mathcal{N} \iff$ x $\perp y|Z$ en el conjunto de nodos $\mathcal{N}'$

- Z separa a x e y en la red bayesiana $\mathcal{N} \iff Z$ separa a x e y en el conjunto de nodos $\mathcal{N}'$

- Sea X el conjunto de todos los nodos no separados de x por Z

- Sea Y todos los nodos que no están ni en X ni en Z

Por la proposición 3, como X, Y, Z son conjuntos disjuntos y Z separa X e Y entonces $X \perp Y|Z$.
Como $X \perp Y|Z$, por la proposición 2, deben existir dos funciones f, g tal que

$$P(X, Y, Z) = f(X, Z)g(Y, Z) \qquad (1).$$

Se puede observar que $x \in X$ e $y \in Y$. Sean los conjuntos $X' = X \setminus \{x\}$ e $Y' = Y \setminus \{y\}$ tenemos:

$$\begin{aligned} P(X', x, Z, Y', y) = P(X, Y, Z) &\underset{\text{usando (1)}}{=} f(X, Z)g(Y, Z) \\ &= f(X', x, Z)g(Y', y, Z) \qquad (2) \end{aligned}$$

Consecuentemente queremos demostrar que $P(x, y, Z) = f(x, Z)g(y, Z)$ ya que con esto y la proposición 2 sabríamos que $x \perp y|Z$.

$$\begin{aligned} P(x, y, Z) &\underset{\text{por la definicion de inferencia marginal}}{=} \sum_{X',Y'} P(x, X', Z, y, Y') \\ &\underset{\text{usando (2)}}{=} \sum_{X',Y'} f(X', x, Z)g(Y', y, Z) \\ &= \sum_{X'} f(X', x, Z) \sum_{Y'} g(Y', y, Z) \\ &= f'(x, Z)g'(y, Z) \end{aligned}$$

Con lo cual tal y como hemos dicho antes como $P(x, y, Z) = f(x, Z)g(y, Z)$ sabemos gracias a la proposición 2 que $x \perp y|Z$.

3.6. Recubrimiento de Markov (*Markov Blanket*)

En una red bayesiana, el recubrimiento de Markov de un nodo x es el conjunto de nodos formado por:

- Los padres de x
- Los hijos de x
- Los padres de los hijos de x

3.7. Independencia en una red bayesiana dado el recubrimiento de Markov

Corolario 1. En una red bayesiana, una variable x es condicionalmente independiente de todas las demás variables dado su recubrimiento de Markov

Demostración:
Por la propiedad global de Markov sabemos que si $MB(x)$ separa x de todos los nodos entonces $x \perp$ todos los nodos $|MB(x)$, donde $MB(x)$ es el recubrimiento de Markov del nodo x.
Para demostrar que $MB(x)$ separa x de todas las demás variables de la red bayesiana, es suficiente demostrar que cualquier variable que no está en $MB(x)$ es independiente de x condicionado a $MB(x)$. Es decir cualquier variable y que no esté en $MB(x)$, debe cumplir que $P(x|MB(x), y) = P(x|MB(x))$. Para demostrar esto utilizaremos la definición de condicionamiento en una red bayesiana y la regla de la cadena de probabilidad:

$$P(x|MB(x), y) \underset{\text{Por la regla de Bayes}}{=} P(x, y|MB(x))/P(y|MB(x))$$

Por la regla de la cadena de probabilidad:

$$P(x, y|MB(x)) = P(y|x, MB(x))P(x|MB(x))$$

Sustituyendo en la regla de Bayes:

$$P(x|y, MB(x)) = P(y|x, MB(x))P(x|MB(x))/P(y|MB(x))$$

Como $y \notin MB(x)$, se cumple que x e y son independientes dado $MB(x)$, por lo tanto $P(y|x, MB(x)) = P(y|MB(x))$ y volviendo a sustituir en la regla de Bayes tenemos:

$$P(x|y, MB(x)) = P(x|MB(x))$$

Y esto es lo que queríamos ver, ya que nos lleva a que $MB(x)$ separa x de todas las demás variables, por lo que $x \perp$ todos los nodos $|MB(x)$.

3.8. Propiedad local de Markov

Este resultado nos ayuda con la demostración del teorema 2, ya que demuestra directamente la equivalencia entre las afirmaciones (2) $\rightarrow$ (3). Para realizar esta demostración usaremos la propiedad global de Markov y lo dividiremos en dos casos, cuando $z \in Pa(x)$ y cuando $z \notin Pa(x)$ en los cuales utilizaremos lo visto en el capitulo 2 en la sección 2.2.

Corolario 2. En una red bayesiana, una variable X es independiente de todos sus no descendientes dado sus padres.

Demostración:
Por la propiedad global de Markov es suficiente ver que $Pa(x)$ separa x de los no descendientes de x.

Consideramos un camino entre x y un no descendiente y. Sea z un vecino de x en el camino (es decir z es un padre o un hijo de x).

- Caso 1: $z \in Pa(x)$
  Vamos a tener $z \rightarrow x$ y, por ende, el camino hacia y no será un camino activo, por lo cual todos los caminos entre x e y estarán bloqueados por los $Pa(x)$.

- Caso 2: $z \notin Pa(x)$
  Nos moveremos hacia abajo de z hasta que encontremos un nodo convergente. Este nodo y sus descendientes no están en $Pa(x)$ ya que, si así fuese, estaríamos hablando de un grafo dirigido *cíclico* y nos encontramos en un grafo dirigido acíclico. Por ende, como el nodo convergente

y sus descendientes no pertenecen a $Pa(x)$, no están observados y no se activa la estructura la conexión convergente, por tanto el camino entre x e y está bloqueado.

3.9. Equivalencia entre probabilidades y el grafo

Este teorema es el que estamos intentando demostrar durante todo el capitulo, para hacerlo utilizaremos tanto el teorema 1 como el corolario 2 tal y como vimos anteriormente.

Teorema 2. Sea $P(V)$ una probabilidad conjunta y G un grafo dirigido acíclico sobre un conjunto de variables V. Las siguientes afirmaciones son equivalentes:

1. $P(V)$ se factoriza de acuerdo con G.
2. $P(V)$ cumple la propiedad global de Markov según G.
3. $P(V)$ cumple la propiedad local de Markov según G.

