STAT4001 Data Mining and Statistical Learning

Final group project

Academic Year 2019 ~ 2020

1. The final group project contains two part. Members in the same group will receive the same score for this project.

   - A 8-minute group presentation, scheduled on the last lecture (Nov.25 2019). Due to the time limit, your presentation will be stopped strictly after 8 minutes. Please send your slides to TAs before 11:00am Nov.25 2019, no later slides will be accepted. Every group only need to submit ONE softcopy of slides.
   - A written report which can demostrate the whole analysis clearly and comprehensively.
   - Submission: please send the following softcopies to TAs through email before Dec.2 2019 (Monday). Every group only need to submit ONCE.

     (a) R code file which can reproduce all the results

     (b) The written report

   All the interpretations in your presentation and the written report should be understandable by layman!

2. Instructions
   You are required to perform two real data analysis, one on regression, another one on classification. We prepare two data sets correspondingly, but if you want to find and use your own dataset, please feel free to do so.
   The followings are "Must to do". You will receive bonus if you include new methods, new ideas. (If you use other data sets, you still need to include the following analysis in your project.)

   (a) Regression
   Ask a home buyer to describe their dream house, and they probably will begin with the budget. With several explanatory variables describing many aspects of residential homes in Ames, Iowa, you are going to predict the final price of each home using "House.csv".

      i. Split the dataset into training dataset and testing dataset.

    ii. Perform your analysis by linear regression, ridge regression, LASSO, principal components regression, partial least squares, polynomial regression, regression with step functions, regression splines, local regression, generalized additive model, regression tree, bagging, random forests, boosting,(14 methods in total, but **at least choose 9 methods to implement** ☺) and so on. Use cross-validation to tune the tuning parameter if there's any.

    iii. Interpret the results and report the cross-valiation error for each method.

    iv. Use the testing dataset to predict the sale prices and report the test error for each method.

    v. Draw your conclusion.

  (b) Classification

  The sinking of the Titanic is one of the most infamous shipwrecks in history. On April 15, 1912, the widely considered unsinkable RMS Titanic sank after colliding with an iceberg. Unfortunately, there weren't enough lifeboats for everyone onboard, resulting in the death of 1502 out of 2224 passengers and crew.

  While there was some element of luck involved in surviving, it seems some groups of people were more likely to survive than others. Now, it is your turn to predict which passengers survived the Titanic shipwreck using "Titanic.csv".

    i. Split the dataset into training dataset and testing dataset.

    ii. Perform your analysis by KNN, logistic regression, LDA, QDA, classification tree, bagging, random forests, boosting, support vector machine,(9 methods in total, but **at least choose 7 methods to implement** ☺) and so on. Use cross-validation to tune the tuning parameter if there's any.

    iii. Interpret the results and report the cross-valiation error for each method.

    iv. Use the testing dataset to predict survived or did not survive the Titanic shipwreck for each passenger, and report the test error for each method.

    v. Draw your conclusion.

  (c) In your presentation, you just need to choose 3 methods from regression part you used, and another 3 methods from classification part (in total 6 methods) to present.

3. Reference of datasets: https://www.kaggle.com/competitions

- Then, Good Luck! -