Project Part 1

The goal of this part of the project is to get some familarity with how to approach a machine learning data science project. You will be working with the data science cookie cutter template. This will be an individual project but you will be exploring data sets you may eventually work with as part of your project.

Overview of Steps

Step 0: Keep Data out of Repo but everything else shoudl be there

Warning: Never commit large data files in a github repository.

It is important that you never commit large files into your repository. It is very very hard to delete files in a git repository. Sure, you can delete the file and commit it but a copy of the file ends up living in the hidden .git directory. You can't just delete it there. If you do you risk corrupting the whole repository. Instead you should use .gitignore, as in the blog post Gitignore, What is it and How to Add to Repo. To manage large data file either use git Git Large File Storage if the size is under 2GB, or learn how to keep your data in dropbox, microsoft one drive or google drive and link (shortcut) the file to the repository. You still may need to download it separately but CCNY/CUNY have large storage allowances for one drive and dropbox in particular.

Step 1: Keep all your work in a repo and commit often

Open a git repository using the classroom link. The template should be created for you using Cookiecutter Data Science. You will need to manually go through the files and change some of the variable names. Follow the README.md. Please read the cookiecutter documentation and use the structure of the template for this project.

Keep in mind your experiments should go into the folder notebooks and your report should be in the reports folder. If you can factor out code that you will import into notebooks you should put that code in the src folder. Aagain read the Cookiecutter Data Science documentation.

It is very important that you commit often. If there are not frequent commits I will have to assume the code was copied and pasted and you will recieve little credit. The work should show your thinking process over time and not just at the last minute

Step 2: Selecting Data

Here are a number of resources for finding data sets

General Data Set Aggregators

- kaggle data sets and
- kaggle challenges
- awesome public data sets
- OpenML
- Datahub.io Collections
- Paperswithcode Datasets
- VisualData Discovery
- Registry of Open Data on AWS
- Google Datasets

Public government datasets

- Data.gov
- Data USA
- data.europa.edu
- US Healthcare Data
- US Education Data
- NYC Open Data
- Los Angeles Open Data
- Chicago Open Data
- Houston Open Data
- UK Data Service
- UK Government Data
- United Nations

Machine Learning Datasets for Finance and Econ

- Financial Times
- quandl
- IMF Data
- JPMorgan
- American Economic Association
- World Bank

Image Datasets for Computer Vision

- MIT CSAIL LabelMe
- Imagenet
- Kinetics
- LSUN: Construction of a Large-scale Image Dataset using Deep Learning with Humans in the Loop
- MSCOCO
- COIL-100
- VisualGenome
- Open Images Dataset
- YouTube-8M
- Labeled Faces in the Wild
- Indoor Scene Recognition
- xView
- CelebFaces
- Stanford Dogs Datasets
- Places
- CityScapes
- CIFAR
- Visual Question Answering Datasets (VQA)
- IMDB-WIKI
- MPII Human Pose Dataset
- Google Open Images V5

Natural Language Processing Datasets

- The Big Bad NLP Dataset
- Enron Email Dataset
- Google NGrams

- The Wikipedia Corpus
- SMS Spam Collection Datasets
- Yelp Open Dataset
- Blog Authorship Corpus

Sentiment Analysis Datasets for Machine Learning

- Multi-Domain Sentiment Dataset
- Stanford Sentiment Analysis
- Sentiment140
- IMDB Movie Reviews Dataset
- Twitter US Airline Sentiment
- OpinRank Review Dataset Data Set
- Amazon Review Data
- Sentiment Lexicons for 81 Languages
- Paper Reiviews

Text Datasets for Natural Language Processing

- Jepordy Datasets
- 20 Newsgroups
- Legal Case Reports Data Set
- Microsoft Research WikiQA Corpus
- Public Databases

Audio Speech and Music Datasets for Machine Learning Projects

- Common Voice
- Google AudioSet
- LibriSpeech
- The Spoken Wikipedia Corpora
- VoxForge
- Free Music Archive (FMA)
- Ballroom

Autonomoust Vehicles Datasets

- Berkeley DeepDrive BDD100K
- comma.ai driving dataset
- Oxford’s Robotic Car
- Laboratory for Intelligent & Safe Automobiles
- Baidu ApolloScape: Dataset for Autnomous Driving
- Google-Landmarks-v1, Google-Landmarks-v2
- PandaSet
- NuScenes
- Open Waymo Dataset

Step 3: Plan for you future project

Please use this refrerence in terms of your future project: Stanford Univerisity's CS229. It also has some good discussion of what you might do for a project and where you may find data. Make sure that you are looking at a data set that is not a simple quick execise. This should not be a small data set like the titanic data set or the

MNIST data set. It should either by large in terms of the number of data points (at least several thousand “rows”) or large in the number of features. For example images and audio files are typically complex with many dimensions in terms of number of features.

Think about different hypothesis you could investigate on your data set. Take notes in a markdown file indicating the kind of hypothesis you expect to see when you look at your data. As you do EDA you should revisie these.

Step 4: Data Wrangling, Data Cleaning, Preliminary EDA

How much missing data is there? Determine a prelimanary EDA to look at basic statistics and to see if there is to much missing data? What are the means, standard deviations, and other statisics on feaures. What happends when you filling in missing data will effects the statistics? There should be figures and discussion. Please use notebooks to do your analysis but when you submit, you must summerise your analysis. Do NOT just fill notebooks with output of some kind of runs without discussion or figures.

Step 5: EDA and Data Visualization

Here you should show again the overall statistics of features or potential features. Besides summary statistics, you should be showing how potential features relate to each other using measures such as correlation or mutual information. One of your goals should be either regression or classification and you should explore what features are correllated with your target label or variable. This should be visible in the plots. Again, in your experimental notebooks this may be a bit disorganized, but there should be a notebook where the explanation is clearly spelled out and summerized. Whenever you make a run, take the time to explain not just what the figure is but what it ended up showing. Each diagram figure and result tells you something, explain what.

Step 6: Invesigate some models and the importance of features.

Here you should run classification models or/regression models to see how strong a signal there is in your data set. Your data should **always** have training, validation and testing data subsets. You should make your exploration based on the training and validation sets. The key question is, again what features and normalizations provide the strongets results from simple models such as KNN, linear models (eg logistic regression), and decision trees. You should pick the models appart, and generate some visualizations to understand exactly how these models do and why. To the extent that there is some data on which these models do not work well, you should discuss why by looking at the original data for those data and trying to determine if those points are special in anyother way.

Step 7: Provide evaluation and conclusions

Your result should include peformance metrics, classification reports, accuracies, confusion matrices, and ROC curves. Again you should explain your results. If you have a model with high “accuracy” but unbalanced classes, then your accuracy is meaningless and you should balance first. Make sure you understand what is the reason for unbalanced false postives or false negatives, and how you can fix it. Again, when you present things do not just simply show but explain WHAT it shows and what the impact is. Commit all these into your repo and submit that repo to submission in blackboard.