LECTURE - 52

SUPPORT VECTOR MACHINE (SVM):

It is a machine learning algorithm which can be used for both classification outlier detection and regression.

→ Mostly used in classification problems.

→ Support vectors are simply the co-ordinates of an individual observation.

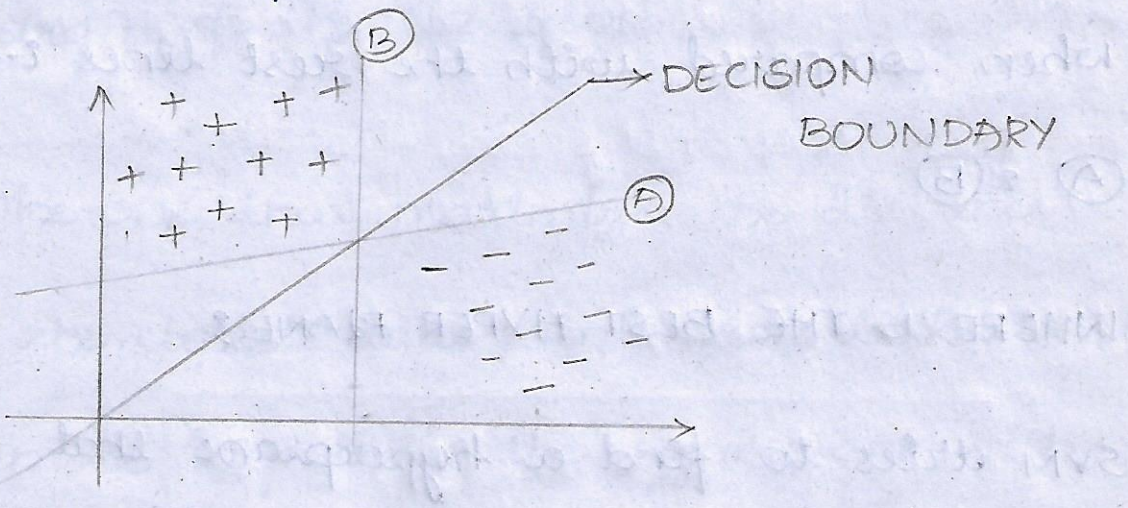
→ The SVM are particular linear classifiers which are based on the margin maximization principle.

→ They perform structural risk minimization, which improves the complexity of the classifier with the aim of acheiving excellent generalization performance.

→ A linear SVM classifier works by drawing a straight line between two classes.

PURPOSE OF SVM:

→ It uses a technique called the kernel trick to transform the data and then based on these transformations it finds an optimal boundary between the possible outputs.



DECISION BOUNDARY: Seperation of classes.

A decision boundary is the region of a problem space in which the output label of a classifier is ambiguous.

↳ If the decision surface is a hyperplane, then the classification problem is linear and the classes are linearly seperable.

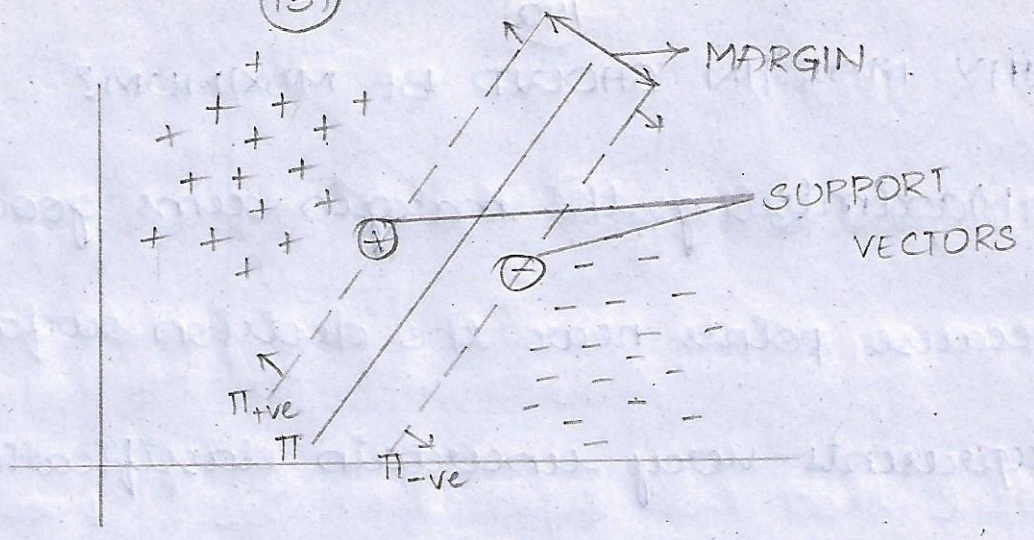
* ↳ Decision boundaries are not always clear cut.

