Resumen IMD

Preguntas exámenes

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

La sensibilidad tiene como objetivo 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.

- **¿En qué consiste el sobre-aprendizaje (overfitting) en la construcción de un clasificador? ¿Es posible evitarlo?**

El sobre-aprendizaje 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 sí hay técnicas para tratar de atenuar su efecto, como el uso de modelos más simples o la detención prematura del entrenamiento mediante validación cruzada.

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

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

- **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 una serie de métodos estándar sobre un conjunto de N problemas para ver si es mejor que todos ellos.

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

- **¿Qué tipo de clústers tiende a generar un método de clustering particional como por ejemplo k-medias?**

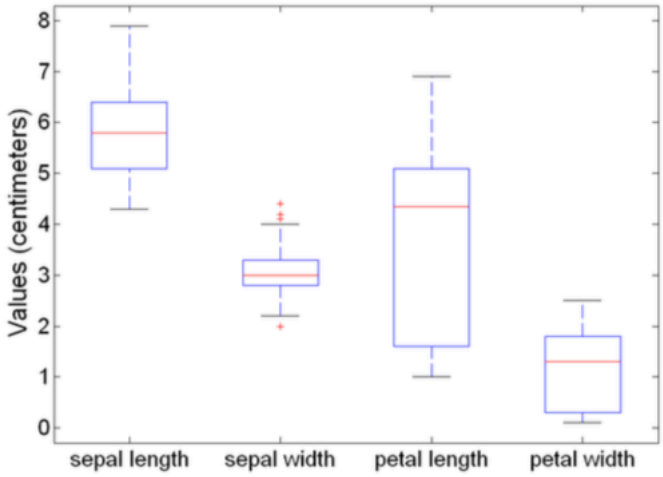
Genera normalmente clústers homogéneos y de forma globular, es por ello que funciona pobremente si nuestros clústers no corresponden a esta forma.

- **Indique dos formas de gestionar valores perdidos (missing values) en un conjunto de datos. Indique también qué ventajas e inconvenientes ve en cada una de ellas.**

Hay dos formas fundamentales, ignorar valores perdidos o estimarlos. Si se ignoran se puede hacer a nivel de instancia, ignorando todas las instancias con al menos un valor perdido, o a nivel de variable, ignorando aquellas variables que tienen al menos un valor perdido en una instancia. El problema de esta aproximación es que podemos perder mucha información.

En el segundo enfoque se estiman los valores perdidos, usando modas, medianas, medias, estimaciones estadísticas o vecinos más cercanos. En cualquier método que usemos tenemos el problema de que introducimos ruido en la muestra.

- **Considere el box plot de las cuatro variables del problema iris mostrado en la siguiente Figura. ¿Qué información puede obtener de esa representación respecto al comportamiento de las variables?**



Según el gráfico, podemos observar que la variable petal length tiene una gran dispersión al igual que, en menor medida, las variables sepal length y petal width. Por el contrario, sepal width tiene valores muy homogéneos entre los diferentes patrones. Respecto a la importancia de estas variables en la clasificación, de un gráfico tipo box plot NO podemos deducir nada, ya que una variable más homogénea puede ser más discriminante que una variable con mayor dispersión.

- **Indique un aspecto positivo y otro negativo de cada uno de los tres siguientes métodos de clasificación: un árbol de decisión, una máquina de vectores soporte y el método de vecino más cercano.**

Árboles de decisión:

- Positivo: Capaces de tratar con problemas muy grandes, muy rápidos en la clasificación, interpretables cuando son pequeños. Buena relación coste/rendimiento, inestables.
- Negativo: Menor rendimiento que otros métodos, inestables.

Máquinas de vector soporte:

- Positivo: Pueden ser muy eficientes con conjuntos de datos con miles de variables, muy buen rendimiento, robustos ante la presencia de ruido, estables.
- Negativo: Son muy costosos computacionalmente, muy sensibles a los parámetros de entrenamiento, estables.

Vecino más cercano:

- Positivo: No necesitan entrenamiento, buen rendimiento, estables ante variaciones en el conjunto de instancias, inestable ante variaciones en el conjunto de variables.
- Negativo: Necesitan almacenar el conjunto de entrenamiento completo por lo que tienen problemas de escalabilidad, estables ante variaciones en el conjunto de instancias, inestable ante variaciones en el conjunto de variables.

