MIS 545 Suggested Project List

Below are descriptions of two datasets and suggested idea that you can use for your projects. More datasets and their URLs are provided on page 2. You can also use alternative datasets not on the list.

1. City of Chicago crimes data

This dataset reflects reported incidents of crime (with the exception of murders where data exists for each victim) that occurred in the City of Chicago from 2001 to present, minus the most recent seven days. Data is extracted from the Chicago Police Department's CLEAR (Citizen Law Enforcement Analysis and Reporting) system. In order to protect the privacy of crime victims, addresses are shown at the block level only and specific locations are not identified.

This dataset is useful for tasks such as classification and prediction. More information about the data and the download link can be found at:
https://catalog.data.gov/dataset/crimes-2001-to-present-398a4

2. IMDB Movie database and user rating data

IMDB is a movie database consists of many different attributes about movies. For example, movie title, genre, actors/actresses, directors, company, year, etc. This dataset is useful for tasks such as classification and clustering. One idea is predicting the user rating based on movie information.

The dataset can be downloaded at: http://www.imdb.com/interfaces .

3. Internet advertisement dataset

This dataset represents a set of possible advertisements on Internet pages. The attributes encode the geometry of the image (if available) as well as phrases occurring in the URL, the image's URL and alt text, the anchor text, and words occurring near the anchor text. There are two class labels: advertisement ("ad") and not advertisement ("nonad"). Among the 3279 observations, 458 are advertisements and 2821 are not.

This dataset is useful for tasks such as classification. It can be downloaded at:
http://archive.ics.uci.edu/ml/datasets/Internet+Advertisements

-- See Next Page --

Other available data sets.

- KDNuggest: jobs, data, software
  - http://www.kdnuggets.com/datasets/index.html (good starting point!_
- Kaggle competitions often include links to open datasets
  - https://www.kaggle.com/datasets
  - E.g., data about San Francisco https://data.sfgov.org/
- UCI Machine Learning Repository, http://archive.ics.uci.edu/ml/datasets.html
- Open data portals,
  - http://dataportals.org/
  - https://github.com/openlab/OGDI-DataLab
- World Bank: http://data.worldbank.org/
- Google public data: http://www.google.com/publicdata/directory
- Google Trends data: https://www.google.com/trends/
- Amazon's public data sets: http://aws.amazon.com/public-data-sets/
- Social Media data: https://dev.twitter.com/streaming/public
  https://www.instagram.com/developer/
  https://www.tumblr.com/docs/en/api/v2
- The government
  - Crime, justice: http://www.bjs.gov/index.cfm?ty=dca
  - Census: http://census.gov/data/data-tools.html
  - http://us-city.census.okfn.org/
  - http://www.data.gov/
- Linguistic Data Consortium
  - https://www.ldc.upenn.edu/data-management
- Movies and Music
  - Million song database, http://labrosa.ee.columbia.edu/millionsong/
  - https://github.com/sidooms/MovieTweetings
  - http://developer.rottentomatoes.com/
  - Open movie DB, OMDB: http://www.omdbapi.com
  - Film stats: http://www.statista.com/topics/964/film/
  - Lyrics: https://developer.musixmatch.com/
- Sports
  - http://fantasydata.com/products/real-time-sports-data-api.aspx?gclid=CJbV2vyWt8oCFYRDaQodW1sBsQ
- Crime
  - Different datasets: https://catalog.data.gov/dataset?tags=crime
- Data Science at Microsoft

    - http://research.microsoft.com/en-US/projects/data-science-initiative/datasets.aspx
- Wikipedia: download, API
    - https://www.mediawiki.org/wiki/Developer_hub
- Reddit: https://github.com/reddit/public-data-sets
- Ad hoc
    - http://datahub.io/dataset
    - http://pages.stern.nyu.edu/~adamodar/New_Home_Page/datacurrent.html (small)
    - https://www.yelp.com/academic_dataset
    - http://linkeddata.org/data-sets
    - https://archive.org/index.php