ANÁLISIS DE DATOS

Ricardo Aler Mur

1. ¿Qué se entiende por particiones estratificadas? ¿Para qué tipo de problemas puede ser interesante utilizarlas?
2. ¿Qué es la "maldición de la dimensionalidad"?
3. ¿Cuál es la principal desventaja de PCA si lo usamos para seleccionar atributos en clasificación?
4. Describir brevemente lo que hacen los dos tipos de técnicas (edición y condensación) que existen para seleccionar instancias para el algoritmo de k-vecinos
5. Una matriz de confusión de dos clases es de 2x2=4 componentes, pero es suficiente con dos valores para definirla completamente (por ejemplo, es suficiente con TP y FP). ¿Cuántos valores son necesarios para definir de manera completa una matriz de confusión de 3 clases? ¿Y una de N clases? ¿por qué?
6. ¿Es un clasificador con TP = 0.2 y FP = 0.2 trivial? ¿Por qué?
7. ¿Cuál es la diferencia entre árboles de modelos y árboles de regresión?
8. ¿Qué efecto tienen los atributos irrelevantes sobre los árboles de decisión? ¿y sobre el algoritmo del vecino más cercano?
9. Dibujar un dominio de datos en dos dimensiones y dos clases que requiera construir un árbol de decisión con muchos nodos

1) **Respuesta:** dos particiones de datos (entrenamiento y test) están estratificadas, si por ejemplo, la proporción de datos positivos y negativos en la partición de entrenamiento es la misma que en la partición de test. Este tipo de particionamiento es útil en general, puesto que la partición de test tiene que ser representativa de la partición de entrenamiento. Pero es obligatorio utilizarlas en problemas de muestra desbalanceada puesto que si por ejemplo, hay muy pocos datos positivos y el particionamiento de los datos se hace de manera aleatoria, puede llegar a ocurrir que todos los datos positivos acaben en entrenamiento y ninguno en test, o al contrario.

2) **Respuesta:** a medida que crece el número de dimensiones (atributos de entrada), el número de datos que son necesarios para hacer el aprendizaje crece, en el peor caso, de manera exponencial (nota: recordar el caso de la superficie de las hiperesferas).

3) **Respuesta:** PCA es un método de transformación y selección de atributos que identifica los componentes más importantes para explicar la variabilidad de los datos. Por ejemplo, el primer componente de PCA es el eje que explica la máxima cantidad de varianza. El segundo componente es el siguiente que explica la mayor cantidad de varianza, y así sucesivamente. El mayor problema de PCA en clasificación es que es un método de transformación y selección no supervisado. Es decir, en ningún momento tiene en cuenta la clase de los datos con lo que en algunas ocasiones puede ocurrir que los componentes más importantes no sean los que estén más correlacionados con la clase.

4) Las técnicas de edición (como la de Wilson) eliminan aquellos datos "raros" en su entorno cercano, eliminando de esta manera los datos con ruido y suavizando las fronteras. Las técnicas de condensación eliminan aquellos datos redundantes (o sea, de la misma clase que los de su entorno), por tanto innecesarios para la clasificación y reduciendo grandemente el número de datos que es necesario almacenar en el algoritmo del vecino más cercano.

5) **Respuesta:** Una matriz de confusión M de 3 clases es de 3x3 (= 9 componentes). Por ejemplo, en la primera columna está la proporción de los datos de la clase 1 clasificados correctamente M[1,1], la de los datos de la clase 1 clasificados incorrectamente en la clase 2: M[2,1] y de igual manera con la clase 3: M[3,1]. Pero la suma de M[1,1]+M[2,1]+M[3,1] tiene que ser 1, puesto que se trata de todos los datos de la clase 1 (el 100%). Lo mismo ocurre con las clases 2 y 3: M[1,i]+M[2,i]+M[3,i]. Por tanto, en cada columna, uno de los componentes lo podemos calcular como 1 menos la suma de los otros dos. Por ejemplo, M[3,i] = 1-(M[1,i]+M[2,i]). Eso quiere decir que en cada columna un componente se puede calcular a partir de los otros dos. Por tanto, de los 9 componentes, será suficiente con (3-1)+(3-1)+(3-1) = 6 componentes para representar de manera completa una matriz de confusión con 3 clases. Si hubiera N clases, tenemos N columnas y por tanto será suficiente con N*(N-1) componentes.

6) **Respuesta:** razón superficial: sabemos que todos aquellos clasificadores en el espacio ROC con TP=FP son triviales. Explicación más profunda: la razón es que si tiramos una moneda cuya probabilidad de caer de cara es, digamos, de 0.9 (y de caer de cruz 0.1), ya tendríamos por pura casualidad un clasificador con TP = 0.9 y FP = 0.9, puesto que esa moneda clasificaría como positivos el 90% de los datos positivos (o sea TP = 0.9) y como positivos el 90% de los negativos (luego FP = 0.9).

7) **Respuesta:** Los árboles de regresión tienen números en las hojas (que son la media de todos los datos que llegaron a esa hoja) y los árboles de modelos tienen modelos de regresión lineal en las hojas.

8) **Respuesta:** Los atributos irrelevantes son desastrosos para el algoritmo del vecino más cercano, puesto que falsean completamente las distancias entre datos, pero mucho menos para los árboles de decisión, porque los atributos irrelevantes no están correlacionados con la clase y por tanto no serán elegidos para ponerlos en los nodos del árbol.

9) **Respuesta:** Pensadla ☺ y preguntad si no se os ocurre.