Final Project

Submit Assignment

Due May 7 by 11:59pm **Points** 60 **Submitting** a file upload
Available Apr 6 at 12am - May 8 at 11:59pm about 1 month

DUE May 7, 2020

*****Please read very carefully*****

For this project, we will work on the data set collected from a study of breast cancer:

breast_cancer_train.csv

The original dataset contains expression levels of 24,187 genes for 97 patients, 46 relapse ("status" is 1) and 51 non-relapse ("status" is 0). 78 cases were used as the training set (34 relapse and 44 non-relapse) and 19 (12 relapse and 7 non-relapse) as the test set. The dataset has been preprocessed. We normalized the expressions levels and filtered the genes by a p-value criterion. After this step, 4918 genes remain. For this project, I only upload the training set, which contains 78 cases, with 4918 predictors and a binary response.

Task:

1. Choose an appropriate method we discussed this semester and build a model to predict the patient's statue (relapse or non-relapse).
2. Evaluate the model performance by cross-validation.
3. Wrap your model in a function, which takes gene expression levels as input, and return the prediction of patients' status.

Important dates:

- May 7 project report (**Required**: a zipped folder with an r markdown notebook, and supporting files such as data files.)
  - Report format: should avoid too much output. Please refer to **knitr documentation (http://yihui.name/knitr/options/)** for how to turn off output and messages.

Grading

- Correctness of implementation (30 points)
- Performance of the method (10 points)
- Report (20 points)

Some Rubric

Criteria	Ratings	Points
Correctness of implementation Appropriate analysis tools. Correct interpretation of results. There should a function that can be directly used for prediction.		30.0 pts
Performance of the method We will test your method on an unreleased test data set. A test error lower than 40% can get all 10 points.		10.0 pts
Report Writing, presentation and organization. Be concise.		20.0 pts
		Total Points: 60.0