Aprendizaje Automático I

Ejercicio: Valores SHAP

DSLab

noviembre, 2023

Los modelos predictivos complejos (random forest, xgboost, deep learning) no son fáciles de interpretar.

En modelos predictivos, dada una determinada predicción, responder a esta pregunta: ¿cuál ha sido la influencia de cada variable de entrada para obtener esa predicción?, no siempre es sencillo.

Una técnica reciente para interpretar modelos complejos es: SHAP (SHapley Additive exPlanations) desarrollada por Scott M. Lundberg.

SHAP mide el impacto de las variables teniendo en cuenta la interacción con otras variables. Evalua cómo un ligero cambio en una variable puede cambiar mucho el resultado final. Los valores de Shapley calculan la importancia de una característica comparando lo que un modelo predice con y sin la característica. Sin embargo, dado que el orden en que un modelo recibe las características puede afectar a sus predicciones (piensa en un árbol de decisión), esto se hace en todos los órdenes posibles, para que las características se comparen equitativamente.

Consideremos un modelo de `random forest` entrenado sobre los datos `Caravan` de la librería `ISLR2`. Los datos contienen 5822 registros de clientes reales. Cada registro consta de 86 variables, que contienen datos sociodemográficos (variables 1-43) y propiedad de productos (variables 44-86). Los datos sociodemográficos proceden de los códigos postales. Todos los clientes que viven en zonas con el mismo código postal tienen los mismos atributos sociodemográficos. La variable 86 (Compra) indica si el cliente ha comprado una póliza de seguro de caravana. Puede obtenerse más información sobre las distintas variables en http://www.liacs.nl/~putten/library/cc2000/data.html

1. Estudia el conjunto de datos. ¿Qué puedes concluir de los datos dada la distribución de la variable respuesta?
2. Equilibra las clases para que el 50% de las observaciones de entrenamiento tengan 'Yes' en la variable respuesta, y el restante 50% tengan 'No'.
3. Ajusta un modelo de `random forest` empleando la función `randomForest` de la librería `randomForest`. ¿Qué variables son más importantes en la construcción del bosque?
4. Construye una curva ROC para estudiar el rendimiento del modelo.
5. Empleando la función `explian` de librería `DALEX` y la visualización mediante un `plot` explica la predicción obtenida para las dos primeras observaciones de la muestra de prueba.