11/03/2024

Separar los datos de entrenamiento para tener un dataset de validación (olvidándonos de los datos de tests) para probar diferentes modelos.

Como separar los datos de validación:

* Validación simple: separar un porcentaje en validación y otro en tests. Su principalmente es que el error se puede ver influenciado por la aleatoriedad de la separación.
* Leave One Out Cross-Valdation (LOOCV): Utilizar todos los registros menos uno, y luego hace la media de todos los error.es Su problema principal es su alto coste.
* K-Fold Cross-Validation: Parecido al anterior, pero dividiendo el dataset en un numero n de dataset. Es menos costoso que el anterior
* Repeated K-Fold Cross-Validation: como el anterior, pero repitiendo el proceso varias veces
* Bootstraping: coger grupos del mismo tamaño con reemplazo (es decir, que los grupos se pueden repetir). Se entrena con los grupos elegidos y se valida con los grupos no elegidos. Se repite n veces para evitar que un grupo se escoja varias veces

Sobreajuste vs subajuste

Si el error es muy alto tanto en entrenamiento como en test se considera subajuste e indica que el modelo no es capaz de aprender de los datos. Puede ser debido a un mal análisis exploratorio de datos, falta de registros, falta de atributos en los registros.

Cuando el modelo se ajusta perfectamente a los datos de entrenamiento, pero da mucho error en la validación, indica sobreajuste.