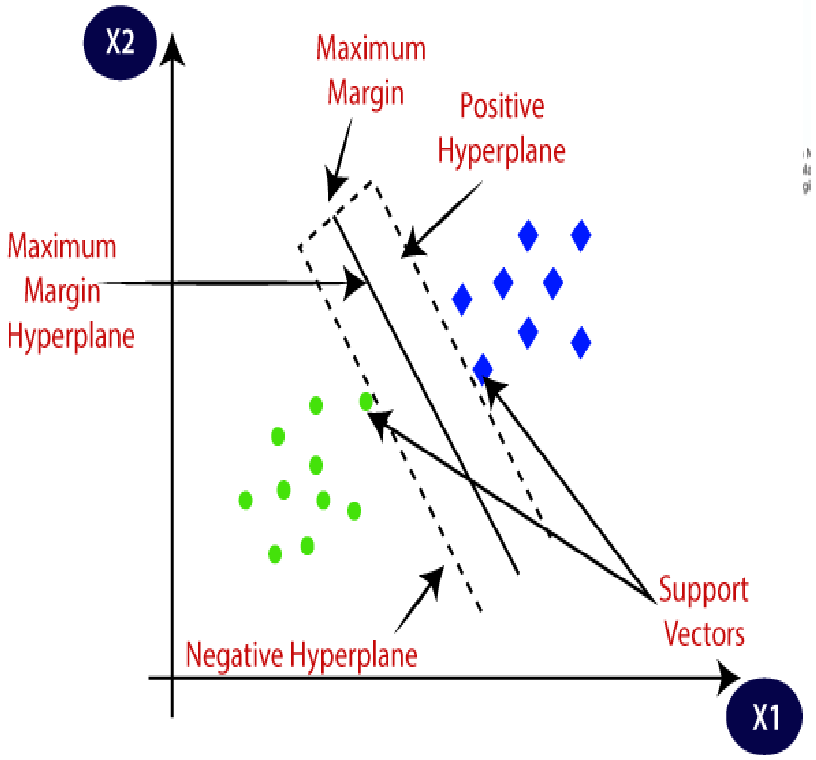
Support Vector Machine

Dr. Kuppusamy .P
Associate Professor / SCOPE

Support Vector Machine (SVM)

image source: javatpoint

- Support Vector Machine is a supervised learning algorithm used for Classification and Regression problems.
- However, primarily used for Classification problems in Machine Learning.



Goal:

- SVM creates the optimal decision boundary that isolate n-dimensional space into classes.
- So, the new data point can be classified into the correct category.
- The optimal decision boundary is called a hyperplane.
- SVM chooses the extreme data points/vectors that help in creating the hyperplane called as support vectors.

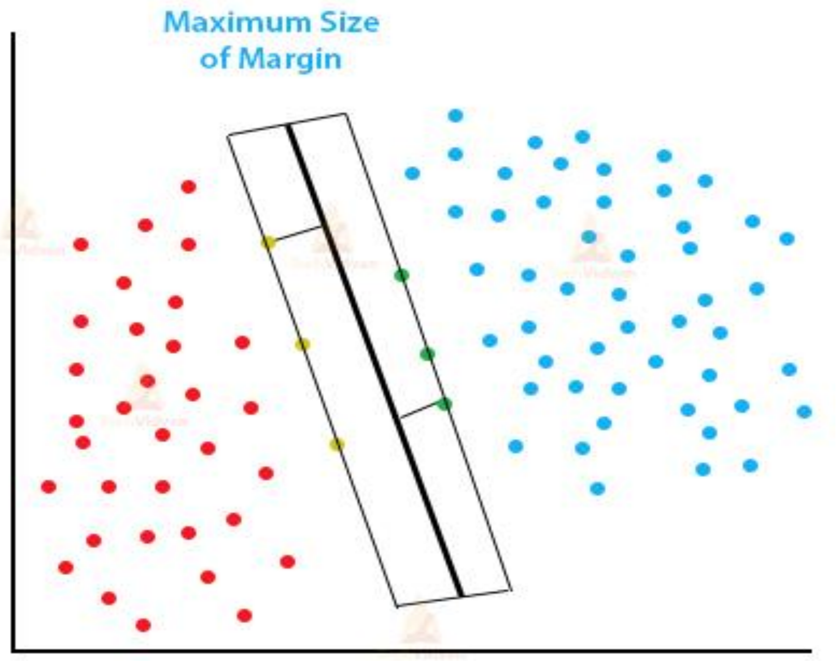
Hyperplane

image source: techvidvan

- The SVM needs hyperplane (central line) that is as far as possible from the closest member of each class.
- The hyperplane is the central line in the plot.
- In this plot, the hyperplane is a line because the dimension is 2-D.
- For 3-D plane, the hyperplane will be a 2-D plane.
- Let's consider a feature space (a blank piece of paper).
- Assume a line is cutting through it from the center. It is called the hyperplane.
- The equation for the hyperplane is a linear equation.

$$h(x) = y = \ w_0 + w_1\ x_1 + w_2\ x_2$$

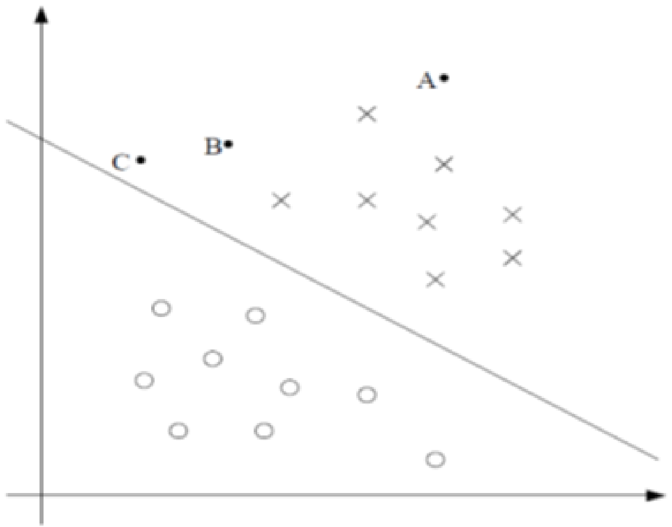
- w_0 is the intercept of the hyperplane. w_1 and w_2 define the first and second axes respectively. x_1 and x_2 are for two dimensions.



Classification

- When data points lie under the hyperplane then y < 0. When they are above the hyperplane then y >=0. This is how we classify data using a hyperplane. In SVM, If y = 1 then data is in class 1. If y = -1 then data is in class -1.

Margins: Intuition



- Let consider Linear SVM model for a binary classification problem with labels y and features x. Output represents as y ∈ {-1, 1} (instead of {0, 1}).
- The hypothesis function for a linear combination of the inputs

$$h_{w,b}(x) = w^T x + b$$

- Predict $$y'^{(i)} = sign(h(x^{(i)}; w, b)) = \begin{cases} -1, & if \;\; h(x^{(i)}; w, b) < 0 \\ 1, & if \;\; h(x^{(i)}; w, b) \geq 0 \end{cases}$$

- b performs the role of θ_0, and w takes the role of $[\theta_1 \ldots \theta_n]^T$
- Let's consider the plot in which 'x's represent positive training examples (y=1), 'o's denote negative training (y=0) examples.
- The decision boundary is the line denotes the equation $w^T x = 0$ (separating hyperplane).
- Three points have been labelled A, B and C.
- The point A is very far from the decision boundary (assume h(x) = 9). During the prediction, the value of y at A estimated as y = 1.
- But, the point C is very close to the decision boundary (assume h(x) = 0.1). The user predicts as y = 1 due to positive value. But a small change to the decision boundary could make a prediction to be y = 0.
- Hence, user has more confident about prediction at data point A than at C.

