Andrew McCann

2/17/16

CS445

Homework #3

Melanie Mitchel

SVM Write-up

Experiment 1: I used scikitlearn's built in SVM package SVC.

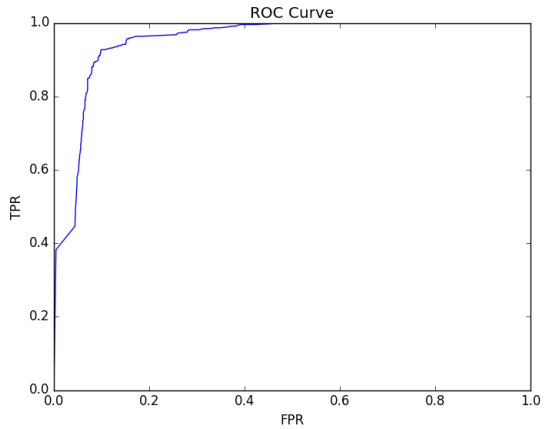
Best C: For my best C value I never got a consistent value as I anticipated. I really think this is a result of a bug within my code because I encountered other strange behavior that I will note later. The value was typically higher than .6. But I don't think I ever saw an instance where 1 was the best. Tended to fall on **.7**

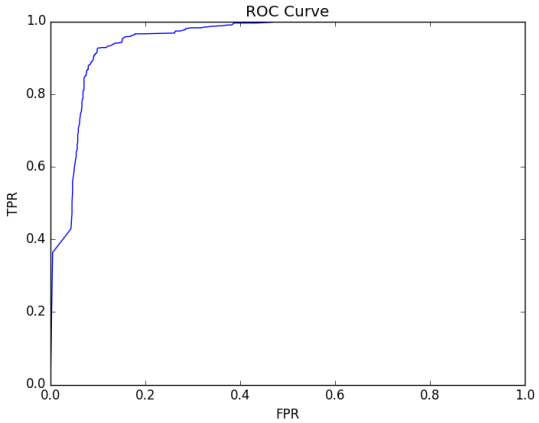
Accuracy: 90.6%

Precision: 90.5%

Recall: 90.7%

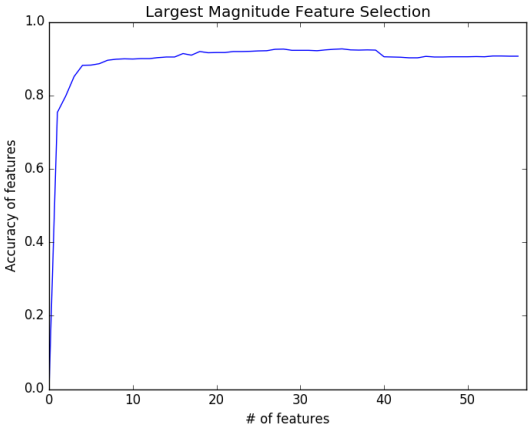
When discussing my results with some classmates it sounds like I'm missing about 3 points on all those values. I couldn't find any source of that other than the way I split my data into k-segments with Numpy's split method. It doesn't drop an exact number of positive and negative instances, but I did not think that would yield a -3point difference consistently.

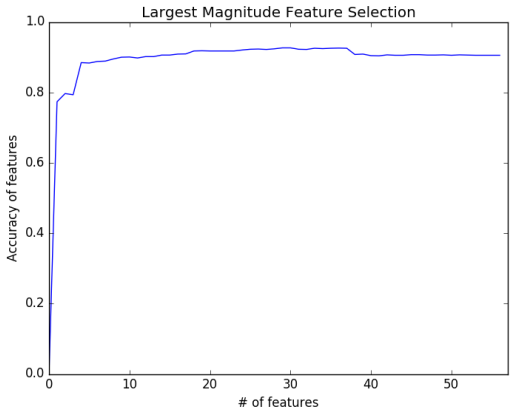




These graphs look almost identical, but upon closer inspection you'll see they are actually different. Virtually every time I ran the experiment it came out like this. It doesn't feel right because of that sharp jump around .4.

Experiment 2:



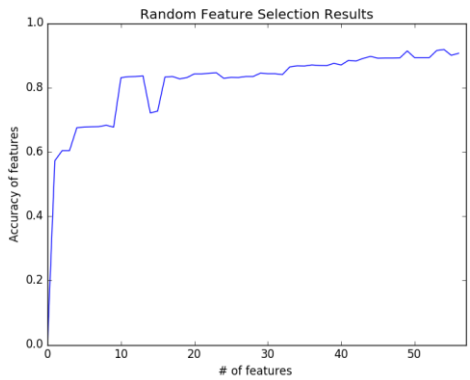


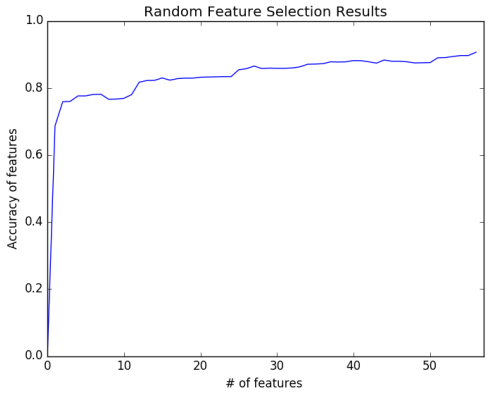
Fairly predictable results, though most of the time I get this more jagged version on the right which doesn't make too much sense. I blame whatever bug is causing the weird behavior in my ROC since that trains everything.

Top 5:[26, 24, 41, 51, 25]: These were consistently the top 5 indicators, with 6 sometimes sneaking in there for a cameo. As long as I can count properly 26 lands on 'george' which seems to make sense since that would likely be spam for anyone outside of the population that is named George. 24 appears to be 'hp'. Seems like that character string wouldn't appear frequently in English, so it might set off some flags. 41 looks to be 'meeting' which would seem odd being repeated in an email. '!' is 51 which makes sense. At least in my own experience the exclamation isn't used all that frequently in day to day email communication. And 25 is 'hpl' another strange string of characters that wouldn't regularly appear in communication.

Feature Selection: It seems like you can create a reliable classifier with few features. In fact in my case (might not be correct because of aforementioned bugs) accuracy actually dropped lower as I included the lower values of features, so there is an obvious benefit (in my case) to only select certain features from the full list.

Experiment 3:





I don't have a great explanation for this bizarre fluctuation in accuracy. There must be some features that just crush my accuracy due to some bug, and when those are included they drag the whole classifier down.

Random Feature Selection vs SVM Weighted: Definitely saw some weirder results with the random selection. It ended up in about the same place which was interesting. But the road there was just steeper, and sometimes weirdly jagged as those poor features are added into the pool. It is interesting to see how some features just have very little impact on the classification at all. This seems consistent with the feature set. Common words within a normal frequency wouldn't be obvious indicators of if the email was spam or not.