ANÁLISIS DE DATOS

Ricardo Aler Mur

AUTOEVALUACIÓN DE ALGORITMOS DE INDUCCIÓN, CON PREGUNTAS Y RESPUESTAS

1) ¿Qué mide la entropía que se usa para construir árboles de decisión?

Respuesta:.Mide la cantidad de información medida en bits que es necesario utilizar para codificar de forma óptima las clases de los datos. Por ejemplo, en un problema de 10 clases, donde una de las clases es muy frecuente y las otras mucho menos, se le podría asignar un código más corto (en bits) a dicha clase más frecuente, y más largo a las otras. Si las diez clases tuvieran una frecuencia similar, sería necesario codificar todas ellas con el mismo número de bits (log2(10)).

2) ¿Qué es más fácil que sobreadapte, un árbol con muchos nodos o un árbol con pocos nodos?

Respuesta: El que tiene muchos nodos, porque le proporciona más grados de libertad al modelo. En el caso específico de los árboles de decisión, sabemos que los nodos cercanos a las hojas se construyen con muy pocos datos, con lo que es posible que el árbol memorice dichos datos y no generalice bien. Esa es la razón por la que se utilizan técnicas de poda. Por el contrario, un árbol con pocos nodos puede subadaptar.

3) ¿Por qué la entropía (o ganancia en información o information gain) no funciona bien cuando un atributo tiene muchos valores?

Respuesta: Pensemos en un caso extremo con un atributo que tenga tantos valores como datos. En ese caso, se podría poner un dato en cada hoja del árbol y los datos quedarían perfectamente clasificados (con una entropía de cero). Pero evidentemente, esa es una clasificación arbitraria que no va a generalizar más allá de los datos de entrenamiento.

4) ¿Cuál es una manera sencilla de aprender reglas a partir de árboles?

Respuesta: Se puede crear una regla para cada camino desde la raíz hasta cada hoja.

5) ¿Qué tipos de árboles para predicción numérica existen?

Respuesta: Árboles de regresión, donde las hojas contienen las medias de los datos que llegan a dichas hojas y árboles de modelos, donde cada hoja contiene un modelo de regresión lineal.

6) ¿Cómo es posible construir modelos de regresión con variables categóricas?

Respuesta: Los modelos de regresión requieren variables numéricas y las variables categóricas no lo son. Pero a partir de cada variable categórica con n valores, podemos crear n variables binarias, con valor 1 si la variable toma el valor correspondiente y 0 en caso contrario. Estas variables ya se pueden utilizar en los modelos de regresión.