Assignment No 3 (Group B)

Problem Statement:- Locate dataset (eg. sample_weather.txt) for working on weather data which reads the text input files and finds average for temperature, dew point and wind speed.

Introduction:-

Support Vector Machine:-

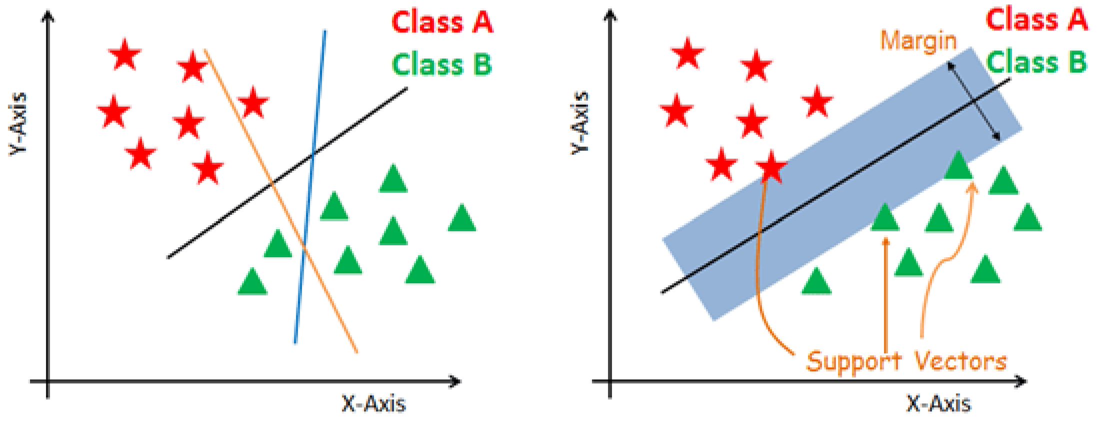
“Support Vector Machine” (SVM) is a supervised machine learning algorithm that can be used for both classification or regression challenges. However, it is mostly used in classification problems. In the SVM algorithm, we plot each data item as a point in n-dimensional space (where n is a number of features you have) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiates the two classes very well

Generally, Support Vector Machines is considered to be a classification approach, it but can be employed in both types of classification and regression problems. It can easily handle multiple continuous and categorical variables. SVM constructs a hyperplane in multidimensional space to separate different classes. SVM generates optimal hyperplane in an iterative manner, which is used to minimize an error. The core idea of SVM is to find a maximum marginal hyperplane(MMH) that best divides the dataset into classes.

How does it work?

The main objective is to segregate the given dataset in the best possible way. The distance between the either nearest points is known as the margin. The objective is to select a hyperplane with the maximum possible margin between support vectors in the given dataset. SVM searches for the maximum marginal hyperplane in the following steps:

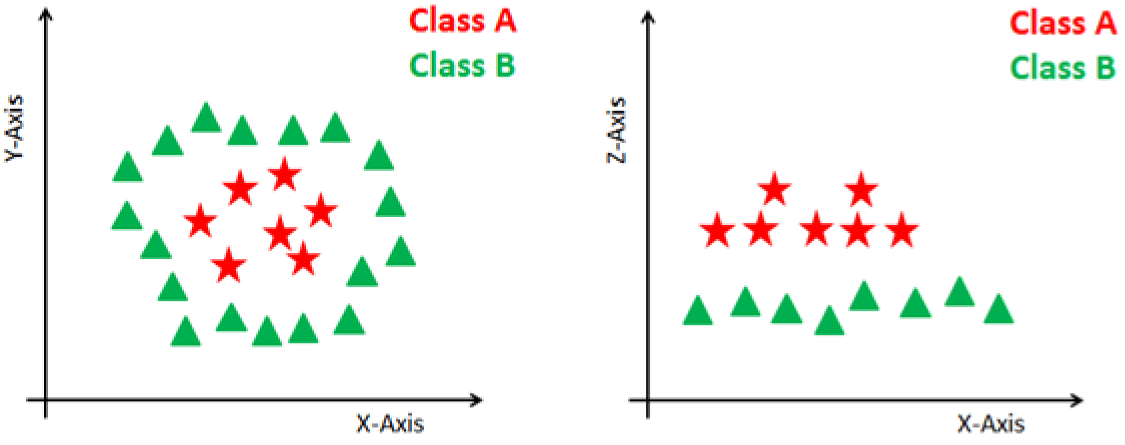
1. Generate hyperplanes which segregates the classes in the best way. Left-hand side figure showing three hyperplanes black, blue and orange. Here, the blue and orange have higher classification error, but the black is separating the two classes correctly.
2. Select the right hyperplane with the maximum segregation from the either nearest data points as shown in the right-hand side figure.



Dealing with non-linear and inseparable planes

Some problems can’t be solved using linear hyperplane, as shown in the figure below (left-hand side).

In such situation, SVM uses a kernel trick to transform the input space to a higher dimensional space as shown on the right. The data points are plotted on the x-axis and z-axis (Z is the squared sum of both x and y: z=x^2=y^2). Now you can easily segregate these points using linear separation.



The advantages of support vector machines are:

- Effective in high dimensional spaces.
- Still effective in cases where number of dimensions is greater than the number of samples.
- Uses a subset of training points in the decision function (called support vectors), so it is also memory efficient.
- Versatile: different Kernel functions can be specified for the decision function. Common kernels are provided, but it is also possible to specify custom kernels.

The disadvantages of support vector machines include:

- If the number of features is much greater than the number of samples, avoid over-fitting in choosing Kernel functions and regularization term is crucial.
- SVMs do not directly provide probability estimates, these are calculated using an expensive five-fold cross-validation (see Scores and probabilities, below).

Implementation:-

Step 1:- Download dataset from below link
https://github.com/shubhomoydas/ad_examples/blob/master/datasets/weather/weather_data.zip
Step 2:- Combine feature and target var in one data frame
Step 3:- Add column name to above dataset(Like Iris flower dataset-add column name)
Step 4:- Find statistics(Mean, Mode, Median) using python code(which we used in previous assignment)
Step 5:- find out missing/NA values
Step 6:-Find Outliers
Step 7:- Use SVM for Prediction(instead of logistic use SVM)

```
# Apply SVM regression
from sklearn.linear_model import SVMRegression
model = SVMRegression()
model.fit(X_train,Y_train)
print('Model Score: ', model.score(X_test, Y_test))
```

Question:-

1) What do know about Hard Margin SVM and Soft Margin SVM?
2) Explain SVM
3) What are Support Vectors in SVMs
4) What is the basic principle of a Support Vector Machine?
5) What happens when there is no clear Hyperplane in SVM
6) Compare *SVM* and *Logistic Regression* in handling outliers
7) When SVM is *not* a good approach?