Lecture Scribe for ML(CS60050) Introduction to SVM

Instructor Prof. Aritra Hazra

Date : 11/02/2021

1 Introduction

We are given with a feature space, say X has attributes x_1 and x_2, so certain points can be plotted as :

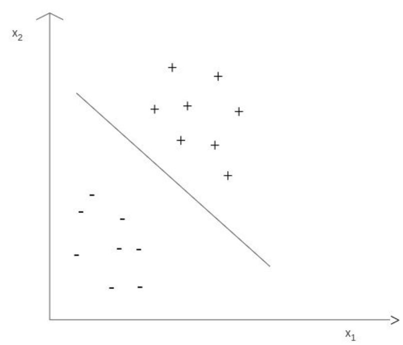

Figure 1:

$$f : X \to Y = +1 or - 1$$

Effectively, it is a supervised learning problem and binary classification problem. Let's say the discriminant used for classifying the data can be a linear model (denoted by red) or by using neural network (denoted by blue) in fig:2

The problem we tackled in terms of classification is called **Discriminant Analysis**. Here, we are to focus on linear discriminant analysis.

2 Linear Discriminant Analysis

Among all the linear discriminator that can be drawn , how to know which one does the best job at classification.[Fig:3] Any linear discriminator can be given

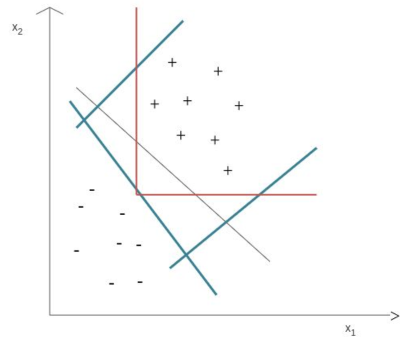

Figure 2:

as $w_1x_1 + w_2x_2 + b = 0$.
For any new point,$x_n =< a_1, a_2 >$

$$w_1a_1 + w_2a_2 + b \geq 0 \rightarrow +1$$
$$w_1a_1 + w_2a_2 + b > 0 \rightarrow -1$$

The concept that is followed is that the discriminator line passing through exactly middle of the both class is best.
As,

$$W^TX + b \geq 0 \rightarrow y = +1$$
$$W^TX + b < 0 \rightarrow y = -1$$

For a point i :

$$y_i(w_1x_{i1} + w_2x_{i2} + b) \geq 0$$

2.1 How is the "middle" defined ?

For a training point x_i, the perpendicular distance of x_i from the discriminator line is given as[Fig 4] :

$$d_i = \frac{|w_1x_{i1} + w_2x_{i2} + b|}{\sqrt{w_1^2 + w_2^2}}$$

The goal is to maximize the minimum distance of the points from the dis-

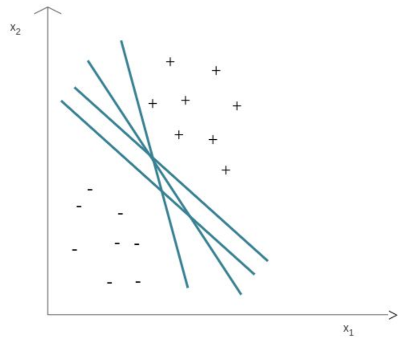

Figure 3:

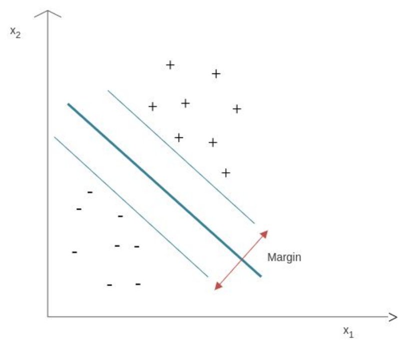

Figure 4:

criminator.

$$MAX[min(d_i)]$$
$$=MAX[min_i \frac{|w_1x_{i1} + w_2x_{i2} + b|}{\sqrt{w_1^2 + w_2^2}}]$$

It is to be noted that the optimization depends only on the numerator.
The w_1, w_2 and b are chosen in such a way that the x_i for which the distance is minimum, the minimum distance (numerator) becomes 1.
Hence the optimization boils down to :

$$Max(\frac{1}{\sqrt{||W||}})$$

where $||W|| = W^T W$

So now the minimum distance for any point from the discriminator is 1 :

$$y_i(w_1x_{i1} + w_2x_{i2} + b) \geq 1$$

3 Primal Optimization problem

$$Minimize(\frac{1}{\sqrt{||W|||}}) \rightarrow MAX(\frac{0.5}{\sqrt{W^TW}})$$

subject to :

$$y_i(W^TX_i + b) \geq 1$$

Now at any point of time, there will be two points that will support the line to stand in the middle which leads to the concept of **Support Vector Machine**.

