Data analysis with R

DM EXAM 30.1.2020

Regression problem (9 points)

Data were generated from some distribution function as $Y = f(X, Z, W, V) + \varepsilon$ where

- $Y \in \mathbb{R}$ is the response variable
- X, Z, W, V are $p = 4$ predictors
- $\varepsilon \sim N(0, \sigma^2)$ with $\sigma = 0.01$
- Training set: $(y_i, x_i, z_i, w_i, v_i)$ for $i = 1, \ldots, n$ with $n = 101$
- Test set: $(y_i^*, x_i^*, z_i^*, w_i^*, v_i^*)$ for $i = 1, \ldots, m$ with $m = 1000$

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

The performance metric is the Root Mean Squared Error $\text{RMSE}_{\text{Te}} = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(y_i^* - \hat{y}_i^*)^2}$

The percent of points is calculated as $\min\left(\frac{0.0617 - x}{0.0617 - 0.015}, 100\%\right)$ where x is your final RMSE_{Te} score.

The benchmark score $\text{RMSE}_{\text{Te}} = 0.0617$ is obtained by the following model:

```
load("trte.RData")
library(randomForest)
set.seed(123)
fit = randomForest(y ~ ., data=tr)
yhat = predict(fit, newdata=te)
head(yhat)
```

```
           1            2            3            4            5            6
 0.0077752642 0.0436895290 -0.0064982453 0.0005289994 0.0011582793 -0.0030871543
```

```
# 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.