DSA4211 Project Description

02/10/2022

Problem

In a synthetic training dataset of size $n = 250$, there are $p = 100$ predictors and one response variable. Your task is to build a regression model and predict the values of the response on a test dataset of size $m = 10000$.

Dataset

- `train-xy.csv`: A CSV file of 101 columns, with the first column being the response and the rest being the predictors.
- `test-x.csv`: A CSV file of 100 columns of predictors.

Deliveries

(a) A CSV file, named by `AXXXXXXXX.csv`, where `XXXXXXXX` should be replaced with your student number, having only one column that contains the predicted values of the response for each row in `test-x.csv`; the column header should be 'Y'; an example file is given by `A0000000A.csv` in the Canvas folder **Project**. Due date: **20:00, October 21, 2022**, submitted via the Canvas Assignment **Project-R1**.

(b) A CSV file in the same format of (a), containing your revised predicted values. Due date: **20:00, October 28, 2022**, submitted via the Canvas Assignment **Project-R2**.

(c) A report in the format of MS Word or PDF, named by `AXXXXXXXX.doc` or `AXXXXXXXX.pdf`, containing 1) at most one page (12pt font size) of executive summary that concisely describes your methods/models, your findings, your reflection, or anything that you deem interesting/important; 2) any number of pages that provide details about how you train, diagnose and validate your models, etc. Due date: **20:00, October 28, 2022**, submitted via the Canvas Assignment **Project-Report**.

(d) A well documented R/Python/MATLAB (or other programming language of your choice) script, such as `AXXXXXXXX.R` or `AXXXXXXXX.py` or `AXXXXXXXX.m`, containing all code you use to train your final model. Due date: **20:00, October 28, 2022**, submitted via the Canvas Assignment **Project-Code**.

Grading

Your mark for the project, capped by 100, is divided into 40% for report, 40% for the prediction accuracy and 20% for the code. For the prediction part, it is calculated according to $40 \times$ (your test R^2)/(maxmimum test R^2 among all submissions). For example, if your test R^2 is 0.7 while the maximum test R^2 is 0.8 (achieved by someone else), then your score for the prediction part will be $40 \times 0.7/0.8 = 35$. In addition, the one with the maximum test R^2 will be scored 40. The R^2 on the test data is calculated in the following way:

$$\text{Test } R^2 = 1 - \frac{\sum_{i=1}^{m}(\hat{Y}_{i,test} - Y_{i,test})^2}{\sum_{i=1}^{m}(Y_{i,test} - \bar{Y}_{test})^2},$$

where $\hat{Y}_{i,test}$ is your predicted response, Y_i is the true response, and $\bar{Y}_{test}$ is the mean of the response in the test dataset.

- By October 21, your test R^2 will be calculated and the score for your prediction will be posted to Canvas gradebook under the name **Project-R1**. The maximum R^2 will also be announced so that you can deduce your R^2 from your score. You can then revise your prediction according to this feedback and optionally submit a new prediction by October 28 as instructed in (b).
- On October 28, your test R^2 for the revised prediction will be calculated and the score will be calculated again.
- Your final score for the prediction part is the maximum of your score on October 21 and score on October 28.
- If you decide not to revise your prediction, then your score will be the one on October 21.
- If You decide not to take the opportunity of feedback and only submit your prediction on October 28, then the score on October 28 will be your final score for the prediction part.

For the prediction, it is extremely important to follow the above description to prepare your CSV files, including how the file is named. This is because I will use a script to automatically produce the test R^2 and the score for you.

For grading the code, I will especially look at the documentation and organization of the script.

Notes

- This is NOT a group project; you need to complete the project independently.
- You can discuss with your classmates about the project; however, you need to code up the script and write the report on your own. Similarity check will be conducted on the submitted codes, prediction results and report.
- You can use any model/method (even not covered in the lectures/textbooks) you deem useful. However, you need to provide a concise description of the method/model you use and the rationale behind your choice.
- The report does not have to be very long/detailed. Please contain only essential information/discoveries/discussions/plots.