Mathematics behind the Functional Margin

Let consider two points x_1 & x_2.

$$w^T x_2 + b = +1 \quad \text{--- (1)}$$

$$w^T x_1 + b = -1 \quad \text{--- (2)}$$

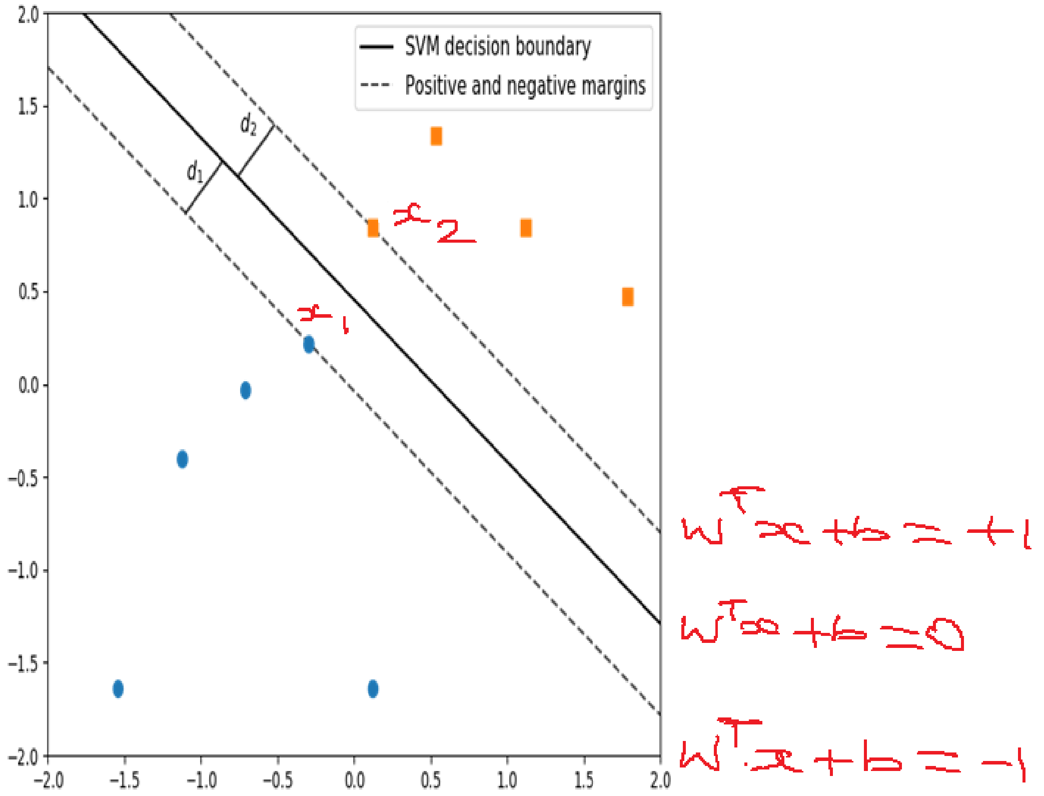
$$w^T (x_2 - x_1) = 2 \qquad \because (1) - (2)$$

$(x_2 - x_1)$ - distance between x_1 & x_2.

w^T is a normal vector. It can be written

$$\therefore \; w^T (x_2 - x_1) = 2$$

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



Here, Maximize the $\frac{2}{\|w\|}$ w.r.t w, b parameters.

Subject to $y^{(i)} \begin{cases} 1, & w^T x + b \geq 1 \\ -1, & w^T x + b \leq -1 \end{cases}$

Functional and Geometric Margin

- This leads to the idea of finding the parameters *(w, b)* that will maximize the values of h when $y^{(i)} = 1$, and minimize the values of h when $y^{(i)} = -1$.
- Goal of SVM is maximizing the minimal value of functional margin i.e., finding largest geometric margin.
- Functional margin provides the information that each point is properly classified or not.
- To maximize the **functional margin** $\hat{\gamma}$,

$$\hat{\gamma}^{(i)} = y^{(i)}\left(\mathbf{w}^T\mathbf{x}^{(i)} + b\right)$$

- Here, the issue is how the predicted class depends only on the sign of h i.e., we can scale (magnitude) the parameters to maximize the margin. Functional margin depends on the coefficient values. So, it varies when coefficient value (scaling) changes.
- E.g., (w, b) → (10w, 10b), without changing the predicted classes. It scales the values of h by a factor of 10 that provides the false idea that our model is 10 times more confident in its predictions.
- This issue is addressed by the geometric margin (scaled version of the functional margin). The geometric margin of $\hat{\gamma}$ is defined as the Euclidean distance of the i^{th} observation to the decision boundary.
- Geometric Margin $\gamma^{(i)} = y^{(i)}\left(\frac{\mathrm{w}^T x^{(i)} + \mathrm{b}}{\|w\|}\right)$
- It identifies the separating line that maximizes the minimum of geometric function.
- Unlike the functional margin, this measure is invariant to the scaling of parameters i.e., not depends on coefficient values.
- It provides the hyperplane defined by $w^T x + b = 0$ is exactly same as defined by $10w^T x + 10b = 0$.

Functional and Geometric Margin Optimization

- It maximizes the margin by adjusting the hyperplane and the decision boundaries to avoid mis-classification of any data point.

$$h(x) = w^T x^{(i)} + b$$

Identify the classification using Functional Margin: $y^{(i)}(w^T x^{(i)} + b)$

$$\max_{w,b} \hat{\gamma} = \frac{2}{\|w\|}$$

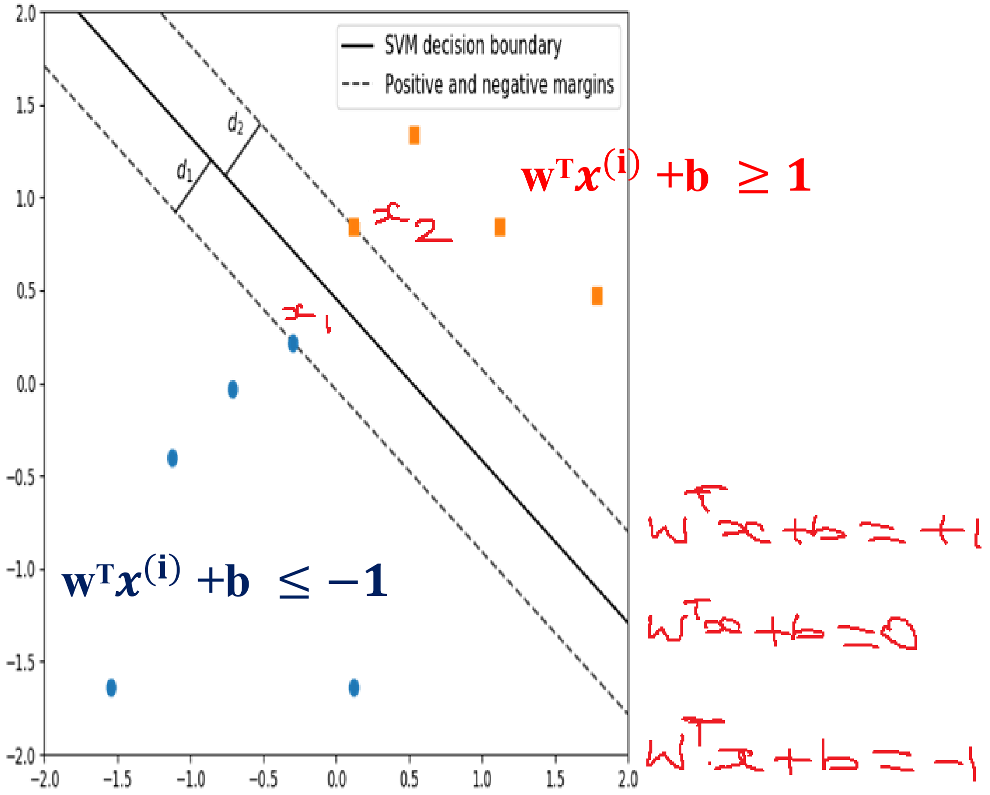
Here, $\|w\|$ is not differentiable at 0.