Demostración:

- 1) $\Rightarrow$ 2) Por el teorema 1
- 2) $\Rightarrow$ 3) Por el corolario 2
- 3) $\Rightarrow$ 1)

Vamos a realizar una inducción con el número de nodos. Supongamos que tenemos n nodos en G. En el caso de 1 nodo es trivial. Supongamos cierto para el caso en el que tengamos $n-1$ nodos.
Consideramos el caso de n nodos: Sea x un nodo hoja en G, $V' = V \backslash \{x\}$. Por la propiedad local de Markov, x es independiente de todos los nodos dado $Pa(x)$

$$P(V) = P(V')P(x|V') = P(V')P(x|Pa(x))$$

Sea G' una red bayesiana obtenida de eliminar x de G. Entonces $P(V')$ obedece la propiedad local de Markov de acuerdo con G' y utilizando la hipótesis inicial ,debido a que hay $n-1$ nodos en V', $P(V')$ factoriza de acuerdo a G'. Con lo cual $P(V)$ factoriza de acuerdo con G

Capítulo 4

Modelización

En este capítulo, vamos a estudiar el conjunto de datos que se nos ha proporcionado, además del modelo que vamos a utilizar para predecir las distintas variables que necesitamos.

4.1. Conjunto de datos

En este apartado, pondremos en práctica todo lo aprendido hasta ahora utilizando un caso concreto de alumnos. El conjunto de datos correspondiente a este caso se denomina "collece", el cual recoge información de 105 grupos de alumnos distintos mientras realizan una actividad propuesta ([Cre13]).

Collece recoge el trabajo propuesto, distintas áreas que evaluar para determinar si el trabajo se aprueba o no, como pueden ser coordinación, colaboración, etc. Estas variables son las que con nuestro modelo intentaremos predecir para saber si el grupo aprobará o no. También recoge el progreso de los alumnos en cinco tiempos distintos a lo largo de la realización del trabajo.

Este progreso lo medimos como el conjunto de acciones que van realizando donde tenemos tres tipos de acciones. En el conjunto de datos no se recogen las acciones como tal, sino la sucesión de acciones que se hacen. Por ejemplo la variable `X_t_i_j` donde t es el tiempo en el que se realiza esta sucesión de acciones (que pueden ser 0, 1, 2, 3 y 4) y las variables i, j son la acción que se realiza en primera instancia y la acción que se realiza después respectivamente donde $i, j \in \{0, 1, 2\}$ e i puede ser igual a j ya que se puede realizar la misma acción dos veces consecutivas.

La variable que recoge el trabajo, cuyo nombre en Collece es W tiene valores entre el 0 y el 3.
Por otro lado las columnas que recogen las distintas áreas a evaluar, que en nuestro caso son trabajo en grupo (GTrabajo), disposición a la experimenta-

ción (DispExperim), Coordinación y Colaboración, tan solo recogen valores entre el 0 y el 1, representando el 0 que ese grupo ha suspendido esta área y 1 que el grupo la ha aprobado. Por ultimo, las variables que recogen el progreso a medida que transcurre el tiempo son las denominadas en collece como `X_t_i_j`, tal y como hemos visto antes. Cabe destacar que `X_t_i_j` $\leq$ `X_(t+1)_i_j` ya que `X_t_i_j` recoge todas las sucesiones de acciones i, j realizadas hasta el tiempo t.

En la figura 4.1 podemos observar cómo se ven en la tabla de datos tanto la variable de trabajo W, como las variables que queremos predecir, las cuales son las distintas áreas que queremos evaluar.

W	Coordinacior	GTrabajo	DispExperim	Colaboracior
1	1	1	1	0
1	1	1	0	1
1	1	1	1	1
1	1	1	1	0
1	0	1	0	0
3	0	0	0	0
3	0	1	1	0
3	1	0	0	1
3	0	0	0	0
3	1	1	0	1
1	1	0	0	1
1	0	0	0	0

Figura 4.1: Tabla de variables que predecir y estáticas

Por otro lado en la figura 4.2 podemos visualizar unas observaciones de las distintas variables que recogen la combinación de acciones para $t = 0$.

X_0_0_0	X_0_0_1	X_0_0_2	X_0_1_0	X_0_1_1	X_0_1_2	X_0_2_0	X_0_2_1	X_0_2_2
6	3	7	2	0	2	8	1	22
5	4	3	2	0	1	5	0	20
3	2	0	0	0	2	1	0	28
5	2	2	1	0	1	3	1	34
3	2	3	0	0	2	5	0	14
1	3	1	3	0	0	1	0	0
5	3	0	3	0	0	0	0	0
5	3	0	1	0	2	1	0	0
2	1	0	0	0	0	0	0	0
9	4	1	3	0	2	2	0	12
2	1	1	2	0	0	1	1	0
4	5	0	5	0	0	0	0	0

Figura 4.2: Tabla de variables dinámicas

4.2. Red bayesiana del experimento

El modelo que vamos a utilizar para realizar las predicciones es el siguiente:

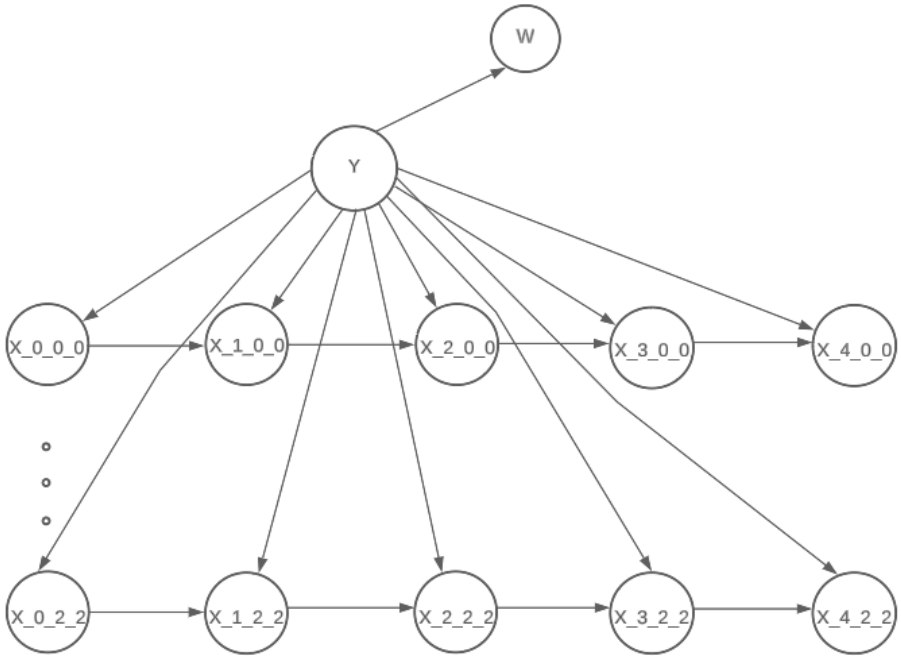


Figura 4.3: Red bayesiana del experimento

Como podemos observar, la variable que queremos predecir Y influye en todas las variables de la red bayesiana. Esta variable será desconocida durante todo el experimento.

