Support Vector Machine on Student Performance Data

Shubham Barudwale
School of Computing Sciences and Engineering, VIT Chennai, Tamilnadu, India 600127
Email: barudwaleshubham.dinesh2015@vit.ac.in

Abstract-

Support Vector Machine (SVM) is the model of classification of data using superwised learning approch. The SVM can be applied in both linearly seperable and linearly n seperable data using the maximum margin classificaion line. In higher dimensions it is done by hyperplanes by changing the dimensions using transformations.

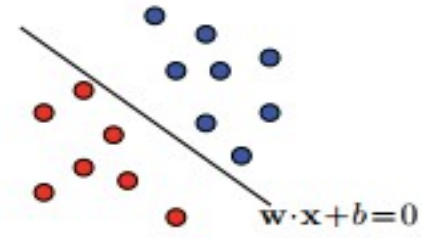
In this experiment the SVM is applied on the Student performance data to analyse and compare the classification with other methods. The data here gives the accuracy about 94% which is good accuracy. We can see the results of the various parameters in the given paper.

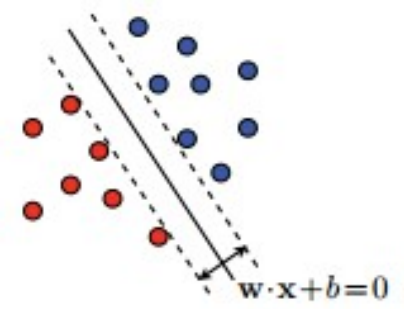
1. Introduction

Support Vector Machine (SVM) is the superwised learning algorithm. In perceptron we have seen that the differentiation line/ classifier line can be any line between the two groups. It can cause some errors in the test data by classifying the data incorrectly. That is generalization of the data can be not correct always.

To overcome this the SVM is the best for getting the clear and optimal classification line. Which is equidistant from both the classes and can classify the data optimally.

The given images shows that how perceptron works in classification and how SVM gives better results on the same data with maximum margin to accomodate the test data.
Image 1: Perceptron, image 2: SVM





2. Algorithm and Methodology

SVM can be used in the conditions where the training points are less than 1000 because the only poits matters to draw the SVM classifier line is the Support vectors. In SVM the main goal of the algorithm is to maximize the margin between the points and the classifier line.

Let line or hyperplane divide the data. We can consider the pair (w,b) as the canonical hyperplane. And the distance between the points and hyperplane is defined as

$$|w.X_0 + b| / ||w||$$

For canonical hyperplanes the margin is given by

$$\rho = \min (|w.X_0 + b| / ||w||) = 1/ ||w||$$

To create the larger margin we have to reduce the value of the ||w||. So one part of optimization is this part.

Other part of the optimization is Hinge loss function.

$$c(x,y,f(x)) = 1 - y * f(x)$$

where $c(x,y,f(x)) = 1$ if classified correctly
$= 0$ if classified incorrectly
x -> input
f(x) -> output
y -> expected output

to increase the accuracy of the SVM we need to increase the number of c.

Hence the combined function to optimize is:

$$[\lambda||w||^2 + \Sigma (1-y_i (w_i * x^T))]$$

after differentiation with respect to the weights we get:

$$2 \lambda ||w|| - y_i * x_i$$

After differentiating we can get two conditions.

Correctly classify : $\Delta w = 2 \|w\| \lambda * ή$
Misclassified: $\Delta w = (2 \|w\| \lambda - y_i x_i)* ή$

At the end the following relation takes place that the less is the generalization, more is the regularization.

When the data is not linearly seperable the the transformation takes place and the data is transferred to the higher dimensions and then it is classified. The method uses the various transform functions. Which is called the Kernal trick.

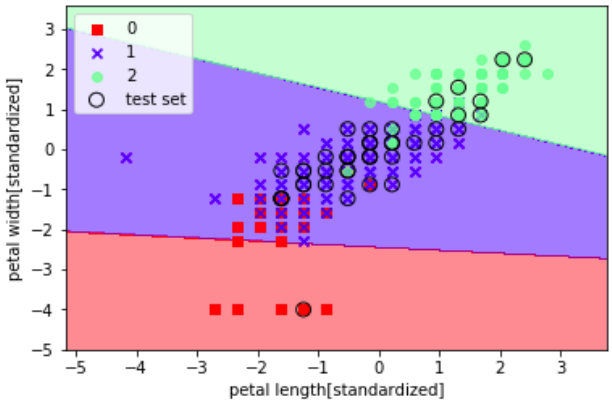
3. Data

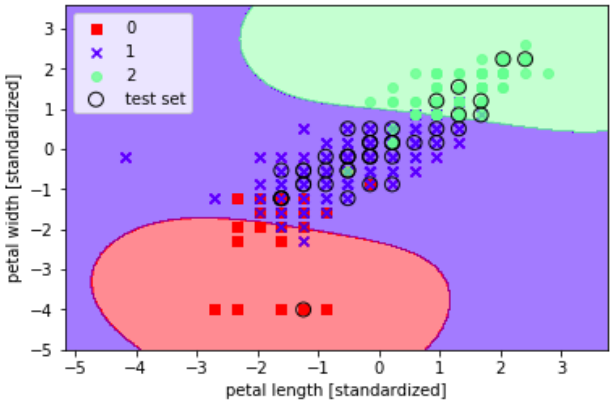
Student performance dataset is the dataset from which we are trying to gain some information from 650 student's academic performance data and various attributes which are affecting the performance. Attributes include sex, age, study time, free time, family information etc.

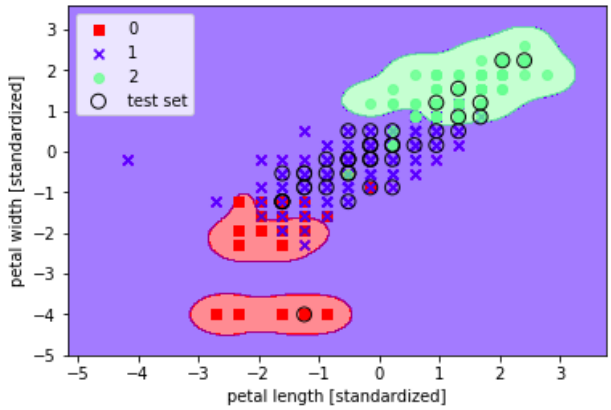
We will be using the Logistic regerssion method to learn and predict the performance. Out of 650 I have taken 455 data entries for lerning and rest for prediction i.e testing the performance of the algorithm.(70:30). The labels or target functions are mentioned in the 'new' field which is added manually. bad(0-7), medium(8-14), good(15-20) .

5. Experiments

We have used the algorithm for training the students data. And the accuracy on test data was very good as much as 93.66%. After 1000 eopches the cost is reduced to almost zero and then was constant as we can see from the graph.







6. Conclusion

Algorithm of this algorithm gave very good results. Accuracy was about 93.6% while using SVM Which is the better accuracy algorithm.

7. Referances

1) Python Machine Learning by Sebastian Raschka Chapter 12

2) SVM , Mehryar Mohri