Ejercicios examenes
Métodos Avanzados en Aprendizaje Automático

January 15, 2025

Bloque 1

Ejercicio 1. Se dispone de un conjunto de datos de 1000 patrones y 10 000 atributos, que presentan en muchos casos una alta correlación enre ellos. Dicho conjunto tiene algunas particularidades, por ejemplo:

1. Tiene una primera columna con un identificador único que representa cada patrón.
2. Otra columna indica una medida de peso, pero algunas instancias lo tienen registrado en gramos, otras en kilogramos y otras directamente no disponen de dicha información.
3. Una tercera columna incluye el color que puede tomar uno de estos tres valores: $'black'$, $'white'$ o $'grey$.
4. El resto de atributos contienen información relevante para el modelo y no presentan peculiaridades a destacar.

Especificar los pasos necesarios para procesar esta información y que pueda ser utilizada por un modelo de **regresión lineal**. Para cada paso del proceso indicar qué técnica utilizar y el proceso que conllevaría utilizarla.

Solución: Los métodos para procesar los datos que ya han sido recolectados son los siguientes:

- Limpieza de los datos.
- Transformación de los datos.
- Reducción de la dimensionalidad.

A continuación, se indica el método que se realiza para los datos particulares que se indican arriba:

1. La columna de identificación se puede eliminar dado que no es relevante sobre la clase a evaluar. Sólo se usa para identificar patrones.
2. La columna de pesos debería unificarse en una única unidad de medida. Para ello, se puede preguntar al experto para cambiar a gramos o kilogramos y en rellenar los datos faltantes o en su eliminación si se diera el caso.

3. La columna color se puede eliminar previa transformación a tres nuevas columnas binarias que fueran $'black'$, $'white'$ o $'grey'$ mediante one-hot encoding que hace que todos los colores estén a distancia 1 los unos de los otros.

4. El resto de las columnas, al no presentar peculiaridades, no es necesario hacer nada.

Ejercicio 2. Se considera el método del Discriminate Lineal de Fisher (LDA) para un problema de clasificación binaria. Para cada una de las siguientes afirmaciones, indicar si son verdaderas o falsas junto a una breve justificación.

a) LDA proyecta los datos de dimensión p a dimensión 1 y, en función de un determinado umbral, determina la etiqueta a predecir.

b) LDA es un método más apropiado para datos linealmente separables que para aquellos con una frontera no lineal.

c) El principal objetivo de LDA es transformar los datos a un espacio donde los puntos resultantes tengan mínima varianza intraclase pero máxima varianza inter-clase.

d) LDA funciona muy bien cuando la información contenida en los datos reside en su varianza.

Solución:

a) Verdadero. LDA proyecta los datos de dimensión p a dimensión 1. El proceso sigue os siguientes pasos:

1. LDA encuentra una dirección óptima en el espacio de dimensión p, que está representada por un vector de pesos $\mathbf{w}$, que maximiza la separación entre las dos clases, es decir, **maximiza la distancia entre las medias de las clases** y minimiza la **varianza intra-clase** en esa dirección.
2. Una vez encontrado $\mathbf{w}$, cada punto a clasificar x se proyecta sobre la dirección mediente:
$$z = \mathbf{w}^T x$$
3. Una vez proyectados los puntos en la nueva dimensión, se determina un umbral z_0, en la nueva dimensión, para clasificar los puntos. Este umbral suele estar ubicado entre las medias de las proyecciónes de las dos clases.

b) Verdadero. LDA es más apropiado para datos que son aproximadamente **linealmente separables**, porque asume que las clases pueden ser separadas por una **frontera lineal** y que las clases siguen distribuciones normales con la misma matriz de covarianza.

c) Verdadero. El principal objetivo de LDA es proyectar los datos a un espacio de dimensión menor de tal manera que:

1. Minimice la varianza intra-clase, es decir, los puntos de la misma clase deben estar lo más agrupados posible después de la proyección para que los puntos dentro de cada clase sean similares entre sí.

2. Maximice la varianza inter-clase, es decir, deben estar lo más separadas posible en el nuevo espacio para facilitar la distribución entre las clases, ya que las medias de las clases proyectadas estarán alejadas entre sí.

El objetivo formal de LDA es encontrar un vector de proyección $\mathbf{w}$ tal que la siguiente relación sea maximizada:

$$J(\mathbf{w}) = \frac{\mathbf{w}^T S_B \mathbf{w}}{\mathbf{w}^T S_W \mathbf{w}}$$

Donde:

- S_B es la matriz de dispersión inter-clase, que mide la dispersión entre las medias de las clases proyectadas.
- S_W es la matriz de dispersión entra-clase, que mide la varianza dentro de cada clase.

d) Verdadero. LDA funciona muy bine cuando la información relevante en los datos está contenida en la varianza. Esto es porque el objetivo principal es encontrar una proyección que maximice la **varianza inter-clase** y minimice la **varianza intra-clase**.

Ejercicio 3(O). Para los siguientes ejemplos, indicar qué tipos de métodos no supervisados son más apropiados para resolver cada problema, así como un ejemplo de algoritmo concreto que se podría utilizar en cada caso.

a) Para realizar un análisis de sus ventas, un supermercado desea estimar la probabilidad de compra de cada artículo.

b) Una tienda quere realizar un análisis de los compradores de los últimos tres años y separarlos en función de los items que compran.

c) Una empresa de venta online quiere mostrar a sus usuarios los items «comprados juntos habitualmente».