A su vez, tan solo tenemos una única W ya que tenemos un trabajo propuesto que es la variable estática, que conocemos antes de empezar el experimento.

Por último, tenemos las variables X, que son las variables dinámicas, las cuales están agrupadas por combinaciones de acciones, donde tan solo dependen de la variable Y y de las variables con la misma combinación de acciones pero con un tiempo anterior. En el caso de observar la variable Y, las distintas filas de combinaciones de acciones se vuelven independientes unas de otras.

Para hacer el estudio, utilizaremos 4 redes bayesianas como la que acabamos

de ver, cada una de ellas se diferencian en la variable Y, ya que tendremos una red bayesiana para Coordinación, otra para GTrabajo, DispExperim y por ultimo Colaboración. Por tanto tendremos que realizar las predicciones en 4 redes bayesianas distintas.

4.3. Definición de variables

- w= variables estáticas (no cambian) y asumimos que las conocemos antes de empezar el experimento. Son m variables estáticas.

- x= variables dinámicas (series temporales). Contamos con n variables dinámicas.

- y= variable que predecir, asumimos que es discreta (en este caso la cogeremos binaria), es desconocida durante todo el experimento.

- t= son los instantes de tiempo que estudiamos durante el experimento.

4.4. ¿Porque nos interesa este modelo?

Primero veamos que:

$$P(x, y, z) = P(y) \prod_{i=1}^{m} P(w_i|y) \prod_{j=1}^{n} (P(x_j^{(1)}|y) \prod_{k=2}^{t} P(x_j^{(k)}|x_j^{(k-1)}, y))$$

Ahora vemos por qué nos interesa este modelo, ya que con él podemos resolver la pregunta que nos hacemos en cada instante de tiempo τ, la cual es $P(y \mid w, x^{(\leq \tau)})$ es decir queremos predecir la variable y observando tanto la variable estática w como las variables dinámicas recogidas hasta ese instante de tiempo $x^{(\leq \tau)}$.

Desarrollemos esta expresión hasta que tengamos algo conocido para así poder calcular esta expresión cuando sea y^0 o cuando sea y^1. Por ello a la hora de realizar este desarrollo utilizaremos la letra k para referirnos tanto al 0 como al 1.

$$
\begin{aligned}
P(y^k \mid w, x^{(\leq\tau)}) &\underset{(1)}{=} \frac{P(y^k, w, x^{(\leq\tau)})}{P(w, x^{(\leq\tau)})} \\
&\underset{(2)}{=} \frac{P(w, x^{(\leq\tau)} \mid y^k)P(y^k)}{P(w, x^{(\leq\tau)})} \\
&\underset{(3)}{=} \frac{P(w \mid y^k)P(x^{(\leq\tau)} \mid y^k)P(y^k)}{P(w, x^{(\leq\tau)})} \\
&\underset{(4)}{=} \frac{(a)}{(b)}
\end{aligned}
$$

Donde tenemos:

$$\prod_{i=1}^{m} P(w_i \mid y^k) \prod_{j=1}^{n} (P(x_j^{(1)} \mid y^k) \prod_{s=2}^{t} P(x_j^{(s)} \mid x_j^{(s-1)}, y^k))P(y^k) = \qquad ((a))$$

$$\sum_{l=0}^{1} \prod_{i=1}^{m} P(w_i \mid y^l) \prod_{j=1}^{n} (P(x_j^{(1)} \mid y^l) \prod_{k=2}^{t} P(x_j^{(k)} \mid x_j^{(k-1)}, y^l))P(y^l) = \qquad ((b))$$

Expliquemos las diferentes igualdades:

1 Usando la definición de probabilidad condicional ($P(x|y,z) = \frac{P(x,y,z)}{P(y,z)}$)

2 Esta igualdad se debe a la regla de Bayes ($P(x,y,z) = P(y,z \mid x)P(x)$)

3 Si observamos y^k entonces como podemos ver en el grafo se bloquean los caminos entre w y x, con lo cual son independientes.

4 Para explicar esta igualdad procederemos incrementalmente:

$$P(w \mid y^k) = \prod_{i=1}^{m} P(w_i \mid y^k) \qquad (*)$$

Ya que cada variable estática es independiente de las demás, al momento de observar y se bloquean los caminos entre las variables estáticas.

$$P(x^{(\leq\tau)} \mid y^k) = \prod_{j=1}^{n} (P(x_j^{(1)} \mid y^k) \prod_{s=2}^{t} P(x_j^{(s)} \mid x_j^{(s-1)}, y^k)) \qquad (**)$$

Esto se debe a que cada serie temporal es independiente, si observamos y pero dentro de cada serie temporal, la variable x no es independiente de su momento de tiempo anterior.

Comencemos viendo el numerador que con lo visto anteriormente es inmediato.

$$P(w \mid y^k)P(x^{(\leq\tau)} \mid y^k)P(y^k) \underset{\text{usando (*) y (**)}}{=} (a)$$

Para el denominador se usa una estrategia similar pero añadiendo el uso de la definición de marginalización y la regla de Bayes.

$$\begin{aligned} P(w, x^{(\leq\tau)}) &\underset{\text{(por la def de marginal)}}{=} P(w, x^{(\leq\tau)}, y^0) + P(w, x^{(\leq\tau)}, y^1) \\ &\underset{\text{(por la regla de Bayes)}}{=} P(w \mid y^0)P(x^{(\leq\tau)} \mid y^0)P(y^0) + P(w \mid y^1)P(x^{(\leq\tau)} \mid y^1)P(y^1) \\ &\underset{\text{usando (*) y (**)}}{=} (b) \end{aligned}$$

De esta manera hemos conseguido transformar la expresión que buscábamos en algo conocido para nosotros.

$$P(y^k \mid w, x^{(\leq\tau)}) = \frac{(a)}{(b)}$$

con $k \in \{0, 1\}$.

