CS170A — Course Project Starting Points — Fall 2016

A course project requires submission of a report (in PDF format) including: (1) a description of a dataset that you constructed; (2) a log of the analyses you attempted (including models and methods covered in CS170A); (3) a list of results you obtained, (4) a summary of effort put in, results gotten out, and lessons learned. There is no way to guarantee data mining will find results, but the log can record your explorations.

The report can include visualization output, descriptive statistics, algorithms or models developed, performance results, etc., along with the (e.g., Matlab) scripts used for these things. It is usually about 15-20 pages, but that is easy to fill with all this output. The report will be graded based on specific criteria described below.

Elements of a Successful Project

Only work on data that you know something about, is interesting to you, and that you care about. The more interesting the data, the more interesting the story or results that you can find in it. Once you have data you care about, it is fun to put work into analyzing it, and it is easier to find patterns.

The project is for people who want to put energy into something creative. It is something to put on your resume. It is not a homework or an easy alternative to the Final; it will end up requiring more time than just taking the Final.

The project needs to be clearly novel. This is easiest if the data is 'new', and compiled from different sources by you; and novelty can come from integration of datasets you can find on the web. Alternatively there can be novelty in the analysis.

General examples of past projects: analysis of strategies in baseball (also soccer, etc.), what kinds of movies win the Golden Globe awards, pairs trading stock market analysis, how countries try to win gold medals in the Olympics, basic game engine design in Matlab, which kinds of cars get EPA approval, seasonality of global tourism, rock music classification.

Actual titles from earlier projects: Sport Attendance, March Madness, Identifying Galaxies, Effect of Finance on Education, Improving Eigenfaces, Voice Spectra, Reddit Submissions, Kalman Filters, NBA Schedules, Developing Nations, Soccer Performance, YouTube Statistics, HIV Treatment, Crime Rate Analysis, Distinguishing Guitar Style, Performance Cars.

If you are not sure where to find data that is 'novel', but know people that you are interested in working with (and who have data — e.g., UCLA faculty, or friends at Google or Yahoo, a company you want to do an internship with, etc.), you could try starting with that.

Final Deliverables: Project Report and Analysis Scripts

In addition to a dataset and set of analysis scripts (e.g., Matlab scripts), the project should produce a PDF report (usually 15-20 pages) containing specific things:

- a description of a dataset you used, and how it was obtained;
- a step-by-step log of analyses you attempted;
  (There is no way to guarantee you will find results, but the log can record your exploration.)
- a summary of how much effort went into each of the grading criteria below.
- a summary of experiences, insights, and lessons learned.

Think of the report as something to show in job interviews, and something to put on your resume.

What to Not Do

Do not just grab a dataset that has been heavily analyzed already, e.g. from the UC Irvine archive.
These datasets make it very hard to do something novel.
The data should have at least thousands of observations, but the dataset should NOT be in the gigabyte range.
No toy datasets with 100 observations, but no gigantic datasets either. With large datasets, also, it often may work better to start on samples of the data, and scale up to the full dataset when the analysis pipeline is working.

Groups of students cannot work on a project together without a very strong justification.
In particular, they will need to make it explicit who will do what and why more than one student is necessary to do these things. Undergraduate group projects under time pressure are high-risk at best, so do not propose a group project unless you have very strong reasons.

Some data analysis is needed, not just data acquisition and visualization.
Do not pick data that could take 4 weeks to accumulate, since that will leave no time to analyze it. Analysis must use more than just exploratory plotting or basic statistics. Some nontrivial matrix analysis relevant to CS170A must be involved.

Grading Criteria

The final grade will be a sum of scores for the following grading criteria:

1. Data familiarity/interest (Why is it that are you knowledgeable about/interested in this data?)
2. Data acquisition effort (How much effort is required to construct the dataset?)
3. Data novelty (In what ways is looking at this kind of data new or innovative?)
4. Analysis effort (How much work is needed to analyze the data?)
5. Analysis methods (Which methods are most unusual, novel or worthwhile?)
6. Project difficulty (In what ways is the project most challenging?)
7. Project relevance to the course (How does the project relate to matrix methods and modeling?)
8. Project novelty (Which aspects of the project are most unusual or creative?)
9. Report (Which aspects of the report require the most effort?)
10. Overall effort (Which aspects of the project require the most effort?)

The project report should make a summary of strengths of the project for each of these criteria.

Project Proposal

To do a course project, a proposal must be submitted before the deadline: November 2.
The proposal should be a one-page description of the project that includes:

- a detailed description of the data to be explored
- a less detailed description of the project, that explains — for each of the grading criteria above — why the project is a good project.

I will get back to the proposers immediately, either approving or not approving the propopsal.

Sample Data Sources — Some Places to Shop for Ideas

- Kaggle — interesting data and data mining contests: http://www.kaggle.com
- d3.js — a wonderful site for ideas about interactive data analysis on the web: http://d3js.org (Gallery)
- NY Times Labs http://nytlabs.com
- Los Angeles Times Data Desk http://datadesk.latimes.com
- Guardian.co.uk/data http://www.guardian.co.uk/data
- Google Public Data Server http://www.google.com/publicdata/directory
- California State Datasets http://www.ca.gov/data/state_data_files.html
- Google Trends, Google Flu Trends, and Google Correlate.
- An amazing epidemic: http://www.cdc.gov/obesity/data/adult.html
- KD Nuggets — a hub for data mining: http://www.kdnuggets.com
  e.g.: http://www.kdnuggets.com/datasets/
- InfoChimps — once a good starting point for finding data: http://infochimps.org
  e.g.: http://www.infochimps.com/tags/techcrunch
- Federal data clearinghouse http://www.data.gov
- National Bureau of Economic Research http://www.nber.org/data
  (many interesting datasets: Macroeconomics, industry, trade, demographics, hospital, patents, ...)
- Tracking the U.S. Congress http://www.govtrack.us/developers/data.xpd