Large Scale Machine Learning - Assignment 3
Abhishek Sinha, Arun Sai, Ashish Bora

March 29, 2016

1 Q1

- Model: $\ell 1$ regularized logsitic regression
- Private Score: 0.896
- Code : $hw3p1.py$. Predictions on test data: $hw3p1.csv$

For this question we generated additional features from the original categorical features. All pairs and triples of the original categorical features were generated. Also used one-hot encoding on all the resulting categorical features.

2 Q2: XGBoost

- Model: Boosted Decision Trees (trained using XGBoost).
- Private Score: 0.883
- Code : $hw3p2.py$. Predictions on test data: $hw3p2.csv$

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. *Doesn't* use one-hot encoding. Parameters found using 5 fold cross validation: $learning\_rate = 0.2, n\_estimators = 100, colsample\_bytree = 0.1, max\_depth = 6$ (we did a greedy search to find the parameters, so these may not the the best possible parameters)

3 Q2: XGBoost with one-hot encoding

- Model: Boosted Decision Trees (trained using XGBoost).
- Private Score: 0.8848

Boosted Decision Trees with one-hot encoding on the original categorical features. Doesn't use any other features.

4 Q3: Ensemble

- Model: ensemble of $\ell 1$ regularized logsitic regression, XGBoost with one-hot encoding, XGBoost without one-hot encoding (*these three are described in previous sections*) and linear SVM, random forests trained with 'entropy' criterion. The predictions of the ensemble are computed as a weighted average of the above 5 classifiers, with weights direclty proportional to the performance of individual classifiers.
- Private Score: 0.9090
- Predictions on test data: $hw3p3.csv$. Code for random forests: $hw3p3\_randomForest.py$
- Screen shot of submission to kaggle : $best.png$, $best\_2.png$

Best parameters for Random Forest
The Random Forest was trained on the original features. One hot-encoding was not used for categorical features. The best parameters obtained using 5-fold cross validation were– $n\_estimators = 270, max\_features = 4, max\_depth = 23, min\_samples\_leaf = 2, min\_samples\_split = 8, criterion = entropy$. The private score only with Random Forest was 0.8762.

SVM
Linear SVM with one-hot encoding on the original categorical features. Doesn't use any other features.