Análisis Discriminante Lineal, Análisis Discriminante No Lineal, Naïve Bayes

Máster en Data Science y Big Data en Finanzas

Javier Nogales – PhD Matemáticas
Catedrático, Estadística e IO, UC3M
fcojavier.nogales@uc3m.es 2022

Final Exercise

Breast cancer patients have been subject to initial diagnosis for at least 5 years

After that time, relapse or non-relapse of metastases may appear

- Objective: Predict probability of relapse after initial diagnosis
- Using genetic profile as predictors
- The data file is in Campus Virtual
- Goal: you need to predict the probability of relapse and non-relapse for each patient, based on their genetic information

Final Exercise

- The original paper with the details is here: https://www.nature.com/articles/415530a
- Microarray data are based on DNA sequencing and contains expression levels for genes
- Specifically, the data file contains the expression levels for 97 patients (rows) and 24481 genes (features). The last column of the data file contains the labels
- Out of this 97 patients: 46 patients developed metastases within the 5 years (labeled as relapse), while 51 remained healthy (labeled as non-relapse) after the 5 years
- Expressions with value=100 mean NAs

Final Exercise

- Use simple ideas to deal with NAs, but note the number of observations is small (and expensive)
- One important challenge is the high-dimensionality of microarray data: thousands of features (genes) but less than hundreds of samples (patients)
- But most of the genes (features) are redundant to predict the outcome. Hence, a previous feature selection is required
- Note a good feature selection will imply better performance in developed classification tools

Final Exercise

- Because the sample is small, use 5-fold cross-validation to train and validate the models
- Focus on maximizing kappa, but also focus on reducing the false negatives (non-relapse prediction but then a real relapse)
- You can take inspiration from other analysis, but always cite the source
- Upload to Campus Virtual: notebook (Rmd or Jupyter), and compiled version (html)

Grading

- Data cleaning and pre-processing: 1 point
- Feature selection: 2 points
- Statistical classification tools: 6 points
- Report: 1 point

Afi Escuela de Finanzas