Capítulo 5

Experimentación y resultados

En este capítulo pondremos en práctica todo lo visto en el capítulo anterior utilizando el conjunto de datos que nos han proporcionado. Además, analizaremos los resultados obtenidos en cada una de las variables a predecir. Toda la información sobre curvas ROC y el área bajo la curva, esta extraído de [Mur22].

En la fase experimental del trabajo observaremos la capacidad predictiva del modelo en los distintos instantes de tiempo en los que se recogen los datos y para cada una de las variables Y (Coordinación, GTrabajo, DispExperim y Colaboración).

Predeciremos Y con todos los datos disponibles en el instante de tiempo que nos encontremos. Tan solo podremos empezar a predecir la variable Y a partir del instante de tiempo $t = 1$.

Ahora veamos la finalidad de un modelo y el método que utilizaremos para calcular las predicciones.

5.1. Modelo

La finalidad última de un modelo es predecir la variable respuesta en observaciones futuras o incluso en observaciones que el modelo aún no ha visto antes. El error mostrado por defecto tras entrenar un modelo suele ser el error de entrenamiento, el error que comete el modelo al predecir unas observaciones que ya ha visto. Aunque estos errores son útiles para comprender cómo esta aprendiendo el modelo, no es una estimación de cómo se comportará el modelo ante observaciones que no ha visto nunca, ya que este error de entrenamiento es demasiado optimista. Para conseguir una estima-

ción mas certera, hay que recurrir a un conjunto test o emplear estrategias de validación.

Los métodos de validación son estrategias que permiten estimar la capacidad predictiva de los modelos cuando se aplican a nuevas observaciones, haciendo uso únicamente de los datos de entrenamiento. La idea en la que se basa es la siguiente: el modelo se ajusta empleando un subconjunto de las observaciones del conjunto de entrenamiento y se evalúa con las observaciones restantes, viendo qué tan buenas han sido las predicciones de estas últimas observaciones. Este proceso se repite múltiples veces y los resultados se agregan y promedian. Gracias a las repeticiones se compensan las posibles desviaciones que pueden aparecer por el reparto aleatorio de las observaciones. La diferencias entre los distintos métodos suele ser la forma en la que se generan los subconjuntos de entrenamiento.

En este caso utilizaremos la validación cruzada *leave-one-out* (LOOCV por sus siglas en inglés).

5.2. Validación cruzada *leave-one-out*

La validación cruzada *leave-one-out* es una técnica de evaluación de modelos en el campo del aprendizaje automático y la estadística. Es una forma especial de validación cruzada donde se realiza la partición de los datos en conjuntos de entrenamiento y prueba de manera que cada observación se utiliza como el conjunto de prueba una vez, dejando las demás observaciones para entrenar el modelo.

Si se emplea una única observación para calcular el error, este varía mucho dependiendo de qué observación se haya seleccionado. Para evitarlo, el proceso se repite tantas veces como observaciones disponibles, excluyendo en cada iteración una observación distinta, ajustando el modelo con el resto y calculando el error con dicha observación. Finalmente, el error estimado por el LOOCV es el promedio de todos los errores calculados.

El proceso de validación cruzada *leave-one-out* se realiza de la siguiente manera:

1. Se divide el conjunto de datos en N partes, donde N es el número total de observaciones en el conjunto de datos.

2. Para cada iteración, se entrena el modelo utilizando $N - 1$ partes de los datos y se evalúa con la observación restante que se dejó fuera.

3. Se repite el proceso para todas las N observaciones.

4. Al finalizar, se obtiene un promedio de las métricas de evaluación (como la precisión, el error cuadrático medio, etc.) de todas las iteraciones para tener una estimación del rendimiento del modelo.

El método LOOCV permite reducir la variabilidad que se origina si se divide aleatoriamente las observaciones únicamente en dos grupos. Esto es así porque al final del proceso de LOOCV se acaban empleando todos los datos disponibles tanto como entrenamiento como validación. Al no haber una separación aleatoria de los datos, los resultados de LOOCV son totalmente reproducibles.

La principal desventaja de este método es su coste computacional. El proceso requiere que el modelo sea reajustado y validado tantas veces como observaciones disponibles (N) lo que en algunos casos puede ser muy complicado. Esta técnica es considerada más rigurosa en comparación con otras formas de validación cruzada, ya que, como hemos dicho anteriormente, utiliza todas las muestras en el conjunto de datos para evaluar el modelo y proporciona una estimación menos sesgada del rendimiento del mismo.

En resumen, la validación cruzada *leave-one-out* es una técnica de validación que utiliza cada observación individualmente como conjunto de prueba y las demás muestras como conjunto de entrenamiento para evaluar el rendimiento del modelo de manera rigurosa y sin sesgo.

5.3. Curva ROC

Consideremos un problema de clasificación binaria mediante la aplicación de un conjunto de diferentes umbrales a la probabilidad usando un valor τ, derivando el coste relativo de un falso positivo y un falso negativo. Ahora discutiremos el rendimiento resultante.

Para cualquier umbral fijo τ, consideramos la siguiente regla de decisión:

$$\hat{y}_\tau(x) = \mathbb{I}(p(y = 1 \mid x) \geq 1 - \tau)$$

donde $\mathbb{I}$ es la función indicatriz.
Podemos calcular el número empírico de falsos positivos (FP) que surgen en un conjunto de N ejemplos etiquetados de la siguiente manera:

$$FP_\tau = \sum_{n=1}^{N} \mathbb{I}(\hat{y}_\tau(x_n) = 1, y_n = 0)$$

De manera similar, podemos calcular el número empírico de falsos negativos (FN), verdaderos positivos (TP) y verdaderos negativos (TN). Almacenaremos estos resultados en matriz de confusión de clases 2×2, donde el elemento

		Estimate		Row sum
		0	1	
Truth	0	TN	FP	N
	1	FN	TP	P
Col. sum		$\hat{N}$	$\hat{P}$	

Figura 5.1: Matriz de confusión de clases para un problema de clasificación binaria ([Mur22]). *Abreviaciones:* TP es el número de verdaderos positivos, FP es el número de falsos positivos, TN es el número de verdaderos negativos, FN es el número de falsos negativos, P es el número real de positivos, $\hat{P}$ es el número predicho de positivos, N es el número real de negativos, $\hat{N}$ es el número predicho de negativos.

