SPAM AND HAM CLASSIFIER

Here, we are supposed to build a Bayesian classifier which correctly identifies spam and ham emails. In order to build the Bayesian classifier, we first extracted all the data from the training set and divide it into two separate classes for spam and ham. Then for each email present in either spam or ham we then count the instances of each word and store it in a dictionary. Similarly, for the Ham emails we store the words present in the email.

Now to classify whether a given email is spam or ham we calculate the conditional probability of each word present in the email for both the classes and multiplied them to assign a particular weight to the given email for each class. Whichever class has higher weight will give us the label of that email.

While considering the probability of each word present in the email since the values are actually small we take the logarithm of the values so as to avoid underflow. We have applied Laplacian smoothing to help solve the problem of unknown words being present in the test set email.

The probability that a given email is Spam or Ham is given below

$$Pr(Spam) = \frac{Number\ of\ Spam\ Emails}{Total\ Number\ of\ emails}$$

$$Pr(Ham) = \frac{Number\ of\ Ham\ Emails}{Total\ Number\ of\ emails}$$

The probability that a word is present in the Spam email is given by

$$Pr(W \mid Spam) = \frac{Number\ of\ times\ word\ W\ occurs\ \ in\ Spam\ emails}{Total\ Number\ of\ words\ present\ in\ all\ the\ spam\ emails}$$

However, this probability turns out to be really small and cannot handle unknown words present in the new emails. We make use of Laplacian smoothing which gives us the value of

$$Pr(W \mid Spam) = \frac{Number\ of\ times\ word\ W\ is\ present\ in\ Spam\ emails\ +\ alpha}{Total\ Number\ of\ words\ present\ in\ al\ the\ spam\ emails\ + (Distinct\ words\ present * \ alpha)}$$

Here, alpha is the Laplacian smoothing parameter and we have considered it as 10. We have tried alpha values as 0.01, 0.1, 1 ,10, 100 in order to identify improvements in the predications if any. Here are the values that we got.

Alpha	**Accuracy**	**F1 Score**
0.01	90.2%	0.874680
0.1	90.2%	0.874680
1	90.2%	0.874680
10	90.4%	0.876606
100	90.4%	0.874015

We can see that the F1 Score increases from alpha=1 to alpha=10 and then drops again from alpha=10 to alpha=100. Hence, we can consider alpha = 10 as our optimal Laplacian Smoothing parameter.

Now according to Bayes theorem, we have

$$Pr(Spam|W) = \frac{Pr(W|Spam) * Pr(Spam)}{Pr(W)}$$

By applying the chain rule, we then obtain the cumulative probability whether a given email is spam or ham.

Thus, by using the above equation we can calculate whether a given a email is spam based on whether the W is present in it or not.

The confusion matrix generated for the given test data file.

Actual **Predicted**	**Ham**	**Spam**
Ham	341	79
Spam	17	563

Accuracy obtained on the given test dataset is 90.4% with an F1 score of 0.876606.