STA141C: Big Data & High Performance Statistical Computing

Final Project Proposal

Cho-Jui Hsieh
UC Davis

April 4, 2017

Final Project

- Form a group of (up to) 3 students
- Choose a topic
- Submit the final project proposal

  Due 11:59pm PST, May 19
- I'll have individual meetings with each group

  Maybe May 23 (Tuesday)
- Submit the final project report

  Due 11:59pm PST, June 12 (tentative)

Final Project Proposal

- Up to 1 page (single column)
- Mainly for me to understand your topic and give some suggestions (grading will be based on your final project)
- Include the following information in the proposal:
  - Your names
  - Email address
  - Project topic
  - Proposed work (what are you planning to do)
  - Important references
  - What's the difficulty of the proposed work (so I can provide some suggestions)

Topic for Final Project

Potential topics:

- Work on one of the current Kaggle competition
- Take some of the datasets released from earlier Kaggle competition and try to solve the problem
- Take some of the other datasets (e.g., KDDCup data, Yahoo Webscope data `https://webscope.sandbox.yahoo.com/`), and try to solve the problem
- Compare different algorithms on some datasets (e.g., compare classification algorithms, compare regression algorithms, compare clustering algorithms, etc)
- Implement one of the algorithm described in this course, and try to scale it to large datasets.
- 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/computational statistics

Topics for Final Project

Examples: Datasets/Problems from Kaggle

- Question pairs prediction (homeworks) (ongoing)
- House price prediction challenge (ongoing): `https://www.kaggle.com/c/house-prices-advanced-regression-techniques` `https://www.kaggle.com/c/sberbank-russian-housing-market`
- Cancer screening (harder, ongoing) `https://www.kaggle.com/c/intel-mobileodt-cervical-cancer-screening`
- Count sea lions in the photo (harder, ongoing) `https://www.kaggle.com/c/noaa-fisheries-steller-sea-lion-population-count/`
- Credit card fraud detection (classification) `https://www.kaggle.com/dalpozz/creditcardfraud`
- Many others

Topics for Final Project

Other examples:

- 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
  - ...

Topics for Final Project

You can also implement/compare existing algorithms for some applications.

- Compare algorithms for classification:
  - SVM, logistic regression, XGBoost/LightGBM, random forest, Deep learning, $\cdots$
  - Datasets can be found in LIBSVM data or UCI data:
    `https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/`
    `https://archive.ics.uci.edu/ml/datasets.html`
- Compare algorithms for regression:
  - Linear regression, kernel regression, random forest, XGboost/LightGBM, $\cdots$
- Compare algorithms for clustering:
  - Kmeans, spectral clustering, metis, $\cdots$
  - Think about different ways to evaluate.
- Compare algorithms/packages for word2vec:
  - Glove, Google W2V, PPMI-SVD, Implicit Matrix factorization, $\cdots$
  - Think about how to evaluate.

Topics for Final Project

You can also implement one of the algorithm in this course and try to scale to large datasets (maybe using multi-core):

- SVM, Logistic regression for large-scale datasets
- Clustering algorithms, for large-scale sparse data
- Matrix factorization

Topics for Final Project

- Choose a research paper and try to implement it. You can try to (1) reproduce the results, (2) try on different datasets (3) apply to other applications
- Any other related topic will be good