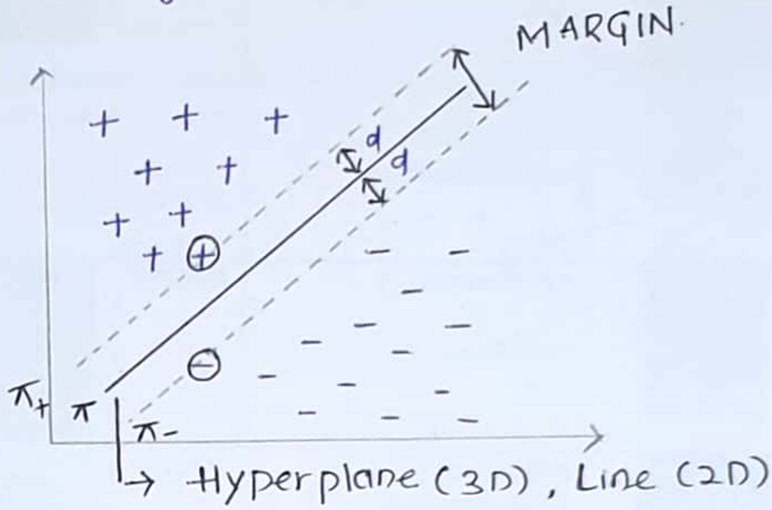
Support Vector Machine Algorithm :-

it is also called mostly S.V.M In this Our task is to find a line that best Separate the positives points from negative points. as widely as possible (Simply make a maximum margin Line), it is a Set of Supervised learning methods. Used for classification, regression and Outlier detection, the advantages of SVM's are _ Effective in high dimentional Spaces.



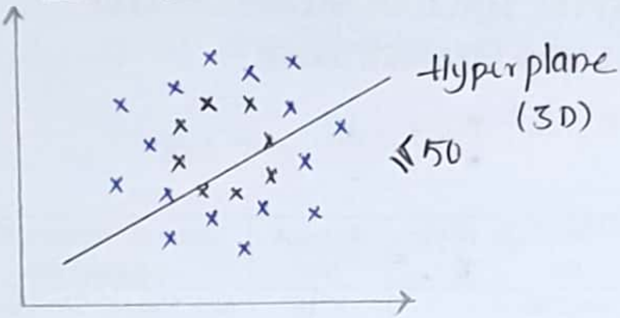
from the figure we can Say that a line (2D) which Separates two data points in to two classes as +ve's, -ve's, and also makes two marginal lines and this widely Separated parallel lines having some distances. this lines are passing through One of the nearest points may be positive or negative, the difference between logistic regression and svm is both are Separate the points but Svm has Some widely deffrenced distances.

the distance between (π_+) line to hyperplane is "d+" and distance between (π_-) line to hyper plane is "d-".

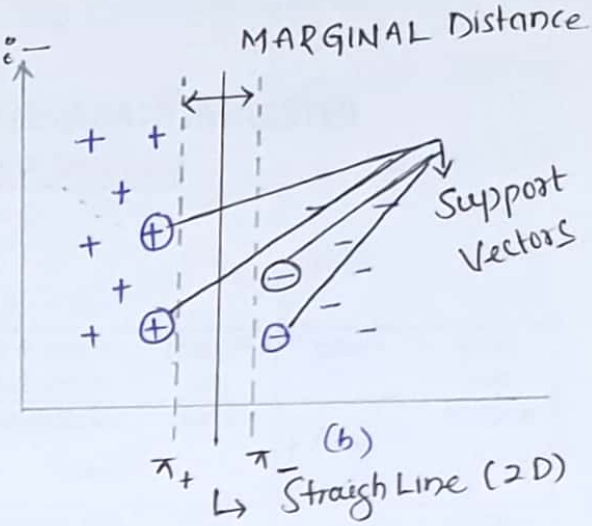
Summation of this two distances is called as "Margin" ②

The role of ~~marg~~ margin is to devided a class in better way to ~~[illegible]~~ tell us the class of a data points for suppose if a new data point comes and lying in between hyperplane and one of margin them we simply understood the point belonging to that class.

