Final Project Handout

Issued: October 15, 2017

An important part of this course are the final projects. These are group projects and we have made the group assignments to ensure similar interests and a combination of skills. Your tasks for the final project are: find a data set of your choice, pose a research question, and answer it using a combination of analysis methods presented in this course and possibly further methods that you read up on.

The timeline including important deadlines is outlined below. There will be fewer problem sets in the second part of the course to ensure that you have enough time to work on your final projects. So make sure to start working on the project in time, and have fun!

Timeline

Now Connect with your project partners and decide on a topic, data set and question. Start writing the project proposal (details below).

Nov 5 Project proposals due.

Nov 26 Intermediate status report due: one paragraph on what you have tried so far, what remains. If it turned out that some of your approaches were not fruitful or you ran into problems and need to change your plans, report that too (and, if need be, talk to us early).

Nov 5-Dec 13 Apply the methods you laid out in the proposal.

Last week of classes Project presentations.

Dec 13 Project report is due.

Data

You are free to pick a topic and data set of your choice. For inspiration, we have included a list of possible datasets at the end of this handout.

Project Proposal

The project proposal should be about 2 pages (without references) and contain the following:

1. the names of all team members
2. the question you are going to address, and a motivation for it
3. a description of your data

4. what analysis methods you are going to use to address this question and a justification for your choices. These should consist of methods you learned during the course and you may read up on additional methods if need be. Will these methods apply directly? Will you need to adjust anything? Does your data need pre-processing (cleaning)?
5. for graduate students: an overview of some related work
6. a realistic timeline for the project. Make sure you plan in sufficient time for computation and exploration; this tends to take more time than expected.

Final Project Presentation

Every group gives a poster presentation about their data set, questions and findings. The other students give written feedback on a subset of presentations. In addition, each student gives feedback on the group process. This will be used to evaluate the projects.

Final Project Report

The final project report should be 6-8 pages long and contain the following:

1. the names of all team members
2. the question you are going to address, and a motivation for it
3. for graduate students: an overview of some related work
4. a description of your data
5. the analysis methods you have been using, including any preprocessing; if you had to deviate from your initial plan, briefly explain why
6. a summary and interpretation of your results

Evaluation

The evaluation of the projects will take into account the proposal, poster presentation, and final report. We will evaluate the analysis (breadth/depth of methods), the presentation, and the interpretation of your findings. If you are a graduate student, you may treat the project as leading to a research paper.

If you run into any problems with the project, come talk to us as soon as possible.

Suggestions for Data Sets

Here are some suggestions for data sets. These are for inspiration; you may use one of them or find your own.

- Mice Protein Expression Data Set (UCI repository)
- Breast Cancer data (UCI repository)

- Greenhouse gas observing network (UCI repository)
- El Nino data (UCI repository)
- Air Pollution data: USA: EPA data https://www.epa.gov/outdoor-air-quality-data
  Europe:
  http://www.eea.europa.eu/data-and-maps/data/airbase-the-european-air-quality-database-8
- Climate data sets and Challenge Problems: http://www.climateinformatics.org/?q=data-sets
  and https://sites.google.com/site/1stclimateinformatics/materials
- Enron email data set (http://www.cs.cmu.edu/~enron/)
- S&P/other financial time series (Yahoo Finance)
- Billion Prices Data sets: http://www.thebillionpricesproject.com/datasets/
- Hubway data http://hubwaydatachallenge.org/
- NYC taxi data: http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml
- Uber data: https://github.com/fivethirtyeight/uber-tlc-foil-response
- Congressional voting data
- Twitter data (https://dev.twitter.com/streaming/public)
- Netflix data
- Presidential election data (primary results, polling data, etc.)
- NFL play-by-play data
- NBA player tracking data (http://stats.nba.com/tracking/#!/player/)
- Stanford Large Networks Dataset collection https://snap.stanford.edu/data/
- list of many more data sets from various areas:
  https://github.com/caesar0301/awesome-public-datasets