

Extraordinaria-2021-resuelto.pdf



destherhada



Inteligencia Artificial II



3º Grado en Ingeniería Informática



Facultad de Informática
Universidad Complutense de Madrid

WUOLAH + BBVA

Hazte **cliente de BBVA** y ...
ahórrate 6 meses
de suscripción



1/6
Este número es indicativo del riesgo del producto, siendo 1/6 indicativo de menor riesgo y 6/6 de mayor riesgo.

Ahora, si te abres una Cuenta Online en BBVA, te reembolsaremos uno de estos suscriptores durante 6 meses (hasta 9,99€/mes) al pagarlos con tu tarjeta Aqua Débito.

NETFLIX **Spotify** **HBOmax**

Disney+ **PlayStation Plus** **DAZN**

Promoción solo para nuevos clientes de BBVA. Válida hasta el 30/06/2023. Estas empresas no colaboran en la promoción.

[Abre tu cuenta](#)



Hazte cliente de BBVA y... ahórrate 6 meses de suscripción

WUOLAH
+ BBVA

NETFLIX

Spotify

HBO max

Disney+

PlayStation Plus

DAZN

Ahora, si te abres una Cuenta Online en BBVA, te reembolsamos una de estas suscripciones durante 6 meses (hasta 9,99€/mes) al pagarla con tu tarjeta Aqua Débito

Promoción solo para nuevos clientes de BBVA. Válida hasta el 30/06/2023. Estas empresas no colaboran en la promoción.

1/6

Este número es indicativo del riesgo del producto, siendo 1/6 indicativo de menor riesgo y 6/6 de mayor riesgo.

BBVA está adherido al Fondo de Garantía de Depósitos de Entidades de Crédito de España. La cantidad máxima garantizada es de 100.000 euros por la totalidad de los depósitos constituidos en BBVA por persona.

Examen de Inteligencia Artificial II

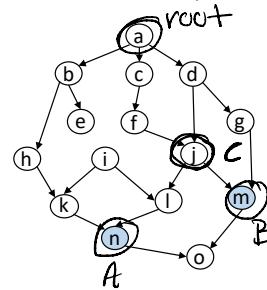
Convocatoria ordinaria - 14 de junio de 2021

Duración: 2 horas

X **Pregunta 1 [1 punto].** Escribe una consulta SPARQL que recupere los títulos y las fechas de estreno de todas las películas dirigidas por Spielberg de menos de 90 minutos de duración. Las películas deben aparecer ordenadas alfabéticamente por título. La fecha de estreno puede no estar disponible en todas las películas. La información se recupera a partir de una base de conocimiento RDF con las siguientes características (los identificadores de las entidades y relaciones aparecen en cursiva):

- Entidades: *Spielberg*, *Film*.
- Relaciones:
 - *instanceOf*: Instance → Class
 - *subClassOf*: Class → Class
 - *title*: Film → string
 - *duration*: Film → int
 - *directedBy*: Film → Person
 - *releaseDate*: Film → date
- Las películas se representan mediante instancias directas o indirectas de la clase *Film*.

X **Pregunta 2 [1 punto].** En una de las prácticas del curso usamos una medida de similitud semántica basada en el cálculo del *Least Common Subsumer* (LCS). Define qué es el LCS y qué representa, escribe la fórmula de la medida de similitud y úsala para calcular la similitud entre los conceptos **n** y **m** en la siguiente taxonomía donde las flechas conectan cada concepto con sus subconceptos.



Pregunta 3 [2 puntos]. Representa en Prolog, siguiendo los convenios explicados en clase, una base de conocimiento a partir de la siguiente información. Escribe un comentario por cada predicado que utilices explicando su significado.

- El Metro está compuesto por distintas líneas de conectan estaciones. Por ejemplo, la línea amarilla en dirección sur conecta las estaciones de Moncloa y Argüelles, Argüelles y Ventura Rodríguez, etc. Las estaciones de Moncloa y Argüelles también están conectadas por la línea gris.
- Siempre que se va a un sitio en Metro es posible regresar haciendo el camino inverso.
- Las líneas de este Metro no tienen ciclos (por ejemplo, no existen líneas circulares). Tampoco tienen sentido los trayectos que contengan ciclos.
- Dos estaciones de la misma línea están a distancia N si para llegar de una a otra debo pasar por N-1 estaciones intermedias.
- Dos estaciones están próximas si es posible viajar de una a otra pasando por 3 estaciones intermedias o menos, y cambiando como máximo una vez de línea.