So, minimize the above function as

$$\min_{w,b} \gamma = \left(\frac{1}{2}\|w\|^2\right);$$

Subject to Condition

$y^{(i)}(w^T x^{(i)} + b) >= 1$; *for i =1 …. m*



Functional and Geometric Margin Optimization

- It maximizes the margin by adjusting the hyperplane and the decision boundaries to avoid mis-classification of any data point.

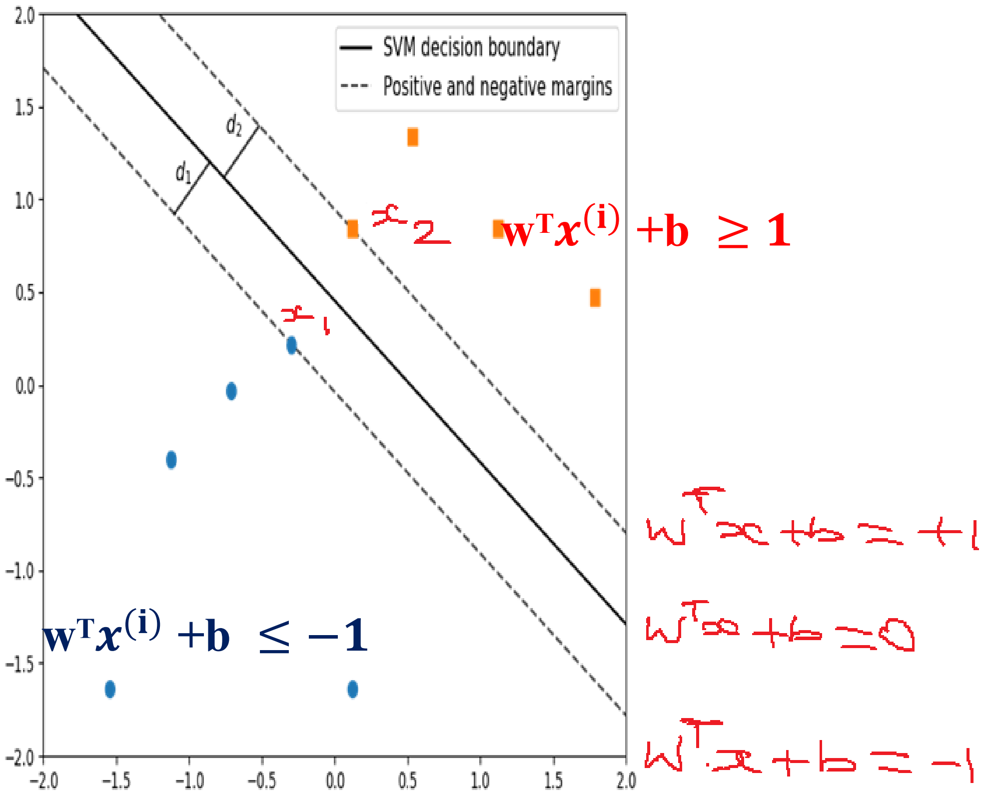
$$h(x) = w^T x^{(i)} + b$$

Identify the classification using Functional Margin: $y^{(i)}(w^T x^{(i)} + b)$

$$\max_{w,b} \hat{\gamma} = \frac{2}{\|w\|}$$

Here, $\|w\|$ is not differentiable at 0.

So, minimize the above function as



$$\underset{w,b}{\textit{Min}}\, \gamma = \left(\frac{1}{2}\|w\|^2\right);$$

Subject to Constraint $y^{(i)}(w^T x^{(i)} + b) >= 1$; *for i =1 .... m*

- Primal problem (a constrained minimization problem) can be expressed a dual problem (a constrained maximization problem).
- The solution to the dual problem provides a lower bound to the solution of the primal problem.

Maximum margin classifier (Hard Margin)

- It maximizes the margin by adjusting the hyperplane and the decision boundaries that ensures the classifier does not misclassify any data point.
- The hard margin works on the assumption that the data is **linearly separable.** It forces the model to correctly classify every data point on the training set.
- The i^{th} data point is correctly classified, if its functional margin is greater than zero:

$$\hat{\gamma}^{(i)} = y^{(i)}(w^T x^{(i)} + b) > 0$$

- From this equation, observes predicted classes depend only on the sign of *h*, and geometric margins are invariant to the scaling of parameters.
- Geometric and the functional margins are equal when $||w|| = 1$ to write the optimization objective of hard margin:

$$\underset{w,b}{MAX} \quad \gamma = y^{(i)}\left(\frac{w^T x^{(i)} + b}{||w||}\right)$$

Subject to Condition $\quad \boldsymbol{y^{(i)} w^T x^{(i)} + b >= \gamma}$, *for i =1 .... m* and

$$\boldsymbol{||w||=1}$$

- The first constraint $\boldsymbol{y^{(i)} w^T x^{(i)} + b >= \gamma}$ forces every data point to be correctly classified.
- The second constraint $\boldsymbol{||w||=1}$ forces γ to not only be a lower bound for the functional margin, but also for the geometric margin.
- It emphasizes the hard margin to maximize the minimum geometric margin without any misclassifications.

Maximum margin classifier (Hard Margin)

- If it provides best result, we could stop here. But unfortunately, the $||w|| = 1$ is a non-convex constraint.
- So we will need to make some changes to get this problem into a more friendly format.
- So, dividing the objective function by the norm i.e., if γ is a lower bound for the functional margin, then $\gamma/||w||$ is a lower bound for the geometric margin.
- So, $\underset{w,b}{MAX} \quad \gamma = y^{(i)}(\frac{w^T x^{(i)} + b}{\|w\|})$ can be written as:

$$\underset{w,b}{MAX} \;\; \gamma = \;(\frac{\hat{\gamma}}{\|w\|}) \;\; ;$$

Subject to Condition $\quad \boldsymbol{y^{(i)}(w^T x^{(i)} + b) >= \hat{\gamma},} \quad$ for $i = 1 \; .... \; m$

- Now it is the objective function that is non-convex, but we are one step closer.
- we can add arbitrary constraints to the parameters.
- So we can impose $\gamma = 1$ and it does not change the model and it can be satisfied by simply rescaling the parameters.

