Master BeNeFri in Computer Science

Course: Statistical Learning Methods
Spring 2016

Exercise #12. Classification: Logistic regression & *k*-NN

Download from folder Exercise#9 on ILIAS website the dataset `Vertebral` dataset (filename: VertebralData.2C.txt) containing various biomedical variables that can be used to predict the orthopedic class of the patient (variable `Status`: `Normal / Abnormal`) and read file VertebralDescription.pdf. You have 310 observations (patients) with six predictors. We have no missing value.

Remind that to use *k*-nn method you first need to define a distance between two observations (e.g., using L1 or L2 norm).

```
# Standardized the values (Z score)
#
means <- lapply(aDataFrame, mean)   # means per variable
sd    <- lapply(aDataFrame, sd)     # sd per variable
usefulData <- (aDataFrame - means) / sd
summary(usefulData)          # check if the mean = 0
```

The data frame `usefulData` contains only standardized values and you can compute the distance between two observations (over all predictors).

1. Apply the *k*-nn strategy to the `Vertebral` dataset.
Using the *k*-nn classifier, you need to predict the category `Status (Normal / Abnormal)` according to the possible predictors. You are free to define the most appropriate value for *k* and to use only a subset of the possible predictors.

Then estimate the error rate of your model using 10-fold cross validation.

2. Compare the predictions you obtained with *k*-nn strategy with the logistic regression.

Use a fair methodology to compare the two classifiers (and explain your choice). Can you estimate the error rate for both strategies? Which classifier is the best? Why?