Project 2 Description COMP 4442

Probability and Statistics for Data Science II, Spring 2019

For the second project, I would like you to find a “sufficiently rich” data set and analyze it, utilizing at least one technique not yet covered in class (with the exception of Principal Component Analysis, which will be allowed). Some potential options are listed below. In addition to commented code, I would like you to write up a paper that summarizes the work you did. The primary goal is for you to tell a story here. If possible, pick a data set that relates to something you care about. I would like you to

Rubric for paper and code

* Data source picked by Sunday April 2nd w/potential technique 10
* Appropriate exploratory data analysis performed 10
* Executive summary explains research question and conclusion 10
* Method for addressing research question explained 10
* Primary method explained in principle 10
* Application of primary method explained 10
* Data satisfaction of requirements of method demonstrated 10
* Primary new method applied and interpreted correctly 10
* Supporting visualizations provided 10
* Body of paper appropriate for entry level professional 10

Possible methods for use in project:

Robust Regression

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

-Regression trees

-Principle Component Analysis

-Partial Least Squares

-Random forests

` -Factor analysis

Clustering

- Latent Class Analysis

- DBSCAN

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

Hierarchical Linear Models

ROC curves

Propensity Scores

Methods for data with dependent error structures (geospatial and longitudinal data)

SEM (Structural Equation Modeling)

IRT (Item Response Theory)

Logit models (discrete choice analysis)

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