Linear Separable & Non Linear Separable :-



(a)



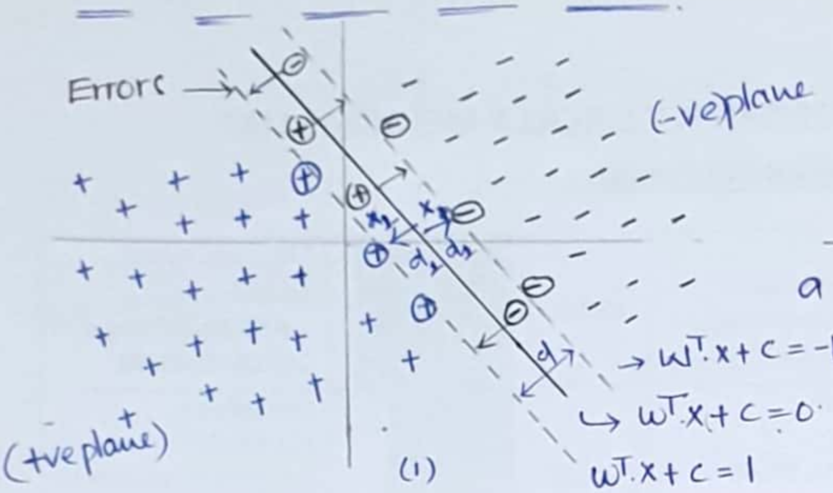
(b)

If we observe fig (a) all the points are overlapping with each other there we can't separate data points by making a plane or Hyper plane. So we can say this is a non-linear separable planes, if solve like this probleme we have to use kernel distribution.

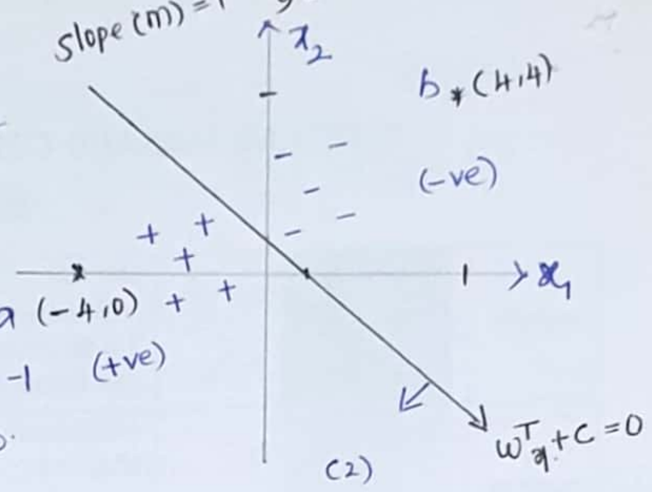
from figure (b) we easily separate the data points by making a straight line this is called as linear separable line.

After classified data points the nearest data points from any class passes through marginal lines are called as support vectors as shown in figure if new data point come that is also.

SVM MATHS INTUITION :-



(1)



(2)

the main diff between logistic regression and svm classification is marginal lines this marginal lines how we created we discuss here

from fig (2) if we observe a line we are making that separates having slope = -1 and the line eqn is $w^T_x + C = 0$ and we know $y = w^T x + c$ as shown in graph.

let consider two points (4, 4) and (-4, 0) as shown in graph.

for suppose point (a) i want 'y' value —

So here $m = -1$; constant (c) = 0 because passes through origin.

Then — $y = \underline{w^T_x} + 0 \Rightarrow$

$$y = \underset{w^T}{\begin{bmatrix} -1 \\ 0 \end{bmatrix}} \underset{x = 'a'}{\begin{bmatrix} -4 & 0 \end{bmatrix}}$$

∴ Dot matrix

$y = 4$, ∴ when ever we do Dot Matrix we get Scalar Value only.

