CS170A – Mathematical Models & Methods for CS – Spring 2013

Some Course Project Pointers

A course project is a term paper including (1) a description of a dataset that you obtained or constructed; (2) a log of the analyses you attempted; (3) a list of results you obtained. There is no way to guarantee data mining will find results, but the log can record your explorations.

The results can include visualization output, descriptive statistics, algorithms or models developed, performance results, etc. However they must involve your dataset, include some ideas about methods that could find something interesting in the data, and ideally include some lessons learned.

Work on data that is interesting 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 could be 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.

If you are not sure where to find interesting data, 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.

1 General Data Sources

- Google Public Data Server http://www.google.com/publicdata/directory
- KDD Nuggets — a hub for data mining: http://www.kdnuggets.com
  e.g.: http://www.kdnuggets.com/datasets/
- InfoChimps — a good starting point for finding data: http://infochimps.org
  e.g.: http://www.infochimps.com/tags/techcrunch
- Kaggle — interesting data and data mining contests: http://www.kaggle.com
- An amazing epidemic: http://www.cdc.gov/obesity/data/adult.html
- Amazon EC2 Public Data Sets (huge datasets for genomics, Wikipedia, economics, astronomy, etc.):
  http://aws.amazon.com/datasets
- The Million Song Dataset: http://labrosa.ee.columbia.edu/millionsong/
  e.g.: http://labrosa.ee.columbia.edu/millionsong/pages/matlab-introduction

2 Economic and Financial Data

- Los Angeles Times Data Desk (thanks Byron!) http://datadesk.latimes.com
- Guardian.co.uk/data http://www.guardian.co.uk/data
- NY Times Labs http://nytlabs.com
- 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, ...)
- Federal Reserve Data Economic Research & Data http://www.federalreserve.gov/econresdata/default.htm
  (including data about mortgage defaults, interest rates, exchange rates, industrial production, ...)
- Federal Statistics Data Access Tools: http://www.fedstats.gov/toolkit.html (see also: http://data.gov)
- California State Datasets http://www.ca.gov/data/state_data_files.html
- Tracking the U.S. Congress http://www.govtrack.us/developers/data.xpd