Práctica 1: Clasificación

Grupo XX

Curso 2021/2022

Esta práctica debe entregarse en formato pdf, incluyendo el código R utilizado, las correspondientes salidas y los comentarios (o interpretaciones de los resultados) pertinentes (para ello se recomienda emplear RMarkdown, a partir de un fichero *.Rmd* o un fichero *.R* mediante spin).

Se empleará el conjunto de datos `CollegeXX` almacenado en el archivo *CollegeXX.RData*, donde *XX* es el número de grupo, que contiene estadísticas de 500 universidades públicas y privadas de EE.UU. (modificación del conjunto de datos `ISLR::College` con un ligero preprocesado y submuestreo), publicadas en la edición de 1995 de "US News and World Report". Se considerarán como respuestas las variables `Private`, factor indicador de universidad privada ("Yes") o pública ("No"), y como predictores el resto de variables del conjunto de datos:

Predictores	Descripción
Apps	Número de solicitudes de ingreso recibidas (en escala logarítmica)
Accept	Número de solicitudes aceptadas (escala logarítmica)
Enroll	Número de nuevos estudiantes matriculados (escala logarítmica)
Top10perc	Porcentaje de nuevos estudiantes en el 10% de los mejores de la clase de ciencias sociales
Top25perc	Porcentaje de nuevos estudiantes en el 25% de los mejores de la clase de ciencias sociales
F.Undergrad	Número de estudiantes universitarios a tiempo completo (escala logarítmica)
P.Undergrad	Número de estudiantes universitarios a tiempo parcial(escala logarítmica)
Outstate	Número de estudiantes de otro estado (en miles)
Room.Board	Gastos de alojamiento y comida (en miles de dólares)
Books	Estimación del coste de los libros (en cientos de dólares)
Personal	Estimación del gasto personal (en miles de dólares)
PhD	Porcentaje de profesores con doctorados
Terminal	Porcentaje de profesores con grado terminal
S.F.Ratio	Razón de estudiantes por profesor
perc.alumni	Porcentaje de ex-alumnos que donan
Expend	Gasto institucional por estudiante (en miles de dólares)
Grad.Rate	Tasa de graduación

Se debe establecer la semilla igual al número de grupo multiplicado por 10 mediante la función `set.seed()` (también se recomienda hacerlo antes de ajustar cada modelo) y se considerarán el 80% de las observaciones como muestra de aprendizaje y el 20% restante como muestra de test.

Ejercicios

1. Obtener un árbol de decisión que permita clasificar las observaciones (universidades) en privadas (`Private="Yes"`) o públicas (`Private="No"`).

   a. Seleccionar el parámetro de complejidad de forma automática, siguiendo el criterio de un error estándar de Breiman et al. (1984).

   b. Representar e interpretar el árbol resultante.

   c. Evaluar la precisión, de las predicciones y de las estimaciones de la probabilidad, en la muestra de test.

2. Realizar la clasificación anterior empleando Bosques Aleatorios mediante el método `"rf"` del paquete `caret`.

   a. Considerar 300 árboles y seleccionar el número de predictores empleados en cada división `mtry = c(1, 2, 4, 6)` mediante validación cruzada, con 10 grupos y empleando el criterio de un error estándar de Breiman.

   b. Representar la convergencia del error en las muestras OOB en el modelo final.

   c. Estudiar la importancia de las variables y el efecto de las principales empleando algún método gráfico (para la interpretación del modelo).

   d. Evaluar la precisión de las predicciones en la muestra de test y comparar los resultados con los obtenidos con el modelo del ejercicio anterior.

3. Realizar la clasificación anterior empleando SVM mediante la función `ksvm()` del paquete `kernlab`,

   a. Ajustar el modelo con las opciones por defecto.

   b. Ajustar el modelo empleando validación cruzada con 10 grupos para seleccionar los valores "óptimos" de los hiperparámetros, considerando las posibles combinaciones de `sigma = c(0.01, 0.05, 0.1)` y `C = c(0.5, 1, 10)` (sin emplear el paquete `caret`; ver Ejercicio 3.1 en *03-bagging_boosting-ejercicios.html*).

   c. Evaluar la precisión de las predicciones de ambos modelos en la muestra de test y comparar también los resultados con los obtenidos en el ejercicio anterior.