Solución:

a) Para estimar la probabilidad de compra de cada artículo en un supermercadoel más apropiado puede ser **Modelos de mezcla gaussiana (GMM)**, que permite modelar los datos como una combinación de varias distribuciones. Esto es porque:

- El supermercado puede tener patrones complejos en las ventas que no son fácilmente visibles y **GMM** permite capturar la distribución de probabilidades entre clases sin requerir etiquetas previas.
- **GMM** asume que los datos son una mezcla de múltiples distribuciones gaussianas, que puede reflejar distintos patrones de compra según el tipo de producto, precio, etc.
- Los componentes generados pueden ser interpretados como grupos de productos con comportamento similares. Luego, las probabilidades asignadas a cada componente pueden interpretarse como la "probabilidad de compra" para cada artículo.

Un ejemplo sería el siguiente:

- Suponiendo que el supermercado tiene datos de ventas históricos, se pueden utilizar estos atributos para modelar la probabilidad de compra.
- Después de aplicar **GMM**, se pueden obtener diferentes clusters, cada uno con una probabilidad asociada a la compra de un grupo de productos en función de las características mencionadas. Así, un producto con una alta probabilidad en un determinado componente puede ser considerado como probable de ser comprado.

b) El método no supervisado adecuado sería el **Clustering K-means**. Las razones son las siguientes:

- El algoritmo es útil para agrupar a los compradores en diferentes categorías según sus patrones de compra. Cada cluster representa un grupo de compradores con comportamientos de compra similares.
- Es una técnica simple y eficiente cuando se trabaja con características claras, como cantidades y tipos de ítems que compran.
- Permite identificar grupos de clientes que tienen hábitos de compra similares, lo que puede ser útil para personalizar campañas de marketing o ajustar el inventario según el tipo de cliente.

Un ejemplo del algoritmo sería:

- La tienda podría tener una base de datos con los productos que cada cliente compró en los últimos tres años. Al aplicar K-means, los clientes podrían agruparse en función de los ítems que compran con mayor frecuencia.
- Un cluster podría representar a compradores que prefieren artículos de tecnología, otro a los que compran productos de moda, y así sucesivamente.

c) Para mostrar a los usuarios los ítems que son "comprados juntos habitualmente", el método no supervisado más apropiado sería usando el **algoritmo Apriori**. La razón es la siguiente:

- Este enfoque está diseñado específicamente para descubrir relaciones y patrones frecuentes entre conjuntos de ítems en grandes bases de datos.
- El objetivo del análisis de reglas de asociación es identificar combinaciones de productos que se compran juntos con mayor frecuencia. Se traduce en reglas como "si el cliente compra el ítem A, entonces tiene una alta probabilidad de comprar el ítem B".
- Esto resulta particularmente útil para recomendaciones cruzadas y para implantar sistemas de recomendación tipo "comprados juntos frecuentemente".

Como ejemplo se podría poner:

- Supongamos que la empresa tiene datos históricos de transacciones, donde cada transacción tiene un conjunto de ítems que fueron comprados juntos. aplicando el algoritmo, se pueden generar reglas del tipo:
  - {Leche, pan} $\rightarrow$ {Mantequilla} (Los clientes que compran leche y pan suelen comprar también mantequilla).
  - {Teléfono movil} $\rightarrow$ {Funda protectora}.

El algoritmo calcula la **soporte** (frecuencia de aparición del conjunto de productos en todas las transacciones), la **confianza** (probabilidad de que si se compra el ítem A, también se compre el ítem B, y la **elevación** (relación entre la frecuencia conjunta y las frecuencias individuales), lo que ayuda a identificar las reglas más relevantes.

Ejercicio 3(E). Se dispone de un conjunto de datos pequeño de 6 patrones al que se le aplica un modelo no supervisado con la intención de obtener información sobre la distribución de los datos. En varias fases se obtienen los siguientes resultados:

$$U = [0\ 0\ 0\ 1\ 1\ 1]\ \ V = [0\ 0\ 1\ 1\ 2\ 2]$$

Calcular las métricas de clustering *RandIndex* y *Mutual Information* sobre estos conjuntos. ¿Se puede concluir que el resultado de clustering es bueno?
Nota. *Recuerde que las fórmulas de estas métricas vienen dadas por:*

$$\mathrm{RI} = \frac{a+b}{\binom{N}{2}}, \quad \mathrm{MI}\left(U,V\right) = \sum_{i=0}^{|U|}\sum_{j=0}^{|V|} P\left(i,j\right)\log\left(\frac{P\left(i,j\right)}{P\left(i\right)P'\left(j\right)}\right)$$

siendo:

$$P\left(i\right) = |U_i|/N, \quad P'\left(j\right) = |V_j|/N, \quad P\left(i,j\right) = |U_i \cap V_j|/N$$

Solución:

- **Mutual Information**:

$$|U_1| = 3; |U_2| = 3; |V_1| = 2; |V_2| = 2; |V_3| = 2;$$

$|U_1 \cap V_1| = 2; |U_1 \cap V_2| = 1; |U_1 \cap V_3| = 0; |U_2 \cap V_1| = 0; |U_2 \cap V_2| = 1; |U_2 \cap V_3| = 2;$
$MI(U,V) = 2/6\ln(2) + 1/6\ln(1) + 0 + 0 + 1/6\ln(1) + 2/6\ln(2) = 4/6\ln(2) = 0.46$

- **RandIndex**: Se hará por pasos:

1. Listamos todos los pares posibles de índices para un vector de longitud $n = 6$ con $i \neq j,\ (i,j) = (j,i)$:

$$(0,1), (0,2), (0,3), (0,4), (0,5),$$
$$(1,2), (1,3), (1,4), (1,5),$$
$$(2,3), (2,4), (2,5),$$
$$(3,4), (3,5),$$
$$(4,5)$$

2. Ahora, seguimos las siguientes condiciones para las parejas (i,j):
   - Si $U[i] = U[j]$ y $V[i] = V[j]$: Sumamos 1 a a
   - Si $U[i] = U[j]$ y $V[i] \neq V[j]$: Sumamos 1 a b
   - Si $U[i] \neq U[j]$ y $V[i] = V[j]$: Sumamos 1 a c
   - Si $U[i] \neq U[j]$ y $V[i] \neq V[j]$: Sumamos 1 a d
3. aplicamos la fórmula

$$RI = \frac{a+d}{a+b+c+d} = \frac{2+8}{2+4+1+8} = 0.67$$

Bloque 2

Ejercicio 4(O). Dado el modelo de regresión logística 3-dimensional definido por los parámetros $\theta = \{b = 1, w_1 = 0, w_2 = 1, w_3 = 0\}$, y dado el siguiente conjunto de datos de clasificación binaria, compuesto por 3 patrones y 3 características (x_1, x_2, x_3):

$x_{i,1}$	$x_{i,2}$	$x_{i,3}$	t_i
0	1	0	1
0	-2	1	0
1	2	1	0

a) Calcular la probabilidad de que cada patrón pertenezca a la clase positiva C_1, y de que pertenezca a la clase negativa C_0.

b) Calcular la tasa de acierto (*accuracy*) del modelo sobre este conjunto de datos.

c) ¿Indica la estructura del modelo lineal algo acerca de qué características son relevantes?

Nota. *La transformación logística se define como:*

$$\sigma\left(\mathrm{x}|\mathrm{w},\mathrm{b}\right) = \frac{1}{1 + e^{-(\mathrm{w}^{\mathrm{T}}\mathrm{x}+b)}}$$

Solución:

a) Se tiene que la probabilidad de que un patrón pertenezca a a C_1 es:

$$\mathrm{p}\left(C_1|\tilde{x};\tilde{w},b\right) = \frac{1}{1 + e^{-(\mathrm{w}^{\mathrm{T}}\mathrm{x}+b)}}$$

$$\mathrm{p}\left(C_0|\tilde{x};\tilde{w},b\right) = 1 - \mathrm{p}\left(C_1|\tilde{x};\tilde{w},b\right)$$

Ahora hacemos el cálculo de los vectores:

$x_{i,1}$	$x_{i,2}$	$x_{i,3}$	$p\left(C_1\right)$
0	1	0	0.8808
0	-2	1	0.2689
1	2	1	0.9525

Por lo que se predice que los vectores pertenecen a las clases $1, 0, 1$ respectivamente.

b) La tasa de acierto del modelo sobre el conjunto de datos es de 66.67%

c) La estructura del modelo lineal indica que la única característica que importa es w_2, además de que tiene un cierto sesgo o *bias*.

Ejercicio 4(E). Sea el siguiente conjunto de datos, compuesto por 3 patrones y cuatro características (x_1, x_2, x_3, x_4):

$x_{i,1}$	$x_{i,2}$	$x_{i,3}$	$x_{i,4}$
1	1	1	1
2	1	2	1
1	2	1	2

Dado el modelo lineal definido por los parámetros $\theta = \{b = 0, w_1 = 1, w_2 = 0, w_3 = 1, w_4 = 0\}$.

a) Asumiendo que el modelo lineal es un modelo de regresión, calcular la salida predicha para los 3 ejemplos.

b) Asumiendo que el modelo lineal es un modelo lineal de clasificación (regresión logística):

- Calcular la probabilidad de que el primer ejemplo pertenezca a la clase negativa C_0.
- Calcular la probabilidad de que el segundo ejemplo pertenezca a la clase positiva C_1.

c) ¿Indica la estructura del modelo lineal algo acerca de qué características son relevantes?