*→ Decision boundary has the maximum distance for both +ve and -ve classes when compared with the rest lines i.e., (A) & (B)

WHERE IS THE BEST HYPER PLANE?

SVM tries to find a hyperplane that seperates +ves and -ves as widely as possible.

→ Also called as "MARGIN MAXIMIZING HYPERPLANE".



MARGIN:

The distance from the decision surface to the closest data point determines the margin of the classifier.

↳ The one that maximizes the distance to the closest data points from both classes then we say it is the hyperplane with maximum margin.

WHY MARGIN SHOULD BE MAXIMUM?

→ Maximizing the margin seems good because points near the decision surface represents very uncertain classification decisions i.e.,

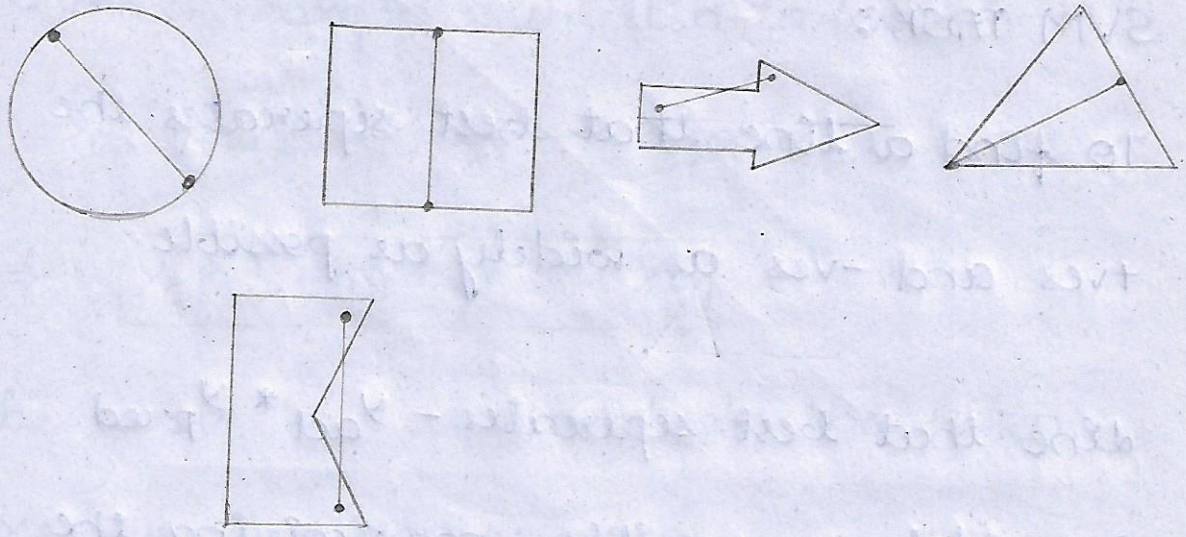
there is a 50% chance of the classifier deciding either way.

* ↳ A classifier with a large margin makes no low certainity classification decisions.

MATHEMATICALLY,

$$\text{Margin} = \text{distance}(\Pi_{+ve}, \Pi_{-ve})$$

So, as the margin increases, generalization accuracy increases.



HOW TO ACTUALLY FIND THE MARGIN MAXIMIZING HYPERPLANE ?

→ In order to understand it, we need have the idea behind the CONVEX HULL.

HULL: Closed figure

From the above figures,

Draw any two points any where and join them by a line. This is called as a Convex Hull.

→ If the line goes out of the boundary then it is not a convex hull.

SVM TASK:

To find a line that best seperates the +ves and -ves as widely as possible.

line that best seperates - $y_{act} * y_{pred}$.

As widely as possible - maximizing the margin.

STEPS FOR SVM ALGORITHM:

STEP-1: Draw the convex hull for +ves and -ves

STEP-2: Find the shortest line connecting those hulls.

STEP-3: Bisect the line i.e., break them into two equal parts.

So, If $y_{act} * y_{pred} > 0$ → SIGNED DISTANCE IS POSITIVE & CORRECT CLASSIFICATION.

⇒ max. margin subjected to (s.t.)