∴ y value is $\underline{\underline{+ve}}$

<u>Note</u>: when ever we get a line the below mentioned points all are always to be postive and Similarly the Opposite Co-ordinate values always to be Negative.

$$\begin{bmatrix} -1 \\ 0 \end{bmatrix} \begin{bmatrix} 4, 4 \end{bmatrix}$$

$y = -8$

after find out y value we get two groups that are (+ve), (-ve) classes, if i consider line below group as "+1" becouse it gives positive points value may be differe and above as "-1" for easy calculation.

from figure (1) let consider nearest points to the line (ie shortest distance from line to point on below side, above side of a line. and make a dot point of line parallel to 'y' line that indicates marginal lines. and distance b/w them is marginal distance.

if we observe clearly it, create the positive plane & Neg. plane

So from that the line eqn's are - $y = w^T.x + c = 0$.

below hyper plane is - $w^T.x + c = 1$

above hyper plane is - $w^T.x + c = -1$

In graph I mentioned points as (x_1, x_2) which are nearest points in diffrent planes of different marginal lines. and the distances are (d_1, d_2) respectivly. -

$$w^T x_1 + \cancel{c} = -1 \quad — ①$$
$$w^T x_2 + \cancel{c} = +1 \quad — ②$$
$$- \quad - \quad -$$
$$w^T(x_2 - x_1) = 2$$

from this two eqn we can get total distance b/w two parallel planes to plane., now i get the distance but I want to remove the w^T. In order to do that we have to use ‖w‖ length of vector becouse we know direction.

Now We have to devide Eqn by $\|w\|$ both Side

$$\boxed{\frac{w^T}{\|w\|}}(x_2 - x_1) = \boxed{\frac{2}{\|w\|}}$$

→ This is Our Optimization Eqn. and we have to Increase this value. (Maximize it)

↳ w^T. Magnitude goes out only direction there

$$\boxed{\begin{aligned} &\text{Simply} \Rightarrow (w^*, c^*) = \text{Max} \frac{2}{\|w\|} \\ &\text{st } y_i \begin{cases} +1 & w^T x + b \geq 1 \\ -1 & w^T x + b \leq -1 \end{cases} \end{aligned}}$$

→ optimization function. — ①

HARD MARGIN SVM ←

$\Rightarrow \boxed{y_i * w^T x_i + c_i \geq 0}$

from this ~~eq~~ eq when ever i get $w^T x + b$ value above 1 always I have to Consider it as +ve value Similarly below '−1' always consider as Negative value.

from that eqn we can say the value is always graterthan "1"

from that eqn we can say the value is always graterthan "1"

if it is not, then we understand there is a mis classification.

if it is not, then we Understand there is a mis classification.

if it is not, then we have to minimize the distances ~~for~~ for from the "SVM" Eqn we have to do reciprocal of it better optimization In Order to do it we have to

$$\boxed{\text{Max } f = \text{Min } \frac{1}{f}}$$

⇒ for minimum value

The Eqn will be — $w^*, c^* = \min \dfrac{1}{2/\|w\|}$

$$\boxed{w^*, c^* = \text{Max} \frac{2}{\|w\|} \Rightarrow w^*, c^* = \min \frac{\|w\|}{2}}$$

⇒ S.V.M — ②

for better optimization we have to Include C value and Zeeta value where 'c' → Indicates how many errors will be model Included, ξ → value of the error.

Here final eqn will be –

$$\boxed{w^*, c^* = \text{Min} \left\{ \frac{\|w\|}{2} + C_i \sum_{i=1}^{n} \xi_i \right\}} \quad - (3)$$

Where 'c' is Hyper parameter and it represent error's, which we are Considered ie for Suppose if the test data values above the marginal lines that Should be Consider as Error, here we have to do treatment on it, and the errors will be low in value Counts, So we can easily low the values of error ie neglizible by hyper parameter tuning. 'c', So intially in Eq (2) model has issue of Overfitting we applied hyper parameter as Shown in eq (3) that tune the model, In Order to get 'best model"

So here Eq (1) represent Hard Margin S.V.M formulation

Eq (2) represent Simple S.V.M formulation.

