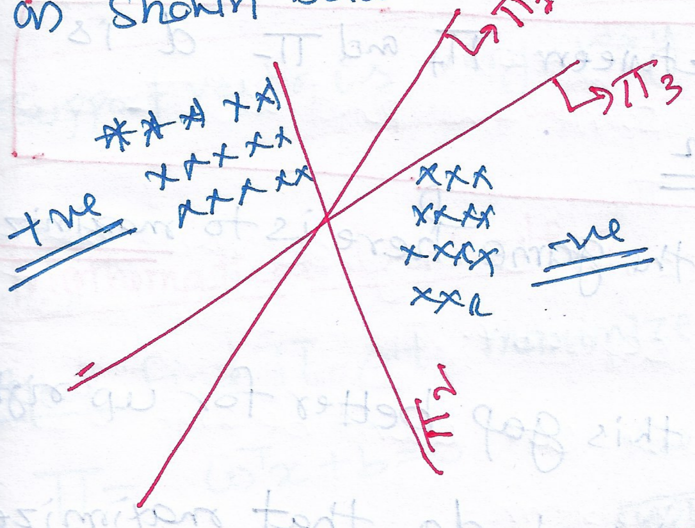
Support Vector Machine (SVM)

First we will discuss the Geometry Behind the SVM

I.E. We'll understand the concept with the help of geometry.

Suppose we've bunch of +ve and -ve points as shown below:



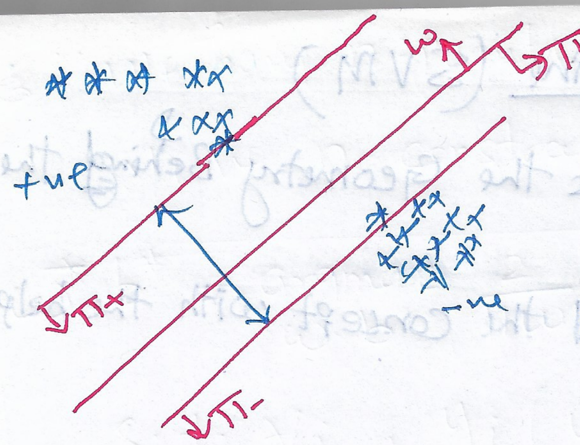
From this we can see that there are many planes that can seperate these points.

> Note: SVM you can use for both classification and Regression.

The key idea of SVM, to find a Hyper Plane that seperates the positive point from negative point as wide as possible.

π_1 is better than π_2 ∵ it seperates the point w.r.t. π_1 are far.

Such a plane k/as : "Margin Maximizing HyperPlane"



→ If we draw a plane parallel to π in both the direction. We will get the plane which will intersect +ve and -ve points

Let's the planes are π_+ and π_-

π_+ and π_- are parallel to π

The distance between π_+ and π_- d is k/as **Margin**

So, the aim of the game here is to maximize this margin.

I.E. the wider this gap better for us right

→ what SVM tries to do, that maximizes the margin (d).

As margin ↑ses, the generalization accuracy ↑. I.E. accuracy towards unseen data set increases.

OK, so far so good. Let's try to understand ~~the~~ another concept k/as Support Vector

The points through which Π_+ or Π_- pass through are Support vector.

I.E. those points which are closest to the hyper-plane, as you go parallel.

$\Pi_+ \rightarrow$ +ve Hyper plane.

$\Pi \rightarrow$ seperating Hyperplane.

$\Pi_- \rightarrow$ -ve Hyper Plane.

Support vector is very-very important in SVM

Mathematical Derivation

Goal: Find Π s.t. maximize the margin.

$\Pi = w^T x + b = 0$ $w \rightarrow$ vector $\perp$ to Π, Π_+, Π_-

$$\left.\begin{aligned} \Pi^+ &= w^T x + b = 1 \\ \Pi^- &= w^T x + b = -1 \end{aligned}\right\} \text{if}$$

Margin: $\boxed{d = \frac{2}{\|w\|}}$

$$\boxed{(w^*, b^*) = \underset{w,b}{\operatorname{argmax}} \frac{2}{\|w\|} \text{ s.t. } y_i(w^T x_i + b) \geq 1 \quad \forall x_i}$$

$\therefore$ We have n-points

Hard margin SVM

$\therefore$ We've n-constraints.

$$\boxed{\begin{aligned} w^*, b^* &= \underset{w,b}{\operatorname{argmax}} \frac{2}{\|w\|} \\ &\text{s.t. } \forall i,\ y_i(w^T x_i + b) \geq 1 \end{aligned}}$$

This work only when my data is linearly seprable.

What if, we have some (few) +ve point at π_- side and -ve point π_+

I.E. Dataset is not Linearly seperable but almost linearly seperable. Then we can't seprate with Hyperplane.

So, in such a case, we won't be able to solve our optimization problem.

→ In this case we introduce an extra term

$\xi \rightarrow$ (zeta)

~~$\xi_i = 0$ if point is above π_+ and below π_-~~

For any point lies b/w π_+ and π_-

~~$\xi_i = 0$ if $y_i(w^T x_i + b) \geq 1$~~

~~$\xi_i > 0$, equal to some units of distance away from correct hyper-plane in incorrect direction.~~

$$w^*, b^* = \underset{w, b}{\operatorname{argmin}} \frac{\|w\|}{2} + C \cdot \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i$$

$$\text{s.t. } y_i(w^T x_i + b) \geq 1 - \varepsilon_i \ \forall i \quad \varepsilon_i \geq 0$$

Minimizing errors = minimize miss classified

$= \min \sum \varepsilon_i$

$C \Rightarrow$ Hyper Parameter

~~$w^*, b^* = \underset{w, b}{\operatorname{argmin}} \frac{\|w\|}{2} + C \cdot \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i$~~

$\frac{\|w\|}{2} \rightarrow 1/\text{margin}$

$\frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \rightarrow$ Avg distance for miss classified point

As $C \uparrow$ your tendency to make mistake $\downarrow$
$\Rightarrow$ overfitting $\Rightarrow$ high variance model

As $C \downarrow$ tendency to underfit $\Rightarrow$ high bias model.

This is k/as soft margin SVM.
↳ ∵ It allows error, but at the same time minimize the error.

As $\epsilon_{ei} \uparrow$ the point is farther away from the correct hyper plane, in the in-correct direction.

So for all the point x_i, we are having ϵ_{ei} s.t.

$\epsilon_{ei} = 0$ if $y_i(w^T x_i + b) \geq 1$ i.e. if the points are correctly classified.

$\epsilon_{ei} > 0$ and equal to the some unit of distance away from the correct hyper plane in in-correct direction.