Inteligencia de Negocio, Curso 2023-2024
Grado en Ingeniería Informática, Universidad de Granada

Convocatoria Ordinaria, Examen de Teoría

17 de enero de 2024

Apellidos, Nombre:

.....

Tiempo disponible: *2,5 horas*

1. **(1,5 ptos.)** Justifica en qué orden aplicarías los siguientes métodos de procesado de datos: reducción del ruido, balanceo de clases por muestreo, selección de características e imputación de valores perdidos.

2. **(2 ptos.)** Compara justificadamente ventajas y desventajas de los siguientes algoritmos de *clustering*: K-Means, DBSCAN, Mean Shift, Jerárquico Aglomerativo Ward y BIRCH.

3. **(1,5 ptos.)** Explica qué inconvenientes puede presentar la *confianza* para valorar la fiabilidad de una regla de asociación y propón una alternativa que palíe esos inconvenientes.

4. **(2 ptos.)** Disponemos del siguiente conjunto de datos con cuatro variables numéricas:

id	x_1	x_2	x_3	x_4
1	40	2	1	0
2	50	3	2	3
3	50	4	2	2
4	65	4	1	3
5	50	1	2	1
6	40	1	2	0

Centroides iniciales

Cluster	x_1	x_2	x_3	x_4
A	50	2	2	1
B	40	2	2	2

Escoge la medida de distancia más apropiada y traza paso a paso el algoritmo K-Means para agrupar los datos en dos *clusters*, siendo sus centroides iniciales al comienzo del algoritmo los indicados arriba. Completa las tablas de la siguiente página con los resultados.

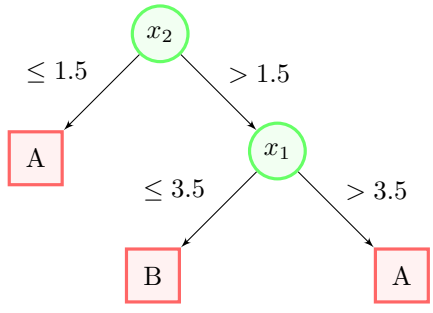
5. **(3 ptos.)** Dado el conjunto de datos de más abajo, escoge la medida de distancia más apropiada y aplica el algoritmo SMOTE con $k = 2$ para igualar la frecuencia de las dos clases. Escoge los números aleatorios que necesites en el orden de arriba a abajo que se indica en la tabla con la columna *random*.

   A continuación, dado el árbol de decisión de más abajo, compara los resultados de predicción con F_1-score sobre la clase 'A' entre el conjunto de datos original y añadiendo a ese conjunto los nuevos datos sintéticos creados. ¿Qué opción predice mejor y por qué?

   Completa las tablas de la siguiente página con los resultados.

id	x_1	x_2	x_3	x_4	***clase***
1	4	2	1	0	A
2	3	2	2	0	B
3	7	1	1	3	B
4	3	1	2	1	A
5	4	2	2	0	A
6	2	4	2	1	B
7	4	1	1	2	A
8	3	1	2	0	A
9	4	1	0	1	A
10	3	1	0	1	A
11	3	2	1	0	B
12	4	2	0	0	A

random
0.2
0.8
0.6
0.1
0.3
0.5
0.9
0.2
0.7
0.5
0.4
0.3
...



Ej. 4: Agrupamiento final

id	***Cluster***
1	
2	
3	
4	
5	
6	

Ejercicio 4: Centroides finales de los *clusters*

Cluster	x_1	x_2	x_3	x_4
A				
B				

Ejercicio 5: Datos sintéticos generados por SMOTE

x_1	x_2	x_3	x_4	***clase***

Ej. 5: Matriz de confusión sobre el conjunto original

Real \ **Predicho**	A	B
A		
B		

F_1-score (sobre clase 'A') =

Ej. 5: Matriz de confusión añadiendo datos sintéticos

Real \ **Predicho**	A	B
A		
B		

F_1-score (sobre clase 'A') =

Ejercicio 4. Dado que los atributos tienen rangos dispares, deberemos realizar alguna normalización para que la medida de distancia sea representativa. Podemos emplear esta medida: $d(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{\sum_{k=1}^{4} \left(\frac{x_{jk} - x_{ik}}{\max x_k - \min x_k} \right)^2}$.