Nota. *La transformación logística se define como:*

$$\sigma\left(\mathrm{x}|\mathrm{w},\mathrm{b}\right) = \frac{1}{1 + e^{-(\mathrm{w}^{\mathrm{T}}\mathrm{x}+b)}}$$

Solución:

a) La salida predicha para los ejemplos del enunciado, asumiendo modelo de regresión, es la siguiente:

$x_{i,1}$	$x_{i,2}$	$x_{i,3}$	$x_{i,4}$	θ
1	1	1	1	2
2	1	2	1	4
1	2	1	2	2

b) Asumiendo que el modelo lineal es un modelo de clasificación, la probabilidad de que un dato pertenezca a la clase C_1 se calcula de la siguiente manera:

$$\mathrm{p}\left(C_1|\tilde{x};\tilde{w},b\right) = \frac{1}{1 + e^{-(\mathrm{w}^{\mathrm{T}}\mathrm{x}+b)}}$$

$$\mathrm{p}\left(C_0|\tilde{x};\tilde{w},b\right) = 1 - \mathrm{p}\left(C_1|\tilde{x};\tilde{w},b\right)$$

Teniendo esto en cuenta:

- La probabilidad de que el primer ejemplo pertenezca a C_0 es:

$$\mathrm{p}\left(C_0|x_1,b\right) = 1 - \mathrm{p}\left(C_1|x_1,b\right) = 1 - \frac{1}{1 + e^{-(\mathrm{w}^{\mathrm{T}}\mathrm{x}+b)}} = 0.1192$$

- La probabilidad de que el segundo ejemplo pertenezca a C_1 es:

$$\mathrm{p}\left(C_1|x_2, b\right) = \frac{1}{1 + e^{-(\mathrm{w^T x}+b)}} = 0.8808$$

c) El modelo lineal indica que las características relevantes son X_1 y x_3 y que no hay presencia de sesgo o *bias*.

Ejercicio 5(O). Dada una Máquina de Vectores Soporte de regresión (Support Vector Regression, SVR), definida por el siguiente problema de optimización primal, y su problema dual equivalente:

$$\min_{\mathbf{w}\in\mathbb{R}^n, b\in\mathbb{R}, \xi\in\mathbb{R}^N} \left\{ \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{N}(\xi_i + \xi_i^*) \right\}$$

$$\text{s.a.} \quad \begin{cases} \mathbf{w}^\top \mathbf{x}_i + b - y_i \leq \epsilon + \xi_i, \\ y_i - \mathbf{w}^\top \mathbf{x}_i - b \leq \epsilon + \xi_i^*, \\ \xi_i, \xi_i^* \geq 0, \\ 1 \leq i \leq N, \end{cases}$$

$$\Longleftrightarrow$$

$$\min_{\alpha,\alpha^*\in\mathbb{R}^N} \left\{ \frac{1}{2}(\alpha^* - \alpha)^\top \mathbf{X}\mathbf{X}^\top(\alpha^* - \alpha) + c(\alpha^* + \alpha)^\top \mathbf{1} - (\alpha^* - \alpha)^\top \mathbf{y} \right\}$$

$$\text{s.a.} \quad \begin{cases} (\alpha^* - \alpha)^\top \mathbf{1} = 0, \\ 0 \leq \alpha, \alpha^* \leq C. \end{cases}$$

a) Explicar el significado del parámetro ϵ. ¿Valores más grandes de ϵ producen mayor o menor tendencia a sobre ajuste?

b) Explicar qué es un vector soporte (*support vector*) en este caso.

c) ¿Qué ventajas supone el hecho de utilizar el problema de optimización dual, frente al primal?

Solución:

a) En el contexto de las SVR, el parámetro ϵ establece un margen de tolerancia alrededor de las predicciones del modelo. Esto significa que si el valor predicho de una muestra cae dentro de una franja de $\pm\epsilon$ alrededor del valor real, entonces no se incurre en ninguna penalización en la función de costo. Con un valor de ϵ mayor, más alta es la tolerancia en la cual los errores dentro de ese margen no son penalizados. Pero un valor demasiado alto de ϵ podría hacer que el modelo ignore variaciones significativas, resultando en un ajuste insuficiente y aumentando el error de entrenamiento.

b) En el caso de una SVR, un vector soporte es una muestra de los datos de entrenamiento cuya predicción por parte del modelo cae **exactamente** en el limite del margen de tolerancia ϵ o **fuera** de este margen. Estos puntos son los únicos que afectan directamente al modelo porque influyen en la función objetivo y los parámetros $\mathbf{w}$ y b.

c) El problema dual en SVR permite el uso de kernels, mejora la eficiencia en problemas de alta dimensionalidad y facilita un modelo más esparso, lo cual hace que sea generalmente preferible frente al problema primal en escenarios de regresión y clasificación con SVM.

Ejercicio 5(E). Dada una Máquina de Vectores Soporte (Support Vector Machine, SVM) de clasificación, definida por el siguiente problema de optimización primal, y su problema dual equivalente:

$$\min_{\mathbf{w}\in\mathbb{R}^n, b\in\mathbb{R}, \xi\in\mathbb{R}^N} \left\{ \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{N}\xi_i \right\} \quad \text{s.a.} \quad \begin{cases} y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 - \xi_i, \\ \xi_i \geq 0, \\ 1 \leq i \leq N, \end{cases}$$

$$\Longleftrightarrow$$

$$\min_{\alpha\in\mathbb{R}^N} \left\{ \frac{1}{2}\alpha^\top \mathbf{X}\mathbf{X}^\top \alpha - \alpha^\top \mathbf{1} \right\} \quad \text{s.a.} \quad \begin{cases} \alpha^\top \mathbf{y} = 0, \\ 0 \leq \alpha \leq C. \end{cases}$$

a) Explicar qué es un vector soporte (*support vector*) en este caso.

b) Explicar en qué consiste en este caso el truco del núcleo (*kernel trick*). No es necesario introducir formalismos, basta con una explicación intuitiva.

c) El problema dual de las SVMs se suele resolver con el algoritmo de Sequential Minimal Optimization (SMO). Explicar cuántos coeficientes duales se actualizan en cada iteración de SMO y por qué.

