Final Project Description, COMP 4441

Probability and Statistics for Data Science I, Winter 2019

The final project for COMP 4441 consists of a detailed analysis of a dataset using at least one method not covered in class. The assignment includes R code for the analysis and a paper explaining the analysis method or methods and the results. When consistent with ethical obligations, the data set should accompany the project.

The subject you investigate in your final project may be controversial. The methods you use and your data should be a responsible application of data analysis and inference.

The intent of the assignment is for students to assemble a full analysis, including exploratory data analysis, and application of at least one statistical method on or above the complexity of a two sample method, a Chi-square test, regression, or analysis of variance. The analysis should address a question of natural interest regarding the data.

There will be poster sessions 5:00-6:50 Monday, March 18 and 6:00-7:50 Thursday, March 21 in ECS 510. Please plan to be present for both sessions. Boards to accommodate a 17”x33” poster will be available (aka 6 8.5”x11” pages). If you choose a different format, please provide your own backing.

Rubric for poster

* Data source and definitions explained 10
* Main features of data set presented with appropriate graphics 20
* Research question presented 10
* Method for addressing research question explained 10
* Data satisfaction of requirements of method demonstrated 10
* Method applied and interpreted correctly 20
* Format appropriate for poster session: visual impact, concise 10
* Clear, correct explanation provided on request 10

Rubric for paper (maximum of 5 pages)

* Data source and definitions explained 10
* Main features of data set presented with appropriate graphics 20
* Research question presented 10
* Method for addressing research question explained 20
* Data satisfaction of requirements of method demonstrated 10
* Method applied and interpreted correctly 20
* Appropriate format, including necessary citations, used 10

Possible methods for use in project:

Logistic Regression (fitting categorical responses)

Robust Regression

Linear Discriminant Analysis (fitting categorical responses)

High Dimensionality methods (model building when many explanatory variables are present)

-lasso & ridge regression

-Regression trees

-Principle Component Analysis

-Partial Least Squares

-Random forests

` -Factor analysis

Clustering

- Latent Class Analysis

Non linear methods

- further study of generalized linear models or general additive models

-Spline estimation

-principal curves

Bayesian Analysis (Big topic, scratch the surface)

Sample Mean Tests for multivariate continuous responses

Time series data

Mixed models

ROC curves

Propensity Scores

From the Book

* Contrasts (Chapter 11)
* Count Data (Chapter 13)
* Proportion Data (Chapter 14)
* Binary Response Variable (Chapter 15)
* Death and Failure Data (Chapter 16)

Data sources:

"datasets" package in R using: require(datasets), help(package=datasets), then help for individual data set. You can use the data set directly by name. To have it in your environment, use data('data set name')

<http://topcoder.co>m, requires sign in

<http://www.bls.gov/nls/nlsy79.htm> , National Bureau of Labor Statistics NLSY79 Longitudinal Survey

<http://archive.ics.uci.edu/ml/datasets.html>, University of California Irvine machine learning repository

<http://www.kaggle.com/>

<http://www.amstat.org/publications/jse/jse_data_archive.htm>, a collection of data sets curated for statistics education by the American Statistical Association

<https://cloud.google.com/bigquery/public-data/>

<https://www.reddit.com/r/bigquery/wiki/datasets#wiki_datasets_publicly_available_on_google_bigquery>

<https://research.stlouisfed.org/fred2/> Federal Reserve Data

<http://www.kdnuggets.com/datasets/index.html> Collected data sets for data analysis and data mining

<http://community.amstat.org/stats101/home>

<http://wise.cgu.edu/helpful-links/data-sources/> a master list of possibilities

<https://stat.ethz.ch/R-manual/R-devel/library/datasets/html/00Index.html> index of data sets provided with R

<http://www.nhtsa.gov/FARS> FARS (Fatal Analysis Reporting System), National Highway Traffic Safety Administration : summary data and raw data for US traffic fatalities, 1975-present.

[www.broad.mit.edu/cgi-bin/cancer/datasets.cgi](http://www.broad.mit.edu/cgi-bin/cancer/datasets.cgi) gene expression data sets (more for data mining?)

<https://toolbox.google.com/datasetsearch> a search tool for data sets

<https://www.ipums.org/> large collection of US and international census and survey data. Requires login and data extraction

<https://dasl.datadescription.com/> collection of data sets searchable by method.

[https://data.cms.gov/](https://data.cms.gov/Medicare-Physician-Supplier/Medicare-Provider-Utilization-and-Payment-Data-Phy/utc4-f9xp/data) healthcare data