CS3244 Curated Datasets

This document describes the datasets that have been curated by the CS3244 staff. Select one to use your project requirements. You are allowed to use other data or create your own data *in addition to* using the dataset that you select for your project. We have divided the datasets into several categories to make it easier to browse and select, and your group members are welcomed to discuss with us on the respective #dataset-xx channels in Slack.

Note that datasets often allow different types of problems to be applied, of which a general scenario may be mentioned within the descriptions below. You can do supervised, unsupervised as well as other variants of ML on most datasets. A notional difficulty is given for the general scenario that the dataset facilitates. If in doubt about whether your idea is feasible, please ask your tutorial group leader in your respective #tg-xx channel or the class-wide #projects channel. You may also want to consult the Project FAQ pinned on the #projects channel, which will be updated periodically.

Tabular Datasets: Channel #dataset-tabular

1. **Resale Flat Prices**
   https://data.gov.sg/dataset/resale-flat-prices
   Size: 212K
   Estimated difficulty: Easy

   Resale transacted prices in Singapore from 1990 to present, managed by the Housing and Development Board (HDB).

2. **Guns Incident Data**
   https://www.kaggle.com/datatattle/guns-incident-data
   Size: 7.41 MB
   Estimated difficulty: Easy

   This dataset contains a list of data relating to gun incidents. You may try to predict the cause of an incident (suicide, homicide, etc.) based on race, date, education level and police involvement.

3. **Fraud Detection in Electricity and Gas Consumption Challenge**
   https://zindi.africa/competitions/ai-hack-tunisia-4-predictive-analytics-challenge-1/data
   Size: 135,493 Instances
   Estimated difficulty: Medium

   The Tunisian Company of Electricity and Gas (STEG) is a public and a non-administrative company, it is responsible for delivering electricity and gas across

Tunisia. The company suffered tremendous losses in the order of 200 million Tunisian Dinars due to fraudulent manipulations of meters by consumers.

Using the client's billing history, the aim of the challenge is to detect and recognize clients involved in fraudulent activities. The solution will enhance the company's revenues and reduce the losses caused by such fraudulent activities.

4. **Credit Card Approval Prediction**
   https://www.kaggle.com/rikdifos/credit-card-approval-prediction
   Size: 439K
   Estimated difficulty: Medium

   Classification of credit records and application records into 'good'/'bad' clients so as to help the credit card approval process. The definition of 'good'/'bad' is not given, so there is a need for further analysis to self-define the labels.

5. **Facebook Comment Volume Dataset**
   https://archive.ics.uci.edu/ml/datasets/Facebook+Comment+Volume+Dataset
   Size: 40K, with 54 features
   Estimated difficulty: Medium

   Instances in this dataset contain features extracted from Facebook posts. The task associated with the data is to predict how many comments the post will receive. The dataset contains 5 variants of the dataset. This dataset only has numeric or binary quantities; there is no text included in the dataset.

6. **End ALS Kaggle Challenge**
   https://www.kaggle.com/alsgroup/end-als
   Size: 71 GB
   Estimated difficulty: Hard

   This dataset is composed of a curated collection of 134 people with ALS or motor neuron disease (MND) and a 30-person healthy control population. It includes genomics, transcriptomics, and clinical data.

Text Datasets: Channel #dataset-text

1. **Quora Question Pairs**
   https://gluebenchmark.com/tasks and https://www.kaggle.com/c/quora-question-pairs
   Size: 404K
   Estimated difficulty: Medium

Question similarity evaluation. Given a pair of questions determine whether they are semantically similar (both of them have the same meaning). The Q&A website Quora released this dataset to build classifiers to help Quorians avoid duplicate effort.

2. **Corpus of Linguistic Acceptability**
   https://nyu-mll.github.io/CoLA/
   Size: 10.6K sentences
   Estimated difficulty: Medium

   Consists of 10,657 sentences from 23 linguistics publications, expertly annotated for acceptability (grammaticality) by their original authors.

3. **Sarcasm Detection**
   https://www.kaggle.com/danofer/sarcasm
   Size: 364.15 MB
   Estimated difficulty: Medium

   This dataset contains 1.3 million Sarcastic comments from the Internet commentary website Reddit. The dataset was generated by scraping comments from Reddit (not by me :)) containing the \s (sarcasm) tag. This tag is often used by Redditors to indicate that their comment is in jest and not meant to be taken seriously, and is generally a reliable indicator of sarcastic comment content.