Ej. 4: Agrupamiento final

id	***Cluster***
1	A
2	B
3	B
4	B
5	A
6	A

Ejercicio 4: Centroides finales de los *clusters*

Cluster	x_1	x_2	x_3	x_4
A	43.333	1.333	1.667	0.333
B	55.000	3.667	1.667	2.667

Ejercicio 5. Hay que añadir 4 datos de la clase B, así que se genera uno sintético a partir de cada uno de los 4 ejemplos de esa clase. Empleamos la misma medida de distancia del ejercicio 3. Con esta distancia, los dos ($k = 2$) vecinos del ejemplo id_2 son el id_6 y el id_{11}, los de id_3 son id_2 e id_{11}, los de id_6 son id_2 e id_{11}, y los de id_{11} son id_2 e id_6. Empezamos por el primer dato de la clase B (id_2), y como el primer número aleatorio es el 0.2 (< 0.5), se escoge el primero de los dos vecinos (id_6). Por lo que se genera un dato sintético entre el dato (3,2,2,0) y su vecino (2,4,2,1) empleando el siguiente número aleatorio (0.8). Pasamos ahora al segundo dato de la clase B, y entre sus dos vecinos, escogemos el segundo (o sea, id_{11}) porque ahora el número aleatorio es 0.6. Se genera un nuevo dato sintético entre el dato (7,1,1,3) y su vecino (3,2,1,0) empleando el número aleatorio 0.1. Igual se hace con el tercer y cuarto dato de la clase B hasta generar los cuatro datos de la siguiente tabla.

Ejercicio 5: Datos sintéticos generados por SMOTE

x_1	x_2	x_3	x_4	***clase***
2.2	3.6	2.0	0.8	B
6.6	1.1	1.0	2.7	B
2.5	3.0	2.0	0.5	B
2.8	2.4	1.2	0.2	B

Ej. 5: Matriz de confusión sobre el conjunto original

Real \ **Predicho**	A	B
A	8	0
B	1	3
F_1-score (sobre 'A') = 0.94118, prec. = 0.88889, recall = 1		

Ej. 5: Matriz de confusión añadiendo datos sintéticos

Real \ **Predicho**	A	B
A	8	0
B	2	6
F_1-score (sobre 'A') = 0.88889, prec. = 0.8, recall = 1		

De los cuatro ejemplos sintéticos generados, tres se clasifican acertadamente con la clase 'B', pero uno de ellos se clasifica erróneamente con la clase 'A'. Por tanto, F_1-score sobre la clase 'A' resulta peor si se añaden al cálculo los datos sintéticos. Sobre la clase 'B', precision es 1 y recall 0.75 en ambos casos. En definitiva, dado que el error en clasificar la clase 'B' es 0.25 antes y después de añadir los datos sintéticos, no se produce mejora en la predicción del árbol. No olvidemos que en un caso real, SMOTE se usa para generar datos y entrenar de nuevo el algoritmo, que generaría un nuevo árbol que previsiblemente podría predecir mejor nuevos casos.

Inteligencia de Negocio, Curso 2022-2023
Grado en Ingeniería Informática, Universidad de Granada

Convocatoria Ordinaria, Examen de Teoría

17 de enero de 2023

Apellidos, Nombre:

.....

Tiempo disponible: *2 horas*

1. **(3 ptos.)** En un problema de clasificación con tres clases, donde la clase minoritaria es 10 veces menos frecuente que la clase mayoritaria, responde justificadamente a las siguientes cuestiones:

   a) ¿Qué medida emplearías para valorar el acierto de un clasificador?
   b) ¿Qué método usarías para comparar la eficacia de diferentes algoritmos de clasificación?
   c) ¿Qué solución propondrías para mejorar la capacidad predictiva de los clasificadores generados por un algoritmo de aprendizaje de árboles de decisión?
   d) ¿Y de un algoritmo *ensemble learning*?

2. **(2 ptos.)** Explica en qué consiste la clasificación multi-etiqueta, el aprendizaje multi-instancia y el aprendizaje semisupervisado. En cada caso, propón un ejemplo de resolución de un problema apropiado para él.

3. **(3 ptos.)** Disponemos del siguiente conjunto de datos para clasificar setas en venenosas y comestibles:

conjunto	id	color	altura	rayas	textura	clase
entrenamiento	1	Púrpura	Alto	Sí	Rugoso	Venenosa
	2	Azul	Alto	No	Peludo	Comestible
	3	Púrpura	Alto	Sí	Suave	Comestible
	4	Rojo	Bajo	Sí	Peludo	Comestible
	5	Azul	Bajo	No	Suave	Comestible
	6	Púrpura	Alto	Sí	Peludo	Venenosa
	7	Púrpura	Bajo	No	Peludo	Venenosa
	8	Azul	Bajo	Sí	Peludo	Venenosa
	9	Rojo	Alto	No	Peludo	Comestible
	10	Azul	Alto	Sí	Suave	Venenosa
prueba	11	Azul	Bajo	Sí	Rugoso	Venenosa
	12	Rojo	Bajo	No	Suave	Comestible
	13	Azul	Bajo	Sí	Suave	Venenosa
	14	Rojo	Alto	Sí	Peludo	Comestible
	15	Púrpura	Alto	No	Rugoso	Comestible
	16	Púrpura	Alto	No	Suave	Comestible

   Traza paso a paso el algoritmo ID3 con índice Gini para generar el árbol de decisión a partir de los *datos de entrenamiento*. Traduce el árbol obtenido a un conjunto de reglas de clasificación. Finalmente, mediante el modelo generado, obtén la matriz de confusión sobre el *conjunto de prueba* y calcula los valores TPR y TNR (considera como positiva la clase 'Venenosa'). Completa las tablas de la siguiente página con los resultados.