Solución:

a) Los vectores soporte son los puntos de entrenamiento asociados a multiplicadores de Lagrange no nulos ($\alpha_i > 0$), y son los únicos puntos que influyen en la solución óptima de la SVM.

b) El *kernel trick* es una técnica en las SVM que permite trabajar con datos en espacios de características de alta dimensión sin tener que calcular explicitamente esas características adicionales. Esto es útil porque, en muchos casos, los datos con son linealmente separables en su espacio original, pero sí en un espacio de mayor dimensión.

c) En cada iteración del algoritmo **SMO** se actualizan exactamente **dos coeficientes duales**: α_i y α_j. Esto se hace para simplificar el problema de optimización, ya que optimizar más de dos coeficientes al mismo tiempo implicaría resolver un problema más complejo, de múltiples variables en cada paso.

Ejercicio 6(O). Considerando las Redes Neuronales Profundas (Deep Neural Networks, DNN) estudiadas durante el curso:

a) ¿En qué consiste el aumento de datos (*data augmentation*)? ¿Cuál es su objetivo?

b) Indicar de forma razonada qué arquitectura especializada se usaría para abordar los siguientes problemas:

a) Generar dígitos manuscritos artificiales que simulen ser reales.

b) Reconstruir registros de voz eliminando el ruido.

c) Predecir la evolución de una serie temporal.

d) Distinguir el tipo de objeto que aparece en una imagen.

c) Se quiere entrenar una DNN para que determine la ruta óptima que debe seguir un robot en un cierto entorno. ¿Qué tipo de paradigma de Aprendizaje Automático se puede usar y por qué?

Solución:

a) El aumento de datos o *data augmentation* consiste en la generación de datos nuevos a partir de los datos que ya se tienen. El objetivo de esta técnica es mejorar la generalización de los modelos.

b)
- Para generar dígitos manuscritos artificiales que simulen ser reales es mejor una Red Generativa Antagónica o **GAN**.
- La arquitectura más adecuada para la reconstrucción de registros de voz es la arquitectura **Autoencoder**.
- Para predecir la evolución de una serie temporal la mejor arquitectura es una **Red Neuronal Recurrente**.
- Para distinguir el tipo de objeto que aparece en una imagen la mejor arquitectura es una **Red Neuronal Convolucional**.

c) El paradigma que se puede usar es el **Aprendizaje por Refuerzo Profundo** debido a su capacidad para manejar decisiones secuenciales, explorar soluciones y adaptarse a entornos complejos.

Ejercicio 6(E). Considerando las Redes Neuronales Profundas (Deep Neural Networks, DNNs) estudiadas durante el curso:

a) ¿En qué consiste el aprendizaje por transferencia (*transfer learning*)?

b) ¿En qué consiste la técnica de Dropout para regularizar una DNN?

c) ¿De qué manera se entrena un Autoencoder?¿Cuál sería la salida objetivo (*target*) correspondiente a cada patrón?

d) Poner un ejemplo de problema real que se pueda resolver utilizando aprendizaje por refuerzo (*reinforcement learning*).

Solución:

a) El aprendizaje por transferencia consiste en usar los pesos de un modelo entrenado satisfactoriamente en un problema similar en un nuevo modelo.

b) La técnica del *Dropout* es una técnica de regularización que desactiva aleatoriamente un porcentaje de neuronas durante el entrenamiento para prevenir el sobreajuste y mejorar la generalización

c) Un *Autoencoder* se entrena usando un dato x como entrada y *target*. El objetivo es aprender una representación comprimida de los datos que permita reconstruirlos con precisión.

d) Un ejemplo de problema real que se pueda resolver utilizando aprendizaje por refuerzo sería un robot a la hora de encontrar una ruta óptima en un cierto entorno.

Bloque 3

Ejercicio 7(O). Se debe calcular $P\left(D|G=g^0,S=s^0\right)$ usando el algoritmo de eliminación de variables para la red bayesiana de la **Figura 1(O)**. Para ello realizar los siguientes pasos:

a) Escribir la factorización de la distribución de probabilidad $P\left(D,G,I,S,L\right)$ de la red.

b) Escribir $P\left(D|G=g^0,S=s^0\right)$ a partir de la fórmula anterior (sin cálculos numéricos).

c) A partir de la fórmula anterior eliminar primero la variable L calculando los factores intermedios.

d) Finalmente, calcular los valores numéricos finales para $P\left(D|G=g^0,S=s^0\right)$ mediante la eliminación de la variable I y normalizando.