Maximum margin classifier (Hard Margin)

- New optimization function is then to maximize 1/||w||, which is equivalent to minimize ||w||.
- Since ||w|| is not differentiable at 0, instead we'll minimize (1/2)*||w||², whose derivative is just *w*.
- Optimization algorithms work much better on differentiable functions.
- Finally, we define the hard margin optimization function as:

$$\underset{w,b}{Min}\,\gamma = \left(\frac{1}{2}\|w\|^2\right) ;$$

Subject to Constraint $\quad y^{(i)} * (w^T x^{(i)} + b) >= 1$; *for i =1 .... m*

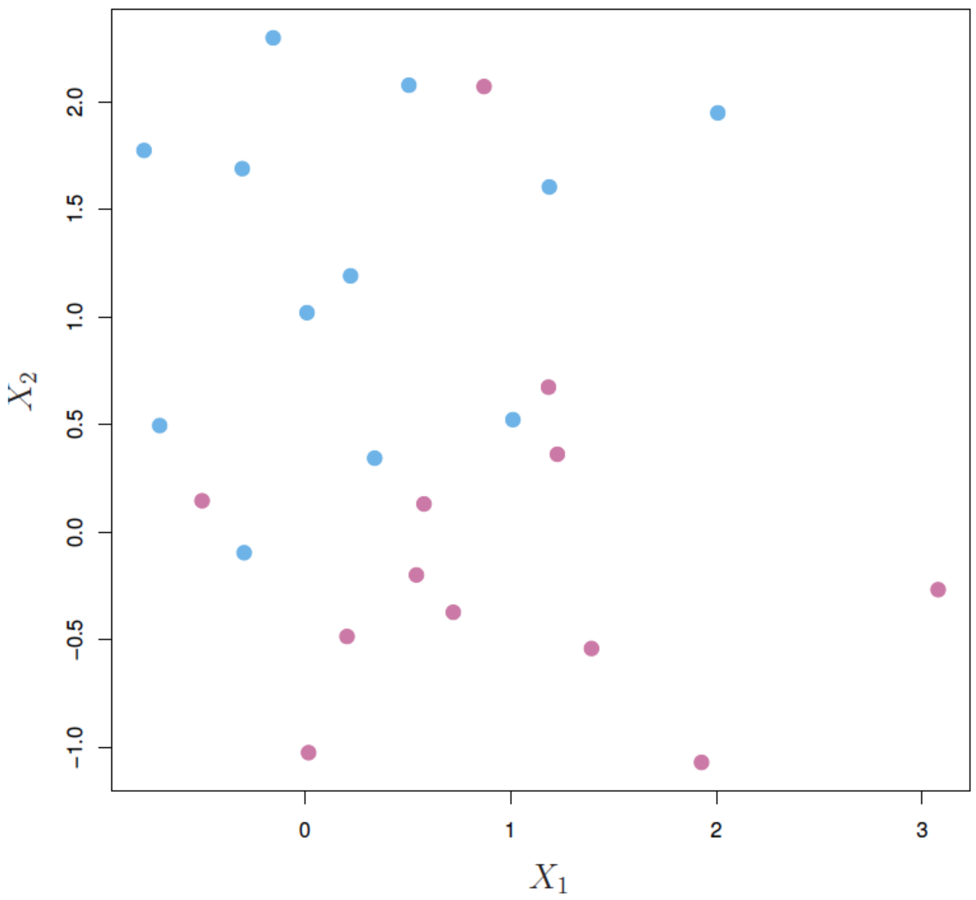
- The objective function to be minimized w and b. The constraint represents always function must return **>=1**.
- It emphasizes the hard margin to maximize the minimum geometric marg.in **without any misclassifications.**

Limitations of Hard Margin

- Hard Margin is infeasible if condition is not satisfied i.e., line cannot separate the non-linear data.
- It is sensitive to the outliers i.e., noisy data.
- So, need of Soft Margin approach.

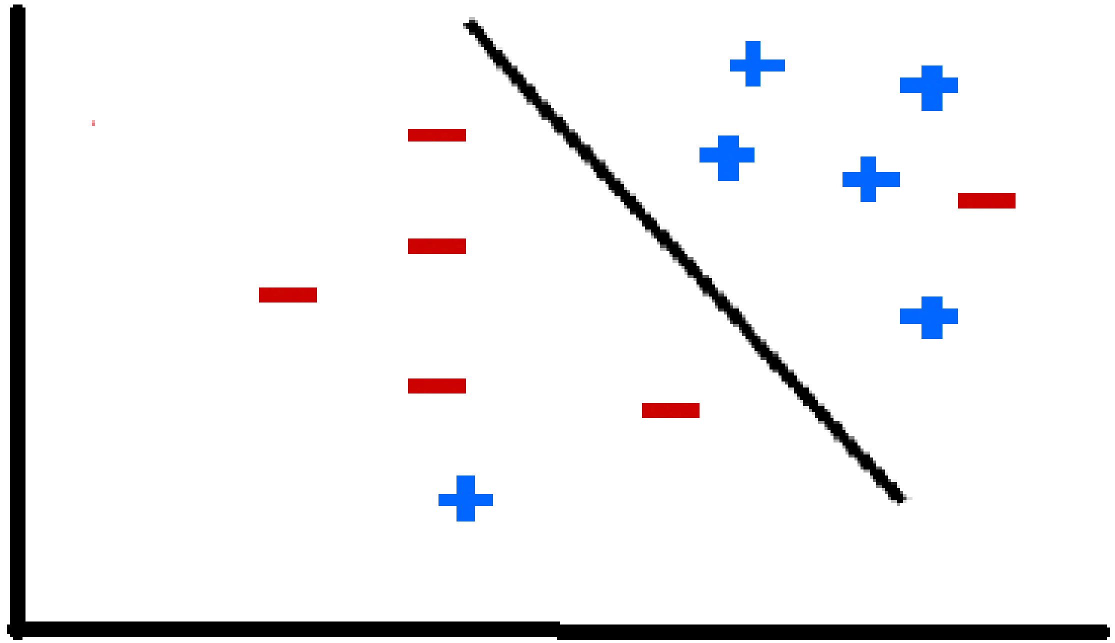
Maximum margin classifier (Hard Margin)

- Hard Margin is infeasible if condition is not satisfied i.e., line cannot separate the non-linear data.