4. **(2 ptos.)** Dado el conjunto de datos del ejercicio anterior, traza paso a paso el algoritmo Naïve Bayes con estimación por máxima verosimilitud sobre el *conjunto de entrenamiento*. Una vez construido el modelo, aplícalo para predecir la clase del *conjunto de prueba*, incluyendo el grado de certeza (probabilidad de predicción). Completa las tablas de la siguiente página con los resultados.

Ejercicio 3: Reglas a partir del árbol de decisión								
Antecedente 1		**Antecedente 2**		**Antecedente 3**		**Antecedente 4**		**clase**
Var.	Valor	Var.	Valor	Var.	Valor	Var.	Valor	Valor

Ejercicio 3: Matriz de confusión		
Predicho / **Real**	Venenosa	Comestible
Venenosa		
Comestible		

TPR =

TNR =

Ejercicio 4: Modelo de Naïve Bayes														

Ejercicio 4: Predicción Naïve Bayes		
id	clase	probabilidad
11		
12		
13		
14		
15		
16		

Ejercicio 3: Reglas a partir del árbol de decisión								
Antecedente 1		**Antecedente 2**		**Antecedente 3**		**Antecedente 4**		**clase**
Var.	Valor	Var.	Valor	Var.	Valor	Var.	Valor	Valor
color	Púrpura	textura	Rugoso					Venenosa
color	Púrpura	textura	Peludo					Venenosa
color	Púrpura	textura	Suave					Comestible
color	Azul	rayas	Sí					Venenosa
color	Azul	rayas	No					Comestible
color	Rojo							Comestible

Ejercicio 3: Matriz de confusión			
Predicho / **Real**	Venenosa	Comestible	
Venenosa	2	0	TPR = 1
Comestible	1	3	TNR = 0,75

Ejercicio 4: Modelo de Naïve Bayes														
clase			color			altura			rayas			textura		
	Ven.	Com.		Ven.	Com.		Ven.	Com.		Ven.	Com.		Ven.	Com.
	5	5	Púrpura	3	1	Alto	3	3	Sí	4	2	Rugoso	1	0
			Azul	2	2	Bajo	2	2	No	1	3	Peludo	3	3
			Rojo	0	2							Suave	1	2

Ejercicio 4: Predicción Naïve Bayes		
id	clase	probabilidad
11	Venenosa	1
12	Comestible	1
13	Venenosa	0.5
14	Comestible	1
15	Venenosa	1
16	Comestible	0.66

Inteligencia de Negocio
Grado en Ingeniería Informática, Universidad de Granada
Curso 2021-2022

Convocatoria Ordinaria

Examen de Teoría

12 de enero de 2022

Tiempo disponible: *2,5 horas*

1. **(1 pto.)** Describe brevemente los modelos básicos de *ensemble learning* Bagging y Boosting explicando los aspectos relevantes de cada uno de ellos y destacando sus diferencias.

2. **(0,5 ptos.)** Indica ventajas e inconvenientes de la selección de características de tipo filtro y envolvente.

3. **(0,5 ptos.)** Explica brevemente las similitudes, diferencias y qué retos abordan el aprendizaje incremental y la minería de flujo de datos.

4. **(1 pto.)** Disponemos de la siguiente base de datos con 6 transacciones:

id	**artículos comprados**
t_1	salchicha, pan, ketchup
t_2	salchicha, pan
t_3	refresco, patatas, salchicha
t_4	patatas, refresco
t_5	patatas, ketchup
t_6	refresco, salchicha, patatas

   Suponiendo el umbral mínimo de soporte al 30 %, obtén todos los conjuntos frecuentes usando el algoritmo Apriori. Traza en detalle los conjuntos candidatos y conjuntos frecuentes generados en cada paso.

   A continuación, obtén las reglas de asociación con un *lift* (empuje) mayor que 1. Considera solo reglas que tienen un ítem en el consecuente.

