uc3m | Universidad Carlos III de Madrid

OPENCOURSEWARE

APRENDIZAJE AUTOMÁTICO PARA EL ANÁLISIS DE DATOS

GRADO EN ESTADÍSTICA Y EMPRESA

Ricardo Aler

APRENDIZAJE AUTOMÁTICO (GR. ESTADÍSTICA Y EMPRESA). 2017-18
9 preguntas, 1.5 horas de duración.

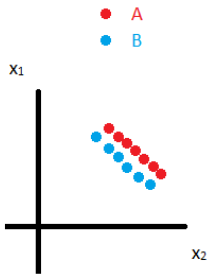
---

1. Supongamos que queremos entrenar un modelo para un problema de clasificación a partir de unos datos disponibles. Si podemos elegir entre estos dos métodos, train/test y validación cruzada, ¿de cuál de ellos se espera que nos proporcione un modelo con una tasa de aciertos mayor?

   Respuesta: Train/test y validación cruzada son métodos para estimar el comportamiento futuro de los modelos, pero no para obtener modelos, por lo que la respuesta sería que ninguno de los dos. En ambos casos, para aprender el modelo final, usaremos todos los datos disponibles, por lo que en ambos casos el modelo obtenido sería idéntico.

2. Describir un ejemplo de problema de clasificación en el que un método de selección de atributos de tipo *filter* no funcionaría bien, pero uno *wrapper* sí.

   Respuesta: En general, cualquier problema de clasificación en el que dos atributos por separado no predigan la clase, pero en conjunto si. Un ejemplo concreto sería el siguiente:



3. Supongamos que un método de selección de atributos *wrapper* está estudiando la combinación de los atributos a_1 y a_5, teniendo la tabla de datos 10 atributos (a_1 hasta a_{10}). ¿Qué procedimiento seguiría un método wrapper para evaluar conjuntamente dichos atributos a_1 y a_5?

Respuesta: los métodos wrapper usan métodos de aprendizaje automático para evaluar subconjuntos o grupos de atributos. Por ejemplo, para evaluar a1 y a5, un método wrapper construiría un modelo (un árbol por ejemplo) usando sólo esos dos atributos, y evaluaría dicho modelo con un conjunto de validación. El resultado de dicha validación sería la evaluación conjunta de esos dos atributos.

4. Describir el procedimiento mediante el cual Random Forest ordena los atributos de entrada de acuerdo a su importancia.

Respuesta: De manera resumida, RF evalúa un atributo calculando el error out-of-bag usando una tabla de datos modificada, donde el atributo que queremos evaluar toma valores arbitrarios. Si el error es grande, eso quiere decir que el atributo es importante.

Los pasos concretos son estos:

1. Se calcula el error out-of-bag $\hat{e}$
2. Recordemos que un atributo a_k no es más que una columna de valores en la tabla de datos
3. Reordenamos aleatoriamente los valores de a_k en la tabla de datos de entrenamiento y calculamos el nuevo error out-of-bag $\hat{e}_{ak}$
4. Ordenamos a las variables por diferencia de error: $\hat{e}_{ak} - \hat{e}$. Aquellas que tengan mayores diferencias son más importantes para la predicción

5. Definir el RMSE y enumerar y explicar dos problemas que tiene (Root Mean Square Error) a la hora de evaluar modelos de regresión, y cómo pueden solucionarse.

Respuesta: RMSE es la raíz cuadrada de la media de los cuadrados de los errores de las instancias. Tiene dos problemas:

1. Aquellas instancias muy mal predichas por el modelo (error grande) tienen demasiado peso en el error final porque el error se eleva al cuadrado.
2. Su valor es relativo a la escala de la variable de respuesta. Es decir, multiplicar la variable de salida por 1000 haría que el RMSE fuera aproximadamente 1000 veces mayor, sin que eso implique que el modelo sea 1000 veces peor.

6. Describir el procedimiento “one-versus-one”. ¿En qué situaciones se usa?

Respuesta: Para aplicar one-versus-one es necesario construir varios modelos, uno para separar cada pareja de clases. La clase de una nueva instancia es la clase mayoritaria de dichos modelos (es decir, se vota). En caso de empate entre clases se puede clasificar la nueva instancia como la clase cuya frontera está más lejos de la instancia. Se puede utilizar para abordar problemas con más de 3 clases cuando los modelos que queremos usar, sólo son capaces de separar dos clases (como las máquinas de vectores de soporte).

7. ¿Qué representan las variables de holgura (slack variables) y por qué es necesario utilizarlas?

Respuesta: la distancia desde las instancias mal clasificadas al margen correspondiente a la clase que pertenece dicha instancia mal clasificada. Márgen (distancia de separación entre las dos clases) y número de errores, son objetivos típicamente contrapuestos. Así, las variables de holgura permiten a las SVMs cometer algunos errores, gracias a lo cual el margen puede ser mayor.

8. Supongamos que creamos un ensemble de 100 árboles con dos métodos, Bagging con árboles de decisión y Random Forests. ¿Tardarían aproximadamente el mismo tiempo en entrenar? ¿Por qué?

   Respuesta: puede parecer que como ambos son ensembles de árboles de decisión, tardarían un tiempo similar. Pero como Random Forest sólo necesita evaluar la entropía de m atributos (seleccionados aleatoriamente de entre los M posibles y siendo típicamente m << M), mientras que Bagging necesita evaluar la entropía de todos los M atributos, RF tardaría, en principio, menos.

9. Supongamos que estamos haciendo boosting con árboles para resolver un problema de regresión. ¿Cuál sería el primer modelo que se obtendría? ¿Cómo se obtendría el segundo modelo? Explicar con detalle.

   Respuesta: en boosting se aprenden modelos de manera secuencial. El modelo n+1 intenta predecir correctamente las instancias mal predichas por el modelo n. Esto es cierto para clasificación y para regresión. Concretando para regresión, el primer modelo que se obtendría sería la media de la variable de respuesta de los datos de entrenamiento. El segundo modelo sería un árbol entrenado para resolver un conjunto de datos, cuyas entradas serían los atributos originales y cuya salida sería el pseudo-resíduo, que el caso del segundo modelo sería la diferencia entre la media (modelo anterior) y la variable de respuesta original.