1

SPARQL

SELECT ?titulo ?fechaEstreno

WHERE

?film instanceOf /subClassOf* Film; *recuperamos todas las películas*

directedBy Spielberg; *de esas las que hayan sido dirigidas por Spielberg*

title ?titulo; *preguntamos por el título*

duration ?duration. *y la duración*

OPTIONAL ?film releaseDate ?fechaEstreno.

usamos un optional para que nos aparezca la consulta aunque no tenga este valor.

FILTER *(?duration > 90)* *filtramos la búsqueda por*

duración de la película

ORDER BY (?titulo)

y finalmente ordenamos por título.



**Hazte cliente de BBVA y ...
ahórrate **6 meses**
de suscripción**

Ahora, si te abres una Cuenta Online en BBVA, te reembolsamos una de estas suscripciones durante 6 meses (hasta 9,99€/mes) al pagarla con tu tarjeta Aqua Débito

NETFLIX**HBOmax**

Spotify®

Disney+

PlayStation Plus

DAZN



2

La medida basada en distancia LCS mide la similaridad entre dos conceptos como la razón entre las siguientes longitudes:

$$\text{sim}(A, B) = \frac{\delta(\text{root}, c)}{\delta(\text{root}, c) + \delta(c, A) + \delta(B, c)}$$

$\delta(A, B)$ = n° mínimo de aristas que conecta A y B

$$\text{sim}(n, m) = \frac{\delta(a, j)}{\delta(a, j) + \delta(j, n) + \delta(m, j)} = \frac{2}{2 + 2 + 1} = \frac{2}{5} //$$

3

???

Hazte cliente de BBVA y... ahórrate 6 meses de suscripción

WUOLAH
+ BBVA



Ahora, si te abres una Cuenta Online en BBVA, te reembolsamos una de estas suscripciones durante 6 meses (hasta 9,99€/mes) al pagarla con tu tarjeta Aqua Débito

Promoción solo para nuevos clientes de BBVA. Válida hasta el 30/06/2023. Estas empresas no colaboran en la promoción.

- Las estaciones pueden o no tener distintos servicios. Por ejemplo, Moncloa tiene tiendas, Bibliometro, cajeros automáticos, etc. Pero Ventura Rodríguez no tiene ninguno de esos servicios.
- Las estaciones exclusivamente comerciales son las que tienen tiendas y cajeros automáticos, y no tienen Bibliometro.

1/6

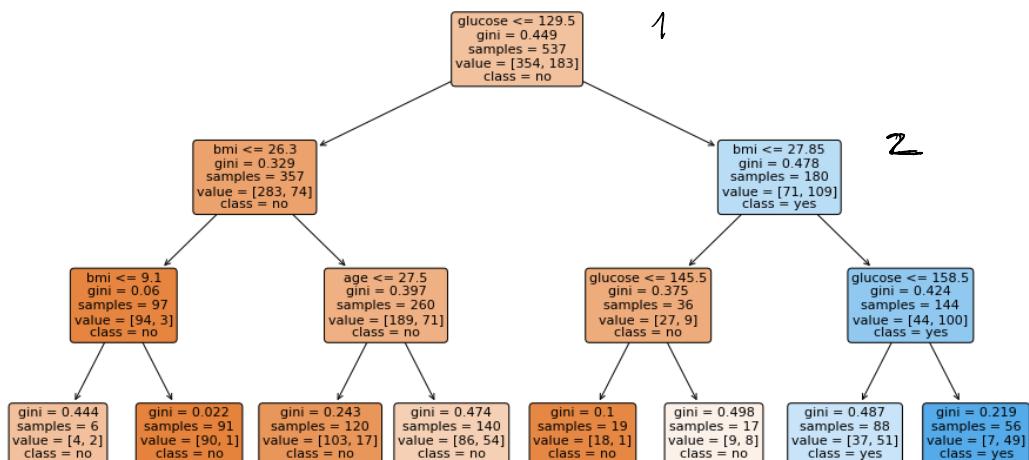
Este número es indicativo del riesgo del producto, siendo 1/6 indicativo de menor riesgo y 6/6 de mayor riesgo.