5. **(1 pto.)** Traza el algoritmo PRISM, indicando paso a paso cómo genera las distintas reglas de recomendación de lentillas, en base al conjunto de datos mostrado en la página siguiente. Considera el uso de una regla por defecto.

id	edad	prescripción de gafas	astigmatismo	tasa de producción de lágrimas	lentilla recomendada
1	joven	miope	no	reducida	ninguna
2	joven	miope	no	normal	blanda
3	joven	miope	sí	reducida	ninguna
4	joven	miope	sí	normal	dura
5	joven	hipermétrope	no	reducida	ninguna
6	joven	hipermétrope	no	normal	blanda
7	joven	hipermétrope	sí	reducida	ninguna
8	joven	hipermétrope	sí	normal	dura
9	pre-présbita	miope	no	reducida	ninguna
10	pre-présbita	miope	no	normal	blanda
11	pre-présbita	miope	sí	reducida	ninguna
12	pre-présbita	miope	sí	normal	dura
13	pre-présbita	hipermétrope	no	reducida	ninguna
14	pre-présbita	hipermétrope	no	normal	blanda
15	pre-présbita	hipermétrope	sí	reducida	ninguna
16	pre-présbita	hipermétrope	sí	normal	ninguna
17	présbita	miope	no	reducida	ninguna
18	présbita	miope	no	normal	ninguna
19	présbita	miope	sí	reducida	ninguna
20	présbita	miope	sí	normal	dura
21	présbita	hipermétrope	no	reducida	ninguna
22	présbita	hipermétrope	no	normal	blanda
23	présbita	hipermétrope	sí	reducida	ninguna
24	présbita	hipermétrope	sí	normal	ninguna

UNIVERSIDAD DE GRANADA

Departamento de Ciencias de la Computación e Inteligencia Artificial

Inteligencia de Negocio
Grado en Ingeniería Informática
Convocatoria Ordinaria: 17/02/2021
Inicio del examen: 16:00
Duración: **1:30h**

Enunciado del Problema

Proponga tres problemas complejos donde los datos sean el eje central para su resolución y que tengan un claro interés social y/o económico. Los problemas deben de ser uno de los 4 ámbitos siguientes como datos de entrada

(a) predicción de series temporal en el ámbito IoT, b) procesamiento de imágenes con deep learning, c) procesamiento de lenguaje natural para análisis de opinions, d) detección de anomalías)

Desde la perspectiva de la Ciencia de Datos decida cómo los abordaría. Para cada uno de ellos diseñe una metodología de resolución.

Se valorará principalmente:
a) La originalidad de los problemas planteados.
b) La calidad de las propuestas de solución planteadas, y su adecuación al problema correspondiente.
c) La redacción y justificación de lo planteado.

Entrega

La entrega se hará por medio de PRADO, mediante una entrega específica. En caso de haber cualquier problema técnico, se enviará dentro del plazo de entega por email a dmolinac@go.ugr.es

INTELIGENCIA DE NEGOCIO
Convocatoria Ordinaria de Enero
17 de enero de 2020

1. (1 pto.) ¿Por qué (y en qué situaciones) es interesante aplicar un preprocesamiento basado en filtros o un preprocesamiento basado en *ensemble*? Explicar brevemente ambos, pros y contras.

2. (1,25 ptos.)

A) Explicad las etapas de un modelo aprendizaje de análisis de sentimientos.

B) ¿Qué aporta el *machine learning* en el análisis de sentimientos?

3. (1,25 ptos.) Supongamos un problema con clases no balanceadas, 3/4 clase A y 1/4 clase B. Se aplica preprocesamiento (SMOTE) y un clasificador Random Forest y el clasificador en un particionamiento 5fcv obtiene una media de 75% en clasificación. Explicad qué otras características puede tener el problema que justifiquen su mal comportamiento. Enumerarlas y justificarlas.

4. (1,5 ptos.) Disponemos de la siguiente base de datos conteniendo 4 transacciones:

TID artículos comprados

t1 K,A,D,B

t2 D,A,C,E,B

t3 C,A,B,E

t4 B,A,D

Suponiendo los umbrales mínimos de soporte y confianza al 50% y 90% respectivamente, se pide obtener todos los conjuntos frecuentes fijado dicho soporte usando el algoritmo Apriori y las reglas asociadas al nivel indicado de confianza.

Tiempo: 90 minutos

Inteligencia de Negocio. Grado en Ingeniería Informática. Curso 2018-2019

Examen ordinario (14 de enero de 2019)

Tiempo disponible: *2 horas*

1. **(2 ptos.)** Responde a las 30 preguntas tipo test de la hoja siguiente. Solo una de las cuatro respuestas es válida en cada pregunta. Cada respuesta correcta suma $2/30 = 0{,}0667$ puntos, cada fallo resta $2/90 = 0{,}0222$ puntos.

2. **(1 pto.)** Describe brevemente los modelos básicos de multiclasificador Bagging y Boosting. Destaca los aspectos relevantes de cada uno de ellos y explica sus diferencias como multiclasificadores.