		Estimate	
		0	1
Truth	0	TN/N=TNR=Spec	FP/N =FPR=Type I = Fallout
	1	FN/P=FNR=Miss=Type II	TP/P=TPR=Sens=Recall

Figura 5.2: Matriz de confusión de clases para un problema de clasificación binaria normalizada por fila para obtener $p(\hat{y}|y)$ ([Mur22]). *Abreviaciones:* TNR = tasa de verdaderos negativos, Spec = especificidad, FPR = tasa de falsos positivos, FNR = tasa de falsos negativos, Miss = tasa de error, TPR = tasa de verdaderos positivos, Sens = sensibilidad. Tener en cuenta que FNR=1-TPR y FPR=1-TNR.

(i, j) de esa matriz es el número de veces que un elemento con la etiqueta de clase verdadera i fue clasificado como si tuviese la etiqueta j. En el caso de problemas de clasificación binaria, la matriz resultante se verá como la figura 5.1.

A partir de esta tabla, podemos calcular $p(\hat{y}|y)$ o $p(y|\hat{y})$, según si normalizamos en las filas o en las columnas. Podemos derivar varios estadísticos resumidos a partir de estas distribuciones, como se resume en la figura 5.2. Por ejemplo, la tasa de verdaderos positivos (TPR) se define como:

$$TPR_\tau = p(\hat{y} = 1 \mid y = 0, \tau) = \frac{TP_\tau}{TP_\tau + FN_\tau}$$

y la tasa de falsos positivos (FPR) se define como:

$$FPR_\tau = p(\hat{y} = 1 \mid y = 0, \tau) = \frac{FP_\tau}{FP_\tau + TN_\tau}$$

Ahora podemos representar gráficamente la TPR contra FPR como una

función implícita de τ. Esto se conoce como curva de características de operación del receptor o curva ROC.

5.4. Curva ROC en el experimento

En esta sección observaremos las curvas ROC de cada una de las variables (Colaboración, Coordinación, GTrabajo y DispExperim) que han sido predichas en sus diferentes tiempos. A continuación veremos sus gráficas en el orden en el que las hemos nombrado anteriormente.