BBVA está adherido al Fondo de Garantía de Depósitos de Entidades de Crédito de España. La cantidad máxima garantizada es de 100.000 euros por la totalidad de los depósitos constituidos en BBVA por persona.

Escribe a continuación las siguientes consultas:

- Estaciones próximas donde sólo una de ellas tenga Bibliometro.
- Estaciones que no sean exclusivamente comerciales.

Pregunta 4 [1.5 puntos]. Se ha construido un árbol de decisión para tratar de determinar si los pacientes sufren diabetes a partir de distintas variables descriptoras: el índice de masa corporal (*bmi*), la edad (*age*), los niveles de insulina (*insulin*), los niveles de glucosa (*glucose*), la presión sanguínea (*bp*), etc.



Contesta razonadamente a las siguientes cuestiones:

- Indica si las clases a predecir están equilibradas en el conjunto de datos de entrenamiento y cómo debemos tenerlo en cuenta al interpretar el error del modelo.
- Indica cuáles son las variables más relevantes en este problema y cómo podemos saberlo.
- Calcula la entropía total del segundo nivel del árbol (siendo la raíz el primer nivel).
- Indica los valores del parámetro *r* para los que se podaría el nodo “*bmi <= 9.1*” usando poda pesimista.
- Calcula la tasa de aciertos, y la precisión y exhaustividad por clase de un árbol que sólo hiciera una pregunta sobre el mismo conjunto de datos con el que se ha construido.

Pregunta 5 [1.5 puntos]. Explica brevemente la veracidad o falsedad de las siguientes afirmaciones.

- Los árboles de decisión no se ven penalizados si el conjunto de datos contiene variables de entrada (x_i) redundantes o irrelevantes para el problema a resolver.
- En aprendizaje supervisado, una baja correlación entre las variables en entrada (x_i) indica que el problema será difícil de modelar de forma precisa.
- El tiempo de cómputo de un clasificador *k*-NN es proporcional al valor de *k*.
- Un modelo con una exactitud (*accuracy*) de 99% en un problema de clasificación binaria tendrá una precisión (*precision*) y exhaustividad (*recall*) elevadas en ambas clases.

a) Tenemos 537 individuos de los cuales 354 no tienen diabetes y 183 si.

$$Si = \frac{183}{537} = 34'07\%$$

$$No = \frac{354}{537} = 65'92\%$$

Vemos que las clases no están demasiado equilibradas.

Cuando entrenamos un clasificador con clases no equilibradas puede que sobreaprenda sobre una de las clases (la mayoritaria) y que cuando le usemos con un conjunto de datos diferente al que fue entrenado no clasifique bien.

b) Las variables más relevantes son el nivel de glucosa, ya que es la primera variable que se usa para abrir el árbol, que aunque no consigue sacar ningún nodo puro clasifica bastante bien. Además se vuelve a usar en el nivel 3 del árbol en la parte derecha.

La siguiente más relevante sería bmi, ya que se usa en ambos nodos del nivel 2 del árbol. En la parte izquierda (no) del árbol no consigue separar los si es de los no es, pero en la parte derecha, aunque no consigue clases puras consigue clasificar bastantes individuos en distintos nodos.

c) Entropía total del segundo nivel

$$E_{\text{total}} = E(\text{root}) - E(\text{nodo izquierdo}) - E(\text{nodo derecho}) ??$$

$$E(\text{nodo izquierdo}) = -P(\text{si}) \log_2 P(\text{si}) - P(\text{no}) \log_2 P(\text{no})$$

$$P(\text{si}) = 74/357$$

$$P(\text{no}) = 283/357$$

$$E(\text{nodo izquierdo}) = -\frac{74}{357} \log_2 \frac{74}{357} - \frac{283}{357} \log_2 \frac{283}{357} = 0'7362$$

$$E(\text{nodo derecho}) = -P(\text{si}) \log_2 P(\text{si}) - P(\text{no}) \log_2 P(\text{no})$$

$$P(\text{si}) = 109/180$$

$$P(\text{no}) = 71/180$$

$$E(\text{nodo Derecho}) = -\frac{109}{180} \log_2 \frac{109}{180} - \frac{71}{180} \log_2 \frac{71}{180} = 0'9676$$

