STA 221: Project Details

1 Logistics

- Each group must have 2 students: If you want to join a group or if want a member in your group, use Piazza.
- Choose a topic and submit a Proposal: Due 11:59PM PST, May 03. After meeting with the instructor (as mentioned below), you can make changes to this proposal till 11:59PM PST, May 08.
- I'll have individual meetings with each group on **May 04 and May 06 (during class hours)** (more details about the schedule later).
- Final Project Report Due: June 10 (tentative)

2 Project Proposal

- Up to 1 page (single column)
- Include the following information in the proposal:
  - Your names
  - Email address
  - Project topic
  - Proposed work (what and how you are planning to do)
  - Important references, if any.

Before coming to the individual meeting, you must have a good idea of what you want to do for your project (i.e., you may want to have a rough proposal ready). Use the meeting to: (1) fine-tune your proposal (for e.g., the datasets and algorithms/packages to be used), (2) make sure the project, if completed as proposed, would lead to a good grade, and (3) make sure the level/standard of the project is at the level required for this course.

3 Ideas

- Work on some data science competitions: Take some of the datasets released from earlier Kaggle competition and try to solve the problem using your approach. Take some of the other datasets (e.g., KDDCup data, Yahoo Webscope data, and try to solve the problem. You can also get your own data (using pandas) to frame your own data analysis problem.
- Compare different algorithms on several large-scale datasets (e.g., compare classification algorithms, compare regression algorithms, compare clustering algorithms, etc) and derive insights into their working performance.
- Implement one of the algorithm described in this course, from scratch and try to scale it to large datasets and parallelize it. Choose a research paper and implement the algorithm; test on different data/problem, or try to modify the algorithm
- Any other problems related to data analysis/machine learning that you find interesting in your field of interest

3.1 Competition Datasets/Problems

- Recent COVID-19 datasets: `https://cloud.google.com/blog/products/data-analytics/free-public-datasets-for-covid19` and `https://cset.georgetown.edu/covid-19-open-research-da`
- House price prediction challenge: `https://www.kaggle.com/c/house-prices-advanced-regression-tech`
- Google Cloud NCAA ML Competition: `https://www.kaggle.com/c/mens-machine-learning-competitio`
`https://www.kaggle.com/c/womens-machine-learning-competition-2018`
- Toxic Comment Classification Challenge: `https://www.kaggle.com/c/jigsaw-toxic-comment-classific`
`data`
- KKBox's Churn Prediction Challenge: `https://www.kaggle.com/c/kkbox-churn-prediction-challenge`
- KDDCup 2017: Highway Traffic Flow Prediction `https://tianchi.aliyun.com/competition/information.htm?spm=5176.100067.5678.2.8CnCPt&raceId=231597`
- KDDCup 2016 (Academic graph)
- Yahoo Webscore data `https://webscope.sandbox.yahoo.com/`
  - Predict movie or music ratings
  - Learning to rank challenges
- UCI Machine Learning Dataset: `https://archive.ics.uci.edu/ml/index.php`
- Try to find a competition that interests you!

3.2 Comparing Algorithms using Python Commands

In this type of project, you compare the performance of different algorithms and derive insights. For example, you find out that if the data has *some properties*, it is best to use Algorithm `name here`.

- Compare all algorithms for classification:
  - SVM, logistic regression, random forest, Deep learning, ⋯
  - Datasets can be found in UCI data: `https://archive.ics.uci.edu/ml/datasets.html`
- Compare all algorithms for regression:
  - Linear regression, kernel regression, random forest, Deep learning, ⋯
- Compare all algorithms for clustering:
  - Kmeans, spectral clustering, metis, ⋯
  - Think about different ways to evaluate.
- Compare algorithms/packages for word2vec:
  - Glove, Google W2V, PPMI-SVD, Implicit Matrix factorization, ⋯
  - Think about how to evaluate.

3.3 Implementing from scratch

- SVM, Logistic regression for large-scale datasets
- Clustering algorithms, for large-scale sparse data
- Dimension reduction algorithms for large-scale data.
- Choose a research paper on the above algorithms and try to implement it. You can try to (1)reproduce the results, (2) try on different datasets (3) apply to other applications. If you are doing this contact me to get a relevant research paper to start with.

4 Final Report

There will be more details about this soon.