3. **(0,5 ptos.)** Supón un conjunto de datos de clasificación que tiene 4 atributos de entrada, 500 ejemplos y 2 clases. El 25 % de los ejemplos están en la clase positiva y el 75 % en la negativa. ¿Cómo abordarías el problema para aplicar árboles de decisión?

4. **(1,5 ptos.)** Disponemos de la siguiente base de datos con 5 transacciones:

TID	artículos comprados
t_1	Dátiles, Peras, Plátanos, Uvas
t_2	Naranjas, Uvas, Peras, Kiwis
t_3	Peras, Plátanos,
t_4	Peras, Kiwis, Uvas,
t_5	Plátanos, Naranjas, Uvas, Kiwis, Peras

   Suponiendo los umbrales mínimos de soporte y confianza al 50 % y 75 % respectivamente, obtén todos los conjuntos frecuentes usando el algoritmo Apriori. A continuación, genera las reglas asociadas al nivel indicado de confianza. Considera solo reglas que tienen al menos un ítem en el antecedente pero ten en cuenta cualquier número de ítems en el consecuente.

DECSAI
Departamento de Ciencias de la Computación e I.A.
Universidad de Granada

INTELIGENCIA DE NEGOCIO

Convocatoria Extraordinaria, 2 de Febrero de 2018

1. (0.8 ptos.) ¿Qué medidas o criterios conoce para evaluar un clasificador? Comente los diferentes aspectos que se deben tener en cuenta para estimar los diferentes aspectos importantes del rendimiento de un algoritmo de clasificación.

2. (0.8 ptos.) Explique las características de los algoritmos de agrupamiento (o segmentacióon) jerárquicos. ¿Qué tipos de algoritmos de agrupamiento jerárquico hay? ¿Cuáles son sus diferencias? Finalmente, describa qué ventajas presentan frente a algoritmos de agrupamiento simples como K-medias.

3. (0.8 ptos.) Indique al menos 3 medidas usadas para evaluar la precisión de una predicción hecha por un modelo de predicción de series temporales. Explique el objetivo asociado cada medida en el ámbito de las series temporales.

4. (1.3 ptos.) Disponemos de la siguiente base de datos conteniendo 5 transacciones. TID artículos comprados:

t1 K, A, D, B, C
t2 D, A, C, E
t3 C, A, D, E
t4 K, A, D
t5 B, A, K, D

Se pide: 1. Suponiendo los umbrales mínimos de soporte y confianza al 50% y 85% respectivamente, se pide obtener todos los conjuntos frecuentes fijado dicho soporte usando el algoritmo Apriori y las reglas asociadas al nivel indicado de confianza.
2. Considere el interés como medida para evaluar las reglas extraidas en el paso anterior. Si el interés mínimo es de 0.6, ¿qué reglas deberíamos considerar como aptas?

5. (1.3 ptos.) Considere el dataset de clasificaciónde la Tabla siguiente, donde la variable ``*¿Venenosa?*'' representa la clase de cada instancia y el resto (*Color, Altura y Rayas*) corresponden a los atributos de entrada.

Color	Altura	Rayas	Venenosa?
Púrpura	Alto	Si	Si
Púrpura	Alto	Si	Si
Rojo	Bajo	Si	No
Azul	Bajo	No	No
Azul	Bajo	Si	Si
Rojo	Alto	No	No
Azul	Alto	Si	Si
Azul	Bajo	Si	Si
Azul	Alto	No	No
Azul	Bajo	Si	Si
Rojo	Bajo	No	No
Púrpura	Bajo	No	Si
Rojo	Alto	Si	No
Púrpura	Alto	Si	Si
Púrpura	Alto	No	No
Púrpura	Alto	No	No

Se pide: 1. Construir un **árbol de clasificación** utilizando *GainRatio* como criterio de división de los nodos. Muestre los pasos de cálculo para la división de cada nodo, indicando cuál es el atributo elegido en cada caso. Dibuje el árbol de decisión final.
2. Traduzca el árbol obtenido en el punto anterior a un conjunto de reglas de clasificación.

Tiempo: 100 minutos

C/ Periodista Daniel Saucedo Aranda s/n, ETSI Informática y de Telecomunicación, 18071 Granada
Tlf: +34.958.244019 Fax: +34.958.243317

DECSAI
Departamento de Ciencias de la Computación e I.A.
Universidad de Granada

INTELIGENCIA DE NEGOCIO

Convocatoria Extraordinaria
12 de Julio de 2017

1. (1 pto.) Explicad las ventajas e inconvenientes de un selector de características de tipo envolvente y de tipo filtro.

