Part1 – Report

Assignment 4

By, Anup Bharadwaj, Raghuveer Kanchibail, Supreeth Suryaprakash

Naïve – Bayes:

Frequency:

We calculate Probability of each Word being a spam * Probability of Spams. To calculate the probability of each word being a spam we took into consideration Count of each word in Spam Mails divided it by total number of words in Spam Mails. For Probability of Spams, we just took total number of Spam Emails divided total number of Emails.

Binary:

Whereas in Binary we formed a vector that will be either one or zero based on the occurrence of the word in Spam Emails. Remaining was same as above.

Once the training set is run, the model serializes a modelFile using Pickle. Then, when test set is run it deserialies the modelFile and predicts whether it is spam or not. To calculate accuracy we make use of Confusion Matrix.

Below are the results,

The total accuracy taking into account word frequency is 94.5554249902%

Top ten words associated with spam:

1 . td
2 . received
3 . tr
4 . esmtp
5 . jul
6 . p
7 . email
8 . sep
9 . localhost
10 . aug

Top ten words least associated with not spam:

1 . received

2 . sep
3 . esmtp
4 . td
5 . localhost
6 . oct
7 . aug
8 . postfix
9 . tr
10 . ist

The total accuracy for Binary is 94.124559342%

Top ten words associated with spam(Binary):

1 . td
2 . received
3 . tr
4 . esmtp
5 . jul
6 . p
7 . email
8 . sep
9 . aug
10 . helvetica

Top ten words least associated with not spam(Binary):

1 . received
2 . sep
3 . esmtp
4 . td
5 . localhost
6 . oct
7 . aug
8 . postfix
9 . tr
10 . ist

In our case, Naïve Bayes with Frequency worked fine. Well the, accuracy gap wasn't too large.

Decision-Tree

This classifier is a classic predictive modeling technique which may be visually represented as a n-ary tree with a class tagged to every value of the corresponding feature. Each data point from the test dataset is then classified by traversing through the tree nodes. Traverses to right if the condition in the node is true and traverses to left if the condition in the node is false. Information gain, which uses the entropy to assess the importance of the features themselves in classification. The formula for calculating the information gain is as below:

$$Gain(S, A) \equiv Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

where, S is the main set and A is the attribute under consideration.

$$Entropy(S) \equiv \sum_{i=1}^{c} -p_i \log_2 p_i$$

where p is the probability of the class c.

This has been done for both frequency of words and binary. There wasn't too much difference in the accuracy values. We got 97.21 % accuracy for Tree with frequency and 98.23 % accuracy for binary.

When you run the program, you will see the Tree formation for both.

DTree – for Frequency:

```
('received', 13440)
Right  ('returnpath', 2546)
  Right  ('dogmaslashnullorg', 2686)
    Right  ('ist', 3344)
      Right  ('deliveredto', 2701)
        Right  ('received', 13440)
          Right  ('imap', 1814)
            Right  ('precedence', 1115)
              Right  ('errorsto', 948)
                Right  ('bulk', 1202)
                  Right  ('xmailmanversion', 927)
                    Right  ('received', 13440)
                      Right  ('smtp', 1916)
                        Right  ('received', 13440)
                          Right  ('email', 3085)
                            Right  ('contenttype', 2841)
                              Right  ('textplain', 1683)
```

```
                    Right  ('mimeversion', 1892)
                      Right  ('listarchive', 853)
                        Right  ('mailing', 1025)
                          Right  ('list', 2184)
                            Right  ('xoriginaldate', 329)
                              Right  ('esmtp', 7369)
                                Right  ('dogmaslashnullorg', 2686)
                                  Right  ('machine', 110)
                                    Right  {'nonspam': 1}
None
                                    Left  {'spam': 5}
None
None
                                  Left  ('potential', 109)
                                    Right  {'spam': 1}
None
                                    Left  {'nonspam': 30}
None
None
None
                                Left  {'nonspam': 10}
None
None
                              Left  {'nonspam': 9}
None
None
                            Left  {'spam': 1}
None
None
                          Left  ('inline', 146)
                            Right  ('contentdisposition', 157)
                              Right  ('habeas', 131)
                                Right  {'nonspam': 1}
None
                                Left  {'spam': 3}
None
None
                              Left  {'nonspam': 1}
None
None
                            Left  {'spam': 13}
None
None
None
                        Left  ('zzzzlocalhostnetnoteinccom', 112)
                          Right  {'nonspam': 1}
None
                          Left  {'spam': 8}
None
None
None
                      Left  {'spam': 2}
None
None
                    Left  {'spam': 12}
None
None
                  Left  {'nonspam': 11}
```

Result truncated due to it's sheer volume

DTree – for Binary:

```
('received', 13440)
Right  ('returnpath', 2546)
  Right  ('dogmaslashnullorg', 2686)
    Right  ('ist', 3344)
      Right  ('deliveredto', 2701)
        Right  ('zzzzlocalhost', 1264)
          Right  ('precedence', 1115)
            Right  ('errorsto', 948)
              Right  ('xmailmanversion', 927)
                Right  ('smtp', 1916)
                  Right  ('zzzzasonorg', 252)
                    Right  {'spam': 40}
None
                  Left  ('phobos', 475)
                    Right  ('zzzzlocalhostnetnoteinccom', 112)
                      Right  {'nonspam': 34}
None
                      Left  {'spam': 10}
None
None
                    Left  {'nonspam': 18}
None
None
None
                Left  ('claimed', 122)
                  Right  ('httpwwwlinuxiemailmanlistinfoilug', 129)
                    Right  ('please', 1314)
                      Right  {'spam': 8}
None
                      Left  {'nonspam': 21}
None
None
                    Left  {'spam': 1}
None
None
                Left  {'nonspam': 47}
None
None
None
              Left  {'spam': 1}
None
None
            Left  {'nonspam': 52}
None
None
          Left  ('url', 451)
            Right  ('textplain', 1683)
              Right  ('questions', 214)
                Right  {'spam': 5}
None
                Left  {'nonspam': 8}
None
```

```
None
            Left  {'spam': 2}
None
None
          Left  ('zzzzlocalhostnetnoteinccom', 112)
            Right  {'nonspam': 2}
None
            Left  {'spam': 275}
None
None
None
None
        Left  ('phoboslabsnetnoteinccom', 562)
          Right  ('phobos', 475)
            Right  ('yyyylocalhostnetnoteinccom', 406)
              Right  ('xpyzor', 181)
                Right  {'nonspam': 177}
None
                Left  {'spam': 75}
None
None
              Left  {'nonspam': 11}
None
None
            Left  {'spam': 91}
None
None
          Left  ('jul', 2736)
            Right  ('xspamlevel', 1112)
              Right  {'nonspam': 19}
None
              Left  {'spam': 11}
None
None
            Left  {'nonspam': 880}
None
None
None
None
       Left  {'spam': 1}
None
None
     Left  ('jul', 2736)
       Right  ('international', 136)
         Right  ('reserved', 121)
           Right  ('paper', 110)
             Right  {'spam': 3}
None
            Left  {'nonspam': 4}
None
None
          Left  {'spam': 1}
None
None
        Left  {'nonspam': 61}
None
None
       Left  ('deliverydate', 357)
         Right  ('jun', 741)
```

```
Right  ('due', 120)
  Right  ('registered', 120)
    Right  ('direct', 109)
      Right  {'spam': 1}
```

References:

1. Prof. Crandall's Lecture slides
2. https://netmatze.wordpress.com
3. http://www.patricklamle.com/