Eq (3) represent Simple SVM formulation with Regularization term where (C is hyper parameter)

from Hard Marginal SVM formulation we will get dual form of S.V.M

Dual form of S.V.M :- Above we discussed formulation was the Simple form of S.V.M the alternate method is dual form of S.V.M. which uses "Lagrange's Multiplier" to Solve the Constraints Optimization problem

Note :- if $y_i * (w^T x_i + c_i) > 0$ then x_i is a Support vector and when $y_i * (w^T x_i + c_i) = 0$ then x_i is not a Support vector.

Eqn will be –

$$\boxed{\begin{gathered} \text{Max} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \cdot \alpha_j \cdot y_i \cdot y_j \cdot x_i^T \cdot x_j \\ \text{st } 0 \leq \alpha_i \; \& \; \sum_{i=1}^{n} \alpha_i y_i = 0 \end{gathered}} \quad - (4)$$

from this dual-form eq (4) we applying Gradient descent

Eq (4) – represent dual-form of Hard Margin S.V.M formulation

if we observe eq 4 → $\{x_i^T \cdot x_j\}$ represent dot product of familiar object

this notation is called as "kernel" trick.

kernel trick eqn – $\boxed{\text{Max}\left\{\sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \cdot \alpha_j \cdot y_i \cdot y_j \cdot K(x_i \cdot x_j)\right\}}$ – (5)

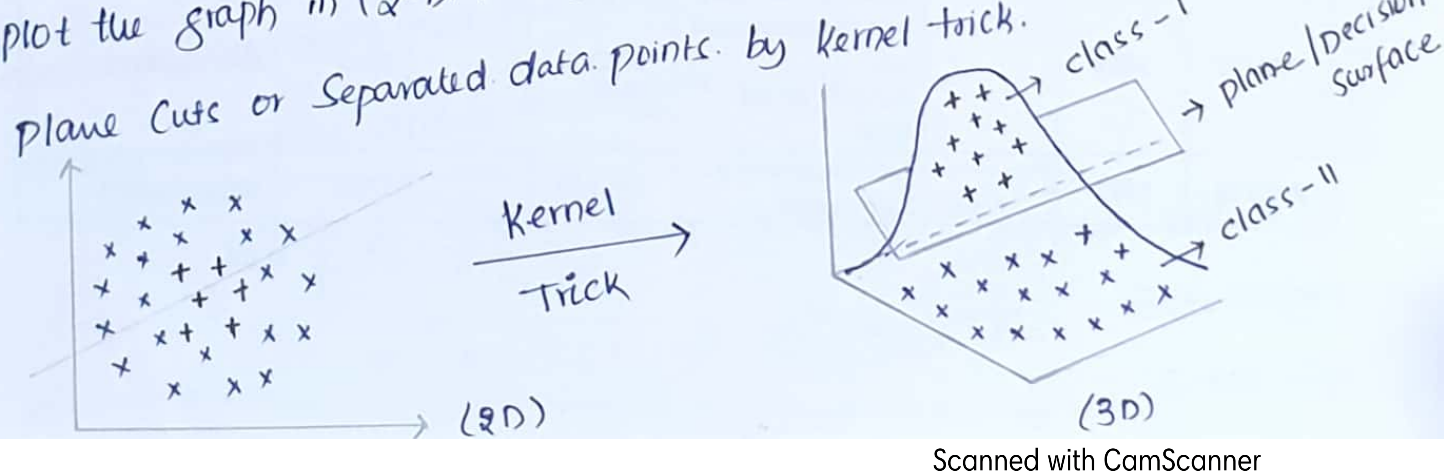
where – $K(x_i \cdot x_j) = x_i^T \cdot x_j$

So from this we can say – $\boxed{K(x_i \cdot x_j) = x_i^T \cdot x_j = \{x_i \cdot x_j\}}$ → This term is represent linear S.V.M, it is nothing but also called as logistic regr. because in linear S.V.M the linear line just classify the data only not mentioned marginal lines. ie simply acted as logistic regression.

So Instead of using simple linear SVM most are prefer polynomial or Quadratic kernel S.V.M which looks like –

$$\boxed{K = (x_i \cdot x_j) = (x_i \cdot x_j)^r} \quad \text{– (6)}$$

So here we just squaring the eqn of linear S.V.M or we can say we apply degree of '2' in polynomial eqn (or) Quadratic kernel S.V.M, Now we plot the graph in (2-D and 3-Dimention) where we understand how the plane cuts or separated data points. by kernel trick.



