

Programming assignment 2 report

Programming assignment 2 is a Naïve Bayes classifier model intended to classify emails as spam or not spam. First, I split the data in half randomly to get a training set and a test set. For each set I then split them into spam and not spam sets. This was to help me use more concise code when creating the confusion matrix as shown in figure 1 below. I also used the split to determine the means and standard deviations of the training spam and not spam emails for each feature. Using these attributes, I created a function that computes the probability of the data if it were spam and another that computes the probability of the data if it were not spam. To classify, I chose whichever class gave higher probability with these functions. Of the 2300 testing examples, 904 were spam and 1396 were not. Of the 1396 non-spam, 861 were classified correctly as non-spam and 535 were classified as spam. Of the 904 spam emails, 859 were correctly classified as spam and 45 were classified as spam. This means that the overall accuracy was $\frac{861+859}{2300} = 74.78\%$. Using not-spam as a positive classification, the precision was $\frac{861}{861+535} = 61.68\%$. The recall was $\frac{861}{861+45} = 95.03\%$.

Figure 1: Confusion matrix with the true class on the left and the class as determined by the algorithm across the top

Class	Not spam	Spam
Not spam	861	535
Spam	45	859

I think the attributes are not independent as assumed by the Naïve Bayes classifier. The classifier does a good job at identifying true spam, but misclassifies many non-spams as spam. I think this is because some of the features used are common across spam and are somewhat related, but may also be present in non-spam. This would result in a model good at classifying spam but not good in classifying non-spam.