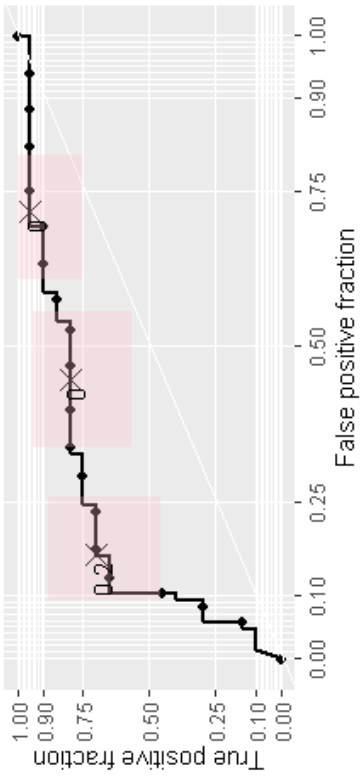


Figura 5.4: Colaboración $t = 2$

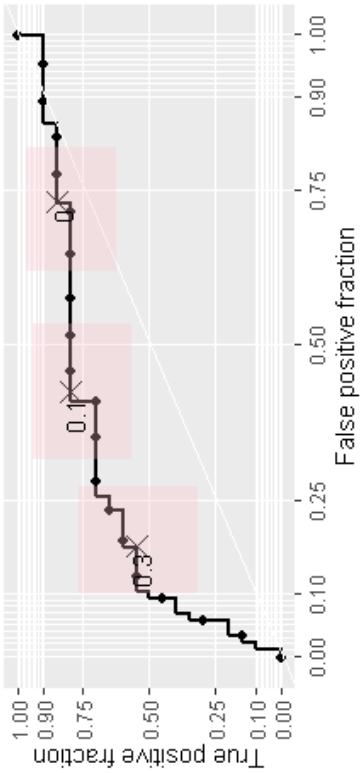


Figura 5.3: Colaboración $t = 1$

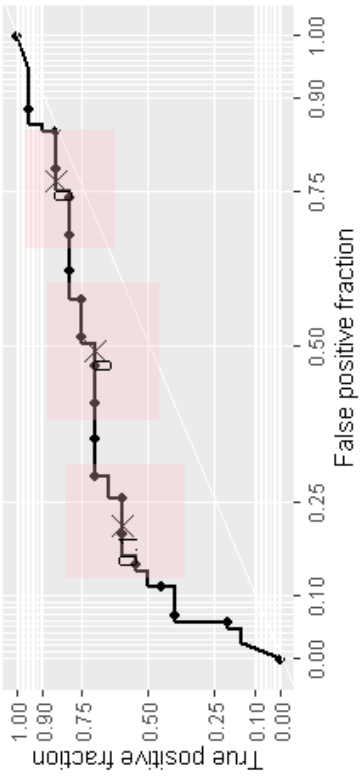


Figura 5.6: Colaboración $t = 4$

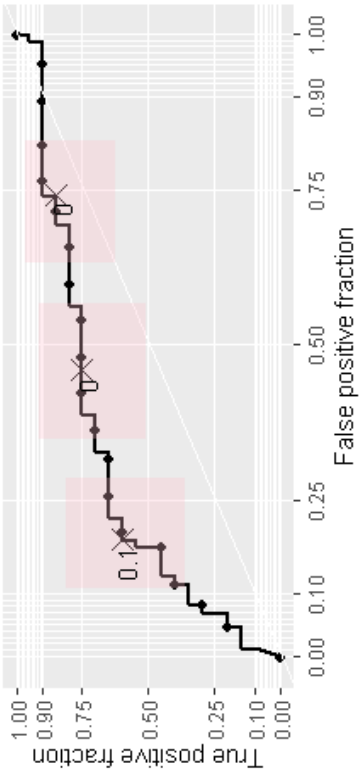


Figura 5.5: Colaboración $t = 3$

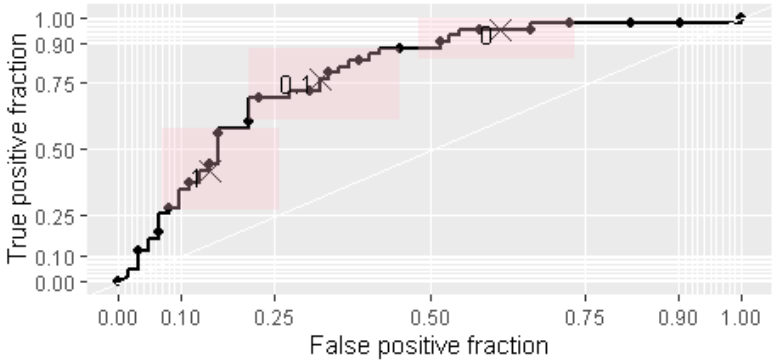


Figura 5.7: Coordinación $t = 1$

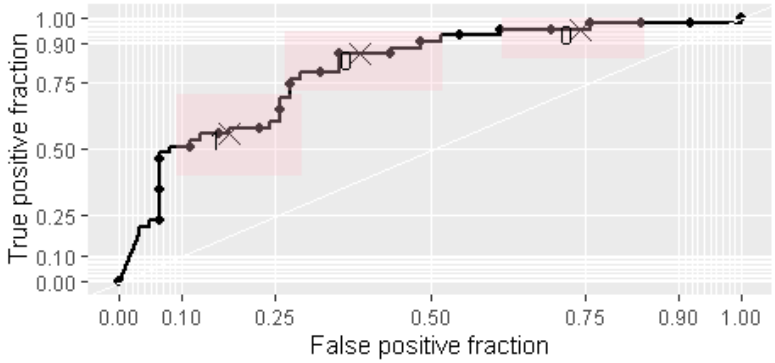


Figura 5.8: Coordinación $t = 2$

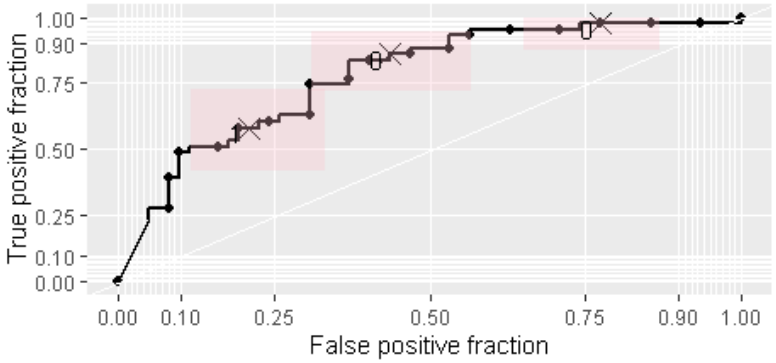


Figura 5.9: Coordinación $t = 3$

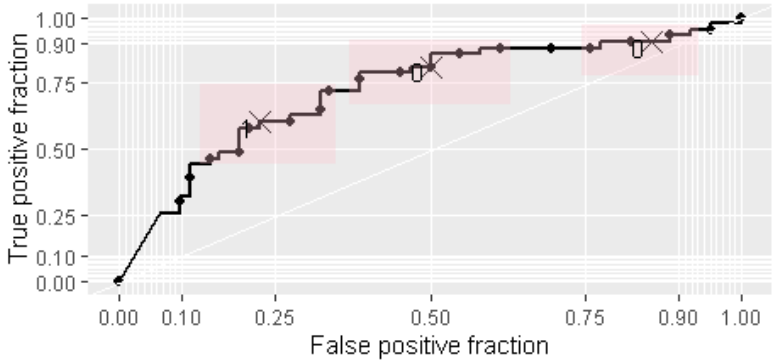


Figura 5.10: Coordinación $t = 4$

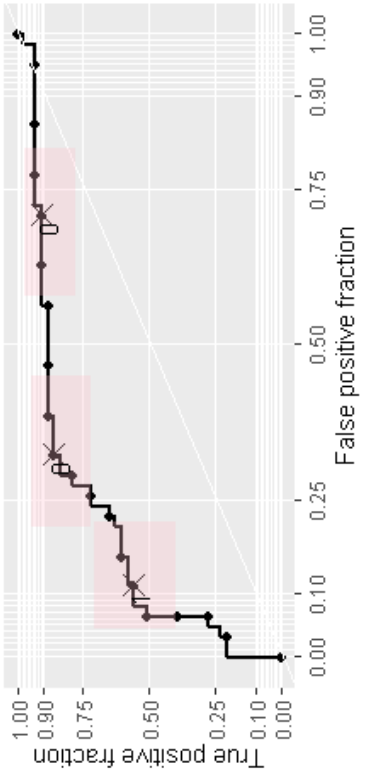


Figura 5.11: GTrabajo $t = 1$

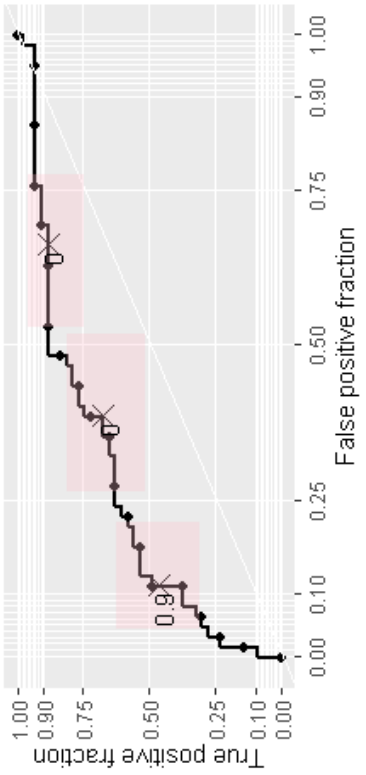


Figura 5.12: GTrabajo $t = 2$

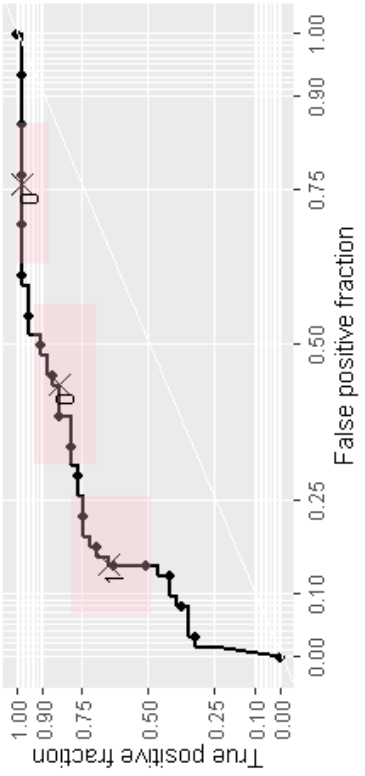


Figura 5.13: GTrabajo $t = 3$

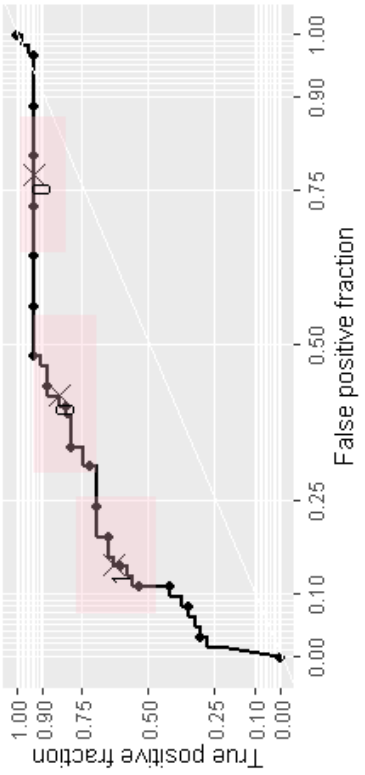


Figura 5.14: GTrabajo $t = 4$

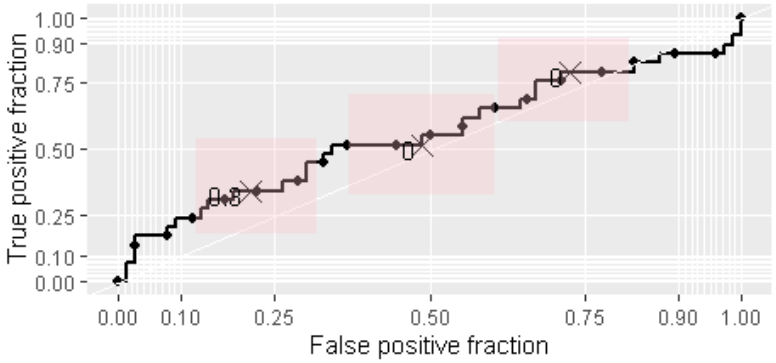


Figura 5.15: DispExperim $t = 1$

