Data analysis with R

DM EXAM 18.11.2019

Regression problem (8 points)

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

- $Y \in \mathbb{R}$ is the response variable
- $X = (V_1, \ldots, V_{10})^\mathsf{T}$ are $p = 10$ predictors
- Training set: $(y_1, x_1), \ldots, (y_n, x_n)$ with $n = 100$
- Test set: $(y^*_1, x^*_1), \ldots, (y^*_m, x^*_m)$ with $m = 2000$

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

$$\mathrm{RMSE}_{\mathrm{Te}} = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(y^*_i - \hat{y}^*_i)^2}$$

The percent of points is calculated as

$$\min\left(\frac{2.89 - x}{2.89 - 1.89}, 100\%\right)$$

where x is your final $\mathrm{RMSE}_{\mathrm{Te}}$ score.

The benchmark score $\mathrm{RMSE}_{\mathrm{Te}} = 2.89$ is obtained by the following model:

```r
load("trte.RData")
fit = lm(y ~ ., data=train)
yhat = predict(fit, newdata=test)
head(yhat)
```

```
##        1        2        3        4        5        6
## 16.85206 14.74805 19.93895 19.43093 11.20774 19.91978
```

```r
# 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 **2 HOURS** 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.