4. Training and Evaluating Models

When we look at the data, we can see that the problem is classification problem. Therefore, I have chosen Decision Tree, SVM and Naïve Bayes supervised learning algorithms. Below I will explain strengthens and weaknesses of these models.

Decision Tree is the first model I experimented. Strengthens of decision tree:

- Training and Prediction time is low
- Decision tree is easy to use and easy to explain to somebody.

Weaknesses of decision tree:

- Extremely sensitive even a small changes in data

Training Set 100 (Decision Tree)			
training time	prediction time	F1 score on training set	F1 score on test set
0.001	0.000	1.0	0.708661417323

Training Set 200 (Decision Tree)			
training time	prediction time	F1 score on training set	F1 score on training set
0.002	0.000	1.0	0.693548387097

Training Set 300 (Decision Tree)			
training time	prediction time	F1 score on training set	F1 score on test set
0.002	0.000	1.0	0.688524590164

Second algorithm I chose is Support Vector Machine (SVM). Advantages of SVM:

- SVM is robust to overfitting. As we see in below comparison tables. When we increase training data size, F1 score for training data does not change much.
- SVM can model non-linear relations.
- Effective in higher dimensional spaces

Weaknesses of SVM:

- More training and prediction time is needed.

Training Set 100 (SVM)			
training time	prediction time	F1 score on training set	F1 score on test set
0.001	0.001	0.86301369863	0.794520547945

Training Set 200 (SVM)			
training time	prediction time	F1 score on training set	F1 score on test set
0.003	0.003	0.863636363636	0.820512820513

Training Set 300 (SVM)			
training time	prediction time	F1 score on training set	F1 score on test set
0.009	0.006	0.861538461538	0.838709677419

Third algorithm I chose is Naïve Bayes. Strengths of Naïve Bayes:

- The model is simpler.
- Computational complexity is less than SVM.
- Fast to train and fast to classify

Weaknesses of Naïve Bayes:

- Assumes that all features are independent [1]

Training Set 100 (Naïve Bayes)			
training time	prediction time	F1 score on training set	F1 score on test set
0.001	0.000	0.815384615385	0.77519379845

Training Set 200 (Naïve Bayes)			
training time	prediction time	F1 score on training set	F1 score on test set
0.001	0.000	0.786764705882	0.772727272727

Training Set 300 (Naïve Bayes)			
training time	prediction time	F1 score on training set	F1 score on test set
0.004	0.002	0.797136038186	0.782608695652

5. Choosing the Best Model

As we see in previous section, each model has strengthens and weaknesses. If we are searching the best model in terms of computationally cost, then Decision Tree is the best model. In terms of accuracy, SVM is the best model. Since data is small and complexity of SVM is not issue here. Therefore, SVM is the best single model for this problem.

Explanation in laymen's terms: Imagine that we have a big land with two types of seeds which each one needs different harvesting technique. You want to draw the best line between these two groups in order not to waste any crop. And you realized that you can find the best separator by having as big as gap on either side of the line. Our algorithm, i.e. SVM is also doing the same thing.

After fine-tune the model, I got F1_score=0.86.

[1] - http://blog.echen.me/2011/04/27/choosing-a-machine-learning-classifier/