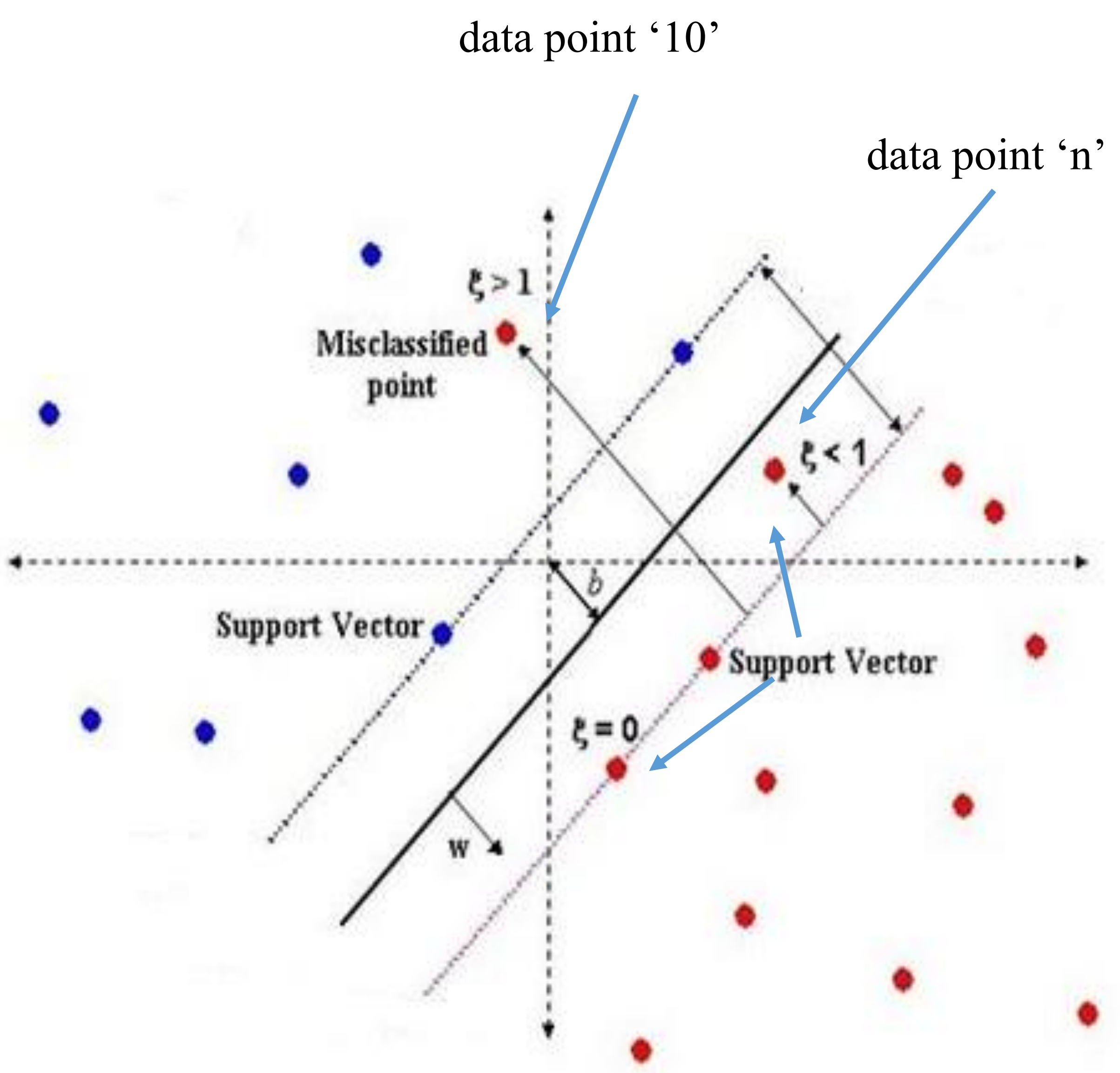
Maximum margin classifier (Hard Margin)

- Hard Margin is sensitive to the outliers i.e., noisy data.



Soft Margin

- The generalized (robust) model may allow few misclassifications, It almost classifies all the data points.
- The generalization of the maximal margin classifier using soft margin is called **Support Vector Classifier (SVC).**
- It can be done by adding **slack variables** to the objective function.
- Every data point (observation) contain its own slack measure that allows few observations to fall on the wrong side of the margin. But penalized by parameter C (cost of misclassification)..
- In soft margin, the data point ‘n’ also act as a support vector (locates in correct side of hyperplane. But wrong side of negative margin).
- **Misclassified data point 10:** Located in Wrong side of hyper plane.



Soft Margin

- New constraint can be rewritten as:

$$y^{(i)} * \left(w^T x^{(i)} + b\right) \geq 1 - \epsilon^{(i)}$$

- Here, function should estimate one extra variable for every observation.
- New objective function also should allow the margins to be as wide as possible, and the slack variables to be as small as possible to prevent margins violations.
- Re-write **soft margin SVM** classifier objective:

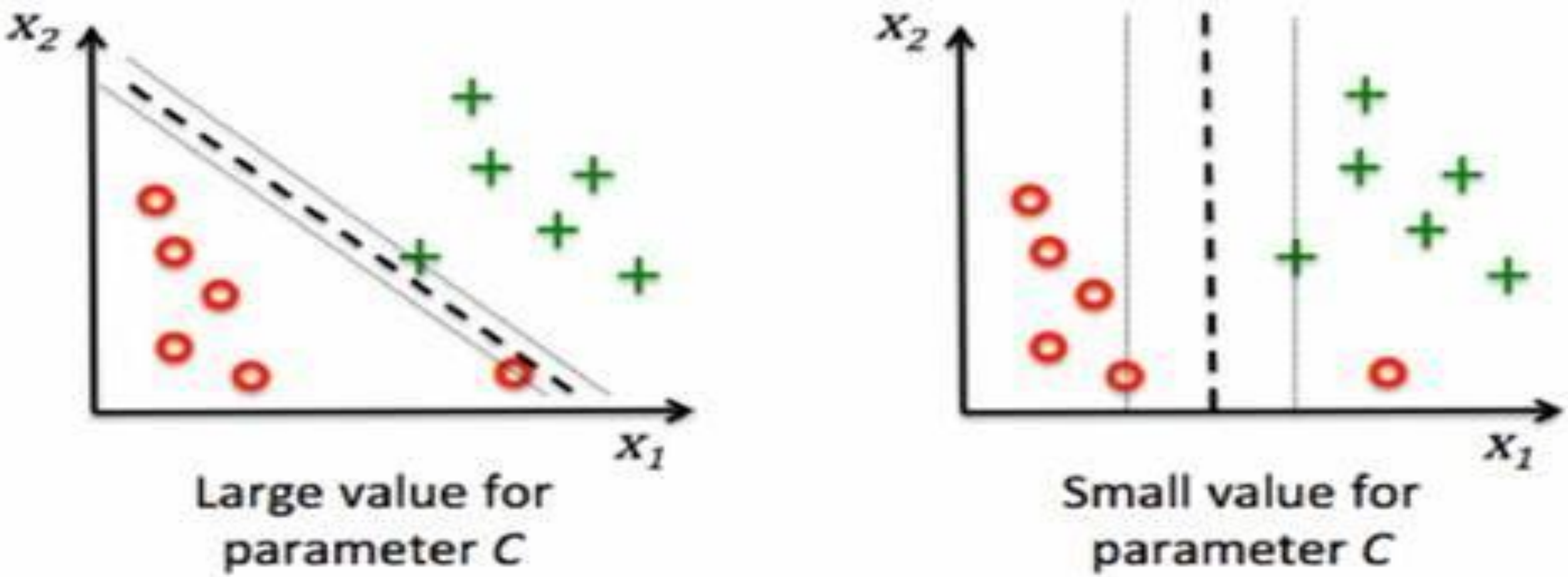
$$\underset{\mathbf{w},b,\epsilon}{\text{minimize}} \quad \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{m}\epsilon^{(i)}$$

$$\text{subject to} \quad y^{(i)}(\mathbf{w}^T\mathbf{x}^{(i)} + b) \geq 1 - \epsilon^{(i)} \;\; \text{and} \;\; \epsilon^{(i)} \geq 0 \;\; \text{for} \;\; i = 1, \ldots, m$$

- C is a penalty hyperparameter that controls the tradeoff between a wider margin and a lower total error penalty.
- When C increases, it forces the optimization algorithm to find smaller values for ϵ.

Soft Margin

- Sum of slack is total distance of the points that are in wrong side of the margin.
- If C is high, narrow margin. Results low bias and high variance
- If C is small, wider margin. Results high bias and low variance.



- Even sometime, Soft Margin SVM will **not provide optimal Solution.** In this case, should use **Kernel function.**

Difference between Hard Margin & Soft Margin

- If data is linearly separable, apply a hard margin. Otherwise, a soft margin SVM is appropriate that allows few misclassifications.
- Sometimes, the data is linearly separable, but the margin is so small that the model becomes prone to overfitting or being too sensitive to outliers. In this case, should choose a larger margin using soft margin SVM in order to generalize the model.