$y_i * \{ mx_i + c \} \geq 0 \quad \forall_i$ (for all)

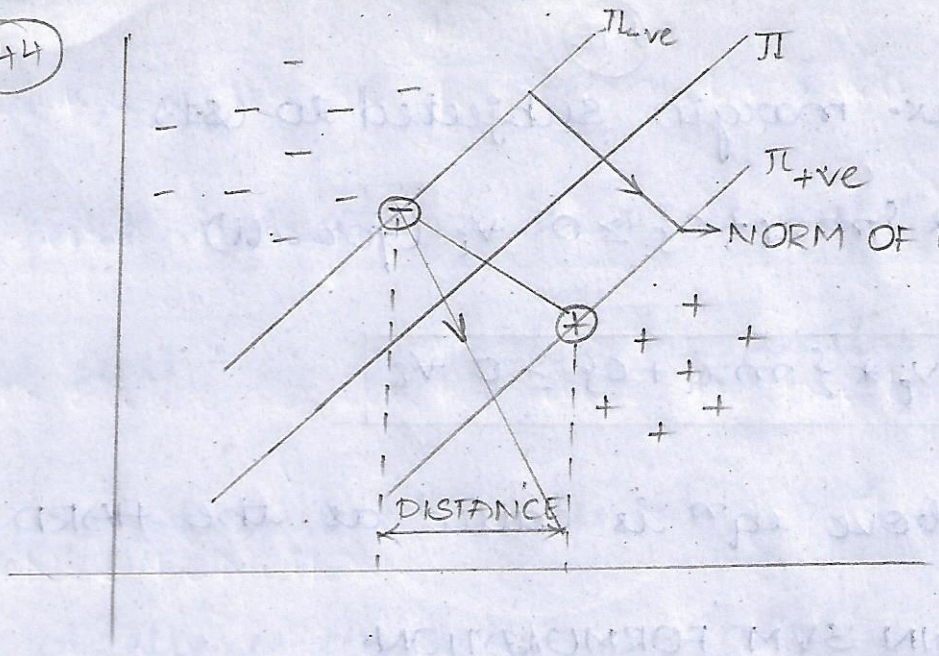
⇒ $\boxed{y_i * \{ mx_i + c \} \geq 0 \quad \forall_i}$

The above eqn is called as the HARD MARGIN SVM FORMULATION.

The above eqn states that for all the data points, it should be True such that every point in the data is correctly classified.

⇒ $y_i * \{ w^T . x_i + w_0 \} \geq 0 \quad \forall_i$ → other form of representation.

Now, lets see the mathematical representation of margin (least distance).



$$\Rightarrow \pi_{-ve} = w^T \cdot x_i + w_0 = -1 \rightarrow (1)$$

$$\Rightarrow \pi = w^T \cdot x_i + w_0 = 0$$

$$\Rightarrow \pi_{+ve} = w^T \cdot x_i + w_0 = 1 \rightarrow (2)$$

Subtracting
~~Adding~~ (1) & (2)

$$\Rightarrow w^T \cdot x_{i_{+ve}} + w_0 = +1 \rightarrow (2)$$

$$\Rightarrow w^T \cdot x_{i_{-ve}} + w_0 = -1 \rightarrow (1)$$

$$\Rightarrow w^T \cdot x_{i_{+ve}} - w^T \cdot x_{i_{-ve}} = 2$$

$$\Rightarrow w^T \cdot (x_{i_{+ve}} - x_{i_{-ve}}) = 2$$

$$\Rightarrow w^T (x_{+ve} - x_{-ve}) = 2$$

$$\Rightarrow \boxed{\frac{w^T}{\|w\|} (x_{+ve} - x_{-ve})} = \frac{2}{\|w\|} \rightarrow \text{MARGIN}$$

↳ PROJECTION DISTANCE ON $\vec{w}$

∴ The mathematical representation of margin is

$$\Rightarrow \max\left\{\frac{2}{\|w\|}\right\} \text{ s.t. } y_i * \{w^T \cdot x_i + w_0\} \geq 0 \ \forall i$$

This is called as Hard margin support vector formulation.

→ There is a dual form of hard margin SVM formulation i.e.,

$$\Rightarrow \max \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i \cdot y_j \cdot x_i^T \cdot x_j \text{ s.t.}$$

$$0 \leq \alpha_i \text{ & } \sum_{i=1}^{n} \alpha_i \cdot y_i = 0$$

To solve the dual form of hard margin we use 'LAGRANGE'S MULTIPLIER' which is a constrained optimisation.

Now, we have a problem i.e.; Dual form of hard margin.

↳ In order to it, we use Kernel Function.

KERNEL FUNCTION:

SVM algorithms use a set of mathematical functions that are defined as the kernel.

↳ The function of kernel is to take data as input and transform it into required form.

→ For example,

linear, non-linear, polynomial, radial basis function (RBF) and sigmoid.