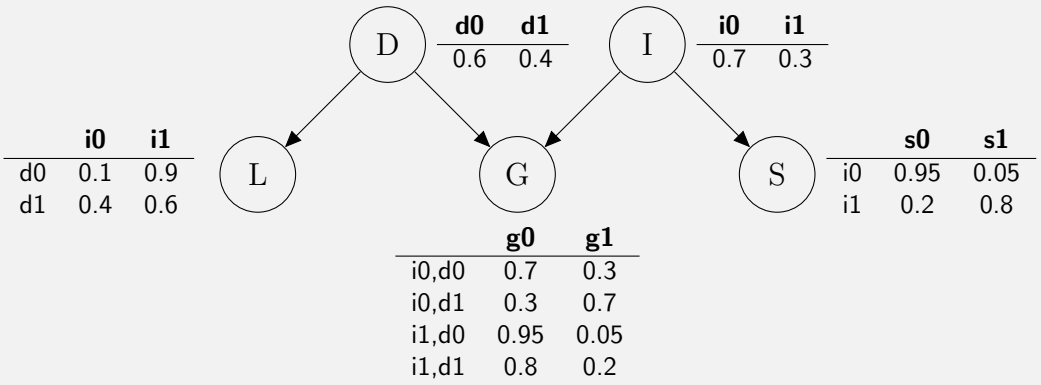

Figura 1(O)

Solución:

a) La factorización de la distribución de probabilidad es la siguiente:

$$P\left(D,G,I,S,L\right)=P\left(L|D\right)P\left(D\right)P\left(G|D,I\right)P\left(I\right)P\left(S|I\right)$$

b) La probabilidad $P\left(D|G=g^0,S=s^0\right)$, usando Bayes, queda de la siguiente manera:

$$P\left(D|G=g^0,S=s^0\right)=\frac{P\left(D,G=g^0,S=s^0\right)}{P\left(G=g^0,S=s^0\right)}$$

Descomponiendo $P\left(D,G=g^0,S=s^0\right)$, usando la factorización, nos queda:

$$P\left(D,G=g^0,S=s^0\right)=\sum_I\sum_L P\left(L|D\right)P\left(D\right)P\left(G=g^0|D,I\right)P\left(I\right)P\left(S=s^0|I\right)$$

El denominador $P\left(G=g^0,S=s^0\right)$ es la probabilidad marginal:

$$P\left(G=g^0,S=s^0\right)=\sum_D\sum_I\sum_L P\left(L|D\right)P\left(D\right)P\left(G=g^0|D,I\right)P\left(I\right)P\left(S=s^0|I\right)$$

Por lo que podemos reescribir la fórmula de la siguiente manera:

$$P\left(D|G=g^0,S=s^0\right)=\frac{\sum_I\sum_L P\left(L|D\right)P\left(D\right)P\left(G=g^0|D,I\right)P\left(I\right)P\left(S=s^0|I\right)}{\sum_D\sum_I\sum_L P\left(L|D\right)P\left(D\right)P\left(G=g^0|D,I\right)P\left(I\right)P\left(S=s^0|I\right)}$$

c) Para eliminar la variable L, nos vamos a aprovechar de que aparece únicamente en $P(L|D)$. Como $\sum_L P(L|D) = 1$, la fórmula se simplifica:

$$P\left(D|G=g^0, S=s^0\right) = \frac{\sum_I P(D)\, P(G=g^0|D,I)\, P(I)\, P(S=s^0|I)}{\sum_D \sum_I P(D)\, P(G=g^0|D,I)\, P(I)\, P(S=s^0|I)}$$

d) El cálculo de $P(D|G=g^0, S=s^0)$ queda de la siguiente forma:

$$P\left(D|G=g^0, S=s^0\right) = \frac{P(D, G=g^0, S=s^0)}{\sum_D P(D, G=g^0, S=s^0)}$$

Ahora vamos con los cálculos de $P(D, G=g^0, S=s^0)$:

- Para $D = d^0$.

$$\begin{aligned} P\left(d^0, g^0, s^0\right) &= P\left(D=d^0\right) \cdot \sum_I P\left(G=g^0|D=d^0, I\right) P(I)\, P\left(S=s^0|I\right) \\ &= 0.6 \cdot \big[P\left(G=g^0|D=d^0, I=i^0\right) P\left(I=i^0\right) P\left(S=s^0|I=i^0\right) \\ &\quad + P\left(G=g^0|D=d^0, I=i^1\right) P\left(I=i^1\right) P\left(S=s^0|I=i^1\right)\big] \\ &= 0.6 \cdot \left[0.7 \cdot 0.7 \cdot 0.95 + 0.95 \cdot 0.3 \cdot 0.2\right] = 0.3135 \end{aligned}$$

- Hacemos lo mismo para $D = d^1$

$$\begin{aligned} P\left(d^1, g^0, s^0\right) &= P\left(D=d^1\right) \cdot \sum_I P\left(G=g^0|D=d^1, I\right) P(I)\, P\left(S=s^0|I\right) \\ &= 0.4 \cdot \big[P\left(G=g^0|D=d^1, I=i^0\right) P\left(I=i^0\right) P\left(S=s^0|I=i^0\right) \\ &\quad + P\left(G=g^0|D=d^1, I=i^1\right) P\left(I=i^1\right) P\left(S=s^0|I=i^1\right)\big] \\ &= 0.4 \cdot \left[0.3 \cdot 0.7 \cdot 0.95 + 0.8 \cdot 0.3 \cdot 0.2\right] = 0.099 \end{aligned}$$

Con los cálculos realizados obtenemos:

$$P\left(D=d^0|G=g^0, S=s^0\right) = \frac{0.3135}{0.3135 + 0.099} = 0.76$$

$$P\left(D=d^1|G=g^0, S=s^0\right) = \frac{0.099}{0.3135 + 0.099} = 0.24$$

Ejercicio 7(E). Se debe calcular $P\left(D|I=i^0, L=l^0\right)$ usando el algoritmo de eliminación de variables para la red bayesiana de la **Figura 1(E)**. Para ello, realizar los siguientes pasos:

a) Escribir la factorización de la distribución de probabilidad $P\left(D, G, I, S, L\right)$ de la red.

b) Escribir $P\left(D|I=i^0, L=l^0\right)$ a partir de la fórmula anterior (sin cálculos numéricos).

c) A partir de la fórmula anterior, eliminar primero la variable S calculando los factores intermedios.

d) Finalmente, calcular los valores numéricos finales para $P\left(D|I=i^0, L=l^0\right)$ mediante la eliminación de la variable G y normalizando.

