Assignment 3 Report

CS5691

PRML

K Akranth Reddy
ME20B100

30 April,2023

Question 1

Note: To run the model in on a folder in test data, add the test folder to location where Jupyter notebook is there and run the last parts of jupyter notebook.

The data set i have used is from Kaggle. I have analysed three algorithms
1. Naive Bayes
2. KNN
3. SVM

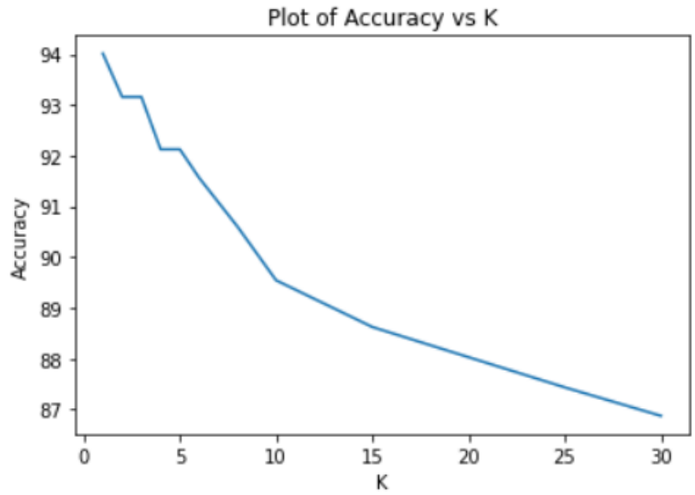
Preprocessing data
The original data set had 5572 emails. Number of spam emails are 747.
The data set is split into two parts training and testing, in 80:20 ratio.
Training data has 4457 samples, Testing data has 1115 samples.
First I have added all the text in emails to NumPy matrix that stores strings, next removed punctuation's and numbers in the data, split them into words. Next I ran a function is a counts the frequency of occurrence of each words. we can see upon running that the top words are common like a,the,i,you which aren't really contribute to spam or ham, so removed these words and included the words that occurred more than 15 time in the whole data set, in test data set there are around 10,000 unique words, I have compressed it into 754 words. A binary 1D array can now represent emails.

Algorithm 1: Naive Bayes
The parameters of the model are calculated from maximum likelihood estimates. These parameters are used to calculate the $\frac{P(Y/X_i^0)}{P(Y/X_i^i)}$, data point is predicted class 0 if the ratio is greater than 1 otherwise class 1 is predicted. $\frac{P(Y/X_i^0)}{P(Y/X_i^i)} = \frac{\hat{p}\prod(\hat{p}_i^o)^{f_i}.(1-\hat{p}_i^o)^{1-f_i}}{(1-\hat{p})\prod(\hat{p}_i^1)^{f_i}.(1-\hat{p}_i^1)^{1-f_i}}$

Errors on Training data.
number of correctly predicted spam 90.6%
number of incorrectly predicted spam 9.3%
Number of correctly predicted ham 99.1%
Number of incorrectly predicted ham 0.8%
Overall accuracy 98.0%

Errors on Test data.
number of correctly predicted spam 85.9%
number of incorrectly predicted spam 14.1%
Number of correctly predicted ham 98.8%
Number of incorrectly predicted ham 1.2%
Overall accuracy 97.1%
Comment on Naive Bayes predictions:
The test data accuracy of the model on predicting ham data correctly is 98.8 which is very good, for a spam data point the accuracy of correctly predicting spam is 85.9 this is reasonable.

Algorithm 2: K-Nearest neighbours



The above plot describes how accuracy of model changes with hyper parameter value k. The model best performs for k=2.
I have considered Manhattan distance to measure the distance between two points.
Manhattan distance between two points is $d(I_1, I_2) = |I_1 - I_2|$
The algorithm is, I have measure the Manhattan distance between all the points and the testing point,For the K nearest points, the majority label value is predicted as class label.
number of correctly predicted spam 42.9%
number of incorrectly predicted spam 57.0%
Number of correctly predicted ham 99.4%
Number of incorrectly predicted ham 0.6%
Overall accuracy 91.8%
Comments on KNN prediction:
By observing the above results we can see that the model does performs better for ham data set, it is worse than random binary classifier for predicting spam data predictions.

Algorithm 3: Support Vector Machines
I have used soft margin support vector machine algorithm from the Sk-learn library. Trained the model on test data.
Error on Training data.
number of correctly predicted spam 94.6%
number of incorrectly predicted spam 5.4%
Number of correctly predicted ham 100.0%
Number of incorrectly predicted ham 0.0%
Overall accuracy 99.3%

Error on Testing data.
number of correctly predicted spam 80.5%
number of incorrectly predicted spam 19.5%
Number of correctly predicted ham 99.9%
Number of incorrectly predicted ham 0.1%
Overall accuracy 97.3%

Comment on SVM predictions:

The test data accuracy of the model on predicting ham data correctly is 99.9 which is very good, for a spam data point the accuracy of correctly predicting spam is 80.5.

Conclusion

SVM and Naive Bayes model performs well for the data set.

KNN perform worse than random binary classifier for predicting a spam data.