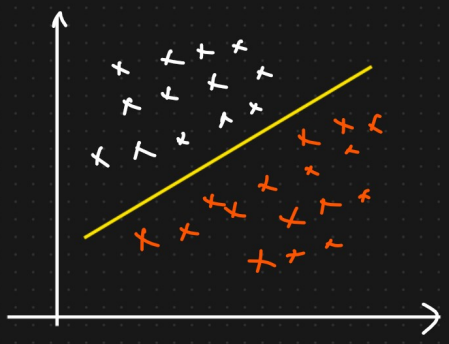
Support Vector Machine [SVM]

SVM is one of the supervised learning algorithm that can be used for both classification and regression problem statements.

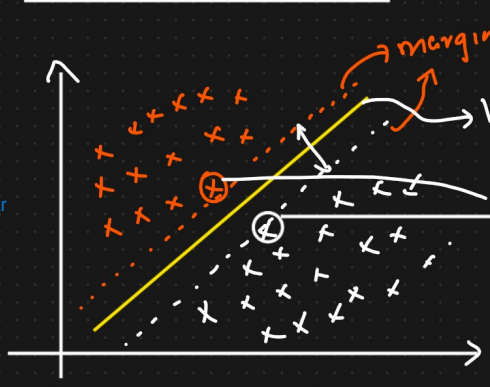
1. Classification (SVC) ⇒ Support Vector Classifier — For classification SVC used
2. Regression (SVR) ⟶ Support Vector Regression — For classification SVR used



In support vector classifier along with best fit line(similar to logistic regression) we are also making the marginal lines/planes in such a way that the perpendicular distance between marginal line/plane and best for line/plane is maximum.

Depending on the use case marginal lines/planes can be:
1. Sott margin
2. Hard margin

① Support Vector Classifier

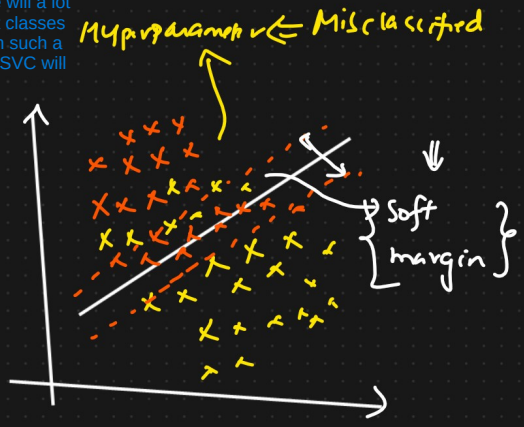


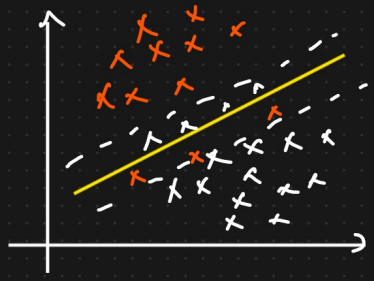
Note: Margin is the perpendicular line b/w marginal line and best fit line.

Also, the points near the marginal line are considered as support vectors using which we are basically drawing the marginal lines

SOFT MARGIN:

In real world scenario there will a lot of overlapping b/w different classes of data points. Therefore, in such a case the margins drawn in SVC will act as soft margin.

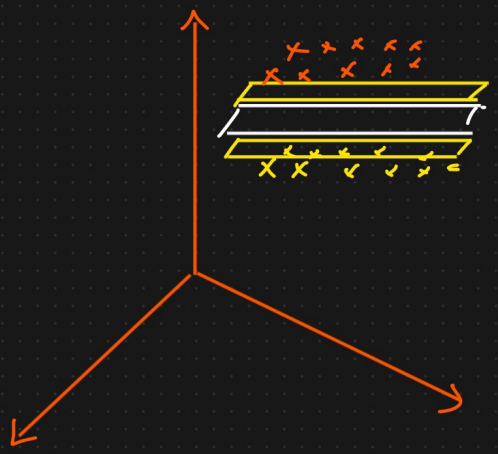




⇒ Hard Margin

HARD MARGIN:

In an ideal case where data points are not overlapping with each other then the margin thus drawn to segregate the different classes of data points will treated as hard margin.



In 2D best fit and margin will be a line.