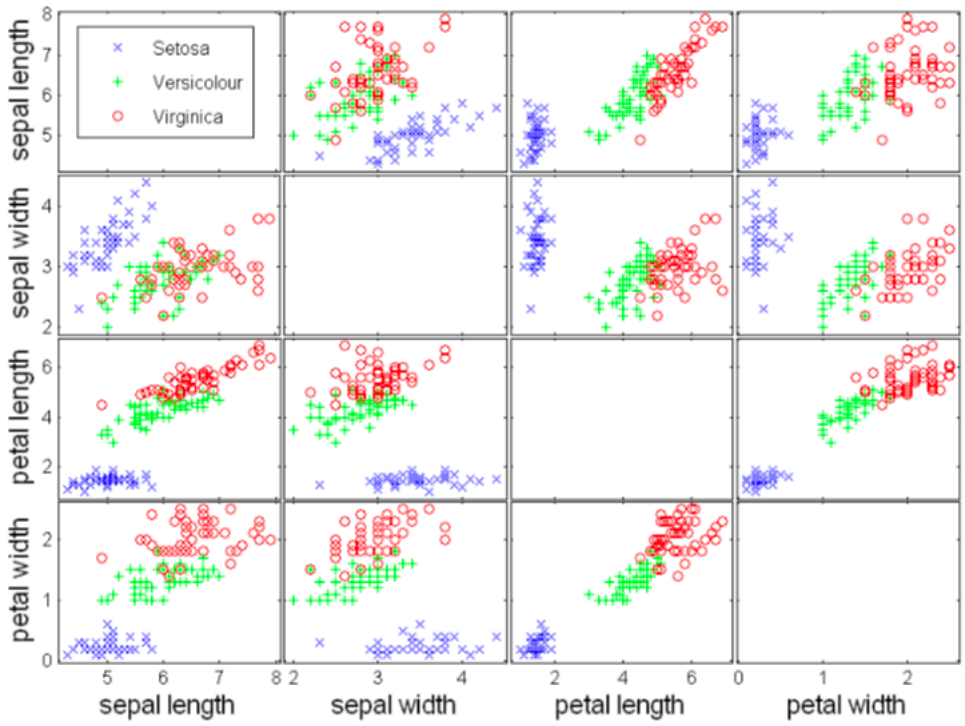
- **En la construcción de reglas de asociación, ¿qué efecto tiene el uso de un soporte mínimo variable según los ítems en un itemset sobre el algoritmo A priori?**

El soporte pierde la propiedad de anti-monotonía y por lo tanto el algoritmo A priori deja de ser aplicable porque está basando en dicha propiedad. Existen diferentes modificaciones del algoritmo para poder seguir siendo aplicado aunque con menos efectividad.

- **Indique dos puntos fuertes del agrupamiento jerárquico con respecto al particional.**

Un punto fuerte es que no asume un número determinado de clústers en el conjunto de datos. Otro punto fuerte es que el resultado es una taxonomía de las instancias que puede ser de mucha utilidad en muchas áreas de conocimiento.

- **Considere el gráfico de dispersión siguiente, que representa las cuatro variables del problema iris. ¿Qué información se puede extraer del problema iris y de cada una de las variables de la gráfica?**



Se pueden obtener diferentes conclusiones. Como idea general podemos afirmar que le problema será fácil de resolver porque hay varios pares de variables que muestran una clara separación entre las clases. Cualquier algoritmo de clasificación será capaz por lo tanto de aprender una clasificación correcta. Por otro lado, respecto a las variables, vemos que sepa-with muestra en general una peor separación entre las clases mientras que las otras tres variables son mucho más discriminantes.

- **Indique qué test de hipótesis sería el más adecuado en cada una de las siguientes situaciones:**

a) Deseamos comparar dos métodos de clasificación aplicados a una serie de problemas de prueba para establecer si uno de los dos es mejor que el otro.

b) Deseamos comprobar si un método que hemos desarrollado es superior a cuatro métodos previos. Hemos probado los cinco métodos en un conjunto de problemas de prueba.

c) Deseamos comprobar si existen diferencias en un conjunto de problemas de prueba entre 6 métodos de clasificación estándar.

