Support Vector Machine

Objectives

- Understanding of linear classification and margin classification.
- Understanding the terms like support vectors, soft margin, hyperplane.
- Kernel and kernel trick.

Support Vector Machine

SVM is one of the popular supervised machine learning methods that can be equally used for classification and regression, but SVM is mostly used for classification. The principle of SVM is to find an hyperplane which can classify the training data points into labelled categories. The input of SVM is the training data and uses this training sample point to predict the class of test points.

Consider the following set of points of two classes shown in the graph.

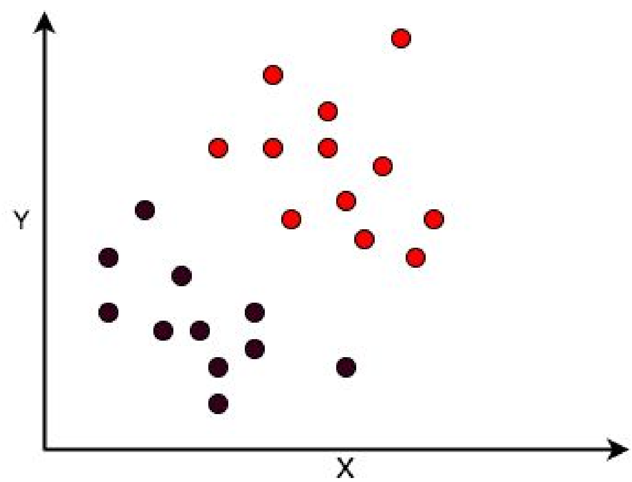


Figure 1 Problem of classification

By looking at the figure we can see that the points can be separated using a hyperplane(line) where + class points are above the line and – class will be below the line. Here we need to remember that there can be many hyperplanes which separate the given points in different ways as shown in figure. Each of the hyperplanes is valid as it separates the given points successfully. But, here our objective is to find the optimal hyperplane.

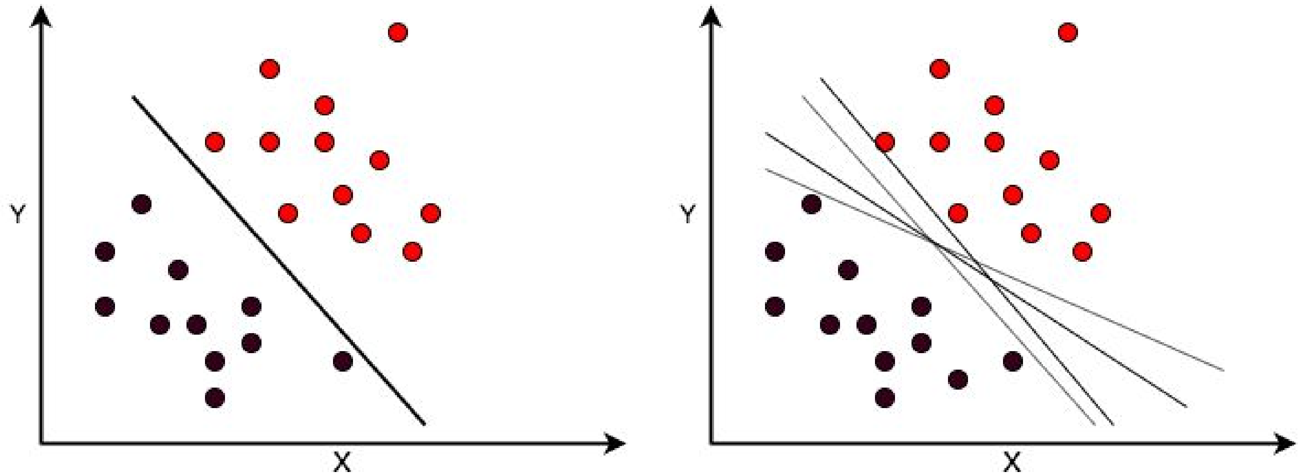


Figure 2 Possible separating hyperplanes

Support vector machine chosen as the best hyperplane is one which is at the maximum distance from the data points from each category. For a given hyperplane, one can compute the distance between the closest data point and hyperplane from both classes. If we double the distance values, we will get a margin**.**

