Support Vector Machine

Stuti Chaturvedi
202161006

Haripriya Goswami
202161003

Yashita Vajpayee
202162012

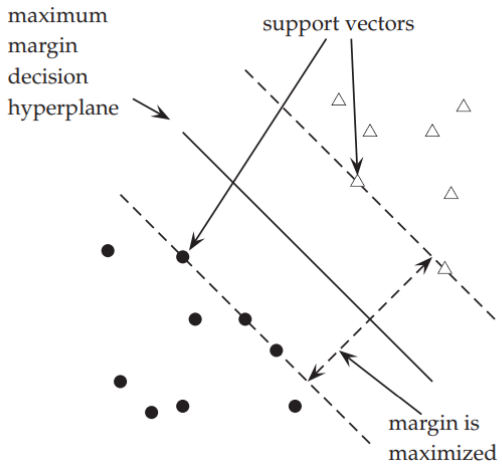
***Abstract*—In this experiment we aim to model SVM classifier for two class dataset and also evaluate the results. We will also use 20 newsgroup dataset and compare its SVM model evaluation results to other classifiers.**

I. Introduction

Support Vector Machines or SVM is one of the most widely used classifier in today's times because its accurate results and easy computation. SVM uses vector space based classification technique where, given a classified training set, SVM determines the class of the test data.

II. Theory

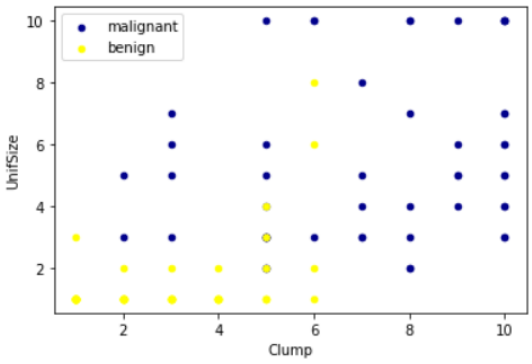
In SVM, a training set which is distributed in the vector space is divided into different classes. A *Margin* can be drawn in this set which divides the data set and has shortest distance from the nearest data points of each of the class. Support vectors are these nearest data points which lie in the boundary of the Margin. A visual representation can be seen below:



The classifier aims to maximize this margin so that the two classes for better classification.

III. Approach

Firstly, for the first dataset, we fetched the data set of some cell which are divided into classes— malign and benign. These cells are classified based on many features, here is the classification based on two features— UnitSize and Clump:



Another dataset which is used is the one we used for text classification of experiment— 20 news-group dataset Training sets are defined and then fitted to the SVM model using *svm* from *sklearn* library. The test data is then predicted against this model and through *metrics* from *sklearn* we evaluate the results by making a classification report and confusion matrix.

IV. Results

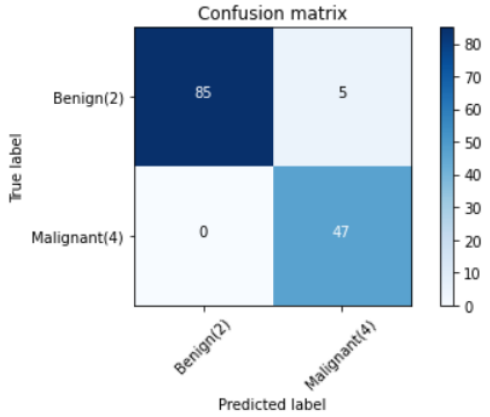
A. Test data I

In the first data set we get the following classification report:

	precision	recall	f1-score	support
2	1.00	0.94	0.97	90
4	0.90	1.00	0.95	47
accuracy			0.96	137
macro avg	0.95	0.97	0.96	137
weighted avg	0.97	0.96	0.96	137

Normalized confusion matrix and simple confusion matrix:

```
Confusion matrix, without normalization
[[85  5]
 [ 0 47]]
```



B. Test data II

From the dataset of 20 news-groups we can see the evaluation as:

```
Newsgroup Categories :  ['sci.med', 'sci.space', 'sci.electronics']
Accuracy : 93.99830938292477%
                precision    recall  f1-score   support

sci.electronics       0.89      0.97      0.93       393
        sci.med       0.95      0.91      0.93       396
      sci.space       0.98      0.94      0.96       394

       accuracy                           0.94      1183
      macro avg       0.94      0.94      0.94      1183
   weighted avg       0.94      0.94      0.94      1183

Confusion Matrix :
 [[381   9   3]
 [ 32 360   4]
 [ 13  10 371]]
```

V. CONCLUSION

Through this experiment, we learnt constructing an SVM model for two class data set and evaluate it. Also, after observing the f-score of the 20 news-group data set, we can say that SVM provides a better classification than other classifiers.

REFERENCES

[1] Introduction to information retrieval by Christopher D Manning Prabhakar Raghavan Hinrich Schutze