Stat 154 Final Project

Overview

Due December 18th at 11:59pm

In this project, you will be asked to explore two different datasets. The first is "fashion mnist", which consists of greyscale images of various clothing items. The second is a crime dataset for various communities in the US. The details of each dataset are contained in the files `crime_and_communities.Rmd` and `fashion_mnist.Rmd`.

Requirements and guidelines

Your task is to provide a comprehensive analysis of each dataset. Using the techniques used in class, lab and homework, you should explore each dataset as though you were a data scientist or researcher working on these problems. We will not describe exactly what methods and techniques you should use; rather, you should use the knowledge you've learned throughout the semester to decide relevant techniques. Throughout this process, it is important that you provide thorough explanations for *what you are doing*, and *why*. These responses will be as important, if not more so, in determining your grade than your actual results. However, your analysis should demonstrate the use of at least 4 of the following (potentially in combination with each other):

- Dimension reduction
- Linear/nearest neighbor/partial least squares regression
- Regularization/penalization (e.g. Lasso, Ridge)
- Logistic regression/classification
- Clustering
- Tree-based methods (e.g. Decision trees, random forrests)

For each dataset, the analysis will be separated into exploration and prediction sections. The crime and communities dataset consists of a regression task, while the fashion mnist dataset consists of a classification task. For both tasks, you should employ model selection/model evaluation methods (such as cross validation) as done on previous assignments. You should report a final test set performance for each task (note: you are responsible for splitting the data into training and testing).

You are welcome to use packages available in R/Python, though preferably those used in homework and lab. You should also be sure to include relevant and informative visualizations in your report, where applicable.

Formatting and submission

Students may work in groups of 2, though *each student should submit their own files* (with both group members' names included in each). **Students should register their groups using this Google Form no later than December 1st**.

Each student should submit a total of 4 files:

- Code files
  - `fashion_mnist_<your_id>.Rmd` and `crime_and_communities_<your_id>.Rmd` (or .ipynb if you are using python)
- Readable files
  - `fashion_mnist_<your_id>.pdf` and `crime_and_communities_<your_id>.pdf` (or .html files, if you prefer)

The template files we give you have basic formatting, but they should be broken down further into meaningful chunks based on which aspects of the data you are exploring/which techniques you are using.

Grading

The assignments will be graded on the following:

- Thorough and accurate description of methods and results
- Correct implementation of techniques
- Readibility of code and report, informative visualizations