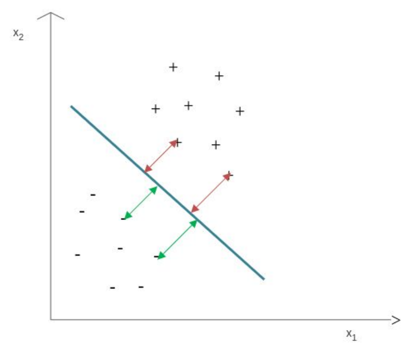

Figure 5:

4 Dual Optimization problem

For and all each $X_i, \exists \alpha_i$ such that constraints: $\alpha_i \geq 0$,

$$MaximizeL = \frac{1}{2}W^TW - \sum_{i=1}^{N} \alpha_i(y_i(W^TX_i + b) - 1)$$

Maximizing

$$\frac{\partial L}{\partial W} = 0$$

$$\Rightarrow W - \sum_{i=1}^{N} y_i\alpha_iX_i = 0$$

$$\Rightarrow W = \sum_{i=1}^{N} y_i \alpha_i X_i$$

$$\frac{\partial L}{\partial b} = 0$$

$$\Rightarrow \sum_{i=1}^{N} \alpha_i y_i = 0$$

4.1 What does Lagrange's multiplier trying to do ?

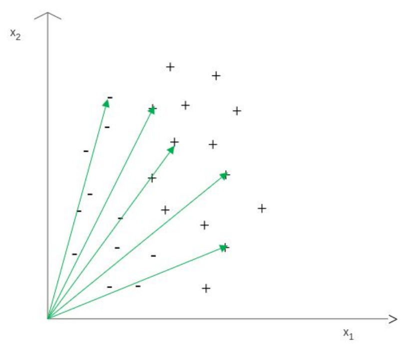

Figure 6:

The points X_is are multiplied with $\alpha_i \geq 0$ and then with y_i(label). It means it is trying to enhance the vectors in all these directions and trying to find the resultant vector from them.

Note that

- Not all $\alpha_i > 0$, only the ones from support vectors are greater than zero, hence it is computationally very efficient.
- Once W is found, finding b is very easy.

$$W^T X + b = 1$$

$$b = 1 - W^T X$$

where W denotes weight and b denotes bias.

So the maximized value of L is :

$$L = \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j (X_i.X_j) - \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i y_i \alpha_j y_j (X_i.X_j) - b \sum_{i=1}^{N} \alpha_i y_i + \sum_{i=1}^{N} \alpha_i$$

As $b\sum_{i=1}^{N}\alpha_i y_i$ is zero , so

$$L = \sum_{i=1}^{N}\alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j (X_i.X_j)$$

The computation of the term $\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j (X_i.X_j)$ is easier if we pre-compute $y_i y_j (X_i.X_j)$. The matrix in which this is pre-computed and stored is called Hessian matrix and is denoted by H.
Therefore, now that we have found $= sum_{i=1}^{N}\alpha_i y_i X_i$ and $\sum_{i=1}^{N}\alpha_i y_i = 0$, let's define the following :

$$\lambda = [\alpha_1\alpha_2...\alpha_N]_{1\times N}$$

$$U = [111...1]_{1\times N}$$

Then L can be written as :

$$L = \lambda U^T - \frac{1}{2}\lambda H \lambda^T$$

This equation is solved using Quadratic programming as L is quadratic w.r.t α. After solving this using quadratic programming, we obtain the values of α_is. By plugging the values of α in the equations the values of W and b are obtained.

$$W = \sum_{i=1}^{N}\alpha_i y_i X_i$$

$$b = 1 - W^T X$$

Scribe made by Anwesha Banerjee [Roll 20CS91R05]