4. **IMDB Spoilers**
   https://www.kaggle.com/rmisra/imdb-spoiler-dataset
   Size: 921.99 MB
   Estimated difficulty: Medium

   User-generated reviews are often our first point of contact when we consider watching a movie or a TV show. However, beyond telling us the qualitative aspects about the item we want to consume, reviews may inevitably contain undesired revelatory information (i.e. `spoilers') such as the surprising fate of a character in a movie, or identity of a murderer in a crime-suspense movie etc. For users who are interested in consuming the item but are unaware of the critical plot twists, spoilers may decrease the excitement regarding the pleasurable uncertainty and curiosity of media consumption. Therefore, a natural question is how to identify these spoilers in entertainment reviews, so that users can more effectively navigate review platforms.

   This dataset is collected from IMDB. It contains meta-data about items as well as user reviews with information regarding whether a review contains a spoiler or not. For more details on the attributes, please check file descriptions.

5. **Twemoji**

https://uvaauas.figshare.com/articles/dataset/Twemoji_Dataset/5822100
Size: 13M tweets
Estimated Difficulty: Medium

Collection of 13M tweets divided into training, validation, and test sets for the purposes of predicting emoji based on text and/or images. The data provides the tweet status ID and the emoji annotations associated with it. In the case of image-containing subsets, the image URL is also listed. The Full, unbalanced dataset consists of a random test and validation sets of 1M tweets, with the remainder in the training set. The Balanced testset is a subset of the test set chosen to improve emoji class balance.The Image subsets are image-containing tweets. Finally, emoji_map_1791.csv provides information regarding the emoji labels and potential metadata.

Vision Datasets: Channel #dataset-vision

1. **Fashion MNIST**
   https://www.kaggle.com/zalando-research/fashionmnist
   Size: 70K
   Estimated difficulty: Medium

   Fashion-MNIST is a dataset of Zalando's article images—consisting of a training set of 60,000 examples and a test set of 10,000 examples. Each example is a 28x28 grayscale image, associated with a label from 10 classes. It serves as a direct drop-in replacement for the original MNIST dataset for benchmarking machine learning algorithms. It shares the same image size and structure of training and testing splits.

2. **DeepWeeds**
   https://github.com/AlexOlsen/DeepWeeds
   Size: 17.5K
   Estimated difficulty: Medium

   The DeepWeeds dataset consists of 17,509 images capturing eight different weed species native to Australia in situ with neighbouring flora.

3. **Covid Chest X Ray Dataset**
   https://github.com/ieee8023/covid-chestxray-dataset/
   Size: ~600 samples.
   Estimated difficulty: Hard

Open dataset of chest X-ray and CT images of patients which are positive or suspected of COVID-19 or other viral and bacterial pneumonias (MERS, SARS, and ARDS.). Somewhat unstructured as it is still an emerging dataset. This is a challenging dataset as it is not very well processed and has significant data skew – it is recommended that your team augment the dataset to have more "normal" training data.

Time Series Datasets: Channel #dataset-timeseries

1. **Huge Stock Market Dataset**
   https://www.kaggle.com/borismarjanovic/price-volume-data-for-all-us-stocks-etfs
   Size: 8.4K companies
   Estimated difficulty: Medium

   Provides data of about 8.4K companies spanning through 2009 to 2017, for all US-based stocks and ETFs trading on the NYSE, NASDAQ, and NYSE MKT. The data (last updated 11/10/2017) is presented in CSV format as follows: Date, Open, High, Low, Close, Volume, OpenInt. Note that prices have been adjusted for dividends and splits.

2. **Recognition of Human Activities Dataset**
   http://archive.ics.uci.edu/ml/datasets/Smartphone-Based+Recognition+of+Human+Activities+and+Postural+Transitions
   Size: 10929 Instances
   Estimated difficulty: Medium

   This dataset consists of smartphone (Samsung Galaxy S II) readings, specifically of the time series variant from the embedded accelerometer & gyroscope of the device. Participants wore the smartphone on the waist while performing six basic activities (standing, sitting, lying, walking, walking downstairs and walking upstairs). Each record of the activity window has an associated label of the activity performed. The objective is to classify new data into its appropriate activity label.

Other Datasets: Channel #dataset-others

1. **OSIC Pulmonary Fibrosis Progression**
   https://www.kaggle.com/c/osic-pulmonary-fibrosis-progression
   Size: 22.35GB
   Modality: Image + Time-series
   Estimated difficulty: Hard

   Predict a patient's severity of decline in lung function based on a CT scan of their lungs. You can determine lung function based on output from a spirometer, which measures the volume of air inhaled and exhaled. The challenge is to use machine learning techniques to make a prediction with the image, metadata, and baseline FVC as input.