Ciencia de Datos 2021

Lista 07

17.mayo.2021

1. Elabore un modelo de árbol de decisión o *random forest* para clasificar un conjunto de datos de emails como *spam* o *no spam*.
   La base de datos corresponde a estadísticas de 4601 emails, principalmente conteos de refuencias relativas de ocurrencia de algunas palabras o símbolos particulares, y se encuentra en el archivo `spambase.csv`.

   Evaluar el desempeño del modelo, mediante métricas de clasificación (*accuracy, F1-score, specificity, sensitivity*), o mediante una matriz de confusión. No se olvide de separar sus datos en un conjunto de entrenamiento, uno de prueba. Si lo desea, puede aplicar estrategias de validación cruzada.

   Más información sobre este conjunto de datos puede hallarse en
   `https://search.r-project.org/CRAN/refmans/kernlab/html/spam.html` y
   `http://archive.ics.uci.edu/ml/index.php`.

2. Los datos en el archivo `winequality-white.csv` corresponden a una evaluación de calidad sobre 4898 muestras de vinos, en función de indicadores y pruebas psicoquímicas.
   El objetivo es modelar el valor de la variable $y =$ *quality*, la última columna de la tabla, en función de las otras variables. Si lo considera conveniente, puede trabajar sobre una submuestra de los datos.

   (a) Haga un modelo de regresión lineal ordinaria (OLS) sobre esta base de datos, para estimar y (no se olvide de agregar una columna de 1's para el término constante).

   (b) Elabore gráficos de diagnóstico y determine si su modelo lineal en (a) cumple con los supuestos del modelo de regresión, y si hace un trabajo adecuado para ajustar los datos. Discuta sus hallazgos.

   (c) En función de lo anterior, proponga una alternativa para mejorar el modelado de y (cualquier alternativa que considere conveniente). Elabore un segundo modelo tomando en cuenta su propuesta, y compare el desempeño del segundo modelo con el realizado en (a). Discuta si su propuesta ayuda a mejorar el entendimiento de los datos.

   Más información sobre este conjunto de datos puede hallarse en
   `https://archive.ics.uci.edu/ml/datasets/Wine+Quality`.