WHAT IS THE PURPOSE OF KERNEL TRICK?

→ It is used to offer a more efficient and less expensive way to transform

data into higher dimensions.

→ With that saying, the application of the kernel trick is not limited to the SVM algorithm.

→ Any computations involving the dot products (x, y) can utilize the kernel trick.

WHAT KERNAL IS USED IN SVM?

→ Use linear SVM (LOGISTIC REGRESSION) for linear problems
  ↳ linear seperation in the data

→ Use Non-linear kernels such as radial basis function kernel for non-linear problems.

So, Now we use the kernel in the dual form of our hard margin

⇒ KERNEL TRICK

$$= \max \sum \alpha_i - \frac{1}{2} \sum \sum \alpha_i \alpha_j \, y_i \cdot y_j \, K(x_i, x_j)$$

For Quadratic, $K(x_i, x_j) = K(x_i \cdot x_j)^2$

For Polynomial, $K(x_i, x_j) = K(x_i \cdot x_j)^d$

where $d = 1, 2, 3, \ldots, d_{n-1}, d_n$

$K(x_i \cdot x_j)^2$ transforms to 3D form.

RADIAL BASIS FUNCTION KERNEL SVM:

A function whose values depends on the distance from the origin (or) from some point.

$$\Rightarrow K(x_i, x_j) = \exp\left\{ \frac{-\|x_i - x_j\|^2}{2\sigma^2} \right\}$$

KERNELTRICK → Take the data $\in$ d transforms to d' where $d' > d$.

↳ In those d' - we figure out linear SVM

→ So, now we have hard margin.

→ Apart from it we have soft margin too.

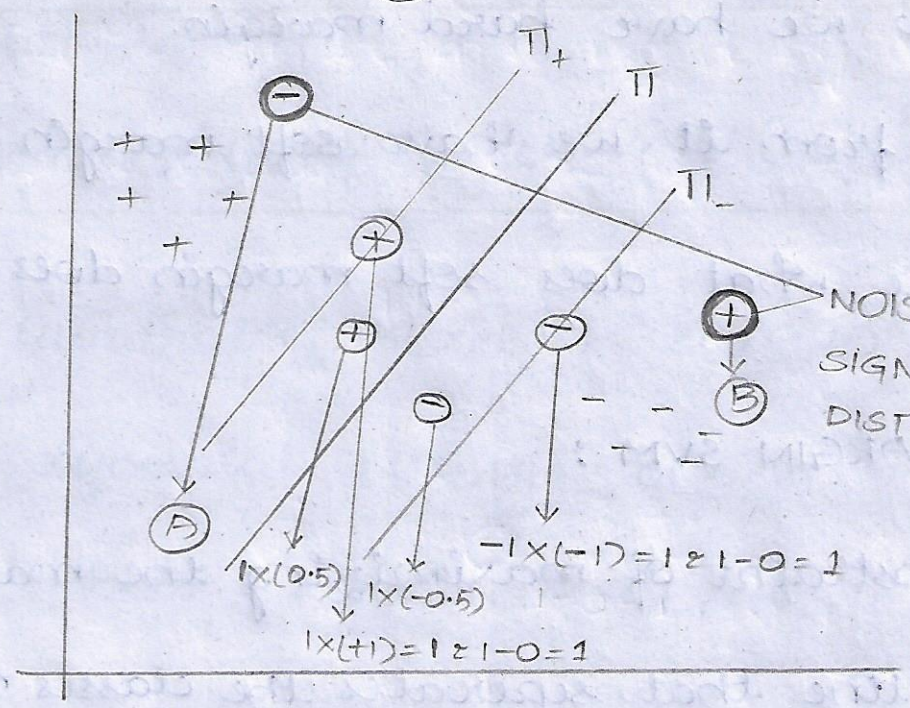
→ lets see what does soft margin does.

SOFT MARGIN SVM :

The constraint of maximizing the margin of the line that seperates the classes must be relaxed. this is often called as soft margin classifier.

→ this change allows some points in the training data to violate the seperating line.

Now lets see the formulation and the rest things (Noise).



(A) $\Rightarrow y_i * \{w^T \cdot x_i + w_0\} = +1 \times 2.5 = -2.5$

(B) $\Rightarrow y_i * \{w^T \cdot x_i + w_0\} = 1 \times -2.5 = -2.5$

NOISE (OR) NOISY DATA :

Noisy Data is a data that has relatively signal-to-noise ratio.

→ The error is referred to as Noise.

→ Noise creates trouble for machine learning algorithms because if not trained Properly, algorithms can think of noise

to be a pattern and can start generalizing from it, which of course is undesirable.

