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 generally dependent on the previous feature to determine a type of feature following it. I don’t think classifier does particularly well on this dataset precisely because the spam filter is learning from a pattern that depends on surrounding types of words. 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 70%. I’m going to try to speak with someone about raising that accuracy before this project gets handed in. Hopefully I remember to update this if I do go that route.

I had a couple peers look at my code to see if there was an obvious reason my recall is so out low. Also likely the reason my accuracy is lower than my peers as well. No one has been able to find anything obvious as of yet,

My output straight from the console:

Accuracy: 73.0551933942

Recall: 35.7221609702

Precision: 89.7506925208

[[ 324 583]

[ 37 1357]]

In table format:

|  |  |
| --- | --- |
| 324 | 583 |
| 37 | 1357 |