CSCI 406 (Data Mining)
Dr. Stephanie Schwartz
Spring 2019
Course Project
30% of final grade

Overview

The aim of this project is to apply data mining and machine learning techniques to an interesting problem of your choosing.

Remember that data mining aims to discover interesting new information (correlations, patterns, trends, relationships, ...) in datasets. Usually this is for either (1) prediction, or (2) understanding. When brainstorming a project idea, ask yourself:

- Is my project implementing a model that I'll use to predict some target variable(s) or is my project aimed at understanding some domain better?
- What predictor features are associated with a target response? What is the (hypothesized) relationship between the predictors and response?

Phase 1 - Project Proposal (10% of project grade)
Due Wednesday, February 20th at 11:59pm (submit as pdf to autolab)

Length: 1-2 page project proposal.

Describe your general research question, as well any specific hypotheses that you have and will investigate (e.g. students who drive a red car have a higher gpa). Also write about the data that you will be using, how you plan to convert it into a form that is able to be processed (if it's not already a single .csv or something similar), and the size of the data. Provide enough detail such that it is clear to me that you have a solid idea of what you want to accomplish.

See the course resources page for links to some freely available data sets. *Feel free to google for other or use data scraping techniques to acquire data! (NOTE: No personal/private data will be allowed to be scraped or used)*

Feedback

I will aim to provide feedback on your research proposal within one week: is your research question appropriate, are there other data sources to look at, other ideas to try, etc. Also feel free to ask me questions in the proposal (e.g. "I have these two ideas in mind, which do you think is better...").

Phase 2 – Approach Document and Data Acquired (10% of project grade)
Due Wednesday, March 20th at 11:59pm (submit as pdf to autolab)

Length: 1-2 page description of approach to project as well as a check during lab period to show me that you've acquired your data.

Describe any modifications to your general research question (based on constraints you've learned about, my feedback, etc.). Then outline your approach to the project: what data mining and analysis techniques will you be applying? Are there things you will conditionally look at based on early findings? Provide enough detail such that it is clear to me that you have a solid idea of what you want to accomplish and that you can be successful in doing so.

Feedback
I will aim to provide feedback on your proposed approach within one week. We will also do peer review at this phase (for some extra points in lab category). I have found that during the project presentations, students have raised some really good points/suggestions so I'm attempting to incorporate this into the earlier stage of the project.

Phase 3 - Oral Presentation (15% of project grade)

What? Project presentations should clearly and effectively motivate and describe your (1) research problem, (2) data, (3) methods used, and (4) analysis. The presentation is short. Use powerpoint, or show code, or draw on the whiteboard, or provide handouts. What I am looking for is clarity and good oral communication of your research. (That said, the powerpoint method is probably safest.)

How Long should the Presentation be? 10 minutes. Example breakdown:

- (1 min) Introduce your research problem. Motivate it.
- (1 min) Introduce your data. Predictor Variables. Target Variables.
- (2 min) Exploratory analysis that shows interesting qualities of the data. Skewness. Outliers. Correlations.
- (1 min) Preprocessing. Handling missing values.
- (4 min) Application of Learning Algorithms. Results.
- (1 min) Conclusion. Outcomes. What did you learn / discover?

When? The presentation dates will be 4/29, 5/1, 5/3. I'll use R's random function to assign the presentation days. *(And depending on your date, you'll either love or hate R even more!)*

Phase 4 - Written Report (65% of project grade)

Length: 6-8 pages
Due Monday, May 6th at 11:59pm (submit to Autolab as R Markdown and html)

Similar to the oral presentation, "tell" the project story in your written report. Include appropriate data visualizations and result tables. The report should not be a "dump" of generated charts and tables, but should contain useful visualizations that are placed into context via the text.

The written report is due *after* the oral presentation so that you can include student and teacher feedback from your presentation (others' perspectives about your data, data mining techniques that were suggested to try, etc.).

The goal of the written report is that someone who is already familiar with data mining, but not *your* topic could be able to read your project report and understand: what your problem is, where the data is from, how it was collected, what data mining techniques you tried, what worked and what didn't work, what conclusions you were able to make, etc.