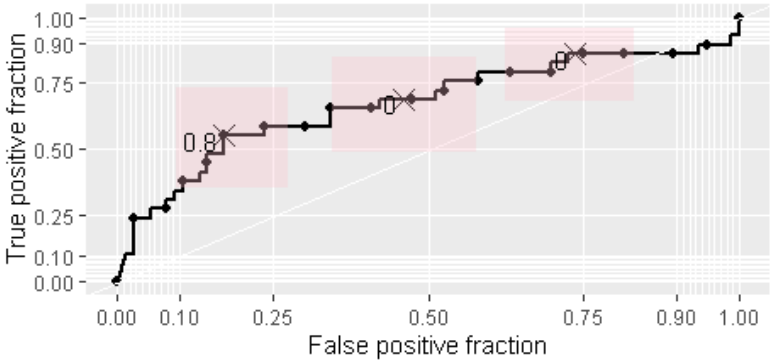


Figura 5.16: DispExperim $t = 2$

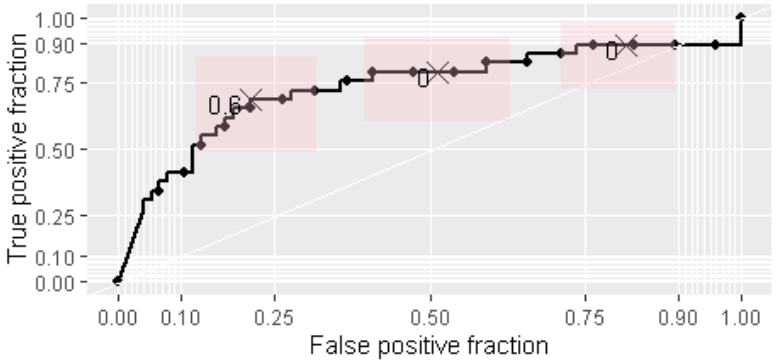


Figura 5.17: DispExperim $t = 3$

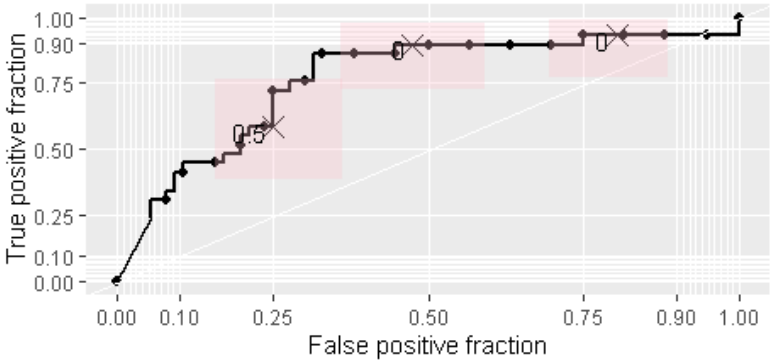


Figura 5.18: DispExperim $t = 4$

5.5. Área bajo la curva

La calidad de una curva ROC se suele resumir en un único número utilizando el área bajo la curva. Los valores de áreas bajo la curva más altos son mejores, siendo el máximo 1.

Otra estadística resumida que se utiliza es la tasa de error igual, también llamada tasa de cruce, que se define como el valor que satisface FPR = FNR. Dado que FNR=1-TPR, podemos calcular la tasa de error igual trazando una línea desde la parte superior izquierda hasta la parte inferior derecha y viendo dónde interseca la curva ROC (ver puntos A y B en la figura 5.19 que se encuentra a continuación). Los valores de la tasa de error más bajos son mejores: el mínimo es 0, correspondiente a la esquina superior izquierda.