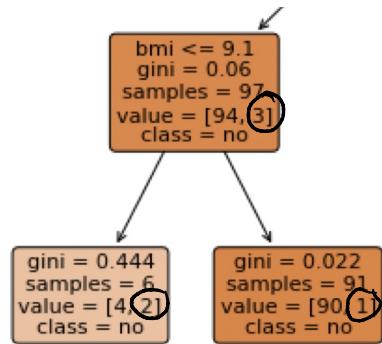
$$E(\text{root}) = -\frac{183}{357} \log_2 \frac{183}{357} - \frac{354}{357} \log_2 \frac{354}{357} = 0'9255$$

$$E_{\text{total}} = 0'9255 - 0'7362 - 0'9676 = -0'7783$$

La entropía se ha negativo y ha habido una disminución de la entropía entre el nivel 1 y el 2.

d) Indica para qué valores de r se podaría en nodo $bmi <= 9.1$ en poda pesimista

Hoja con 3 errores



$$3 \cdot r \leq 2 \cdot r + 1 \cdot r$$

Para cualquier valor de r se podaría el nodo en poda pesimista.

ya que la pregunta no consigue clasificar ninguno de los errores bien. El error de los hijos es mayor o igual al error del nodo padre.

e) Tasa de aciertos = $\frac{VP + VN}{VP + VN + FP + FN} = \frac{109 + 283}{109 + 283 + 71 + 74} = 0'729$

$$\text{Precision} = \frac{VP}{VP + FP} = \frac{109}{109 + 71} = 0'605$$

$$\text{exhaustividad} = \frac{VP}{VP + FN} = \frac{109}{109 + 74} = 0'59$$

clase yes 109 VP 71 FP

clase no 283 VN 74 FN



Hazte cliente de BBVA y... ahórrate 6 meses de suscripción

WUOLAH
+ BBVA

NETFLIX

Spotify

HBO max

Disney+

PlayStation Plus

DAZN

Ahora, si te abres una Cuenta Online en BBVA, te reembolsamos una de estas suscripciones durante 6 meses (hasta 9,99€/mes) al pagarla con tu tarjeta Aqua Débito

Promoción solo para nuevos clientes de BBVA. Válida hasta el 30/06/2023. Estas empresas no colaboran en la promoción.

1/6

Este número es indicativo del riesgo del producto, siendo 1/6 indicativo de menor riesgo y 6/6 de mayor riesgo.

BBVA está adherido al Fondo de Garantía de Depósitos de Entidades de Crédito de España. La cantidad máxima garantizada es de 100.000 euros por la totalidad de los depósitos constituidos en BBVA por persona.

5

VERDADERO

- Si el conjunto de datos a entrenar tiene variables irrelevantes, a la hora de construir el árbol no se utilizarán, así que no, no se ven penalizados.
- VERDADERO, en aprendizaje supervisado buscamos una correlación entre las variables para poder agruparlos, si no tienen demasiada correlación serán más difíciles de agrupar.
- FALSO, el tiempo de cálculo es proporcional a n , siendo n el número de individuos. $O(n)$.
- FALSO, si tiene una tasa de aciertos elevada no tiene porque tener una precisión y exhaustividad elevada. Puede que acierte mucho el clasificador, pero dando muchos falsos positivos, teniendo así una precisión menor.