2. (1 pto.) Describid brevemente las diferencias de C4.5 respecto a ID3

3. (1 pto.) a) Ventajas de Bagging y Boosting frente a otros clasificadores denominados fuertes. b) Diferencias entre Bagging y Boosting

4. (1 pto.) Enumera las ventajas y los inconvenientes que presenta el método de agrupamiento k-medias.

5. (1 pto.) Suponed un conjunto de datos de clasificación que tiene 300 atributos de entrada, 500 ejemplos y 3 clases (cada clase entre 30 y 40% de tamaño). Cien de los atributos de entrada son numéricos en el dominio [1.0, 5.0], 50 son categóricos con 5 valores diferentes y 50 binarios (2 valores categóricos). Queremos aplicar un algoritmo de clasificación. ¿Qué algoritmo aplicarías y por qué? ¿Qué preprocesamiento sería interesante/conveniente para aplicar dicho algoritmo?

6. (1 pto.) ¿Cuál de los siguientes problemas son más adecuados para el enfoque de aprendizaje?
   (i) La clasificación de los números en primos y no primos.
   (ii) La detección de posibles fraudes en los cargos de tarjeta de crédito.
   (iii) La determinación del tiempo que tardaría un objeto que cae en llegar al suelo.
   (iv) Determinar el ciclo óptimo para las luces de un semáforo enciende en un cruce muy concurrido.

   Justificad la respuesta. Describe el procedimiento a seguir para resolver el problema seleccionado.

7. (1 pto.) Suponed un conjunto de datos de clasificación que tiene 40 atributos de entrada, 500 ejemplos y 2 clases. El 10% de los ejemplos están en la segunda clase, y el 90% en la primera. ¿Cómo abordarías el problema para aplicar árboles de decisión?

8. (1 pto.) Describid la técnica de validación bootstraping.

9. (1 pto.) ¿Cómo se abordaría un problema de clasificación con 6 clases? Explica formas de resolver el problema, y sus ventajas e inconvenientes.

10. (1 pto.) ¿Cómo se resolvería un problema de clasificación con 5 millones de instancias y 5 millones características?

Tiempo: 120 minutos

C/ Periodista Daniel Saucedo Aranda s/n, ETSI Informática y de Telecomunicación, 18071 Granada
Tlf: +34.958.244019 Fax: +34.958.243317

DECSAI
Departamento de Ciencias de la Computación e I.A.
Universidad de Granada

INTELIGENCIA DE NEGOCIO

Convocatoria Ordinaria de Febrero
26 de Enero de 2017

1. (1 pto.) Explica las ventajas e inconvenientes de un utilizar un algoritmo de filtrado de ruido.

2. (1 pto.) Diferencias entre Bagging y Boosting. ¿Cómo funcionaría boosting en un problema con ruido de clase?

3. (1 pto.) Enumera 2 características que identifiquen Deep Learning frente a otros algoritmos clásicos de Aprendizaje Automático (incluidas las redes neuronales clásicas).

    a. ¿Qué es big data?

4. En las siguientes situaciones, determina si estamos ante un problema de minería datos y qué tipo de técnicas se podrían aplicar. Razonadlo.
    a. Dividir los clientes de una compañía de acuerdo a su género
    b. Dividir los clientes de una compañía de acuerdo a su fiabilidad
    c. Segmentar los clientes de una compañía con los datos disponibles a los mismos. Predecir el segmento para nuevos clientes.
    d. Calcular las ventas totales de una compañía
    e. Predecir el precio futuro del almacén de una compañía en base a registros históricos
    f. Monitorizar el latido de corazón de un paciente para detectar situaciones anormales
    g. Extraer frecuencias de una señal de sonido.
    h. Predecir si un paciente tiene una enfermedad rara.

Tiempo: 100 minutos

C/ Periodista Daniel Saucedo Aranda s/n, ETSI Informática y de Telecomunicación, 18071 Granada
Tlf: +34.958.244019 Fax: +34.958.243317

DECSAI
Departamento de Ciencias de la Computación e I.A.
Universidad de Granada

INTELIGENCIA DE NEGOCIO

Convocatoria Extraordinaria de Septiembre
15 de Septiembre de 2016

1. (1 pto.) Explica las ventajas e inconvenientes de un selector de características de tipo envolvente y de tipo filtro.

2. (1 pto.) Describe brevemente las diferencias de C4.5 respecto a ID3

3. (1 pto.) Ventajas de Bagging y Boosting frente a otros clasificadores denominados fuertes. Diferencias entre Bagging y Boosting

4. (1 pto.) Enumera las ventajas y los inconvenientes que presenta el método de agrupamiento k-medias.

