Assignment 2

Alanda Cherestal

October 2022

Contents

1 Introduction

In this assignment, I implemented a probabilistic classifier, Naive Bayes to create an email spam filter. The idea for this filter is to take most command words that can be an identification for weather an email is spam or ham. It pretty much turns into a game of trial and error. This assignment is written in python on Google Colab.

2 Data Processing

Every dataset needs to be split into train and test before machine learning can be implemented. Using Numpy, I split the data set into 80% for training and 20% for testing. I took the training set and separated the spam and ham emails for later use. I looked through both the spam and ham to look for words and phrases that can be used for the word dictionary. After finding few phrases, I entered them into a word search function to find the count between in spam and ham. The results are added into a process words dictionary.

2.1 Word Filtering

The word filter takes each email and check if words from the dictionary is present using for-loops with if-statements. The words that are present gets put into an observed words list to eliminate the other words from the email. The observed words enters the SpamHam function I created to determine if the email might be spam or ham.

3 Evaluation

I put the predicted results in a predicted label array to be compared to the true labels from the dataset. I import metrics from sklearn to create a confusion matrix with the results. In total, I've gotten 40 emails correct out of 50 emails from the test split. That results an 80% accuracy.

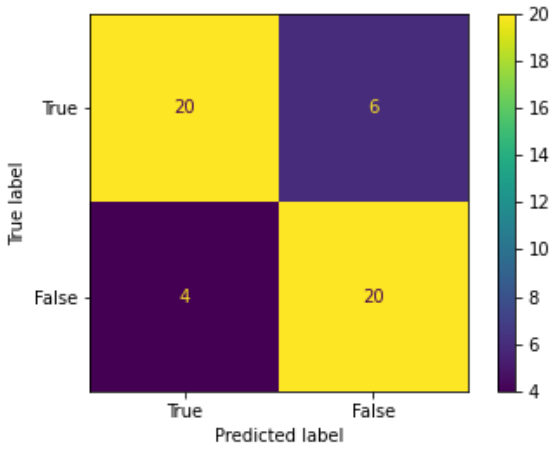

Figure 1: this Figure shows the confusion matrix from code.

3.1 Confusion Matrix

	PP	**PN**
P	**0.40**	0.12
N	0.08	0.40

4 Discussion

In Figure 1, we show the generated confusion matrix from the code. The test split has 50 emails. 40% percent True Positive and 40% True Negative leaving 8% False Positive and 12% False Negative. I can say that the filter can be update to add more phrases and words to decrease the False Negative prediction. The accuracy percentage would also increase.

5 What I've learned

I've learned at from this assignment. I learned that python has a lot of packages you can import. At first, I created the confusion matrix from scratch without knowing there was a package that can implement it with a few arguments. That left me curious to what other packages python has that could have helped in past projects. I also learned that truly understanding the step you need to take for a project takes makes the coding so much easier. The past, I would process everything as I code and simply relied on testers, I wouldn't take my time to understand the big picture in the beginning. I think doing this assignment gave me a better understanding on computer science. The most important thing is that I truly enjoyed this assignment.