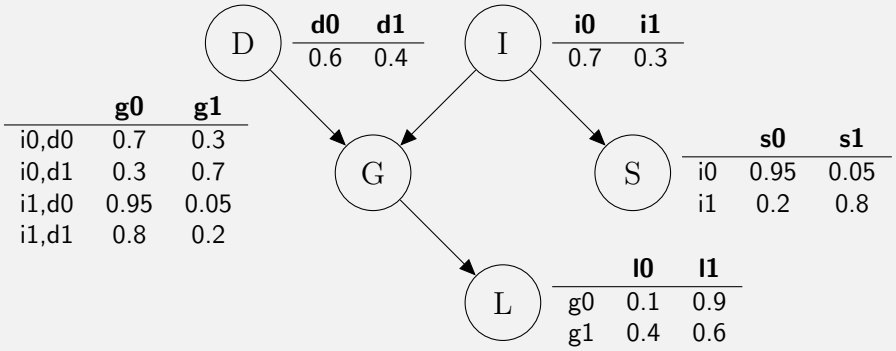

Figura 1(E)

Solución:

a) La factorización de la red bayesiana queda de la siguiente manera:

$$P\left(D, I, G, S, L\right) = P\left(D\right) P\left(I\right) P\left(S|I\right) P\left(G|D, I\right) P\left(L|G\right)$$

b) Para escribir $P\left(D|I=i^0, L=l^0\right)$ usaremos Bayes, quedando la siguiente fórmula:

$$P\left(D|I=i^0, L=l^0\right) = \frac{P\left(D, I=i^0, L=l^0\right)}{P\left(I=i^0, L=l^0\right)}$$

Descomponiendo $P\left(D, I=i^0, L=l^0\right)$, usando la factorización, nos queda:

$$P\left(D, I=i^0, L=l^0\right) = P\left(D\right) P\left(I=i^0\right) \sum_G P\left(G|D, I=i^0\right) P\left(L=l^0|G\right)$$

El denominador $P\left(I=i^0, L=l^0\right)$ es la probabilidad marginal:

$$P\left(I=i^0, L=l^0\right) = \sum_D P\left(D\right) P\left(I=i^0\right) \sum_G P\left(G|D, I=i^0\right) P\left(L=l^0|G\right)$$

Por lo que se puede reescribir la fórmula de la siguiente manera:

$$P\left(D|I=i^0, L=l^0\right) = \frac{P\left(D\right) P\left(I=i^0\right) \sum_G P\left(G|D, I=i^0\right) P\left(L=l^0|G\right)}{\sum_D P\left(D\right) P\left(I=i^0\right) \sum_G P\left(G|D, I=i^0\right) P\left(L=l^0|G\right)}$$

c) En la fórmula anterior ya se ha eliminado la variable S porque no está relacionado con las variables D, L o I en este contexto.

d) Partimos de la fórmula del apartado b):

$$P\left(D|I=i^0, L=l^0\right) = \frac{P\left(D, I=i^0, L=l^0\right)}{P\left(I=i^0, L=l^0\right)}$$

Con la expansión de $P\left(D, I=i^0, L=l^0\right)$ como

$$P\left(D, I=i^0, L=l^0\right) = P\left(D\right) P\left(I=i^0\right) \sum_G P\left(G|D, I=i^0\right) P\left(L=l^0|G\right)$$

Y la expansión de $P\left(I=i^0, L=l^0\right)$ como

$$P\left(I=i^0, L=l^0\right) = \sum_D P\left(D\right) P\left(I=i^0\right) \sum_G P\left(G|D, I=i^0\right) P\left(L=l^0|G\right)$$

Y con los valores que se obtienen de las tablas, los cálculos quedan de la siguiente manera:

- Para $D=d^0$:

$$\begin{aligned} P\left(d^0, i^0, l^0\right) &= P\left(D=d^0\right) P\left(I=i^0\right) \sum_G P\left(G|D=d^0, I=i^0\right) P\left(L=l^0|G\right) \\ &= 0.6 \cdot 0.7\big[P\left(G=g^0|D=d^0, I=i^0\right) P\left(L=l^0|G=g^0\right) \\ &\quad + P\left(G=g^1|D=d^0, I=i^0\right) P\left(L=l^0|G=g^1\right)\big] \\ &= 0.6 \cdot 0.7\big[0.7 \cdot 0.1 + 0.3 \cdot 0.4\big] = 0.0798 \end{aligned}$$