Abre tu cuenta



WUOLAH
+ BBVA

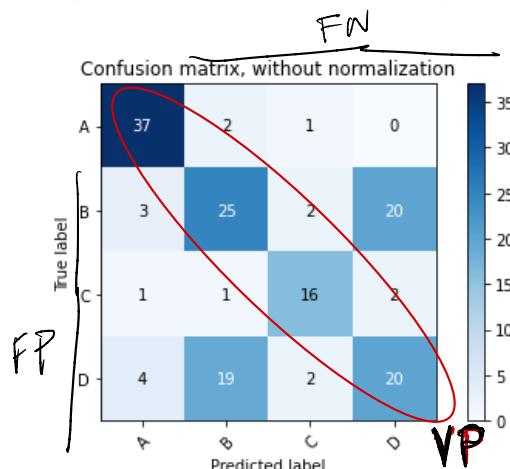
e) VERDADERO, se usa porque no tiene sentido dividir el conjunto de datos en partes mayores como se hace en validación cruzada en k partes. En este caso tomamos $k=n$, siendo n el n° de individuos en el conjunto de datos.

f) VERDADERO en cada paso del algoritmo disminuimos el error hasta llegar a un mínimo local donde el gradiente será 0. Si la función es convexa seguramente encontrar un mínimo global.

- e) El método de validación *leave one out* es recomendable en problemas donde se disponga de pocos datos.
- f) El algoritmo de descenso de gradiente, con una tasa de aprendizaje suficientemente pequeña, encontrará el mínimo global cuando la función a minimizar es convexa.

Pregunta 6 [1 punto]. Explica brevemente 3 estrategias distintas que se pueden utilizar para evitar el sobreaprendizaje al entrenar modelos basados en redes neuronales (las estrategias no tienen por qué ser exclusivas para redes neuronales, pero deben poder usarse con ellas).

Pregunta 7 [1 punto]. Dada la siguiente matriz de confusión, responde razonadamente a las siguientes cuestiones:



- a) ¿Cuántos ejemplos de cada clase contiene el conjunto de entrenamiento? ¿Cuál es la exactitud (*accuracy*) del modelo?
- b) ¿Cuál es la precisión (*precision*) y exhaustividad (*recall*) de la clase C?
- c) ¿Qué clases se distinguen mejor? ¿Qué clases se confunden más?
- d) Indica si las matrices de confusión se usan tanto en problemas de clasificación como de regresión.

Pregunta 8 [1 punto]. Se han analizado 200 críticas positivas y 100 negativas de restaurantes y se ha obtenido la siguiente tabla de frecuencias de aparición de términos.

Términos	Positivas	Negativas
reserva	75	30
servicio	100	80
limpio	50	35
precio	50	65
ruido	5	60
repetir	150	15

- a) Usa un clasificador Naive Bayes para determinar la probabilidad de que un mensaje con los términos *reserva*, *servicio*, *precio* y *repetir* corresponda a una crítica positiva o negativa.
- b) ¿Es Naive Bayes un modelo interpretable o de caja negra? Razona tu respuesta.

[6]

en las diapositivas solo aparece L2 las otras las he tenido que buscar en google

Para evitar el sobreaprendizaje usamos técnicas de regularización

- L2

penaliza pesos de variables demasiado grandes porque la red generalice mejor. Usar una variable que permite controlar cuánto queremos regularizar.

- L1

es muy parecido al L2 pero penaliza menos, la penalización no se eleva al cuadrado sino que toma un valor absoluto. También usa una variable que permite controlar cuánto queremos regularizar.

- dropout

por cada nueva entrada a la red en fase de entrenamiento, se descarta aleatoriamente un porcentaje de neuronas en cada capa oculta, acorde a una probabilidad sc descarte previamente definida.



Hazte cliente de BBVA y... ahórrate 6 meses de suscripción

WUOLAH
+ BBVA

NETFLIX

Spotify

HBO max

Disney+

PlayStation Plus

DAZN

Ahora, si te abres una Cuenta Online en BBVA, te reembolsamos una de estas suscripciones durante 6 meses (hasta 9,99€/mes) al pagarla con tu tarjeta Aqua Débito

Promoción solo para nuevos clientes de BBVA. Válida hasta el 30/06/2023. Estas empresas no colaboran en la promoción.

1/6

Este número es indicativo del riesgo del producto, siendo 1/6 indicativo de menor riesgo y 6/6 de mayor riesgo.