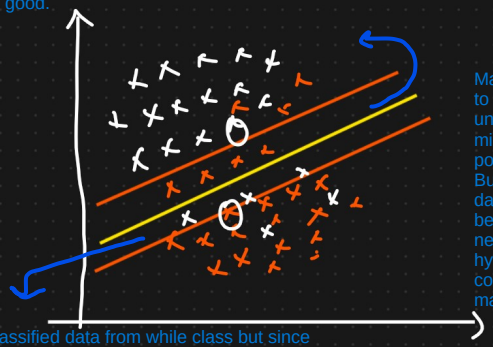
Whereas, in 3D case best fit and margin will be planes.

So what is margin?

Let's say we have a hyperplane — line X(best fit line)
calculate the perpendicular distance from all those 40 dots(data points) to line X, it will be 40 different distances
Out of the 40, the smallest distance, that's our margin!

Soft Margin And Hard Margin In SVC

Miss classified data from orange class but since it's under the margin defined then we should be good.



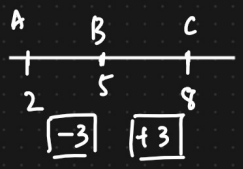
Margin is drawn in order to specify the boundary under which if we get the miss classified data points then we are good. But if we start getting data points above and beyond that, then we need to perform hyperparameter tuning to counter that case(soft margin).

Miss classified data from while class but since it's under the margin defined then we should be good.

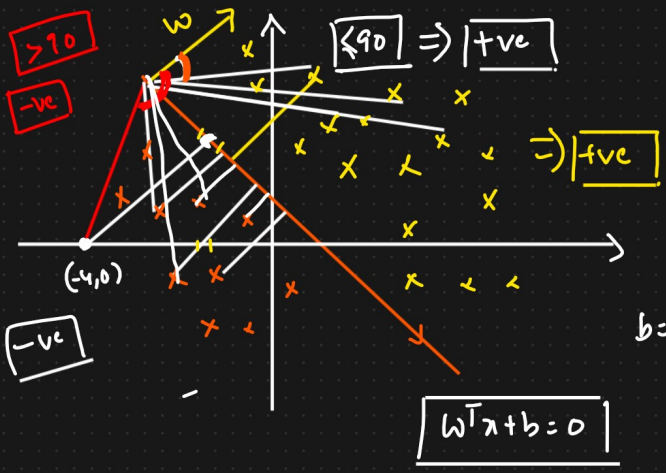
To summarize if asked in interview:
1.We first of all identify the support vectors.
2. then use support vectors to draw the marginal lines in such a way that the perpendicular distance b/w marginal plane and best fit line that splits the data into 2 or more classes is maximum.

If asked that if in logistic regression we were able to so the same using best fit line then why do we need to do extra work by introducing margins?

Ans: Tell them that this is one of the approach. SVM works well with unstructured and semi-structured data like text and images while logistic regression works with already identified independent variables. SVM is based on geometrical properties(since, concept of margin used) of the data while logistic regression is based on statistical approaches(since using sigmoid function).



① SVC Maths Intuition



If the angle between the vector and the points is greater than 90, then distance — — is -ve

w is the vector that is perpendicular to the best fit hyperplane.

If the angle that is formed b/w data point and vector line is greater than 90 then that distance (b/w datapoint and best fit line/hyperplane) will be -ve.

Whereas, if the angle as discussed above is less than 90 then distance will be +ve.

$w^T x = 0$

Deriving equation of best fit plane and marginal planes using equation of straight line

Equation of a Straight Line

$y = mx + c \Rightarrow ax + by + c = 0$

$h_\theta(x) = \theta_0 + \theta_1 x$

$by = -ax - c$

$y = \frac{-a}{b} x - \frac{c}{b}$

$\Downarrow \quad \Downarrow$

$y = m x + -c$

$y = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3$

$y = b + w_1 x_1 + w_2 x_2 + w_3 x_3$

$$w = \begin{bmatrix} w_1 \\ w_2 \\ w_3 \end{bmatrix} \quad x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$

$$w^T = \begin{bmatrix} w_1 & w_2 & w_3 \end{bmatrix} \quad x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$

