DSA4211 Mini Project Description

Purpose

Learn to build a model for high-dimensional data using techniques (e.g., cross validation) learned from the course and beyond, and deepen understandings of various statistical principles such as bias-variance trade-off.

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) (Round 1) 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, April 11, 2025**, submitted via the Canvas Assignment **Project-R1**.

(b) (Round 2) A CSV file in the same format of (a), containing your revised predicted values. Due date: **20:00, April 17, 2025**, submitted via the Canvas Assignment **Project-R2**.

Grading

Your score for this mini project is based on the performance of your model on test data and is calculated according to 100 × (your test R^2)/(maxmimum test R^2 among all submissions in the same round). 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 is $100 \times 0.7/0.8 = 87.5$. In addition, the one with the maximum test R^2 is scored 100. 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 April 11, 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 April 17 as instructed in (b).
- On April 17, your test R^2 for the revised prediction will be calculated and the score will be calculated again.
- Your final score is the maximum of your score on April 11 and score on April 17.
- If you decide not to revise your prediction, then your score will be the one on April 11.
- If You decide not to take the opportunity of feedback and only submit your prediction on April 17, then the score on April 17 will be your final score.

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. The file `validating.R` may help you to check the format of your file.

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 your model and produce the prediction on your own. Similarity check will be conducted on all of the submitted prediction results.
- You can use any model/method (even not covered in the lectures/textbooks) you deem useful.
- **Late submissions in each round are NOT accepted.**