Data analysis with R

DM EXAM 13.2.2020

Classification problem (9 points)

This is a real data set, but predictors are anonymized: you don't know the meaning of any of the predictors.

- $Y \in \{Good, Bad\}$ is the binary response variable
- $X = (X_1, \ldots, X_9)'$ are $p = 9$ quantitative predictors
- Training set: (y_i, x_i) for $i = 1, \ldots, n$ with $n = 1300$
- Test set: (y_i^*, x_i^*) for $i = 1, \ldots, m$ with $m = 5197$

The goal is to predict the response $y_1^*, \ldots, y_m^*$ in the test set.

The performance metric is the Accuracy

$$\text{Acc}_{\text{Te}} = \frac{1}{m} \sum_{i=1}^{m} 1(y_i^* = \hat{y}_i^*)^2$$

The percent of points is calculated as $\min\left(\frac{x - 0.71}{0.77 - 0.71}, 100\%\right)$ where x is your final Acc_{Te} score.

The benchmark score $\text{Acc}_{\text{Te}} = 70.98\%$ is obtained by the following model:

```
library(kknn)
yhat<-kknn(y~., tr, te, k=1)$fitted.values
head(yhat)
# name the .txt file with your badge number, e.g. 2575.txt
write.table(file="2575.txt", yhat, row.names = F, col.names = F)
```

Rules

Training set and test set (file `trte.RData`) are available in the folder "TESTO", along with a template (file `2575.Rmd`) of the reproducible R code.

Within **90 MINUTES** you have to:

1. Upload the **[BADGE].txt** file containing your final predictions in the folder "CONSEGNA"
2. Upload the **[BADGE].html** file (generated by R Markdown) containing the reproducible R code in the folder "CONSEGNA"

Other formats will not be accepted.