Ricardo Aler Mur

AUTOEVALUACIÓN DE Selección y generación de atributos-I, CON PREGUNTAS Y RESPUESTAS

1) Enumerar y justificar varias razones por las que la selección de atributos es importante

Respuesta. 1) la existencia de atributos redundantes: algunos algoritmos de clasificación como el Naive Bayes aprenden mal en presencia de atributos redundantes. 2) la existencia de atributos irrelevantes: aunque los algoritmos de aprendizaje son capaces hasta cierto punto de no utilizar estos atributos, su presencia degrada el funcionamiento del clasificador, con lo que es conveniente eliminarlos en la fase de preproceso. 3) Maldición de la dimensionalidad: a medida que crece el número de dimensiones, el número de datos necesario para ajustar los modelos crece de manera rápida (exponencial en el peor caso). 4) En ocasiones es importante conocer qué atributos son relevantes (por ejemplo, que genes son relevantes a la hora de predecir cierta enfermedad).

2) Si para construir un clasificador lineal tenemos tantas dimensiones como datos de entrenamiento, ¿cuál sería aproximadamente el porcentaje de aciertos en test?

Respuesta: Dado que un clasificador lineal va a ser capaz de separar tantos datos como dimensiones tenga (incluso aunque los datos sean aleatorios), el clasificador que se contruye con dicho conjunto de entrenamiento va a ser prácticamente arbitrario, con lo que el porcentaje de aciertos esperado en test se aproximará al del azar.

3) ¿Por qué la búsqueda exhaustiva no es factible en la mayor parte de los problemas prácticos?

Respuesta: porque el número de subconjuntos que se pueden hacer con n atributos crece de manera exponencial (2^n)

4) ¿Cuál es la principal ventaja de los métodos Wrapper frente a Filter con Ranking? ¿Y su principal desventaja?

Respuesta: Los métodos Wrapper son capaces de detectar atributos que funcionan bien en conjunto pero mal por separado. Los métodos Ranking no pueden hacer esto, puesto que evalúan a los atributos de manera individual. La principal desventaja de Wrapper es el tiempo que tardan, puesto que involucra lanzar un algoritmo de aprendizaje automático por cada subconjunto que hay que evaluar.

5) ¿Cuál es la principal ventaja de Random Projections sobre PCA?

Respuesta: Su rapidez, junto con el hecho de que para dimensionalidades altas, sus resultados son muy similares.