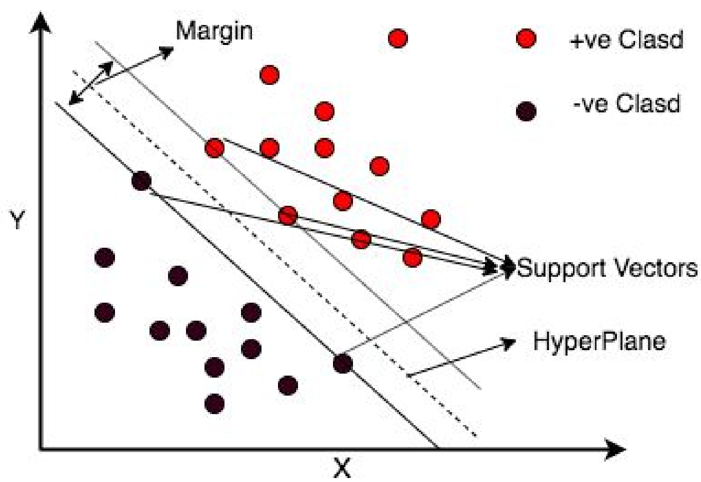


Figure 3 SVM Hyperplane

Margin space is known as no man's land where no data point is present. By looking at the figure we observe that margin width completely depends on how far the points are from the hyperplane.
So, the optimal hyperplane is defined by the biggest margin and the objective of SVM is to find a hyperplane with maximum margin form training data.

Kernel and Kernel Trick

Data classification can be done linear and nonlinear functions as shown in the following figure. The first figure is a rare case and most real-life situations we face a classification problem which is not linearly separable.

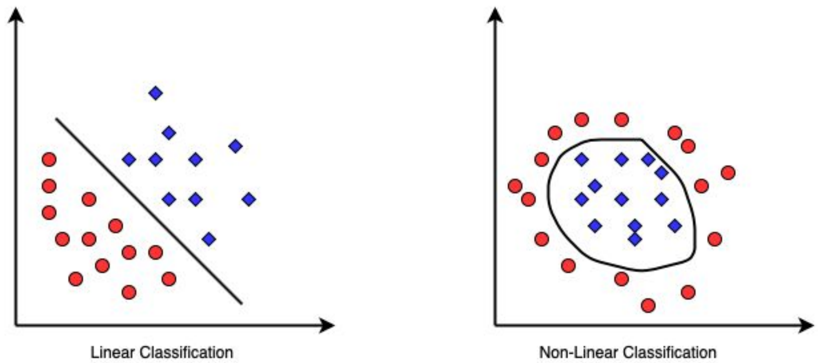


Figure 5: Data Classification

As we know, a support vector machine uses hyper planes to separate data points in the space but with nonlinear data SVM finds it difficult to generate hyperplanes. The solution of this problem is kernel trick. **Kernel function** (Kernel trick) maps the original nonlinear separable data into the higher dimensions which helps to find the suitable hyperplane which separates the data points. It can be also understood as the function which maps a low dimensional space into higher dimensional space for making linear separable data points.

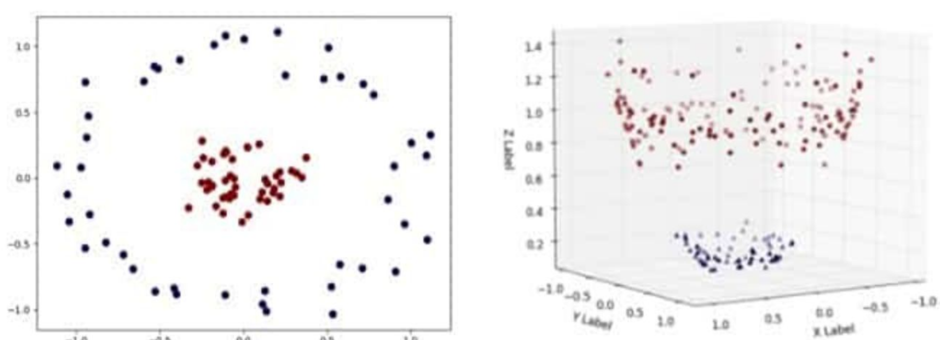


Figure 6: Kernel Trick

As seen from figure 6, that initially the data points are not linearly separable space but after applying kernel trick to it, the data is mapped into higher dimension and become linear separable. Popularly kernel includes three types:

Linear Kernel- It is mostly used kernel specially when data is linearly separable. Data points are separated using a single line or hyperplane. It is widely used in comparison to the other kernels and defined as:

$$K\left(x_i, x_j\right) = x_i . x_j$$

Polynomial Kernel- Polynomial kernel is defined as:

$$K\left(x_i,\ x_j\right) = (x_i . x_j + c)^{\ p}$$

Where, c is an arbitrary constant and p denotes the degree of polynomial. It is clear that linear kernel is a specialized case of polynomial kernel (c =0 and p =1).

Radial basis Kernel- Radial basis kernel is also termed as Gaussian kernel and defined as:

$$K\left(x_i,\ x_j\right) = e^{-\gamma \| x_i - x_j \|^2}$$

Where, γ is hyper parameter controls the variance of model, if γ is large model shows high variance and if γ is small model behaves like linear. $\| x_i - x_j \|$ represent the Euclidean distance between the data point x_i and x_j.

Example-

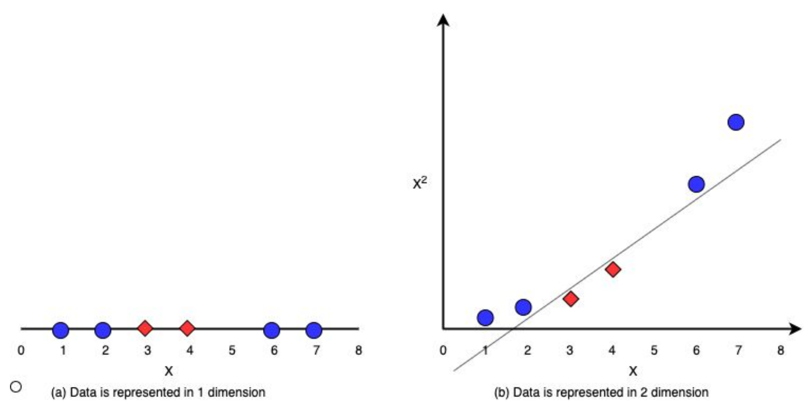
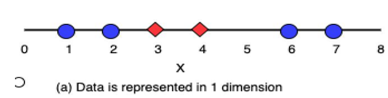


Figure 7 Kernel Trick

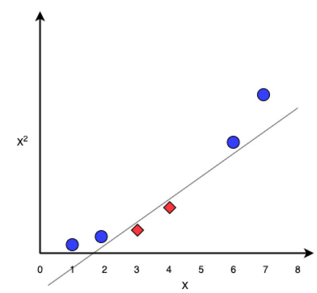
As it can be seen from the figure 7 that initial data is not linearly separable. The given distribution is expressed as:

X	Class
1	Blue
2	Blue
3	Red
4	Red
6	Blue
7	Blue



The given data is in 1 dimension but can be transformed into higher dimension using a kernel trick. Let a new dimension be formed using squaring the variable x. We will get a two-dimensional distribution of given data which is linearly separable and can be classified using a line.

X	X^2	Class
1	1	Blue
2	4	Blue
3	9	Red
4	16	Red
6	36	Blue
7	49	Blue



(b) Data is represented in 2 dimension

Advantages of SVM

- SVM model works well with high dimensional data.
- SVM model equally works well with linear and nonlinear separable data.
- Training model is relatively simple and easy.

Disadvantages of SVM

- Selection of the right kernel and parameters can be computationally expensive.

****