Data source for ML practice

January 21, 2020

0.1 Pedram Jahangiry

0.2 Pandas Datareader

This is a very powerful data source provided by pandas packet. checkout this website.

0.3 scikit learn

This is the best reference for toy datasets already coming with sklearn package. There are also some larger size datasets availabe for download.

0.4 Kaggle

There are lots of machine learning ready datasets available to use for fun or practice on Kaggle's Public Datasets platform. * Use kaggle search

0.5 ML in Finance and Econ

Today, Machine Learning is used in finance for: * Portfolio Management * Algorithmic trading * Fraud detection * Underwriting * Sentiment analysis * Customer Service * Reccomending products/services

Machine Learning is also used in Economics for: * Testing economic models * Collecting new sources of data - Economists have even converted satellite data into estimates of economic growth, and measured neighborhood income levels in Boston and New York using Google Street View * Predicting behaviour of citizens to help inform policy making and predict problem areas

Here is a very good link to a list of available dataset for Finance and Econ

Here is a short list:

- Credit default risk competition
- Credit card fraud dataset
- NYSE dataset
- Lending club dataset

0.6 UCI

- UC Irvine ML repository

Here is a short list of some of mostly cleaned and ready for analysis datasets!

0.7 Binary Classification

- Indian Liver Patient Records
- Synthetic Financial Data for Fraud Detection
- Business and Industry Reports
- Can You Predict Product Backorders?
- Exoplanet Hunting in Deep Space
- Adult Census Income

0.8 Multiclass Classification

- Iris Species
- Fall Detection Data from China
- Biomechanical Features of Orthopedic Patients

0.9 Regression

- Video Game Sales with Ratings
- NYC Property Sales
- Gas Sensor Array Under Dynamic Gas Mixtures

0.10 NLP

- The Enron Email Dataset
- Ubuntu Dialogue Corpus
- Old Newspapers: A cleaned subset of HC Corpora newspapers
- Speech Accent Archive
- Blog Authorship Corpus

0.11 Time Series Analysis

- Cryptocurrency Historical Prices
- Exoplanet Hunting in Deep Space

0.12 Image Processing

- YouTube Faces with Facial Keypoints
- Fashion MNIST

0.13 Mapping and Prediction

- Seattle Police Department 911 Incident Response
- Baltimore 911 Calls
- Crimes in Chicago
- Philadelphia Crime Data
- London Crime

0.14 Large Datasets

- Iowa Liquor Sales
- Seattle Library Checkout Records

0.15 Quandl

Can find many econ and finance free datasets here

0.16 Quantopian

lots of fun with algo trading here