Applying Suppor Vector Machine on Breast Cancer Wisconsin dataset

Shashank Pathak(15BCE1287)*,
*School of Computing Sciences and Engineering, VIT Chennai, Tamilnadu, India 600127
Email: shashank.pathak2015@vit.ac.in*

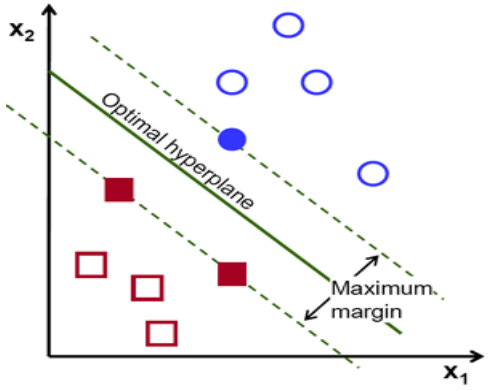
***Abstract*—In machine learning, an SVM is a classifier based on the idea that the best seperating hypeplane to seperate the datapoints belonging to different classes is the hyperplane with the maximum margin and least trainng error.We study the performance of the SVM for the classification of the Breast Cancer Wisconsin Dataset.The system developed performs with 94.63% accuracy for the test sample in a two class classification**

1. Introduction

Support vector machines (SVMs) are a set of related supervised learning methods that analyze data and recognize patterns, used for classification (machine learning)and regression analysis.he standard SVM is a non-probabilistic binary classifier—binary linear classifier, i.e. it predicts, for each given input, which of two possible classes the input is a member of. Since an SVM is a classifier, then given a set of training examples, each marked as belonging to one of two categories, an SVM training algorithm builds a model that predicts whether a new example falls into one category or the other. Intuitively, an SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible

2. Methodology

To understand the basic idea of SVM we uderstand the process of seprating the red and blue datpoint in a dataset.We can separate the red and blue objects with an infinite number of hyperplanes. Which hyperplane is the best? Well, the best hyperplane is the one that maximizes the margin. The margin is the distance between the hyperplane and a few close points. These close points are the support vectors because they control the hyperplane. The graph below illustrates the best hyperplane for the red and blue objects.



This is the Maximum Margin Classifier. It maximizes the margin of the hyperplane. This is the best hyperplane because it reduces the generalization error the most. If we add new data, the Maximum Margin Classifier is the best hyperplane to correctly classify the new data. The Maximum Margin Classifier is our first SVM. But this SVM requires the two classes to be completely linearly separated. In case of Non linear classifier we use kernel trick to go into higher dimensions and then make the data linearly classifiable.

3. Database - Breast Cancer Wisconsin Dataset(Original)

The Wisconsin Breast Cancer datasets from the UCI Machine Learning Repository is used,to distinguish malignant (cancerous) from benign (non-cancerous) samples.This dataset consist of 699 instances and 11 attributes that would help the classiy the data into the two classes.The attributes are described in the fig.1

#	Attribute	Domain
1.	Sample code number	id number
2.	Clump Thickness	1 - 10
3.	Uniformity of Cell Size	1 - 10
4.	Uniformity of Cell Shape	1 - 10
5.	Marginal Adhesion	1 - 10
6.	Single Epithelial Cell Size	1 - 10
7.	Bare Nuclei	1 - 10
8.	Bland Chromatin	1 - 10
9.	Normal Nucleoli	1 - 10
10.	Mitoses	1 - 10
11.	Class:	(2 for benign, 4 for malignant)

Figure 1. Breast Cancer Wisconsin Dataset Attributes

4. Algorithm

For the data which can be separated linearly, we select two parallel hyperplanes that separate the two classes of data, so that distance between both the lines is maximum. The region b/w these two hyperplanes is known as margin maximum margin hyperplane is the one that lies in the middle of them.This is Hard Margin SVM.We now introduce a cost of error(C) into the equation and thus it helps to reduce overfitting.After this the Lagrange euation becomes :

$$\min P(\boldsymbol{w},\, b) = \underbrace{\frac{1}{2}\,\|\boldsymbol{w}\|^2}_{\text{maximize margin}} + \underbrace{C\sum_i H_1[\,y_i\, f(\boldsymbol{x}_i)\,]}_{\text{minimize training error}}$$

where $\mathbf{w}$ denotes the vector of weights, $\mathbf{x}$ is the vector of inputs, C is the cost of error making

.

5. Experiments

The training data of this exercise is formed by a set of Breast Cancer Wisconsin Dataset and each instance belong to one of two different classes; one of the classes is Malignant and the other is Benign.We will use SVM classifier from scikit learn to classify the data int these classes.Now we will set up SVMs parameters.We have to definr the kernel, cost of error(C) and give the training data.For non linear classication we can use different kernels such as Gausian radial basis function('rbf').In such case we have to give a parameter known as gamma to the SVM classifer.Gamma is the RBF kernel.After this we will fit our data in the SVM classifier.Now we will use the test data to test the model's accuracy using score method.

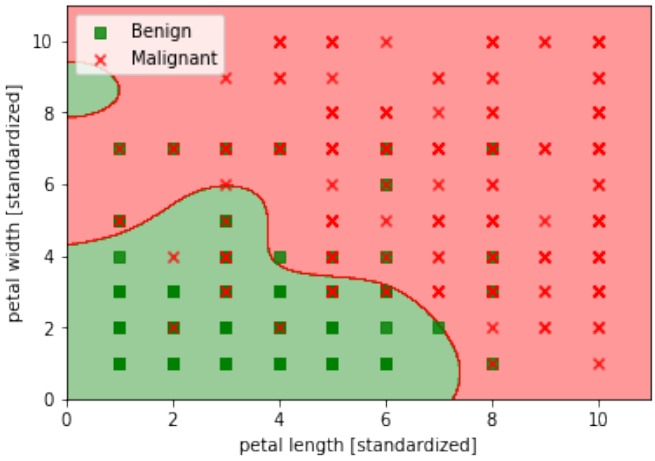


Figure 2.

6. Conclusion

After the classification we can see that the classification was very effective,.This indeed suggest that the SVM model is ver apt to classifify the data into two class Benign and Malignant for the breast cancer dataset The system developed performs with 94.63% accuracy for the test sample in a two class classification.Also we used Pipeling to check on different parameters value.We got a best result of 94.6% for liner kernel and value of C = 10.0