Práctica 2: Regresión

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 respuesta la variable `Accept`, número de solicitudes aceptadas (en escala logarítmica), y como predictores el resto de variables numéricas del conjunto de datos:

Predictores	Descripción
Apps	Número de solicitudes de ingreso recibidas (en 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

Opcionalmente se podría incluir también como predictor la variable `Private`, factor indicador de universidad privada ("Yes") o pública ("No"), aunque seguramente sería recomendable recodificarla como numérica (e.g. 1 = privada, 0 = pública).

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. Ajustar un modelo lineal con penalización *lasso* a los datos de entrenamiento
    a. Seleccionar el parámetro λ de regularización por validación cruzada empleando el criterio de un error estándar.
    b. Obtener los coeficientes del modelo y evaluar las predicciones en la muestra de test (gráfico y medidas de error).
    c. ¿Cuál sería el número de coeficientes distintos de cero si se selecciona λ de forma que minimice el error de validación cruzada?
2. Ajustar un modelo mediante regresión spline adaptativa multivariante (MARS) empleando el método `"earth"` del paquete `caret`.
    a. Utilizar validación cruzada con 5 grupos para seleccionar los valores "óptimos" de los hiperparámetros considerando `degree = 1` y `nprune = c(5, 10, 15, 20)`, y fijar `nk = 30`.
    b. Estudiar el efecto de los predictores incluidos en el modelo final y obtener medidas de su importancia.
    c. Evaluar las predicciones en la muestra de test.
3. Volver a ajustar el modelo aditivo del ejercicio anterior empleando la función `gam()` del paquete `mcgv`.
    a. Incluir los efectos no paramétricos de los predictores seleccionados por el método MARS.
    b. Evaluar las predicciones en la muestra de test y comparar los resultados con los métodos anteriores.