1. ¿Qué significado tienen desde el punto de vista intuitivo las medidas de error sensibilidad y especificidad para problemas de clasificación de dos clases?

RESPUESTA: La sensibilidad tiene como objeto medir el ratio de verdaderos positivos. Mide la capacidad que tiene un clasificador para no errar en la identificación de positivos clasificándolos como negativos. La especificidad hace justo lo contrario, mide la capacidad de no clasificar los datos erróneamente como positivos si son negativos.

2. ¿En qué consiste el sobreaprendizaje (overfitting) en la construcción de un clasificador? ¿Es posible evitarlo?

RESPUESTA: El sobreaprendizaje ocurre cuando un clasificador aprende muy bien el conjunto de entrenamiento a costa de perder su capacidad de generalización. No existen métodos para evitarlo de forma consistente aunque si hay técnicas para tratar de atenuar su efecto, como el uso de modelos más simples o la detención prematura del entrenamiento mediante validación cruzada.

3. ¿Puedo resolver un problema de clasificación de N clases (N > 2) si tengo un método de clasificación que solo puede distinguir entre dos clases?

RESPUESTA: Sí, se puede transformar el problema de N clases en M problemas de dos clases. Métodos conocidos son el one-vs.-one, el one-vs.all o los códigos ECOC.

4. Indique cómo llevaría a cabo la comparación de los métodos siguientes de clasificación:

(a) Comparación de dos métodos sobre un conjunto de N problemas.

(b) Comparación de un método contra un serie de métodos estándar sobre un conjunto de N problemas para ver si es mejor que todos ellos.

RESPUESTA: Para el primer caso tendríamos el test de Wilcoxon. Para el segundo caso aplicaríamos primero un test de Friedman o Iman-Davenport para ver si hay diferencias significativas globales. En caso de que sí las haya podemos aplicar el procedimiento de Holm para comparar nuestro método con cada uno de los métodos estándar pasa a paso.

5. ¿Qué tipo de clústers tiende a generar un metodo de clustering particional como por ejemplo k-medias?

RESPUESTA: Genera normalmente clústers homogéneos y de forma globular, es por ello que funciona pobremente si nuestros clústers no corresponden a esta forma.

6. ¿Qué diferencias hay entre los tipos típicos de clústers generados por los métodos jerárquicos de single link y complete link?

RESPUESTA: Ambos métodos se utilizan para medir la distancia entre dos clusters, single link (enlace simple) toma como dicha distancia a la menor entre todos los puntos de ambos clusters. Por otro lado, complete link (enlace completo) hace justo lo contrario, escoge la mayor de todas.

Enlace simple → Puede manejar correctamente formas no elípticas pero es más sensible al ruido y a los outliers.

Enlace completo → Más robusto frente al ruido, tiende a romper grandes agrupaciones de clusters.

7. ¿En qué se basa el algoritmo a priori para la generación de conjuntos frecuentes de ítems en análisis de reglas de asociación para poder reducir el número de posibles conjuntos de ítems a visitar?

RESPUESTA: Se basa en el principio a priori, que dice que si un conjunto de items es frecuente, todos sus subconjuntos también lo son. Cualquier subconjunto del superconjunto de items, va tener un mayor o igual soporte que el conjunto inicial.

8. ¿Qué efectos puede tener la existencia de ruido o de outliers en un método de boosting de construcción de agrupaciones de clasificadores?

RESPUESTA: Como el método de boosting va adaptando los patrones de entrenamiento según éstos sean bien clasificados o no (aumentando la probabilidad de que estos patrones aparezcan en el conjunto de entrenamiento en cada ronda) se corre el riesgo de aprender el ruido u outlier.

9. Disponemos de tres métodos de clasificación para resolver una serie de problemas. Estos métodos son un árbol de decisión, un máquina de vectores soporte (SVM) y un clasificador por vecino más cercano (1-NN). Tenemos los siguientes tres problemas:

(a) Un problema con un número moderado de variables pero cientos de miles de instancias.

(b) Un problema con un número moderado de instancias pero miles de variables.

(c) Un problema con patrones que contienen mucho ruido.

Indica cuál de los tres métodos anteriores sería el más adecuado para cada problema.

RESPUESTA:
a) Vecino más cercano→ En el problema b) no sería adecuado aplicarlo porque al tener una alta dimensionalidad (muchos/as atributos/variables) las métricas de distancia no serían significativas y por tanto inútiles para clasificar. En el problema c), como el k escogido es muy bajo (igual a 1), sé es demasiado susceptible al ruido, por lo cual es bastante malo también.
Se escoge por descarte la a), que es la única que clasifica bien.

b) Máquina de vectores soporte (SVM) → Las máquinas de vectores soporte están diseñadas para trabajar con datos de alta dimensionalidad (muchos/as atributos/variables), y no son buenas ante la presencia de ruido u outliers (tienden a sobre-entrenar).

c) Árbol de decisión → Gracias a la poda, que se realiza en árboles muy complejos que no generalizan bien. Reduce el número de hojas para prevenir el sobre-entrenamiento debido al ruido presente en los datos.

10. Indica una ventaja del clustering jerárquico con respecto al particional.

RESPUESTA: No es necesario fijar el número de clusters. Es determinista.
Detecta outliers (patrones o valores extraños “no normales”, no se quedan dentro de un cluster, forman ruido).

11. ¿Por qué el error de clasificación no es una buena medida en problemas de clasificación de dos clases con un gran desequilibrio de clases?

RESPUESTA: En problemas de clasificación, el error de clasificación es una medida sesgada a la clase mayoritaria, entonces en problemas desbalanceados es más importante ser capaz de clasificar bien a los patrones de la clase minoritaria.

Ejemplo:
- Clase 0 → 9990 patrones
- Clase 1 → 10 patrones

Un modelo que clasifica en la clase 0 todos los patrones, obtiene un CCR del 9.990/10.000 = 99,9 %. Se está despreciando una parte del problema. Si se pasa una medida de sensibilidad para la clase 1 se obtiene un valor de 0.

(Ejemplo que ponía Chervas de hombres y mujeres en la clase).

12. Si se le da como dato de un problema la matriz de similaridad entre todos los elementos de un conjunto de datos, ¿es posible realizar con dicha información el clustering usando los métodos complete link, single link y average link?

RESPUESTA: Sí, dado que los tres métodos funcionan con similaridad y dissimilaridad. Entonces dada una matriz de similaridad los métodos son aplicables.

(*) Si se pidiera utilizar un método según la distancia entre centroides, sería necesario conocer cada uno de los atributos de los puntos que forman el conjunto.