Large Scale Machine Learning - Assignment 3

Abhishek Sinha, Arun Sai, Ashish Bora

March 29, 2016

We report the models and the results first. For all models, the best hyper-parameters can be found in the Appendix 5.

1 Q1

- Model: $\ell 1$ regularized logsitic regression
- Private Score: 0.896

For this question we generated additional features from the original categorical features. All pairs and triples of the original categorical features were generated. We used One-Hot Encoding (OHE) for the original as well as new features.

2 Q2: XGBoost without OHE

- Model: Boosted Decision Trees.
- Private Score: 0.883

All pairs and triples of the original categorical features and their frequencies (i.e, number of times a particular pair or a triple occurred) were used to train the model.

3 Q2: XGBoost with one-hot encoding

- Model: Boosted decision trees
- Private Score: 0.8848

No additional features were used. OHE gives a 15k dimensional input vector.

4 Q3

The best private score we achieved was 0.90903. We used ensemble of many models which we describe below.

4.1 Models

We used the following models in our ensemble. Numbers in () indicate private scores for individual models. OHE stand for One-Hot Encoding

1. $\ell 1$ regularized logsitic regression with OHE and additioanl features (0.896)
2. XGBoost with OHE (0.8848)
3. XGBoost without OHE (0.883)
4. Random forests trained with entropy criterion without OHE (0.8762)
5. Random Forest trained with gini criterion wihout OHE (0.8744)
6. $\ell 2$ regularized linear kernel SVM with OHE (0.85118)

4.2 Ensembling

Inspired by Adaboost weighing scheme, we tried to weight the models in our ensemble by $\log\left(\frac{privatescore}{1-privatescore}\right)$. These weights are then normalized so that they sum to one and finally, we take the weighted average of the probability estimated given by each model.

We found that this gave too much weight to models with low accuracy, and resulted in worse performance. Instead to boost the weights to high accuracy model, we used weights proportional to $\left(\log\left(\frac{privatescore}{1-privatescore}\right)\right)^8$, where the exponent 8 was chosen empirically. With this scheme, we were able to improve the private score to 0.90903 whereas the best individual model in the ensemble was at 0.896.

5 Appendix : Hyperparameter values

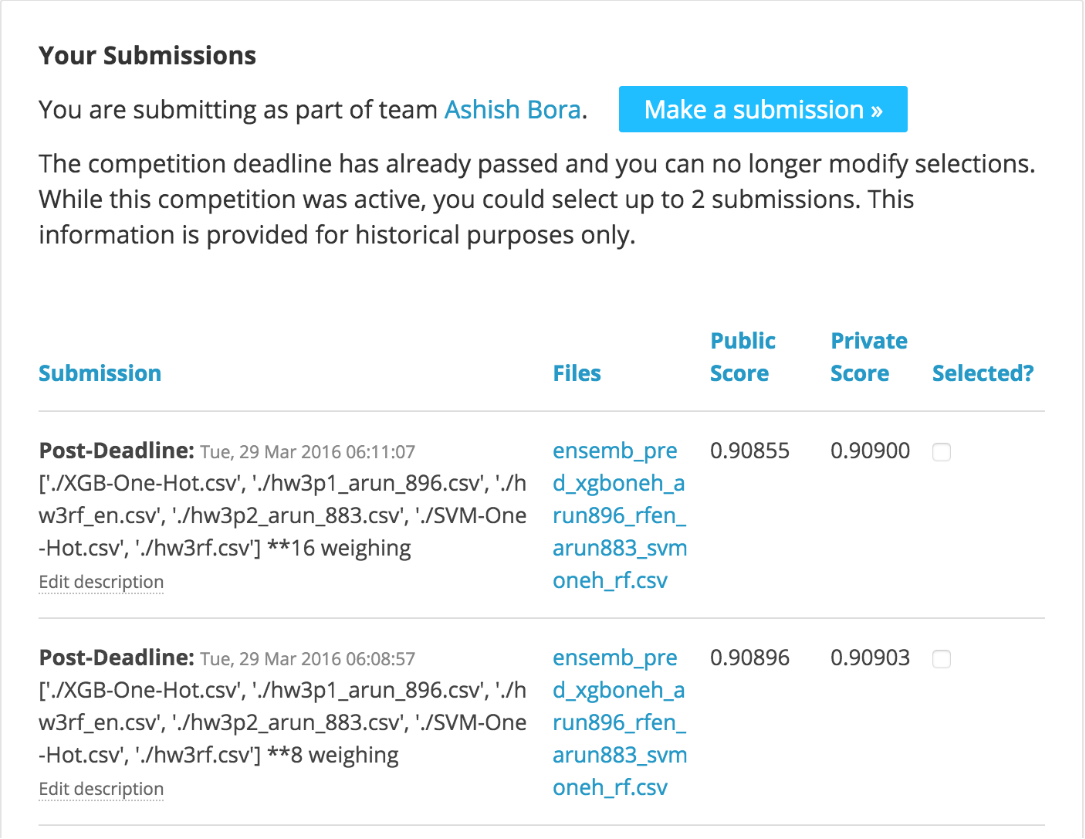
For reproducilibity, we provie the bset set of hyper-parameters found using 5-fold cross validation for all the models. If some paramteres are not mentioned, it is understood that we use the default values.

1. $\ell 1$ regularized logsitic regression
   -
2. XGBoost with one-hot encoding
   - `learning-rate` = 0.2
   - `n-estimators` = 880
   - `colsample-bytree` = 0.22

- `max-depth` = 16
- `min-child-weight` = 0.04
- `max-delta-step` = 4

3. XGBoost without one-hot encoding
    - `learning-rate` = 0.2
    - `n-estimators` = 100
    - `colsample-bytree` = 0.1
    - `max-depth` = 6.
4. Random forests trained with entropy criterion
    - `n-estimators` = 270
    - `max-features` = 4
    - `max-depth` = 23
    - `min-samples-leaf` = 2
    - `min-samples-split` = 8
5. Random Forest trained with gini criterion
6. $\ell 2$ regularized SVM with one-hot encoding
    - `C` = 2

6 Score screenshot



Your Submissions

You are submitting as part of team Ashish Bora.

The competition deadline has already passed and you can no longer modify selections. While this competition was active, you could select up to 2 submissions. This information is provided for historical purposes only.

Submission	Files	Public Score	Private Score	Selected?
Post-Deadline: Tue, 29 Mar 2016 06:11:07 ['./XGB-One-Hot.csv', './hw3p1_arun_896.csv', './hw3rf_en.csv', './hw3p2_arun_883.csv', './SVM-One-Hot.csv', './hw3rf.csv'] **16 weighing Edit description	ensemb_pred_xgboneh_arun896_rfen_arun883_svmoneh_rf.csv	0.90855	0.90900	□
Post-Deadline: Tue, 29 Mar 2016 06:08:57 ['./XGB-One-Hot.csv', './hw3p1_arun_896.csv', './hw3rf_en.csv', './hw3p2_arun_883.csv', './SVM-One-Hot.csv', './hw3rf.csv'] **8 weighing Edit description	ensemb_pred_xgboneh_arun896_rfen_arun883_svmoneh_rf.csv	0.90896	0.90903	□