Para el primer caso, podemos aplicar el test de Wilconxon ya que es un test diseñado para comparar dos métodos sobre un conjunto de problemas. En el segundo caso las comparaciones en parejas usando el test de Wilconxon no sería apropiadas porque acumularíamos los errores. El procedimiento de Holm sería lo recomendable. Finalmente, cuando comprobamos un grupo de métodos entre sí, el test de Nemenyi sería el apropiado siempre y cuando hayamos hecho previamente un test global de diferencias como Friedman o Iman-Davenport.

- **¿Qué función cumple el parámetro C en la construcción de una máquina de vector soporte?**

Los SVM tratan de resolver el problema de clasificación mediante un modelo lineal. En la mayoría de los casos no es posible separa todas las instancias usando un clasificador lineal, en esos casos, C corresponde a la penalización por clasificar incorrectamente un patrón.

- **¿Qué pros y contras encuentra en el algoritmo bisecting k-medias respecto al algoritmo k-medias estándar?**

Como ventajas tiene la no necesidad de fijar los centroides iniciales y que provee como solución final un clustering jerárquico además del particional. Como inconveniente su mayor coste computacional.

- **Considere un problema de clasificación en dos clases, Valor={Bajo, Alto}, con los siguientes atributos:**
  - **Aire acondicionado = {Funcionando, Roto}**
  - **Motor = {Bueno, Malo}**
  - **Kilometraje = {Alto, Medio, Bajo}**
  - **Corrosión = {Sí, No}**

Considere un clasificador basado en reglas con el siguiente conjunto de reglas:
Kilometraje = Alto -> Valor = Bajo
Kilometraje = Bajo -> Valor = Alto
Aire acondicionado = Funcionando, Motor = Bueno -> Valor = Alto
Aire acondicionado = Funcionando, Motor = Malo -> Valor = Bajo
Aire acondicionado = Roto -> Valor = Bajo
Responda la siguientes preguntas:

A) ¿Son las reglas mutuamente excluyentes?

No.

B) ¿Es el conjunto de reglas exhaustivo?

Sí.

C) ¿Es necesario ordenar las reglas?

Sí, porque hay instancias que disparan más de un regla.

D) ¿Es necesario definir una clase por defecto para el conjunto de reglas?

No, porque el conjunto es exhaustivo.

- **¿Qué diferencia hay entre los tipos típicos de clústers generados por la métodos jerárquicos de single link y complete link?**

Ambos métodos se utilizan para medir la distancia entre dos clusters, single link toma como dicha distancia a la menor entre todos los puntos de ambos clústers. Por otro lado, complete link hace junto lo contrario, escoge la mayor de todas.
Enlace simple -> Puede manejar correctamente formas no elípticas pero es más sensible al ruido y a outliers.
Enlace completo -> Más robusto frente al ruido, tiende a romper grandes agrupaciones de clústers.

- **¿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?**

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

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

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.

- **Disponemos de tres métodos de clasificación para resolver una serie de problemas. Estos métodos son un árbol de decisión, una máquina de vectores soporte y un clasificador por vecino más cercano(1-NN). Tenemos los siguientes tres problemas:**
  - **Un problema con un número moderado de variables pero cientos de miles de instancias.**
  - **Un problema con número moderado de instancias pero miles de variables.**
  - **Un problema con patrones que contienen mucho ruido.**

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

A) Vecino más cercano -> En el problema b) no sería adecuado aplicarlo porque al tener una alta dimensionalidad las métricas de distancia no serían significativas y por tanto inútiles para clasificar. En el problema c), como el k escogido es muy bajo, es demasiado susceptible al ruido, por lo cual es bastante malo también. Por descarte, se escoge la a).
B) Máquina de vectores soporte -> Las SVM están diseñadas para trabajar con datos de alta dimensionalidad, 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.

- **Indica una ventaja del clustering jerárquico con respecto al particional.**

No es necesario fijar el número de clústers. Es determinista. Detecta outliers(patrones o valores extraños, no se quedan dentro de un clúster, forman ruido)

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

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 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.

- **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 clústering usando los métodos complete, single y average link?**

Sí, dado que los tres métodos funcionan con similaridad y disimilaridad. Entonces dada una matriz de símilaridad los métodos son aplicables.

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