The University of Texas at Austin
Department of Electrical and Computer Engineering

EE381V: Large Scale Machine Learning — Spring 2016

Problem Set Three

Dimakis/Caramanis Due: Tuesday, March 29, 2016.

Computational Problems

1. **Getting into Kaggle.**
   In this problem we will start playing with Kaggle. We will start with doing submissions to the Amazon Employee Access competition. The challenge is to predict if an Amazon employee will be given access to a requested resource.

   `https://www.kaggle.com/c/amazon-employee-access-challenge`

   Create a Kaggle account. Download the test and training datasets for this competition. Try a vanilla logistic regression. What is the private score you get ?

2. **XGBoost.**
   Install XGBoost. You can follow for example this `https://github.com/dmlc/xgboost/blob/master/doc/build.md`

   Use XGBoost to predict the performance. Use five-fold Cross-Validation to search over XGBoost parameters (Max depth, nestimators, colsamplebytree). What are the best parameters you find ? Make a submission. What private score do you get ?

3. **Winnning.**
   Do your best to get the highest score possible! Try randomizing XGboost random seed, produce many prediction vectors and average them together. Also try using some of the predictors as extra features by adding them in the Xtraining matrix (stacking). Remember to save your predictins to files so that you can easily re-use them later. Read how to make ensembles `http://mlwave.com/kaggle-ensembling-guide` and how to tune XGBoost (see 'The complete guide to parameter tuning in XGBoost') Also read how the winning strategy worked for this competition

   `https://www.kaggle.com/c/amazon-employee-access-challenge/details/winners`

   Send us a screen capture from

   `https://www.kaggle.com/c/amazon-employee-access-challenge/submissions`

   We expect a private score of at least 0.87*.*

For this problem set you may form groups of up to 3 people. Submit your csv submissions, the python code that generates them and a screen capture of your best score from Kaggle. Also submit a short report of the numbers you got for parts 1,2,3 and the best parameters you found. For part 3 describe the strategy you found. Do one homework report per group. Limit the number of Kaggle submissions to 100 per group.