Andrew McCann

2/21/16

CS445

Machine Learning

Homework 4

Gaussian Naïve Bayes

To accomplish this experiment I used Python and NumPy, in conjunction with SciKitLearn (for pulling Metrics easily). It was very simple to implement the Naïve Bayes method in python. I did encounter bizarre divide by zero errors in the process. As mentioned in class it seems to be due to underflow with the tiny values involved in the exp() process. I tried to code in solutions but they seem to diminish my accuracy when implemented and I am unsure why.

I do not think that the values are completely independent in this data set. Especially in the not-spam section. There is a format that English follows so it is going to skew the results as such. Anything that ends up having bizarre outlier counts on types of words would seemingly be classified as spam. I don't think classifier does particularly well on this dataset precisely because the spam filter is learning from a word frequency, which will depend on the structure of the language. From what I understand my peers are trending towards 80% accuracy with this classifier which doesn't seem that great. In my experience it is closer to 75% which isn't ideal, but seems close enough to account for variations in how we identified the test set.

My output straight from the console:

```
Accuracy:  75.1412429379
Recall:  41.3450937155
Precision:  90.3614457831
```

```
[[ 375  532]
 [  40 1354]]
```

In table format:

375	532
40	1354