uc3m | Universidad Carlos III de Madrid

OPENCOURSEWARE
APRENDIZAJE AUTOMÁTICO PARA EL ANÁLISIS DE DATOS
GRADO EN ESTADÍSTICA Y EMPRESA
Ricardo Aler

Metodología (entrenamiento, evaluación, ajuste de hiper-parámetros)

Los objetivos principales de esta clase son:

• Describir por qué es importante evaluar los modelos, explicando el concepto de generalización y sobreajuste (*overfitting*), tanto en clasificación como en regresión.

• Se explican varios métodos de evaluación: entrenamiento / test, entrenamiento / test repetida y validación cruzada (*crossvalidation*), siendo este último el método recomendado.

• La validación cruzada divide el conjunto de datos disponibles en K particiones independientes. Para cada partición P, se entrena un modelo con todas las particiones excepto P, y se lo evalúa con P. La evaluación final es el promedio de las K particiones.

• Entrenamiento / test puede sufrir sesgos en las particiones, especialmente si los datos son escasos. Entrenamiento / test repetido puede tener solapes en las particiones de test. Se recomienda la validación cruzada porque soluciona los dos problemas anteriores y además cada instancia se usa de las dos maneras posibles: como instancia de entrenamiento y como instancia de test.

• Existen muchas medidas de evaluación, tanto para clasificación como para regresión. Dos de ellas se explican con más detalle: tasa de aciertos (o *accuracy*, para clasificación) y error cuadrático medio RMSE (para regresión).

• Se muestra que RMSE depende de la escala de la variable de salida. Para evitar esto, el RMSE se puede normalizar de varias maneras. Una de ellas es el RMSE relativo, donde el RMSE se divide por el error de la media. Los errores relativos varían de 0 a 1, donde 1 significa que nuestro modelo tiene el mismo error de predicción que la media (y por lo tanto, es un modelo trivial).

• Finalmente, se introduce el proceso de ajuste de hiperparámetros.

• En las diapositivas iniciales, se mostró cómo la complejidad del modelo y el sobreajuste está relacionado (cuanto más complejo es el modelo, más probable es sobreajuste, aunque siempre existe la posibilidad de un ajuste insuficiente)

• Se explica que todos los algoritmos de aprendizaje automático tienen algún hiperparámetro que determina el rendimiento del modelo y, específicamente, su complejidad. Por ejemplo, los árboles de decisión tienen la profundidad máxima y el número mínimo de instancias en las hojas. KNN tiene K, que es el número de vecinos a tener en cuenta.

• Esos hiperparámetros deben ajustarse. El proceso de búsqueda *grid-search* prueba todas las combinaciones posibles de valores de hiper-parámetros. Dado que esto puede ser muy costoso, el procedimiento de búsqueda *random search* permite explorar un número límite de posibilidades, eligiéndolas aleatoriamente. Por último, la *sequential model based optimization* (u optimización bayesiana) permite elegir combinaciones de valores de hiper-parámetros basándose en valores que previamente se mostró que funcionaban bien. Es un proceso iterativo, en el que cada iteración intenta mejorar los valores encontrados en la iteración anterior.

Material asociado:

- Diapositivas de la clase y algunos ejercicios (consultar la guía del curso).
- El tutorial de MLR tiene un apartado dedicado al ajuste de hiper-parámetros.
- Una de las prácticas requiere el ajuste de hiper-parámetros.