SUPPORT VECTOR MACHINE EXPERIMENT REPORT

GANESHREDDY ANNAPAREDDY

50442295

ganeshre@buffalo.edu

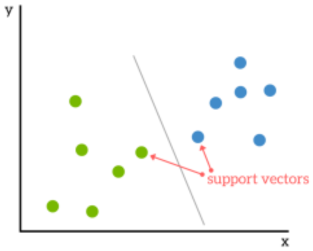
1 SUPPORT VECTOR MACHINE

A supervised machine learning model called a support vector machine (SVM) employs classification techniques to solve two-group classification problems. It can be employed for both classification and regression purposes. SVM model can classify new text after being given sets of labelled training data for each category. They offer two key benefits over more modern algorithms, such as neural networks: greater speed and improved performance with fewer samples (in the thousands). As a result, the approach is excellent for text classification issues, where it's typical to only have access to a dataset with a few thousand tags on each sample.

“””Finding the optimum hyperplane to split a dataset into two classes is the foundation of SVMs.”””

SUPPORT VECTORS :

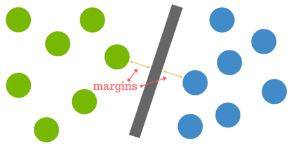
Support vectors are the data points in a data set that, if deleted, would change the location of the dividing hyperplane and are closest to the hyperplane. They may therefore be regarded as the crucial components of a data collection. From the figure on the right side we can see the support vectors on both the sides of a *hyperplane*.



HYPERPLANE :

For a simple classification experiment with only two characteristics (as seen in the figure above), consider a hyperplane to be a line that linearly divides and classifies a collection of data. Intuitively, the further our data points are from the hyperplane, the more certain we are that they have been accurately identified. As a result, we want our data points to be as far away from the hyperplane as feasible while still being on the proper side of it. When fresh testing data is added, the class assigned to it is determined by which side of the hyperplane it arrives on.

The best way to segregate the two classes in a data is done by the distance between the hyperplane and the nearest data point from either set is known as the margin (as we can see from the below image). The objective is to select a hyperplane that has the largest feasible margin between it and any point in the training set, increasing the likelihood that fresh data will be properly categorized.

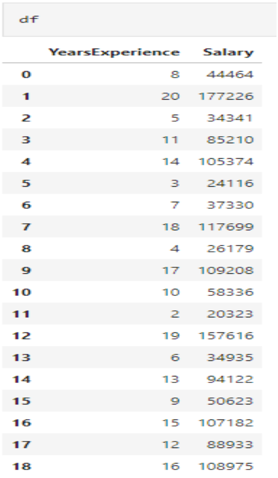


EXPERIMENT :

Dataset :

For the experiment of Support Vector Machine I have created a comma-separated values file with dataset including two rows and eighteen columns in which one of the row is 'YearExperience' and the other is 'Salary'. I have generated random value from 20,000 to 2,00,000 in column 'Salary', and put the 'YearExperience' in the ascending order to the column 'Salary'.

Column	Description
YearsExperience	The amount of years worked
Salary	Amount of salary got by the years of experience



```
df
```

	YearsExperience	Salary
0	8	44464
1	20	177226
2	5	34341
3	11	85210
4	14	105374
5	3	24116
6	7	37330
7	18	117699
8	4	26179
9	17	109208
10	10	58336
11	2	20323
12	19	157616
13	6	34935
14	13	94122
15	9	50623
16	15	107182
17	12	88933
18	16	108975

Training :

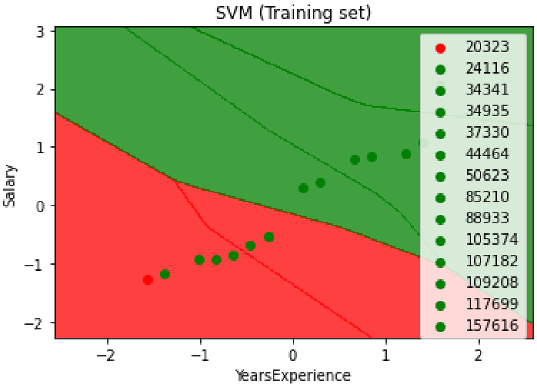
I used the dataset that I have created to train the support vector machine model and see the evaluation result. Python was used for this task on the Jupiter notebook. I have used NumPy, matplotlib, pandas and sk-learns libraries for the experiment.

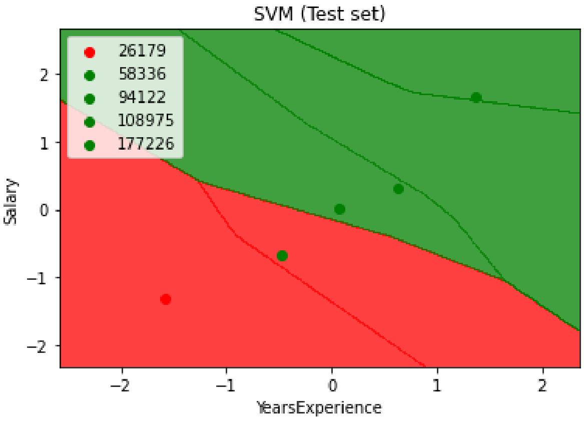
For importing the training model I have used sklearn.model_selection and then used scikit learn's for the standard scalar function where I am training and testing. I am importing the support vector machine model from sklearn.svm.

RESULTS :

Training set:

In the training set we have all the values being showed in which the result can be seen as the graph from the lowest point to highest point in the ascending order, the red dot is the one which has the least salary and experience. But in between the output graph we see a line dividing the red colour and the green colour which is a hyperplane dividing the values 50,623 and 85,210.





Testing set :

In the testing set we can see the number of points have decreased, this is because for us to clearly see the hyperplane on where it is being divided. Here it is divided between 58,336 and 94,122 which are called as support vectors. Even thou there is a margin between 26,179 and 58,336 it doesn't matter as there is clear big margin between 58,336 and 94,122 rather than to the values near to them.