Assignment of Statistical Learning

July 7th, 2017

1 General notes

- The assignment is individual.
- To analyze the data you can use all methods explained in the course (stepwise, ridge/lasso, gam, rpart), mixed by your own abilities. Remember that theoretical statistics is pure science, but data analysis is partially a science and partially an art.
- **Deadline**: September 10th, 2017.

2 Assignment 1

- **Training** data in file `Facebook-Training.txt`.
- **Test** data in file `Facebook-Test.txt`. Here the dependent variable is missing. The reason is explained below.
- Both the training and the test data are separated by tab (`sep = "\t"` in `read.table()`)
- The variables are described in the companion file `Facebook-Notes.txt`.
- The dependent variable is an integer count, so neither categorical nor continuous. However, given that, during the course, we had not enough time to face counting dependent variables, you can treat it as continuous.
- The goal is to predict the dependent variable using the information in the independent variables.
- At the end of your work you have to submit:
  - **Predictions** for the test set. Append them as a final column (named `pred`) to the test data.
  - A **short report**: max 5 pages written using font times 11pt. It may include tables, figures and what you believe useful but not R programs. The report is for the chief of your team: he is not an expert in statistics or data analytics.
  - Your **R script**.
- I will **rank** your predictions using MSE and MAE. Students with best performances will get a bonus.

3 Assignment 2

- **Training** data in file `Spam-Training.txt`.
- **Test** data in file `Spam-Test.txt`. Here the dependent variable is missing. The reason is explained below.
- Both the training and the test data are separated by tab (`sep = "\t"` in `read.table()`)
- The variables are described in the companion file `Spam-Notes.txt`.
- The dependent variable is an binary indicator: 1 = spam, 0 = no spam. The goal is to predict whether an email is spam using the information in the independent variables.
- At the end of your work you have to submit:
  - **Predictions** for the test set. Append them as a final column (named `pred`) to the test data.
  - A **short report**: max 5 pages written using font times 11pt. It may include tables, figures and what you believe useful but not R programs. The report is for the chief of your team: he is not an expert in statistics or data analytics.
  - Your **R script**.
- I will **rank** your predictions using cost of miss-classification. Denoting costs as $c_{tr,cl}$ (tr = truth, cl = classification), I will use the following costs: $c_{0,0} = c_{1,1} = 0$, $c_{1,0} = 1$, $c_{0,1} = 2$ The reason is that false positives, namely to mark a good email as spam, is highly undesirable. Students with best performances will get a bonus.