



# DECSAI

**Departamento de Ciencias de la Computación e I.A.**

Universidad de Granada

## 10 EXÁMENES 2014-2018



# DECSAI

**Departamento de Ciencias de la Computación e I.A.**

Universidad de Granada

## INTELIGENCIA DE NEGOCIO

Convocatoria Ordinaria de Febrero

27 de enero de 2014

- Aunque por abuso del lenguaje hemos hablado de minería de datos y de KDD como sinónimos, indica las diferencias entre ambos términos.
  - Identifica y describe muy brevemente las etapas dentro del proceso de KDD. Haz un gráfico con las etapas.
- Enumera y describe dos problemas abordados en minería de datos. Pon un ejemplo de aplicación real (que no sea el de prácticas) y menciona un algoritmo clásico para cada uno de ellos.
- ¿Por qué es interesante realizar selección de atributos antes de construir un clasificador? Enumera brevemente al menos dos justificaciones para su uso.
- Describe el significado de las medidas de soporte y confianza en reglas de asociación.
- Suponed un conjunto de datos de clasificación que tiene 4 atributos de entrada, 500 ejemplos y 3 clases. Tres de los atributos de entrada son numéricos en el dominio  $[1.0, 5.0]$  y el cuarto es categórico con 4 valores diferentes. ¿Qué preprocesamiento aplicarías a los datos para emplear las técnicas de vecino más cercano?

**Todas las preguntas tienen la misma puntuación**  
**Tiempo: 100 minutos**



### 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 multclasificador: Bagging y Boosting. Aspectos a destacar de cada uno de ellos. Sus diferencias como multclasificadores.
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 de 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**



# DECSAI

**Departamento de Ciencias de la Computación e I.A.**

Universidad de Granada

## INTELIGENCIA DE NEGOCIO

Convocatoria Extraordinaria de Septiembre

4 de Diciembre de 2014

6. (1 pts.) Explica las ventajas e inconvenientes de un selector de características de tipo envolvente y de tipo filtro.
7. (1 pts.) Describe brevemente las diferencias de C4.5 respecto a ID3
8. (1 pts.) Ventajas de Baging y Boosting frente a otros clasificadores denominados fuertes.
9. (1 pts.) Enumera las ventajas y los inconvenientes que presenta el método de agrupamiento k-medias.
10. (1 pts.) 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 pts.) ¿Por qué (y en qué situaciones) es interesante realizar selección de instancias antes de construir un clasificador?
12. (1 pts.) 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 pts.) Describe la técnica de validación bootstrapping.
14. (1 pts.) ¿Cómo abordarías un problema de clasificación con 6 clases? Explica formas de resolver el problema, y sus ventajas e inconvenientes.
15. (1 pts.) ¿Cómo resolverías un problema de clasificación con 30 millones de instancias y 300 características?

**Tiempo: 120 minutos**



# DECSAI

**Departamento de Ciencias de la Computación e I.A.**

Universidad de Granada

## INTELIGENCIA DE NEGOCIO

Convocatoria Ordinaria de Febrero

27 de Enero de 2015

1. (1 pts.) ¿Por qué (y en qué situaciones) es interesante realizar selección de variables antes de construir un clasificador?
2. (1 pts.) Define un árbol de decisión y describe el algoritmo básico de construcción de un árbol de decisión.
3. (1 pts.) Describe la técnica de validación bootstrap y discute qué aporta en el diseño de clasificadores.
4. (2 pts) 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**



# 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 bootstrapping.
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**





# DECSAI

**Departamento de Ciencias de la Computación e I.A.**

Universidad de Granada

## 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 multclasificador: Bagging y Boosting. Aspectos positivos a destacar de cada uno de ellos. Sus diferencias como multclasificadores.
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**





# 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 bootstrapping.
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**



# 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**





### INTELIGENCIA DE NEGOCIO

#### Convocatoria Extraordinaria

12 de Julio de 2017

1. (1 pts.) Explicad las ventajas e inconvenientes de un selector de características de tipo envolvente y de tipo filtro.
2. (1 pts.) Describid brevemente las diferencias de C4.5 respecto a ID3
3. (1 pts.) a) Ventajas de Bagging y Boosting frente a otros clasificadores denominados fuertes. b) Diferencias entre Bagging y Boosting
4. (1 pts.) Enumera las ventajas y los inconvenientes que presenta el método de agrupamiento k-medias.
5. (1 pts.) 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 pts.) ¿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 pts.) 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 pts.) Describid la técnica de validación bootstrapping.
9. (1 pts.) ¿Cómo se abordaría un problema de clasificación con 6 clases? Explica formas de resolver el problema, y sus ventajas e inconvenientes.
10. (1 pts.) ¿Cómo se resolvería un problema de clasificación con 5 millones de instancias y 5 millones características?

**Tiempo: 120 minutos**



### INTELIGENCIA DE NEGOCIO

Convocatoria Extraordinaria, 2 de Febrero de 2018

1. (0.8 pts.) ¿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 pts.) Explique las características de los algoritmos de agrupamiento (o segmentación) 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 pts.) 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 pts.) 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 extraídas en el paso anterior. Si el interés mínimo es de 0.6, ¿qué reglas deberíamos considerar como aptas?

5. (1.3 pts.) Considere el dataset de clasificación de 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**