5. (1 pto.) Suponed un conjunto de datos de clasificación que tiene 5 atributos de entrada, 500 ejemplos y 3 clases (cada clase entre 30 y 40% de tamaño). Tres de los atributos de entrada son numéricos en el dominio [1.0, 5.0], el cuarto es categórico con 5 valores diferentes y el quinto es binario (2 valores categóricos). Queremos un algoritmo de clasificación. ¿Qué algoritmo aplicarías y por qué? ¿Qué preprocesamiento h sería interesante para aplicar dicho algoritmo?

6. (1 pto.)¿Por qué (y en qué situaciones) es interesante realizar selección de instancias antes de construir un clasificador?

7. (1 pto.) Suponed un conjunto de datos de clasificación que tiene 4 atributos de entrada, 500 ejemplos y 2 clases. El 25% de los ejemplos están en la segunda clase, y el 75 en la primera. ¿Cómo abordarías el problema para aplicar árboles de de decisión?

8. (1 pto.) Describe la técnica de validación bootstraping.

9. (1 pto.) ¿Cómo abordarías un problema de clasificación con 6 clases? Explica formas de resolver el problema, y sus ventajas e inconvenientes.

10. (1 pto.) ¿Cómo resolverías un problema de clasificación con 5 millones de instancias y 5 millones características?

Tiempo: 120 minutos

C/ Periodista Daniel Saucedo Aranda s/n, ETSI Informática y de Telecomunicación, 18071 Granada
Tlf: +34.958.244019 Fax: +34.958.243317

INTELIGENCIA DE NEGOCIO

Convocatoria Ordinaria de Febrero
5 de Febrero de 2016

5. (1 ptos.) ¿Por qué (y en qué situaciones) es interesante utilizar las técnicas de imputación de valores perdidos?
   Enumera, describe y justifica el funcionamiento de 2 técnicas de estimación de valores perdidos.

6. (1 ptos.) Explica brevemente los modelos básicos de multiclasificador: Bagging y Boosting. Aspectos positivos a destacar de cada uno de ellos. Sus diferencias como multiclasificadores.

2. (1 ptos.) Enumera las ventajas y los inconvenientes que presenta el método de agrupamiento k-medias.

7. (2 ptos) Enumera y describe cinco problemas abordados en minería de datos. Pon un ejemplo de aplicación real y menciona un algoritmo clásico para cada uno de ellos.

Tiempo: 100 minutos

C/ Periodista Daniel Saucedo Aranda s/n, ETSI Informática y de Telecomunicación, 18071 Granada
Tlf: +34.958.244019 Fax: +34.958.243317

DECSAI
Departamento de Ciencias de la Computación e I.A.
Universidad de Granada

INTELIGENCIA DE NEGOCIO

Convocatoria Extraordinaria de Septiembre
10 de Septiembre de 2015

1. (1 pto.) Explica las ventajas e inconvenientes de un selector de características de tipo envolvente y de tipo filtro.

2. (1 pto.) Describe brevemente las diferencias de C4.5 respecto a ID3

3. (1 pto.) Ventajas de Bagging y Boosting frente a otros clasificadores denominados fuertes.

4. (1 pto.) Enumera las ventajas y los inconvenientes que presenta el método de agrupamiento k-medias.

5. (1 pto.) Suponed un conjunto de datos de clasificación que tiene 5 atributos de entrada, 500 ejemplos y 4 clases. Tres de los atributos de entrada son numéricos en el dominio [1.0, 5.0], el cuarto es categórico con 5 valores diferentes y el quinto es binario (2 valores categóricos). Queremos aplicar k-NN y árboles de decisión. ¿Qué preprocesamiento harías para cada una de las dos técnicas de aprendizaje?

6. (1 pto.)¿Por qué (y en qué situaciones) es interesante realizar selección de instancias antes de construir un clasificador?

7. (1 pto.) Suponed un conjunto de datos de clasificación que tiene 4 atributos de entrada, 500 ejemplos y 2 clases. El 25% de los ejemplos están en la segunda clase, y el 75 en la primera. ¿Cómo abordarías el problema para aplicar árboles de de decisión?

8. (1 pto.) Describe la técnica de validación bootstraping.

9. (1 pto.) ¿Cómo abordarías un problema de clasificación con 6 clases? Explica formas de resolver el problema, y sus ventajas e inconvenientes.

10. (1 pto.) ¿Cómo resolverías un problema de clasificación con 30 millones de instancias y 300 características?

Tiempo: 120 minutos

C/ Periodista Daniel Saucedo Aranda s/n, ETSI Informática y de Telecomunicación, 18071 Granada
Tlf: +34.958.244019 Fax: +34.958.243317

INTELIGENCIA DE NEGOCIO
Convocatoria Ordinaria de Febrero
27 de Enero de 2015

1. (1 pto.) ¿Por qué (y en qué situaciones) es interesante realizar selección de variables antes de construir un clasificador?

2. (1 pto.) Define un árbol de decisión y describe el algoritmo básico de construcción de un árbol de decisión.