→ so, from Ⓐ & Ⓑ, we need to convert into 1 - something.

where something is "ZETA" and represented as 'ζ'

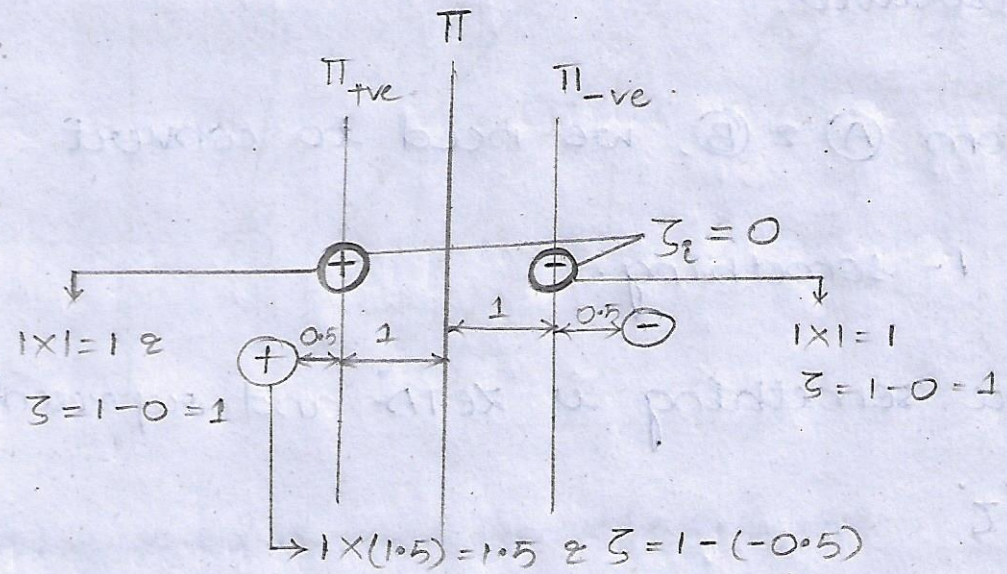
⇒ From Ⓐ & Ⓑ → $1 - \zeta = 1 - (3.5) = -2.5$.

For both +ves & -ves, (Noisy points) the Zeta value is 2.5.

ζ → Adding such a term is called as regularization.

→ Due to which we will make smallest possible error with hyperplane that maximizes the margin.

The eq^{n}'s can be minimized by adding negative values of ζ to it.



From the above figure,

$\zeta_i < 0$ for all the correctly classified classes.

According to $\Pi_{+ve} \geq \Pi_{-ve}$ the classes are misclassified.

→ $\zeta_i \geq 0$ – The point is misclassified according to the $\Pi_{+ve} \geq \Pi_{-ve}$

where ζ_i is called as slack variable.

So, if ζ_i increases, the point is far away in the incorrect direction. i.e., $\pi_{+ve} \geq \pi_{-ve}$

According to π_{+ve} & π_{-ve},

CORRECTLY CLASSIFIED - $y_i * \{w^T \cdot x_i + w_0\} \geq 1$

$\downarrow$

$\zeta_i < 0$

MISCLASSIFIED - $y_i * \{w^T \cdot x_i + w_0\} \leq 1 \rightarrow \zeta_i \geq 0$

NOW we are trying to max. the margin.

$\Rightarrow$ max margin $\simeq$ min. $\frac{1}{\text{margin}}$

[From optimization theory, w.k.t

max. f = min$(-f)$ & max(f) = min(f^{-1})]

$$\Rightarrow \text{max. margin} \simeq \min \left\{ \frac{1}{2} * \|w\| + C \cdot \frac{1}{n} \sum_{i=1}^{n} \zeta_i \right\}$$

$$\text{s.t. } y_i * (w^T \cdot x_i + w_0) \geq 1 - \zeta_i \quad \forall i$$

where C - HYPERPARAMETER OF SVM

If we have noises,

$$\Rightarrow \text{max. margin} = \min\left\{\frac{1}{2} * \|w\| + C \cdot \frac{1}{n}\sum_{i=1}^{n} \xi_i\right\}$$

$$\text{s.t. } y_i * \{w^T x_i + w_0\} \geq 1 - \xi_i \geq 0 \quad \forall i$$

The above eqn is called as the soft margin of support vector machine formulation.

TIME COMPLEXITY OF ALGORITHM:

KNN > SVM > LOGISTIC REGRESSION

↓

training takes a lot of time.

→ If we have points on line then how to do SVM on it?