By above figure shows kernel trick appling data and eq⑥ known as polynomial kernel SVM of degree = 2, we can change degree also in the eqn.

In Generally we use "R.B.F kernel svm" most of the times, this is an important eqn when ever we don't know to use which algorithm of Svm use in that case we go with "R.B.F kernel S.vm".

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

This term 'R.B.F' stands for Radial Basis kernel.

Upto now we discuss S.vm eqn in diff formats such as linear prog Simple S.v.m eqn, Quadratic eqn, polynomial equation with different degrees and finally we using hard margin eqn with kernal tricks to best Optimization

Now we discuss Softmargin eqn, by applying "Lagrange's Multiplier" as Similar to Dual form hard margin eqn, the diffrence is in a hard-margin S.v.m a single outlier can determind the boundary, which makes the classifier, overly sensitive to noise in the data, the result is that soft margin S.v.m could choose decision boundary that has non-zero training error even if data set is linearly separable and is less likely to overfit.

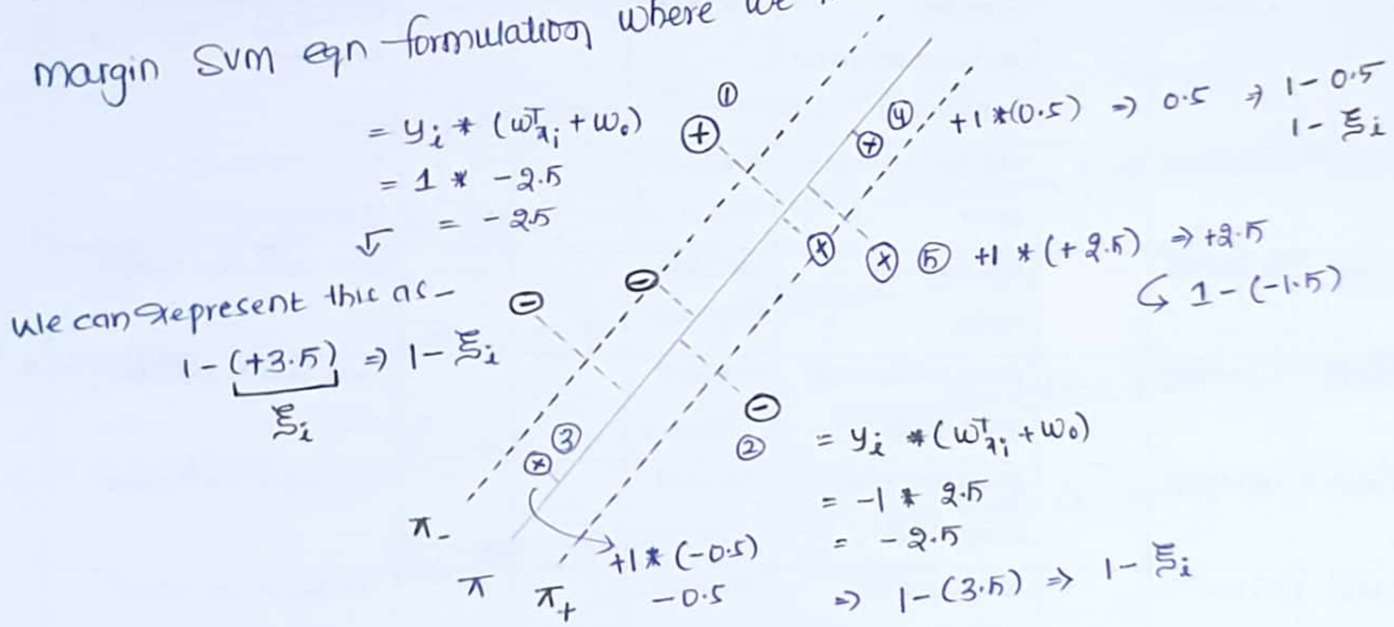
$$\boxed{\text{Max} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \cdot \alpha_j \cdot y_i, y_j, (x_i^T \cdot x_j) \quad \text{st} \quad \boxed{0 \leq \alpha \leq C} \quad \$ \quad \Sigma_{i=1}^{n} \alpha_i y_i = 0}$$

⇒ Called as Soft Margin S.V.M. Eqn.

