CISC 520: Data Engineering and Mining
Course Project Road Map, Summer 2017

1. Overview

This document aims to help you get started with the course project. Project requirements and rules are provided, which you are required to strictly follow. A list of possible datasets for your project is provided. The course project is meant for you to get your hands "dirty" with real world data and tackle a problem of interest to you. Since the project is open-ended, select a topic that is of interest to you, or has relevance in your fields of interest. The project should be data-focused. Although in statistics we usually like to formulate a hypothesis first and then go about collecting data, often in data mining it is the other way around. There are multitudes of interesting datasets out there that can be collected fairly easily.

2. Team

Between 1 and 3 people can collaborate on a project. It is expected that projects with more people will be more ambitious. Send your team members to me as soon as possible.

3. Project

The project might be extremely open-ended. It should consist of the following:

- Find or collect a data set of interest. There are many sources on the web for data sets. I would prefer the data to be of a reasonably large size (this is a data mining class after all), but really large data sets can bog down computers. The R (https://www.r-project.org/) can easily handle data sets in the tens or even hundreds of thousands (depending on your computer). A lower limit for data size should be $n = 1000$ although I will be willing to accept exceptions. See below for places to look for data sets of interest.
- Consider your data carefully. Even if you downloaded it, you should look for information about it. How was it collected? What are the data quality issues? Are there biases inherent in who collected the data or how it was collected? Are there any data preparation needed? And what are these operations? How might this impact the subsequent conclusions?
- Formulate questions that you would like to answer about this data set. You can follow the way the lecture notes listed the question. What is the dependent variable or variables? What are the predictors?
- Implement your analysis using data mining tools. These should have some relation to what we have learned in the class! Are you doing a classification or clustering task? Can the data be expressed as a network of some kind? Are there interesting visualizations to do? How will you evaluate the performance of your model, or choose between competing models?
- Write a report summarizing your data, your question of interest, and your findings. Reference other existing work which has analyzed your data, or addressed similar topics. The report should contain information about the data, exploration of the data set, and an appropriate analysis. Graphs are welcome, but do not overdo them.
- The final report should be no less than 5 well-formatted pages in length (IEEE standard format, 10pt times new roman font, double-column, single line space), including graphs, references, appendix, etc.

4. Software

Use whatever software is comfortable for you.

5. Timeline

You should already start thinking about the project. The final project will be due in the final week of the semester.

6. Proposal

The proposal is the first deliverable of your course project. It should be no less than 2 pages in length (IEEE standard format, 10pt times new roman font, double-column, single line space), including graphs, references, appendix, etc. In your proposal, you are required to answer the following questions to receive full credits:

- What data set is being used? Where does it come from, and what are the characteristics of it (size, missing values, continuous vs. categorical)?
- Is there a reason you picked this data set? Tell me.
- What is the question(s) of interest? Please be specific. Generic questions like "I want to look for patterns in stock prices" are bad. Devise a specific question you can answer with the data.
- What methods do you plan to use? Understanding that this might change and that we have yet to cover many methods in class.

Understand that this is a proposal, and I expect that your question and your approaches will likely change as the semester progresses.

7. Progress report

The progress report could be the initial version (the draft) of the final report, or simply presents what has been done and what will be done for this project.

8. Ideas for projects and datasets

There are lots of data sets available online. If you don't have your own data, pick something that you will enjoy working on, and something where there is a rich source of data available. Take some time in selecting a good data set and feel free to ask me for suggestions. You should take the time to research where and how your data was collected. A good report will dive into biases that may exist or data quality issues.

- Data mining data sets at the KDnuggets website. https://goo.gl/pFIHfG
- Data repository for machine learning at UC Irvine. https://goo.gl/FPhgud
- Data sets for data mining at the University of Edinburgh. https://goo.gl/LSUR8m
- Datasets archive at Statlib. https://goo.gl/kdL1JE
- Ideas for projects from a previous lecturer of this class. https://goo.gl/WFLn0L
- New York City Datasets from CPRC. https://goo.gl/sC8gJP
- Datasets from Chance Magazine. https://goo.gl/vD2Ck8
- Health Data Sets. https://goo.gl/pWtL51
- Disability and Health Data. https://goo.gl/9jQIb2
- Cardiovascular Health Study. https://goo.gl/qDFXE7
- Flowing Data Post on Data sources. https://goo.gl/bHXSGZ
- Data Feeds available as packages. https://goo.gl/uYE8lR
- Stanford Large Network Dataset Collection. https://goo.gl/BrIHVf
- Complete Wikipedia Edit History (18GB!). https://goo.gl/mehv4p
- Bi-Annual Data Exposition. http://stat-computing.org/dataexpo/