Optimization Problems with Constraints

- Optimization function in Linear Regression is MSE i.e., Minimize the loss by varying the parameters *w*.
- Optimization function in Logistic Regression is Binary cross entropy i.e., Minimize the loss by varying the parameters *w*.
- These Linear and Logistic Regression Optimization functions does not have any Constraints.
- But SVM contains the constraints.
- Intuition of **Optimization Problems with Constraints:**
- Let's consider the equation

$$\max_{x,y} z = x^2 y$$

subject to $x^2 + y^2 = 1$

In the plot,

Apply $\max_{x,y} z = x^2 y$

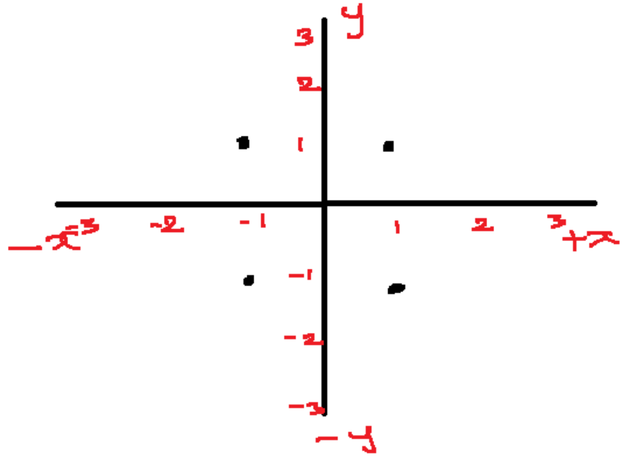
without constraints.

Assume $x = 1$ & $y = 1$, $z = 1$

$x = 1$ & $y = -1$, $z = -1$

$x = -1$ & $y = 1$, $z = 1$

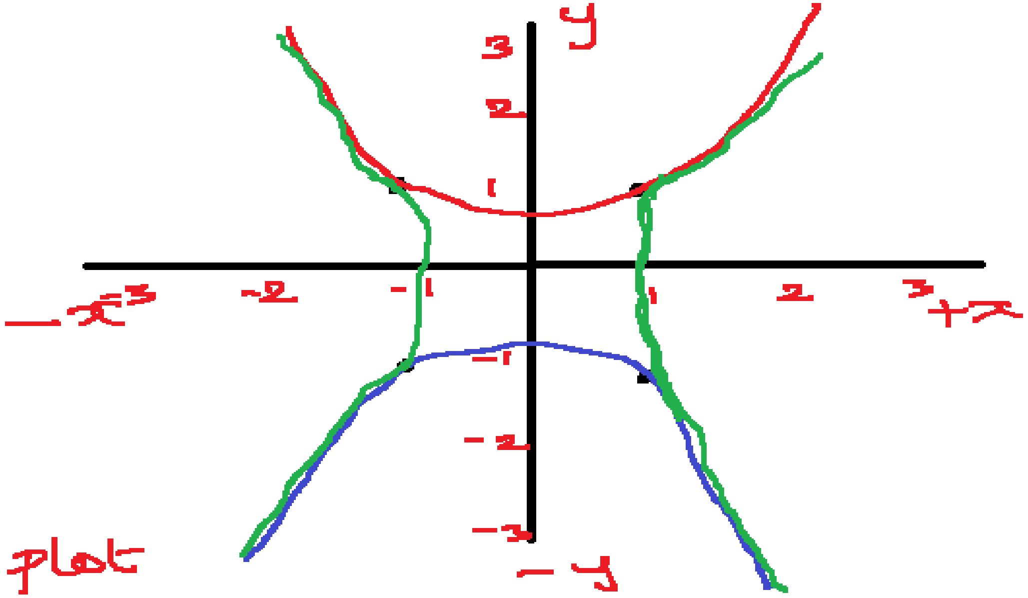
$x = -1$ & $y = -1$, $z = -1$



Intuition of **Optimization Problems with Constraints:**

Need of Constraints

- When x and y values are increases into infinity, function grows into infinity.
- So, need to apply some constraints.



2D plot

If $x = +1$ & $y = +1$, function grows in +ve y direction.

If x & y values are -ve, function grows in -ve direction.

similarly, $x = -1$ & $y = +1$ and $x = -1$ & $y = +1$

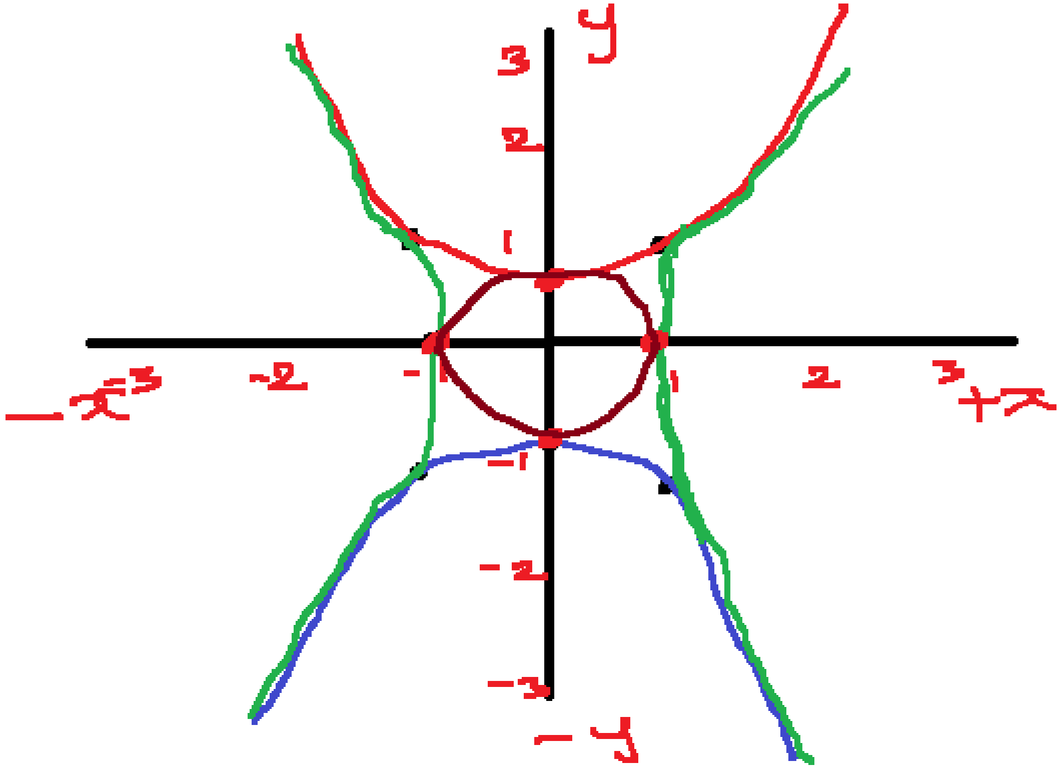
Intuition of **Optimization Problems with Constraints:**

In the plot,

Apply $\underset{x,y}{\text{Max}}\ z = x^2 y$

with constraints

$x^2 + y^2 = 1$ ($\because$ Circle)



- Optimization represents that the points (z value) which are satisfies (touches) the both optimization functions and Constraint functions.
- Maximization represents selects maximum z value among multiple z values. i.e.,, if 50 ‘z’ values are satisfying the both functions, select the maximum values.
- Solve this Optimization Problem with Constraints by applying the Lagrange multiplier.