$$\boxed{y = w^T x + b}$$

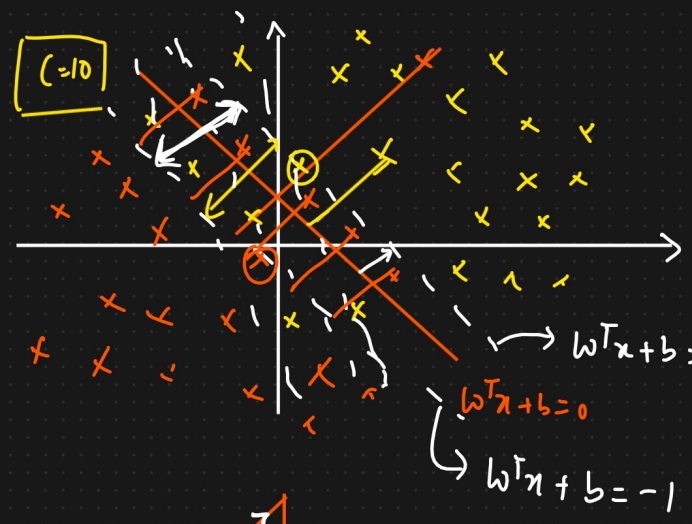
where b is y intercept

Since, Eq^ of best fit line is WTx+b=0, and if go towards positive direction(upwards such that angle b/w vector and closest support vector is less than 90) we will get we will get equation of positive marginal line i.e. wTx+b=+1

Similarly, is we go in -ve direction we will get the equation of negative marginal line i.e. wTx+b=-1

Vector = Magnitude + direction. Eg; traveling 110 km in east direction.

Scalar = Only magnitude and no direction. Eg; traveling 110 km.

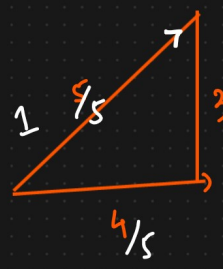


$w^T x_1 + b = +1$

$w^T x_2 + b = -1$

(-) (-) (+)

$\vec{w} \leftarrow \frac{w^T (x_1 - x_2)}{\|w\|} = \boxed{\frac{2}{\|w\|}}$ ⇑⇑ ⇒ Maximum



$\boxed{|v|}$

Unit vector = $|\hat{v}| = \boxed{1}$

Our aim is to maximize the perpendicular distance b/w best fit line and marginal line. This can be re written as our aim is to maximize the distance b/w both marginal lines/planes.

If we subtract both +ve and -ve marginal line equation with each other then we will get the distance b/w then which we ultimately need to maximize.

Above we tried to convert the vector into unit vector and for this have simply divided both LHS and RHS with mod or magnitude of vector

Cost function

Maximinimize $\boxed{\frac{2}{\|w\|}}$ ⇒ Distance between marginal planes.

w, b

Constraint such that

f_1 f_2 f_3 $\boxed{y}$

Actual value

Predicted value

$y_i \begin{cases} +1 & \text{if } w^T x + b \geq 1 \Rightarrow +ve \\ -1 & \text{if } w^T x + b \leq -1 \Rightarrow +ve \end{cases}$

Multiplying actual value and predicted values both in correct and incorrect classification

For all correct classified data point

$\boxed{y_i * [w^T x + b] \geq 1}$

⇓

$\boxed{\text{Predicted point} \Rightarrow \hat{y}}$

Case1: +1 getting correctly classified as +1.
Yi=+ve, wTx+b=+ve

Case2: -1 getting correctly classified as -1.
Yi=-ve, wTx+b=-ve

Case3: +1 getting misclassified as -1.
Yi=+ve, wTx+b=-ve

Case4: -1 getting misclassified as +1.
Yi=-ve, wTx+b=+ve

$\boxed{y_i * \hat{y}_i \leq -1}$

⇓

Incorrect classification

Modified Cost fn

Since, cost function is always represented in minimized term. So, here in order to convert the cost function represented in maximized term into minimized term we will simply take reciprocal (make numerator as denominator and denominator as numerator)

$$\max_{w,b} \frac{2}{\|w\|} \Rightarrow \boxed{\min_{w,b} \frac{\|w\|}{2}}$$

This is the cost function in case of hard margin.

Constraint such that

$$y_i \begin{cases} +1 & \text{if } w^T x + b \geq 1 \\ -1 & \text{if } w^T x + b \leq -1 \end{cases}$$

This is the cost fun^ in case soft margin. Hinge loss is introduced to counter the case of soft margin where we are provided with a lot of overlapped datapoints.