3. (1 pto.) Describe la técnica de validación bootstrap y discute qué aporta en el diseño de clasificadores.

4. (2 ptos) Disponemos de la siguiente base de datos conteniendo 4 transacciones:

TID	artículos comprados
t1	K,A,D,B
t2	D,A,C,E,B
t3	C,A,B,E
t4	B,A,D

Suponiendo los umbrales mínimos de soporte y confianza al 50% y 90% respectivamente, se pide obtener todos los conjuntos frecuentes fijado dicho soporte usando el algoritmo A priori y las reglas asociadas al nivel indicado de confianza.

Tiempo: 90 minutos

INTELIGENCIA DE NEGOCIO

Convocatoria Extraordinaria de Septiembre
4 de Diciembre de 2014

6. (1 pto.) Explica las ventajas e inconvenientes de un selector de características de tipo envolvente y de tipo filtro.

7. (1 pto.)Describe brevemente las diferencias de C4.5 respecto a ID3

8. (1 pto.) Ventajas de Baging y Boosting frente a otros clasificadores denominados fuertes.

9. (1 pto.) Enumera las ventajas y los inconvenientes que presenta el método de agrupamiento k-medias.

10. (1 pto.) Suponed un conjunto de datos de clasificación que tiene 5 atributos de entrada, 500 ejemplos y 4 clases. Tres de los atributos de entrada son numéricos en el dominio [1.0, 5.0], el cuarto es categórico con 5 valores diferentes y el quinto es binario (2 valores categóricos). Queremos aplicar k-NN y árboles de decisión. ¿Qué preprocesamiento harías para cada una de las dos técnicas de aprendizaje?

11. (1 pto.)¿Por qué (y en qué situaciones) es interesante realizar selección de instancias antes de construir un clasificador?

12. (1 pto.) Suponed un conjunto de datos de clasificación que tiene 4 atributos de entrada, 500 ejemplos y 2 clases. El 25% de los ejemplos están en la segunda clase, y el 75 en la primera. ¿Cómo abordarías el problema para aplicar árboles de de decisión?

13. (1 pto.) Describe la técnica de validación bootstraping.

14. (1 pto.) ¿Cómo abordarías un problema de clasificación con 6 clases? Explica formas de resolver el problema, y sus ventajas e inconvenientes.

15. (1 pto.) ¿Cómo resolverías un problema de clasificación con 30 millones de instancias y 300 características?

Tiempo: 120 minutos

C/ Periodista Daniel Saucedo Aranda s/n, ETSI Informática y de Telecomunicación, 18071 Granada
Tlf: +34.958.244019 Fax: +34.958.243317

DECSAI
Departamento de Ciencias de la Computación e I.A.
Universidad de Granada

INTELIGENCIA DE NEGOCIO

Convocatoria Extraordinaria de Septiembre
1 de Septiembre de 2014

1. (1 pto.) a) Aunque por abuso del lenguaje hemos hablado de minería de datos y de KDD como sinónimos, indica las diferencias entre ambos términos.

   b) Identifica y describe muy brevemente las etapas dentro del proceso de KDD. Haz un gráfico con las etapas.

2. (2 ptos.) Explica brevemente los modelos básicos de multiclasificador: Bagging y Boosting. Aspectos a destacar de cada uno de ellos. Sus diferencias como multiclasificadores.

3. (2 ptos.) a) ¿Por qué (y en qué situaciones) es interesante realizar selección de variables antes de construir un clasificador?
   b) ¿Por qué (y en qué situaciones) es interesante realizar selección de instancias antes de construir un clasificador?

4. (3 ptos.) En las siguientes situaciones, determinar si estamos ante un problema de minería datos y qué tipo de técnicas se podrían aplicar. Razonarlo (3 puntos).

   a. Dividir los clientes de una compañía de acuerdo a su género
   b. Dividir los clientes de una compañía de acuerdo a su fiabilidad
   c. Calcular las ventas totales de una compañía
   d. Predecir el precio futuro del almacén de una compañía en base a registros históricos
   e. Monitorizar el latido de corazón de un paciente para detectar situaciones anormales
   f. Extraer frecuencias de una señal de sonido

5. (2 ptos.) Suponed un conjunto de datos de clasificación que tiene 4 atributos de entrada, 500 ejemplos y 2 clases. El 15% de los ejemplos están en la segunda clase, y el 85 en la primera. ¿Qué preprocesamiento se puede aplicar a los datos para emplear las técnicas árboles de decisión?

Todas las preguntas tienen la misma puntuación
Tiempo: 120 minutos

C/ Periodista Daniel Saucedo Aranda s/n, ETSI Informática y de Telecomunicación, 18071 Granada
Tlf: +34.958.244019 Fax: +34.958.243317