Master BeNeFri in Computer Science

Course: Statistical Learning Methods
Spring 2016

Exercise #6: *k*-NN Regression

As an alternative model to the regression, we can use the *k*-nn method. In this case, we must define a distance between two points (or two observations denoted x and y). We can select the Euclidian distance (sqrt(sum((x-y)^2))) or the L1-norm (absolute value or Manhattan distance, sum(abs(x-y))). Many other variants are possible.

In any case, be sure to compute the distance using similar measurements across all the predictors. For example, when comparing two people based on the income and age, the income difference will dominate the result. The age difference will have no effect. To normalize these values, we can subtract the mean and divided the results by the standard deviation (and thus we obtain standardized values). After this, each value is defined in the same measurement. Values larger than the mean are positive (and negative values indicate original values lower than the mean).

In R, you can do the following:

```
myData <- read.table("ComputerData.txt", header=T)
(myData)
#
# We can remove the predictor name, vendor, and ERP
#
usefulData <- myData[, c("MYCT", "MMIN", "MMAX", "CACH", "CGMIN", "CHMAX", "PRP")]
#
# Standardized the values (Z score)
#
means <- lapply(usefulData, mean)   # means per variable
sd    <- lapply(usefulData, sd)     # sd per variable
usefulData <- (usefulData - 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. Write a R function to compute the distance between two observations. As a parameter, we can ask to compute the Euclidian distance (by default) or the L1-norm.

2. Consider the dataset `Computer` dataset (filename: ComputerData.txt) and the dataset `Cars` dataset (filename: Cars2Data.txt). In Exercise #4 and #5, you have proposed a regression model based on a single linear predictor or by taking in account many predictors.
As a new regression model, apply the *k*-nn strategy. You're free to select the value of *k* you want but you need to justify your choice over other possible values for *k*.