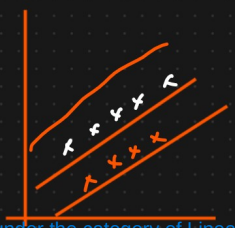
$$\text{Cost fn} = \min_{w,b} \frac{\|w\|}{2} + C_i \sum_{i=1}^{h} \xi_i \Rightarrow \text{Hinge loss}$$

ξ_i pronounced as "Etta"

C_i → Hyperparameter

ξ_i → {Summation of the distance of incorrect data points from marginal plane}.

h → {How many points we can consider for misclassificatn}.



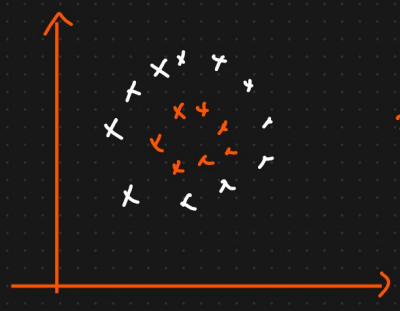
In case of soft margin we will need to add hyperparameter:
1. C of i
2. itta of i=distance from the marginal plane that would have classified the miss classified data point correctly(marginal plane is expected to classify data point correctly).

Whatever, we have discussed above will fall under the category of Linear SVC. Below we have discussed about non linear SVC. To deal with non-linear distribution we use kernels.

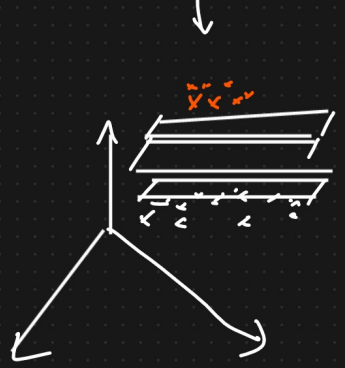
For EG in below example we converted 2D representation into 3D representation using kernels where, we just took a class of data(orange class) and expanded or elongated it across the Z axis. Hence, now in the plane thus formed we can easily make linear classification.

$\lambda_1 \quad \lambda_2$

$$C = \frac{1}{\lambda}$$



$\Rightarrow$ linear SVC $\Rightarrow$ Kernels $\Rightarrow$ RBF

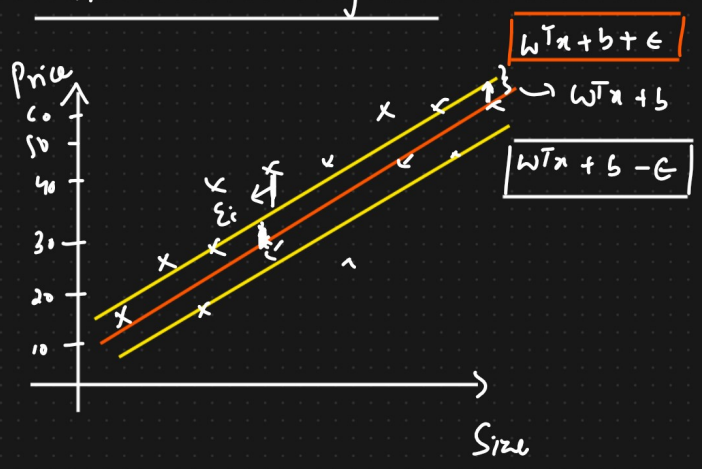


To summarize, in Kernels we increase the number of dimensions so that distinction can be made.

Support Vector Regressor

epsilon

$\epsilon \Rightarrow$ Marginal Error



Cost fn

$$\underset{w,b}{\text{Min}} \quad \frac{\|w\|}{2} + \boxed{C \sum_{i=1}^{n} \xi_i} \Rightarrow \text{Hinge Lors}$$

$\Downarrow$

Constraint

Error $\Rightarrow |y_i - w^T x_i| \leq \epsilon + \xi_i$

$\epsilon \Rightarrow$ Marginal Error

$\xi_i \Rightarrow$ Error above or below the marginal plane.

Can refer to this article for knowing more about SVR:

https://towardsdatascience.com/an-introduction-to-support-vector-regression-svr-a3ebc1672c2