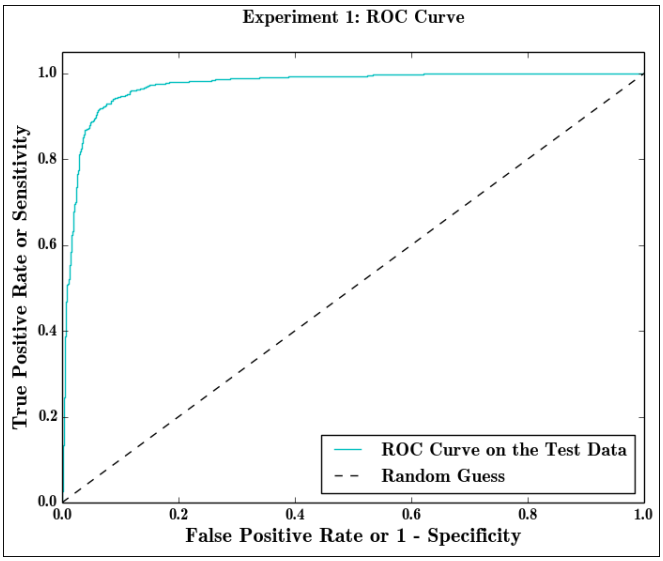
CS 545 Machine Learning
Homework 3: SVMs and Feature Selection
Haomin He

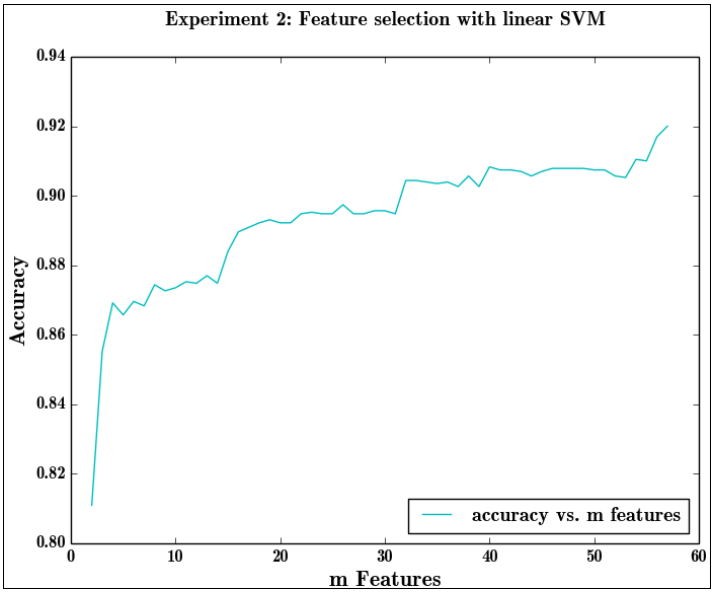
Experiment 1:
– Which SVM package you used
– Accuracy, Precision, and Recall on the test data, using learned model
– ROC Curve
Answer: SVM package used: pandas (pandas data frame), numpy (numpy array), sklearn model_selection (split dataset and scale features), sklearn svm, sklearn.metrics recall_score, precision_score, accuracy_score, roc_curve, matplotlib pyplot (Plot ROC curve).

Experiment 1 accuracy	0.925249891
Experiment 1 precision	0.912457912
Experiment 1 recall	0.896361632



Experiment 2:
– Plot of accuracy (on test data) vs. m (number of features)
– Discussion of what the top 5 features were
– Discussion of the effects of feature selection (about 1 paragraph).
Answer: The top 5 features are: 52, 55, 6, 15, 22 (char_freq_#, char_freq_!, word_freq_your, word_freq_addresses, word_freq_over).

Since we select a subset of m features, we can maximize classification performance and increase accuracy with fewer features.

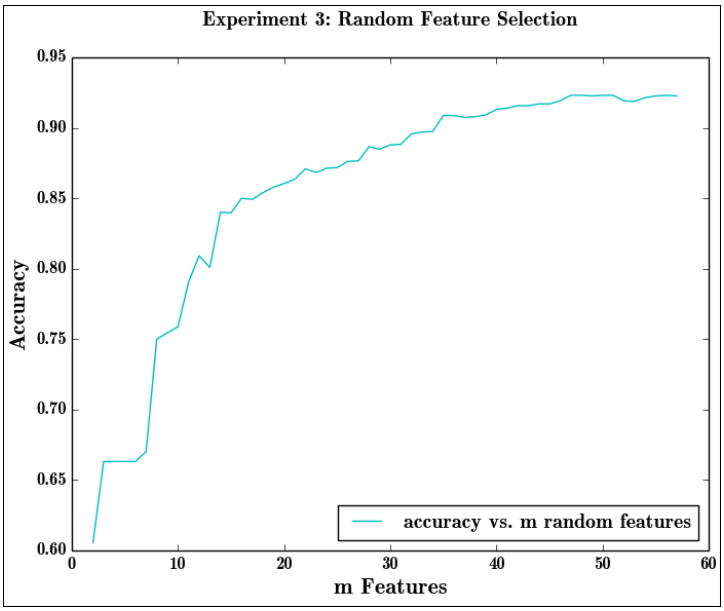


Experiment 3:

– Plot of accuracy (on test data) vs. m (number of features)

– Discussion of results of random feature selection vs. SVM-weighted feature selection (about 1 paragraph).

Answer: At the beginning of the model, results of random feature selection has lower accuracy compares to SVM-weighted feature selection. But in the end, both selections have similar accuracy that is close to 1.


Experiment 3: Random Feature Selection
Accuracy
m Features
accuracy vs. m random features
0.95
0.90
0.85
0.80
0.75
0.70
0.65
0.60
0
10
20
30
40
50
60