Lagrange Multipliers - Optimization Problems with Equality Constraints

- Optimization function that maximizes the distance by minimize the square of denominator provides quadradic equation.
- So, $\underset{w,b}{\mathbf{Min}}\gamma = (\frac{1}{2}\|w\|^2)$; Subject to Constraint $y^{(i)} * (w^T x^{(i)} + b) \geq +1$ *;for i =1 .... m*
- To solve this quadratic programming problem with equality constraints, apply Lagrange multipliers.
- The Lagrange function: $\mathit{Optimization\ function} - \sum_{i=1}^{m} \alpha_i * \mathit{Constrain\ function}_i$

$$\underset{w,b}{\mathbf{Min}}\ \mathcal{L}(\mathrm{w}, \mathrm{b}, \alpha) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{m} \alpha_i \left[y^{(i)} * \left(w^T x^{(i)} + b\right) - 1 \right]$$

- α is Lagrange multiplier
- Solve this by applying $\frac{\partial L}{\partial w} = 0$, $\frac{\partial L}{\partial \alpha} = 0$, $\frac{\partial L}{\partial b} = 0$,

Example - Optimization Problems with Equality Constraints using Lagrange Multipliers

Let's $\max\limits_{w,b} \; 200\, w^{2/3} b^{1/3}$

equality constraint

subject to constraint $\quad 20w + 170b = 20000$

Solve this by convert into Lagrange equation

$$L(w,b,\alpha) = 200\, w^{2/3} b^{1/3} - \alpha\,(20w + 170b - 20000)$$

$$\frac{\partial L}{\partial w} = 200\,\frac{2}{3}\, w^{-1/3} b^{1/3} - 20\alpha = 0$$

$$\frac{\partial L}{\partial b} = 200\,\frac{1}{3}\, w^{2/3} b^{-2/3} - 170\alpha = 0$$

$$\frac{\partial L}{\partial \alpha} = -20w - 170b + 20000 = 0$$

By solving it, $w = 666.66$; $b = 39.12$; $\alpha = 2.59$

$\max\limits_{w,b} L = 51777$

Karush-Kuhn-Tucker (**KKT**)

- KKT theorem solves **Optimization Problems with inequality Constraints.**
- KKT theorem implicitly defines a dual problem.
- Optimization function = $y^{(i)} * y' (x_i) - 1$, and Lagrange parameter is α.
- Karush-Kuhn-Tucker (KKT) constraints given below should be satisfied by optimization problem:
  - $\alpha_i \geq 0$; Lagrange parameter
  - $y^{(i)} * y' (x_i) - 1 \leq 0$; Optimization function
  - $\alpha_i(y^{(i)} * y' (x_i) - 1) = 0$

Interpretation of KKT Constraints:

The KKT conditions dictate that for each data point one of the following is true:

- The Lagrange multiplier is zero, i.e., $\alpha_i = 0$. This point, plays no role in classification. (or)
- $y^{(i)} * y' (x_i) = 1$ and $\alpha_i > 0$: In this case, the data point has a role in deciding the value of *w*.
- Such a point is called a support vector.

Karush-Kuhn-Tucker (**KKT**)

- KKT theorem solves **Optimization Problems with inequality Constraints**
- Steps
- Convert the Maximization equation into Lagrange equations

$$\textit{Optimization function} - \sum_{i=1}^{m} \alpha_i * \textit{Constrain function}_i$$

- Apply the partial derivative with respective to variables w, b, α and equate to 0.
- $\text{i.e., applying } \frac{\partial L}{\partial w} = 0,\ \frac{\partial L}{\partial \alpha} = 0,\ \frac{\partial L}{\partial b} = 0$
- Apply various KKT constraints
  - $\alpha_i \geq 0$; Lagrange parameter
  - $y^{(i)} * y'(x_i) - 1 \leq 0$; Optimization function
  - $\alpha_i(y^{(i)} * y'(x_i) - 1) = 0$
- Find x values to finalize the maximum optimal value.

Karush-Kuhn-Tucker (**KKT) Example**

Max $-x_1^2 - x_2^2 - x_3^2 + 4x_1 + 6x_2$

subject to constraints

$$x_1 + x_2 \le 2$$

$$2x_1 + 3x_2 \le 12$$

here $x_1, x_2 \ge 0$.

step 1:

convert optimization function into Lagrange equation

$$L(x_1, x_2, x_3, \alpha_1, \alpha_2) = -x_1^2 - x_2^2 - x_3^2 + 4x_1 + 6x_2 - \alpha_1(x_1 + x_2 - 2) - \alpha_2(2x_1 + 3x_2 - 12)$$

step 2: Apply partial derivatives

$$\frac{\partial L}{\partial x_1} = -2x_1 + 4 - \alpha_1 - 2\alpha_2 = 0 \quad \text{--- (1)}$$

$$\frac{\partial L}{\partial x_2} = -2x_2 + 6 - \alpha_1 - 3\alpha_2 = 0 \quad \text{--- (2)}$$

$$\frac{\partial L}{\partial x_3} = -2x_3 = 0 \quad \text{i.e } x_3 = 0 \quad \text{--- (3)}$$

Karush-Kuhn-Tucker (**KKT) Example**

Step3: Apply the KKT constraints

(i) $\alpha_1(x_1+x_2-2)=0$ —— (4)

$\alpha_2(2x_1+3x_2-12)=0$ —— (5)

(ii) $x_1+x_2-2\leq 0$ —— (6)

$2x_1+3x_2-12\leq 0$ —— (7)

(iii) $\alpha_1\geq 0$ & $\alpha_2\geq 0$

case 1: $\alpha_1=0$ & $\alpha_2=0$

substitute in (1) & (2)

$x_1=2$; $x_2=3$

Substitute x_1 & x_2 in (6) & (7)

$x_1+x_2-2\leq 0$

$5-2\leq 0$

$3\leq 0$

condition is False. Not satisfied.

$2x_1+3x_2-12\leq 0$

$1\leq 0$; False.

Constraint is not satisfied.

So, select another case $\alpha_1\neq 0$ & $\alpha_2\neq 0$

i.e $x_1+x_2-2=0$

$2x_1+3x_2-12=0$

By solving $x_1=-6$ & $x_2=8$

Substitute in (1) & (2)

$\alpha_1+2\alpha_2=16$

$\alpha_1+3\alpha_2=-10$

$\boxed{\alpha_2=-26}$ constraint Not satisfied.

i.e. α_i should be ≥ 0

Karush-Kuhn-Tucker (**KKT) Example**

Step3: Apply the KKT constraints

(i) $\alpha_1(x_1+x_2-2)=0$ —— ④

$\alpha_2(2x_1+3x_2-12)=0$ —— ⑤

(ii) $x_1+x_2-2\leq 0$ —— ⑥

$2x_1+3x_2-12\leq 0$ —— ⑦

(iii) $\alpha_1\geq 0$ & $\alpha_2\geq 0$

Case 1: $\alpha_1=0$ & $\alpha_2=0$

Substitute in ① & ②

$x_1=2$; $x_2=3$

Substitute x_1 & x_2 in ⑥ & ⑦

$x_1+x_2-2\leq 0$

$5-2\leq 0$

$3\leq 0$

condition is False. Not satisfied.

$2x_1+3x_2-12\leq 0$

$1\leq 0$; False.

Constraint is not satisfied.

