Ricardo Aler Mur

AUTOEVALUACIÓN DE EVALUACIÓN DE TÉCNICAS DE APRENDIZAJE, CON PREGUNTAS Y RESPUESTAS

1) ¿Cuál es el método de evaluación de modelos más comúnmente utilizado cuando se disponen de pocos datos?

Respuesta. El método de validación cruzada. Si se dispone de pocos datos, la evaluación no va a ser significativa y va a ser muy sensible a sesgos aleatorios que pueda contener el conjunto de test. Por eso, la validación cruzada repite el ciclo de entrenamiento-test varias veces (10 veces en el caso de 10-fold crossvalidation) y calcula la media al final, lo que permite que algunos sesgos que aparecen en las particiones de entrenamiento y test de manera aleatoria, se cancelen al calcular la media.

2) ¿Qué son las particiones estratificadas? ¿En qué tipo de problemas es conveniente utilizarlas?

Respuesta: Son particiones en las que se intenta preservar en la partición de entrenamiento y test la proporción de clases que se observa en el conjunto de datos global. El objetivo es que las particiones sean representativas del problema original. Se suelen utilizar en problemas de muestra desbalanceada, puesto que en esos casos puede ocurrir que por casualidad, pocos o ningún dato de la clase minoritaria tengan el número de representantes adecuado en la partición de entrenamiento o test. En casos extremos, si la clase minoritaria tiene muy pocos datos, puede ocurrir que las particiones de entrenamiento o test no contengan ningún dato de dicha clase si las particiones se hacen de manera aleatoria.

3) ¿Qué diferencia hay entre evaluación sensible al coste y aprendizaje sensible al coste?

Respuesta: La evaluación nos permite seleccionar el mejor modelo de entre un conjunto de modelos, de acuerdo a determinadas condiciones operativas. El aprendizaje nos permite aprender un modelo optimizado para dichas condiciones operativas.

4) Decir un algoritmo que permita aprender modelos teniendo en cuenta la distribución de las clases (para problemas desbalanceados)

Respuesta: SMOTE: Synthetic Minority Over-sampling Technique. Esta técnica permite remuestrear los datos de la clase minoritaria, para equipararla en número a las de la clase mayoritaria. No se limita a replicar instancias existentes sino que crea nuevas instancias de manera "razonable" y evita uno de los problemas principales de la replicación pura, que es el sobreaprendizaje.

5) ¿Para qué puede servir una curva ROC?

Respuesta: 1) Permite representar el comportamiento de un clasificador (scorer) para todas las posibles condiciones operativas. 2) Permite seleccionar el mejor threshold de un clasificador (scorer) para condiciones operativas concretas. 3)Permite comparar, seleccionar y descartar clasificadores de un conjunto de ellos, para determinadas condiciones operativas.