# CLASIFICACIÓN: EJERCICIOS DE PRÁCTICAS

# TRABAJO EN EL LABORATORIO (SESIÓN 3)

BOOK: Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani

**An Introduction to Statistical Learning** with Applications in R

Springer, 2013

Chapter 08

# Ejercicio 3.1.

Los modelos basados en GAM sólo pueden abordar problemas de clasificación binaria. Una idea para adaptarlos a la clasificación de n clases, con n>2 consiste en construir n-1 clasificaciones binarios y combinar sus salidas para obtener la verdadera clasificación.

A) Implementa en R esta idea de combinar clasificadores binarios para resolver el problema de clasificación de IRIS (tiene 3 clases) con una aproximación OVA.

B) Haz lo mismo del apartado anterior, pero con una aproximación OVO.

# Ejercicio 3.2.

1. Hemos visto que *“bagging”*, *“randomForest”* y *“boosting”* son técnicas que pueden ser útiles para mejorar la capacidad de predicción de los algoritmos de clasificación. En RWeka no está disponible esta posibilidad para algoritmos basados en reglas. Se pide:

   A. Define una función de R que permita realizar *“bagging”* con RIPPER.

   B. Define una función de R que permita realizar algo similar al *“randomForest”* pero con RIPPER, haciendo uso de la función *“bagging”* definida en la sección anterior.

   C. Realiza un estudio experimental usando al menos 10 bases de datos “arff” de WEKA para clasificación y los algoritmos RIPPER, RIPPER con *“bagging”* (apartado A) y RIPPER con *“randomForest”* (apartado B). ¿Se verifica que se produce una mejora en la capacidad de predicción?

# Ejercicio 3.3.

1. Usa la base de datos *CoordenadasMunicipios.csv*, que contiene las coordenadas GPS de los municipios de Andalucía y la provincia a la que pertenecen. Se trata de determinar la provincia de un municipio a partir de su posición GPS. Visualizar los espacios de decisión de los siguientes clasificadores:
   - (a) un multiclasificador basado en regresión logística usando un modelo no lineal con spline naturales de hasta grado 4. (glm)
   - (b) un árbol de decisión tipo CART (tree)
   - (c) RandomForest con 500 árboles (randomForest)
   - (d) Boosting con 500 árboles (gbm)
   - (e) C4.5 (J48)
   - (f) Ripper (JRip)
2. ¿Qué porcentaje de los espacios de decisión comparten todos los clasificadores anteriores para este problema? Visualiza estas zonas de decisión común.

3. Toma el Script5b_EspaciosDeDecision.R como base para realizar este ejercicio.

# Ejercicio 3.4.

1. Usa la base de datos *DatosSocialesAndalucia.csv*, que contiene información sobre los municipios andaluces. En concreto, las siguientes variables son : Provincia, Población Total, Porcentaje de hombres y de mujeres, Porcentaje de la población entre 20 y 65 años, Incremento de la población en los últimos 10 años, Porcentaje de extranjeros, Porcentaje de Nacimientos y Defunciones, Número de matrimonios de personas de distinto sexo y Número de vehículos particulares. Responde a las siguientes preguntas:

2. Usando las variables anteriores, ¿qué otra provincia andaluza se parece más a la provincia de Granada?

3. ¿Cuál es la variable de entre las involucradas en los datos que permite discriminar más entre los municipios de Sevilla y de Almería?

4. ¿Alguna provincia se puede discriminar claramente del resto?