Assignment 3
Spam - Ham Classification

Bikash Kumar Behera CS19M019
Sanket Neema CS19M055

Introduction:

We build a spam classifier using Naive Bayes algorithm. For these we trains our model using training data set of more than 4000 mails includes both Spam and Ham mails.

A. Dataset:

a. Training Dataset:
We train our model using 3000 Ham-Mails and 1300 Spam-Mail.

b. Testing Dataset:
We test our model on 600 Ham-Mails and on 500 Spam-Mails.

Source: http://nlp.cs.aueb.gr/software_and_datasets/Enron-Spam/index.html -> Enron1

B. Features:

Every different word in the training data is considered and added to the feature set.

C. Procedure:

Step 1 : Data Preprocessing:

- ➔ Tokenization — convert sentences to words
- ➔ Removing unnecessary punctuation, tags
- ➔ Removing stop words — frequent words such as ”the”, ”is”, etc. that do not have specific semantic
- ➔ Stemming — words are reduced to a root by removing inflection through dropping unnecessary characters, usually a suffix.
- ➔ Lemmatization — Another approach to remove inflection by determining the part of speech and utilizing detailed database of the language.

Step 2 : Training using Naive Bayes classifier, with Laplace smoothing:

➔ First we create a dictionary with all the different words present in all the mail (spam or non spam mails). Suppose cardinality of these dictionary is d.

➔ Now for each word in the dictionary (let j^{th} word) compute following parameters.

- $\phi_{j|y=1} = p(x_j = 1|y = 1)$.

  Probability that this word appears in spam mail.

$$\phi_{j|y=1} = \frac{\sum_{i=1}^{m} 1\{x_j^{(i)} = 1 \wedge y^{(i)} = 1\} + 1}{\sum_{i=1}^{m} 1\{y^{(i)} = 1\} + 2}$$

- $\phi_{j|y=0} = p(x_j = 1|y = 0)$.

  Probability that this word appears in non-spam mail.

$$\phi_{j|y=0} = \frac{\sum_{i=1}^{m} 1\{x_j^{(i)} = 1 \wedge y^{(i)} = 0\} + 1}{\sum_{i=1}^{m} 1\{y^{(i)} = 0\} + 2}$$

➔ 3^{rd} parameter is probability of spam mail.

$$\phi_j = \frac{\sum_{i=1}^{m} 1\{z^{(i)} = j\} + 1}{m + k}.$$

Step 3 : Testing/Validating:

Now having all these parameters, to predict weather a mail is spam or not spam, with each different word being feature of these test mail (let x), we will calculate the following probabilities.

1. Posterior Probability that mail is spam, given the features of mails.(Applying Total prob. theorem).

$$p(y = 1|x) = \frac{p(x|y = 1)p(y = 1)}{p(x)}$$

Which further simplifies to:

$$= \frac{\left(\prod_{j=1}^{n} p(x_j|y = 1)\right) p(y = 1)}{\left(\prod_{j=1}^{n} p(x_j|y = 1)\right) p(y = 1) + \left(\prod_{j=1}^{n} p(x_j|y = 0)\right) p(y = 0)},$$

2. Posterior Probability that mail is non-spam, given the features of mails.(Applying Total prob. theorem).

$$P\ (y = 0/x)\ = 1 - P\ (y = 1\ /\ x)$$

Whichever class has the higher posterior probability for given test email, we predict email belong to that class.

D. Results/Accuracy :

The developed model on running on Test Data generates the following confusion matrix:

N = 2900	Actual Spam	Actual Ham
Predicted Spam	1330	156
Predicted Ham	120	1294

Overall Accuracy : 90.48%
Recall : 91.7%
Precision: 89.5 %
F1 Score: 90.58