So, select another case $\alpha_1\neq 0$ & $\alpha_2\neq 0$

Case 2:

i.e. $x_1+x_2-2=0$

$2x_1+3x_2-12=0$

By solving $x_1=-6$ & $x_2=8$

Substitute in ① & ②

$\alpha_1+2\alpha_2=16$

$\alpha_1+3\alpha_2=-10$

$\boxed{\alpha_2=-26}$ Constraint Not satisfied.

i.e. α_i should be ≥ 0

So, select another case $\alpha_1=0$ & $\alpha_2\neq 0$

Karush-Kuhn-Tucker (**KKT**) **Example**

Case 3: $\alpha_1 = 0$ & $\alpha_2 \neq 0$

substitute in 1 & 2

$$-2x_1 + 4 - 2\alpha_2 = 0$$

$$-2x_2 + 6 - 3\alpha_2 = 0$$

By solving $x_1 = \frac{2}{3} x_2$

Substitute $\alpha_1 = 0, \alpha_2 \neq 0$ in 4 & 5

$$2x_1 + 3x_2 - 12 = 0$$

$$\frac{4}{3} x_2 + 3x_2 - 12 = 0$$

$$x_2 = 3 \quad \& \quad x_1 = 2$$

$$x_1 + x_2 - 2 \leq 0$$

$\boxed{5 - 2 \leq 0 \quad \text{False}}$

$$2x_1 + 3x_2 - 12 \leq 0$$

$\boxed{1 \leq 0 \quad \text{False}}$

So, select another case $\alpha_1 \neq 0$ & $\alpha_2 = 0$

Case 4: $\alpha_1 \neq 0$ & $\alpha_2 = 0$ in 1 & 2

$$x_1 = \frac{1}{2} \; ; \; x_2 = \frac{3}{2}$$

$$\alpha_1 = 3 \; ; \; \alpha_2 = 0$$

Substitute

$$x_1 + x_2 - 2 \leq 0$$

$\boxed{0 \leq 0}$ True

$$2x_1 + 3x_2 - 12 \leq 0$$

$\boxed{-13 \leq 0}$ True

This case provides $x_1 = \frac{1}{2}$ & $x_2 = \frac{3}{2}$ to maximize an optimization function.

$\therefore$ Max $z = -x_1^2 - x_2^2 - x_3^2 + 4x_1 + 6x_2$

$$= -\frac{1}{4} - \frac{9}{4} - 0 + \frac{4}{2} + \frac{18}{2} = \frac{-10}{4} + 11$$

$$= \frac{17}{2}$$

Primal Form and Dual Form

- SVM is defined in two different approaches:
  Primal form and Dual form.
- Both provides the similar optimization result and solves quadradic equations. But both approaches are very different.

Primal form

- Primal problem is a **constrained minimization** problem that classifies each data point by transforming from lower dimension to the higher dimension by adding relevant features.
- Duality is defined as optimization problems may be viewed either of two perspectives, the primal problem or the dual problem
- Primal (**minimization**) problem can be expressed aa a dual (maximization) problem.
- Primal mode is preferred when no need to apply kernel trick to the data and the dataset is large but the dimension of each data point is small.

Dual Form

- Dual form is a **maximization** problem.
- Also, It is a convex problem that uses Lagrange multipliers to solve the equation.
- The solution to the dual problem provides a lower bound to the solution of the primal (minimization) problem.
- Solving the dual problem is simpler than solving the primal problem
- Dual form is preferred when data has a huge dimension, and we need to apply the kernel trick.

Primal and Dual Problem for SVM

$\underset{\omega}{\text{Min}}\ f(\omega)$ $f(\omega)$ is optimization fun

S.T $g_i(\omega) \leq 0$; $i = 1 \ldots k \rightarrow$ Inequality condition

$h_i(\omega) = 0$; $i = 1 \ldots l \rightarrow$ equality condition

Frame a Generalized Lagrange function

$$L(\omega, \alpha, \beta) = f(\omega) + \sum_{i=1}^{k} \alpha_i g_i(\omega) + \sum_{i=1}^{l} \beta_i h_i(\omega)$$

Now, Define optimization function $\theta_p(\omega)$

$$\theta_p(\omega) = \underset{\substack{\alpha, \beta \\ \alpha \geq 0}}{\text{Max}}\ L(\omega, \alpha, \beta)$$

$$\therefore \theta_p(\omega) = \underset{\substack{\alpha, \beta \\ \alpha \geq 0}}{\text{Max}}\ f(\omega) + \sum_{i=1}^{k} \alpha_i g_i(\omega) + \sum_{i=1}^{l} \beta_i h_i(\omega)$$

(i) if $g_i(\omega) > 0$ i.e., violates given constraint

The term $\alpha_i g_i(\omega)$ becomes ∞ for large α_i.

$\therefore \theta_p(\omega) = \infty$

(ii) Also, if $h_i(\omega) \neq 0$, i.e. violates the constraint

$\beta_i h_i(\omega) = \infty$ for large $-$ve β_i & $-$ve $h_i(\omega)$

& for large $+$ve β_i & $+$ve $h_i(\omega)$

$\therefore \theta_p(\omega) = \infty$

(iii) if $g_i(\omega) \leq 0$ & $h_i(\omega) = 0$ i.e. satisfies constraints

$$\theta_p(\omega) = \underset{\substack{\alpha, \beta \\ \alpha \geq 0}}{\text{Max}}\ f(\omega) + \sum_{i=1}^{k} 0(g_i(\omega)) + \sum_{i=1}^{l} \beta_i(0)$$

$$\theta_p(\omega) = f(\omega)$$

So,

$$\theta_p(\omega) = \begin{cases} f(\omega) & \text{if constraints satisfied} \\ \infty & \text{if violates the constraints} \end{cases}$$

$$p^* = \underset{\omega}{\text{Min}}\ \theta_p(\omega)$$

Both $\underset{\omega}{\text{Min}}\ f(\omega)$ & p^* offers similar results.

Primal and Dual Problem for SVM

$$p^* = \min_{\omega} \max_{\substack{\alpha,\beta \\ \alpha \geq 0}} L(\omega,\alpha,\beta)$$

Dual Problem:

$$d^* = \max_{\substack{\alpha,\beta \\ \alpha \geq 0}} \min_{\omega} L(\omega,\alpha,\beta)$$

$$= \max_{\substack{\alpha,\beta \\ \alpha \geq 0}} \theta_d(\alpha,\beta)$$

Fact

$$\max \min f(x) \leq \min \max f(x)$$

E.g., $\max_x \min_y \sin(x+y) \leq \min_y \max_x \sin(x+y)$

$$\max_x -1 \leq \min_y 1$$

$$-1 \leq 1$$

$$d^* \leq p^*$$

But under some conditions like Convex

(i) $d^* = p^*$

$\exists\, \omega^* < \beta^*$; ω^* - Solution to primal

$\alpha^* \beta^*$ - Solution to dual.

(ii) $p^* = d^*$

$\omega^* \alpha^* \beta^*$ Satisfies KKT condition.

(i) Apply partial derivative w.r.t $\alpha, \beta, \omega_i = 0$

(ii) $\alpha_i g_i(\omega) = 0$

(iii) $g_i(\omega) \leq 0$

(iv) $\alpha_i \geq 0$

VIT-AP UNIVERSITY

References

1. Tom M. Mitchell, Machine Learning, McGraw Hill , 2017.
2. EthemAlpaydin, Introduction to Machine Learning (Adaptive Computation and Machine Learning), The MIT Press, 2017.
3. Wikipedia
4. https://www.svm-tutorial.com/2016/09/duality-lagrange-multipliers/