Introduction to Machine learning

Ensemble Methods

October 2019

Aim of the practical session

- being able to use the R langage to handle proof of concepts with classication Ensemble methods.
- bagging, random forest, boosting, stacking

I. Simulated data

1. Data set

- Use the library `mvtnorm` to simulate a two dimensional sample composed of a mixture of two gaussians as illustrated in the following figure.

The first group (blue points) contains $n = 100$ observations distributed as $\mathcal{N}(\begin{bmatrix} 2 & 2 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix})$.

The second group (red points) contains $n = 100$ observations distributed as $\mathcal{N}(\begin{bmatrix} 0 & 0 \end{bmatrix}, \begin{bmatrix} 1 & 0.75 \\ 0.75 & 1 \end{bmatrix})$.

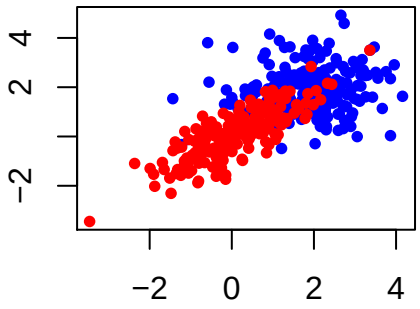

Figure 1: Simulated data

2. Boosting method

The R boosting function of adabag library.

Use the `boosting()`function of the adabag library to implement a classifier with a boosting approach for the previous dataset. With the help of the function, how many trees are by

default generated?

Use the 'boosting.cv()' function to compute the performances using a K fold method. What is, by defaut, the value of K ? Print the performances computed in the confusion matrix.

The R boosting function with naive bayes models.

With the help of the slides, implement a boosting algorithm for simple classifiers as the naive bayes classifiers.

- Compare the results obtained directly using naive bayes classfiers or boosting naive classifiers.

3. Comparison of Ensemble methods

Compare the performances obtained with the following methods: bagging, random forest and boosting.

II. Application on health. Heart Attack data.

- Implement the boosting and stacking methods for the Heart Attack data.