As Similar in hard margin eqn what ever changes we done (like kernal trick, Quadratric eqn, polynomial eqn with different degrees) we do hear also, we can replace that termandgree also.

In Hard Margin s.v.m we are triying to maximize the margin (Signed Distances) to all the points. ie ($y_i * (w^T.x_i + w_0) \geq 0 \ \forall i$ when there is no noicey in data. ie when we plot a graph that should be separated Correctly there is no misclassified points. then only we use hard margin s.v.m. eqn

In Other Case if we got Some misclassified points in graph then we have a problem with that point we have to change or Correctly classify the points In Order to handle this problem. we have to change entair eqn in to Soft margin svm eqn formulation where we treat this problem.



So if we observe the above graph point (1) is misclassified

point(2) – also misclassified

point (3) – Short distance from margin to misclassified point.

point (4) – Short " from " to Correctly classified point

point (5) – Correctly classified point.

Here we Solve the problem of IncOrrectly classified points for that we use Signed distance in that we have actual value and predicted value's are presented. for point (1) we have actual is +ve predicted as (-ve) with Some longer distance as shown in figure shows the entair negative representation ie misclassified point that can be represent with Some (ξ_i) zeeta Value. in point (1) we have value is positive as Similarly if we do with all points –

P(2) → Negative → Misclassified → (ξ_i) positive

P(3) → Negative → Shorter distance → misclassified → (ξ_i) positive

P(4) → positive → Shorter distance → Correctly classified → (ξ_i) Negative.

P(5) → positive → Correctly classified → (ξ_i) Negative.

from the above Summery we can Say Misclassified points having positive (ξ_i) value, for Correctly classified points we have (ξ_i) Negative value. that can be Simply represent as –

when – $\xi_i < 0 \Rightarrow$ (Negative) Correctly classified. (lower than zero)

$\xi_i \geq 0 \Rightarrow$ (positive) Misclassified according to π_+ \$ π_-

ξ_i → called as Slack Variable

Note :– if we increase zeeta value (ξ_i) ↑ → point is far away in the incorrect direction (π_+, π_-)

$y_i(w^T \cdot x_i + w_0) > 1 \Rightarrow \xi_i < 0$ (Correctly classified)

$y_i(w^T \cdot x_i + w_0) \not< 1 \Rightarrow \xi_i \geq 0$ (Misclassified)

So here Our aim is to maximize the margin.

In Order to maximize margin we have to minimize function from that

we can Say $Max(Margin) = Min \frac{1}{Margin}$

from this eqn of Softmargin S.V.m will be –

$$\boxed{\begin{aligned} &\text{Min} \left\{ \frac{1}{2} * \|w\| + c \cdot \frac{1}{n} \sum_{i=1}^{n} \xi_i \right\} \\ &\text{st} \quad y_i * (w^T \cdot x_i + w_0) \geq 1 - \xi_i \quad \forall_i \\ &\xi_i \geq 0 \end{aligned}}$$

⇒ Optimized Soft Margin S.V.m formulation

Where – c is Hyperparameter of S.V.M.

from this eqn we can write formulation of Dual form of S.m. S.V.m.

$$\boxed{\begin{aligned} &\text{Max} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \cdot \alpha_j \cdot y_i \cdot y_j \, K(x_i, x_j) \\ &\text{St} \quad 0 \leq \alpha_i \leq C \;\; \& \;\; \sum_{i=1}^{n} \alpha_i y_i = 0 \end{aligned}}$$

This term represent kernel trick

This entair thing is "Svm" with different formulation, In real time run S.v.m is more cost ie most time Consumining algorithm. after K.N.N

The time Complexity is more for this algorithm, as Compare to K.N.N.

S.V.m is less time Complexity.

KNN. > SVM > Logistic Regression

* K.N.N. is also known as Lazy learner Algorithm.