STATS 415 - Project

Professor Liza Levina
Department of Statistics

Winter 2018

This project is designed to provide an opportunity to apply statistical learning and data mining techniques you have learned in class to analyzing a data set of your choice. You should choose a topic that you are interested in and care about, and then find a data set or collect your own.

Due dates

- Proposal: Marh 13. No more than 1 page, one hard copy per group, turned in either in the lecture or in the GSI mailbox by 5pm.
- Presentations: tentatively between April 10 - April 16, to be announced. Each group will be assigned a date, either in a lab section or in lecture.
- Final project reports: April 17. No more than 8 pages, one hard copy (double-sided printing) per group, turned in either in the lecture or in the GSI mailbox by 5pm.

Teams

The first step is to form a group – part of the point of this project is to learn to do data mining as a team. The teams should be of 3 (preferred) or 4 students; no other team sizes are allowed. It may be easier for the entire group to attend your presentation if you are all in the same lab but **this is not required**.

Project description

Once you settle on a topic and a dataset, you are asked to do the following:

- Do as much data exploration as possible by visualization, summary statistics and perhaps a more sophisticated dimensionality reduction method such as principle component analysis.

- Formulate one or more questions of interest about the data that can be answered with one of the statistical learning techniques covered in this class, supervised or unsupervised. (Unsupervised techniques like clustering will not be covered before winter break but you can plan on using them later). The question should make sense for the dataset (e.g. not have an obvious answer and yet be plausible to answer with the data you have).

- Perform the analysis to answer your question(s) with one or more methods covered in class. You should explain why the methods you chose make sense for your question(s), discuss the data analysis results and their interpretation. If the methods you used need any parameters specified by the user (e.g. K for KNN), you should explain how you made those choices and what implications they have on your results.

Project proposals

Each team submits one hard copy of their project proposal, about 0.5-1 page long. The proposal must include the team members names **and their labs**, the designated presenter, a description of the dataset (source, variables, the number of variables and observations you intend to use in the analysis), the question(s) to be investigated, and the statistical learning tools you intend to use to answer your questions.

Your dataset and questions of interest need instructor's approval, provided as part of your project proposal feedback. You may be asked to pick a different dataset and/or question at that stage if we think the project you are proposing is not realistic.

The proposal is due in class on March 13.

Project presentations

All teams will present their projects in a 10-min presentation at the end of the course. The team should prepare slides for this. Tentatively, the last lab of the term and the last two lectures will be given to project presentations; schedule to be finalized once the teams are formed. The teams should select one person to present as there will not be time to switch speakers. Presentation should be counted as part of that person't contribution to the project, but it cannot be their only contribution.

Project report

Write up your results in a formal report. Unlike homework, the report should not be written in RMarkdown and should not include R code. R output should be included in the form of tables and figures; models should be written as equations (e.g. $y = \beta_0 + \beta_1 x$, not `lm(y   x)`. Please limit your report to **8 pages** including figures and tables, no code. Each report must have a title, a summary/conclusions section, and a paragraph describing the individual contributions of each of the team members.

Each team submits one hard copy of the final project report (printed double-sided please).

The report is due in class on April 17 (last day of classes).

Where can we get data?

If you are struggling to find a question of interest to you and/or data to answer it with, you can also select a dataset from on of the standard data repositories, for example, the UCI Machine Learning Repository (http://archive.ics.uci.edu/ml/) or Kaggle (http://www.kaggle.com). If you do so, please focus on recent data sets, not old ones.

A sample project

On the Canvas project page there is a sample of a successful project from a previous year's class. (Thanks to Annalyn Ng and Ben Charoenwon for sharing this with us). They used their 415 project to participate and won the Yahoo student data mining competition in 2011.