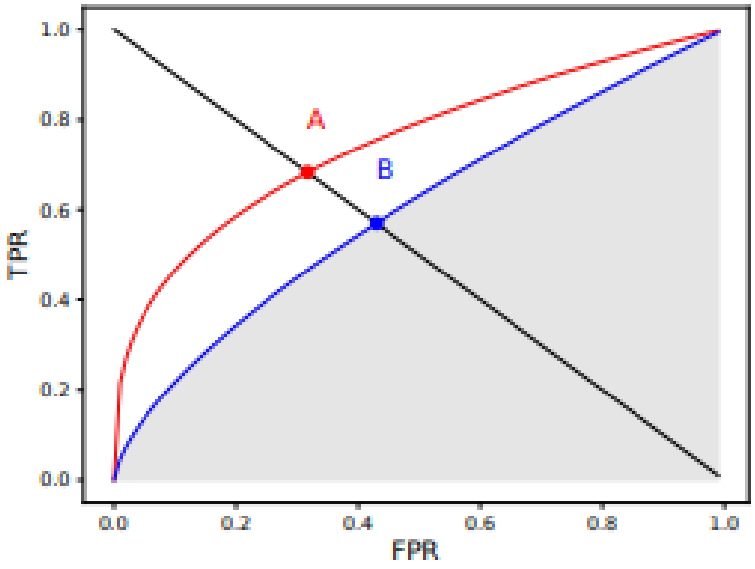


Figura 5.19: Tasa de error igual([Mur22])

5.6. Tabla de áreas en el experimento

Veamos el área bajo las curvas ROC que hemos visto anteriormente para ver qué tan buenas han sido las predicciones realizadas.

Variable que predecir	**Tiempo**			
	$t = 4$	$t = 3$	$t = 2$	$t = 1$
Coordinación	0.74	0.79	0.82	0.81
Colaboración	0.70	0.70	0.78	0.73
GTrabajo	0.81	0.79	0.79	0.74
DispExperim	0.76	0.74	0.67	0.55

Tabla 5.1: Tabla de áreas bajo la curva ROC

Como podemos observar y por lo visto anteriormente los resultados obtenidos son bastante satisfactorios por lo que podemos concluir que nuestro modelo realiza buenas predicciones, ya que el área bajo la curva ROC en casi todos los instantes y para todas las variables que predecir es superior a 0.70.

Capítulo 6

Conclusiones y líneas futuras

En este capítulo nos centraremos en dar un resumen más detallado con una serie de conclusiones, además de las líneas futuras.

6.1. Conclusiones

Comenzamos el Trabajo Fin de Grado ofreciendo una serie de definiciones básicas que resultaron útiles para comprender el concepto de redes bayesianas. Una vez establecidas estas definiciones, nos enfocamos en la representación de las redes bayesianas, destacando dos conceptos fundamentales: los caminos activos y los caminos bloqueados. Estos conceptos desempeñan un papel crucial en el análisis y la interpretación de las redes bayesianas. Además, exploramos algunas de las operaciones realizadas en las redes bayesianas, así como su adaptación a variables continuas. A lo largo de este proceso, utilizamos ejemplos con el fin de brindar una comprensión visual y facilitar la asimilación de los conceptos presentados.

Una vez que hubimos establecido las definiciones de una red bayesiana y hubimos explorado sus diversas características, nos centramos en la demostración del teorema 2. Este teorema establece una serie de equivalencias entre las siguientes afirmaciones:

1. $P(V)$ se factoriza de acuerdo con G.
2. $P(V)$ cumple la propiedad global de Markov según G.
3. $P(V)$ cumple la propiedad local de Markov según G.

Para lograr esto, llevamos a cabo la demostración de varios lemas, proposiciones, corolarios y teoremas. Entre ellos, destacan la propiedad global de Markov y la propiedad local de Markov, que corresponden al teorema 1 y al corolario 2, respectivamente. Estos resultados son fundamentales para establecer las equivalencias enunciadas en el teorema 2.

Posteriormente, procedimos al estudio del conjunto de datos proporcionado, denominado "collece". En este análisis, observamos cada una de las variables presentes en el conjunto y determinamos su tipo. A continuación, construimos la red bayesiana que representa el experimento en cuestión.

Una vez tuvimos la red bayesiana establecida, abordamos la pregunta principal de nuestro estudio: calcular la probabilidad condicional de la variable y dados w y $x^{(\leq\tau)}$, es decir, $P(y \mid w, x^{(\leq\tau)})$. Para facilitar su cálculo, desarrollamos la expresión hasta llegar a una forma que nos permitió utilizar las expresiones conocidas y así obtener la solución requerida.

Finalmente, examinamos los resultados obtenidos en nuestro experimento con el fin de responder a la pregunta planteada anteriormente. Para este propósito, utilizamos las curvas ROC, que también definimos en el análisis. Evaluamos el área bajo la curva para cada una de las variables que estábamos intentando pronosticar en todos los puntos temporales para los cuales disponíamos de datos.

Este análisis nos permitió obtener una medida cuantitativa de la capacidad predictiva de nuestro modelo en cada momento temporal, lo que nos ayudo a evaluar su desempeño y a tomar decisiones fundamentadas con base en los resultados obtenidos.

6.2. Líneas futuras

Por último, establecemos algunas líneas de trabajo futuro:

- Modelizar con otras distribuciones de probabilidad (por ejemplo, de Poisson o multinomial) las variables del conjunto de datos del que disponemos.
- Estudiar la capacidad predictiva de nuestro modelo con otros conjuntos de datos para evaluar como generaliza.
- Explorar la integración de las capacidades de nuestro modelo en otros modelos predictivos (por ejemplo una red neuronal) para alcanzar un equilibrio entre capacidad predictiva e interpretabilidad del modelo.

Bibliografía

[Cal22] Camilo Palazuelos Calderón. *Inferencia probabilística exacta*. Universidad de Cantabria, 2022.

[Cre13] Rafael Duque y Jesus Gallardo Crescencio Bravo. *A groupware system to support collaborative programming: Design and experiences*. 2013. URL: https://www.sciencedirect.com/science/article/pii/S0164121212002439?via%3Dihub.

[Fio20] Mario Fioravanti. *Matemática discreta tercera parte: Grafos*. Universidad de Cantabria, 2020.

[Mur22] Kevin Patrick Murphy. *Probabilistic Machine Learning: An Introduction*. MIT Press, 2022.

[Wik15] Wikipedia. *Red bayesiana*. 2015. URL: https://es.wikipedia.org/wiki/Red_bayesiana.

[Zha08] Nevin L. Zhang. *Introduction to Bayesian Network*. Fall, 2008.