- Para $D=d^1$:

$$\begin{aligned} P\left(d^1, i^0, l^0\right) &= P\left(D=d^1\right) P\left(I=i^0\right) \sum_G P\left(G|D=d^1, I=i^0\right) P\left(L=l^0|G\right) \\ &= 0.4 \cdot 0.7\big[P\left(G=g^0|D=d^1, I=i^0\right) P\left(L=l^0|G=g^0\right) \\ &\quad + P\left(G=g^1|D=d^1, I=i^0\right) P\left(L=l^0|G=g^1\right)\big] \\ &= 0.4 \cdot 0.7\big[0.3 \cdot 0.1 + 0.7 \cdot 0.4\big] = 0.0868 \end{aligned}$$

Con los cálculos realizados, obtenemos

$$P\left(D=d^0|I=i^0, L=l^0\right) = \frac{0.0798}{0.0798 + 0.0868} \approx 0.479$$

$$P\left(D=d^1|I=i^0, L=l^0\right) = \frac{0.0868}{0.0798 + 0.0868} \approx 0.521$$

Ejercicio 8(O). Analizar bajo qué condiciones el conocer D influye a C en la red bayesiana de la Figura 2. Se deben indicar los caminos activos y bajo qué condiciones están activos. Además, se debe indicar bajo qué condiciones no están activos.

Ejercicio 8(E). Analizar bajo qué condiciones el conocer A influye a G en la red bayesiana de la Figura 2. Se deben indicar los caminos activos y bajo qué condiciones están activos. Además, se debe indicar bajo qué condiciones no están activos.

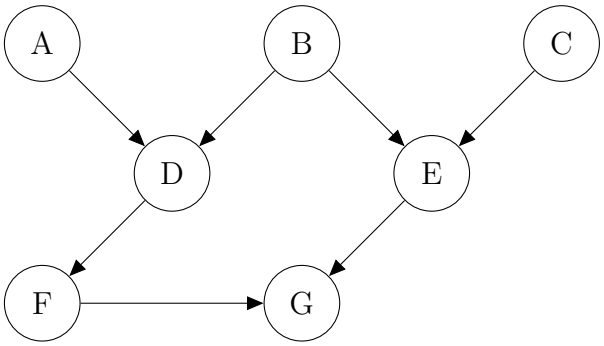

Figura 2

Solución: Para los dos Gráficos vamos a plantear los cuatro tipos de caminos con tres nodos de manera esquemática:

- $A \to B \to C$: Es un camino **casual** y A influye a C mediante B.
- $A \leftarrow B \leftarrow C$: Es un camino **evidencial** y A influye en C mediante B.
- $A \leftarrow B \to C$: Es un camino **Casual y evidencial** y A influye en C mediante B.
- $A \to B \leftarrow C$: Se denomina **camino en V** y A **NO** influye en C mediante B.

Además, dado un camino en el grafo $N_1 - N_2 - ... - N_n$ se dice que está activo con las siguientes condiciones:

- Para todas las **estructuras en V** $N_{i-1} \to N_i \leftarrow N_{i+1}$, el nodo N_i o cualquiera de sus descendientes ha sido observado.
- El resto de sus nodos no ha sido observado.

Teniendo estos caminos y condiciones en mente, ya podemos dar respuesta a las preguntas:

(O) Para que el camino desde D a C podemos encontrar dos posibles caminos:

1. $D \leftarrow B \to E \leftarrow C$
2. $D \to F \to G \leftarrow E \leftarrow C$

Dado que en ambos caminos tenemos presencia de *caminos en V*, en E y G respectivamente, en ningún caso conocer D puede influir a C, a no ser que ya se tenga conocimiento de E, lo que hace que se active el primer camino, o G, lo que activa ambos caminos.

(E) Para llegar a G desde A podemos encontrar dos posibles caminos:

1. $A \to D \to F \to G$
2. $A \to D \leftarrow B \to E \to G$

Podemos observar que en segundo camino tenemos un *camino en V*, en D por lo que por este camino conocer A no influye en G, a no ser que se conozca D o cualquiera de sus descendientes. Por otro lado podemos observar que por el primer camino solo tenemos caminos *casuales*, por lo que conocer A **SI** influye en G.

Ejercicio 9 Escribir esquemáticamente el algoritmo Gradient Boosting en Python para regresión con función de pérdida cuadrática: $L\left(y, F\left(\mathrm{x}\right)\right) = \left(y - F\left(\mathrm{x}\right)\right)^2/2$.

Solución: El algoritmo de Gradient Boosting para regresión con función de pérdida cuadrática $L\left(y, F\left(x\right)\right) = \left(y - F\left(x\right)\right)^2/2$ es el siguiente:

1. Establecemos el valor inicial del modelo $F_0\left(x\right) = \text{media}\left(y\right)$
2. Iteramos para $m = 1, 2, ..., M$, siendo M el número de iteraciones:
   - Calculamos los residuales $r_i^m = y_i - F_{m-1}\left(x_i\right)$
   - Ajustamos el modelo base $h_m\left(x\right) \approx r^{(m)}$
   - Actualizar el modelo: $F_m\left(x\right) = F_{m-1}\left(x\right) + \nu \cdot h_m\left(x\right)$, con $0 < \nu \leq 1$
3. Salida final del modelo $F_M\left(x\right)$ que es la suma de los modelos base ajustados.