Master BeNeFri in Computer Science

Course: Statistical Learning Methods with R
Spring 2022

Exercise #5: Multiple Regression and k-NN with R

1. Download from the ILIAS website the dataset EducationBis (filename: EducationBis.txt) containing the data of Exercise #1 with modifications and without the outliers. Apply the `lm()` function to all the data and interpret the output you obtain with R.

2. Download from the ILIAS website the Computers dataset (filename: Computers.txt, see Exercise #4). The performance of the system (response) is indicated by the variable `PRP`. You must ignore the variable `ERP`. Remove the variables that are not good predictors and use all the remaining ones to build a multiple regression model. Which variables do you use? Does your model explain something? Explain the model building strategy you have applied. Interpret the most important values of the final model you obtain with R.

3. Visualize graphically the relationship found by you model (by considering the most important variable). Do you notice outlier(s) or hard observations to predict with your model?

4. Repeat questions #2 and #3 with the Cars dataset (see Exercise #4). The performance of the system (response) is indicated by the variable `mpg` (miles per gallon).

As an alternative to the regression model, we can use the k-NN algorithm. In this case, we must define a distance between two points (or two observations denoted x and y). We can select the Euclidean distance (sqrt(sum((x-y)^2))) or the L1-norm (absolute value or Manhattan distance, sum(abs(x-y))).

In any case, be sure to compute the distance using comparable measurements for all the predictors. For example, when comparing two people based on the income and age, the income difference will dominate the result, and thus the age gap will have no effect. To normalize these values, we can subtract the mean and divide by the standard deviation (obtaining standardized values). After this operation, each value will be defined in the same measurement scale.

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 can be used to compute the distance between two observations (over all predictors).

5. Write a R function to compute the distance between two observations. As a parameter, we can ask to compute the Euclidean distance (by default) or the L1-norm.

6. Consider both the Computers and Cars datasets. In Exercise #4 and in the previous questions of Exercise #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 algorithm. You’re free to select the value of *k*, but you have to justify your choice over other possible values.