DSA4211 Project Description

Zhenhua Lin

10/12/2021

Problem

In a synthetic training dataset of size $n = 1000$, 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'. Due date: **9AM, Nov 3, 2021**, submitted to the LumiNUS folder **PROJECT-R1**.

(b) A CSV file of the same format of (a), containing your revised predicted values. Due date: **9AM, Nov 10, 2021**, submitted to the LumiNUS folder **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: **9AM, Nov 10, 2021**, submitted to the LumiNUS folder **PROJECT-REPORT**.

(d) A well documented R/Python script, named 'AXXXXXXXX.R' or 'AXXXXXXXX.py', containing all codes you use to train your final model. Due date: **9AM, Nov 10, 2021**, submitted to the LumiNUS folder **PROJECT-REPORT**.

Grading

Your mark for the project, capped by 100, is divided into 40% for report, 40% for the prediction accuracy and 20% for the codes. 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 score 40.

- On Nov 3, your test R^2 will be calculated and the score for your prediction will be posted to LumiNUS 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 Nov 10 as instructed in (b).

- On Nov 10, 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 Nov 3 and score on Nov 10.
- If you decide not to revise your prediction, then your score will be the one on Nov 3.
- If You decide not to take the opportunity of feedback and only submit your prediction on Nov 10, then the score on Nov 10 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 an automatic script to produce the test R^2 and the score for you.

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

Notes

- 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.