Trabajo sobre evaluación de clasificadores

Máster Interuniversitario en Big Data: Tecnologías de Análisis de Datos Masivos

Minería de datos

Curso 2021/2022

El alumno deberá resolver las cuestiones que se plantean. Una vez resueltas, deberá crear una memoria en formato R Notebook/R Markdown con el código incrustado, la salida en formato HTML, en la que quede reflejado el proceso de resolución seguido. Además de la corrección de las soluciones propuestas, se valorará la presentación de la memoria. Recuerda justificar todas las transformaciones que se hagan a los datos utilizados.

El nombre de los archivos a entregar tienen que seguir el formato ApellidosNombre.extesion o ApellidosNombre.extensión. El nombre de alumno debe aparecer claramente identificado en la primera página.

El documento debe contener una primera sección a modo de resumen, en el que se indique:

- Qué procesamiento se ha realizado sobre los datos.
- Qué clasificadores se han probado.
- Qué conclusiones se han alcanzado.

1. Predecir la pérdida de clientes (customer churn)

El fichero customer.csv contiene información sobre los clientes de una determinada compañía de telecomunicaciones. Concretamente, se tiene información sobre algunos aspectos demográficos y los servicios/productos que tiene contratado. En la última columna se indica si el cliente en cuestión abandona la compañía en el mes siguiente. El objetivo del trabajo es construir un clasificador que permita predecir, a partir de la información contenida en el fichero, si el cliente va a abandonar la compañía en el siguiente mes.

2.1. Para una calificación máxima de 5 puntos

- Análisis descriptivo de los datos y aplicación de la técnicas necesarias para limpiar el dataset.
- Un proceso de selección de variables y la generación de los conjuntos de datos con las variables seleccionadas.
- Entrenamiento tres clasificadores (al menos uno debe generar conocimiento explícito), utilizando como medida de calidad el Accuracy y el índice Kappa.
- En el entrenamiento de cada clasificador se debe realizar una búsqueda para determinar qué combinación de hiperparámetros presenta la mejor eficacia.
- Comparación de los clasificadores para determinar cuál es el mejor.
- Evaluación de los clasificadores en los conjuntos de test.

2.2. Para una calificación máxima de 7 puntos

- Incluir lo que se ha mencionado en los puntos anteriores.

- Realizar una búsqueda de hiperparámetros, definiendo los hiperparámetros más adecuado para cada técnica (no se debe restringir la búsqueda de hiperparámetos a los valores por defecto de la librería caret).

2.3. Para una calificación máxima de 9 puntos

- Incluir lo que se ha mencionado en los puntos anteriores.
- Añadir al estudio dos clasificadores más.
- Realizar un proceso de selección de variables adicional. Se deberá indicar en el análisis final si el resultado de aplicar un procesos de selección de variables mejora el resultado obtenido con el data set original.

2.4. Para una calificación máxima de 10 puntos

- Incluir lo que se ha mencionado en los puntos anteriores.
- Repetir el proceso utilizando como medida de calidad el Área bajo la curva ROC.

3. Normas generales

- Se debe garantizar la reproducibilidad de los resultados.
- Se debe aportar toda la evidencia gráfica posible junto con los comentarios pertinentes.
- Se debe comentar el proceso completo, las conclusiones y todas las decisiones tomadas.