BBVA está adherido al Fondo de Garantía de Depósitos de Entidades de Crédito de España. La cantidad máxima garantizada es de 100.000 euros por la totalidad de los depósitos constituidos en BBVA por persona.

7 a)

$$A = 40 \text{ individuos}$$

$$B = 50 \text{ individuos}$$

$$C = 20 \text{ individuos}$$

$$D = 45 \text{ individuos}$$

$$\text{Exactitud/tasa de aciertos} = \frac{VP + VN}{VP + VN + FP + FN} = \frac{98 + 0}{98 + 0 + 57 + 57} = 0'46\%$$

b) C

$$\text{precisión} = \frac{VP}{VP + FP} = \frac{16}{16 + 5} = 0'76\%$$

$$\text{exhaustividad} = \frac{VP}{VP + FN} = \frac{16}{16 + 4} = 0'8\%$$

c)

Las clases que mejor se distinguen son la A y la C

ya que el clasificador deja pocos individuos sin clasificar bien pero clasifica unos pocos individuos de otras clases en estas.

Las clases que más se confunden son la B y la D

vemos claramente en la matriz de confusión que hay muchos individuos de B clasificados en D y viceversa

Abre tu cuenta



WUOLAH
+ BBVA

d) La matriz de confusión se utiliza exclusivamente en problemas de clasificación. En los problemas de regresión se usan medidas de error como el error cuadrático medio o la raíz cuadrada del error cuadrático medio.

8

200 positivas 100 negativas

a)

• Primero calculamos las probabilidades condicionales de presencia y ausencia de cada término.

prob. cond. presencia

$$p(\text{reserva}|\text{positivo}) = 75/200 = 0'375$$

$$p(\text{reserva}|\text{negativo}) = 30/100 = 0'3$$

$$p(\text{servicio}|\text{positivo}) = 700/200 = 0'5$$

$$p(\text{servicio}|\text{negativo}) = 80/100 = 0'8$$

$$p(\text{limpio}|\text{positivo}) = 50/200 = 0'25$$

$$p(\text{limpio}|\text{negativo}) = 35/100 = 0'35$$

$$p(\text{precio}|\text{positivo}) = 50/200 = 0'25$$

$$p(\text{precio}|\text{negativo}) = 65/100 = 0'65$$

$$p(\text{ruido}|\text{positivo}) = 3/200 = 0'025$$

$$p(\text{ruido}|\text{negativo}) = 60/100 = 0'60$$

$$p(\text{repetir}|\text{positivo}) = 150/200 = 0'75$$

$$p(\text{repetir}|\text{negativo}) = 15/100 = 0'15$$

Términos	Positivas	Negativas
reserva	0'375	0'3
servicio	0'5	0'8
limpio	0'25	0'35
precio	0'25	0'65
ruido	0'025	0'6
repetir	0'75	0'15

prob. cond. ausencia

Términos	Positivas	Negativas
reserva	0'625	0'7
servicio	0'5	0'2
limpio	0'75	0'65
precio	0'75	0'35
ruido	0'1975	0'4
repetir	0'25	0'85

1 - presencia

Naive Bayes $\{ \text{reserva, precio, servicio, repetir} \}$

Naive Bayes nos dará la verosimilitud de cada clase.

$$\text{positivo} = \frac{200}{300} \cdot 0'375 \cdot 0'5 \cdot 0'75 \cdot 0'25 \cdot 0'975 \cdot 0'75$$

prob. $\frac{\text{reserva} \cdot \text{servicio} \cdot \text{limpio} \cdot \text{precio} \cdot \text{ruido} \cdot \text{repetir}}{\text{clase}}$

$$\text{negativo} = \frac{100}{300} \cdot 0'3 \cdot 0'8 \cdot 0'65 \cdot 0'65 \cdot 0'4 \cdot 0'15$$

prob
clase · reserva · servicio · tiempo · precio · ruido · repetir

verosimilitud:

positivo: 0'01713

negativo: 0'002028

Según Naive Bayes es más verosímil que el mensaje sea positivo > negativo.

b) Naive Bayes es un modelo interpretable ya que simplifica la realidad asumiendo que todas las variables utilizadas para clasificar son independientes.