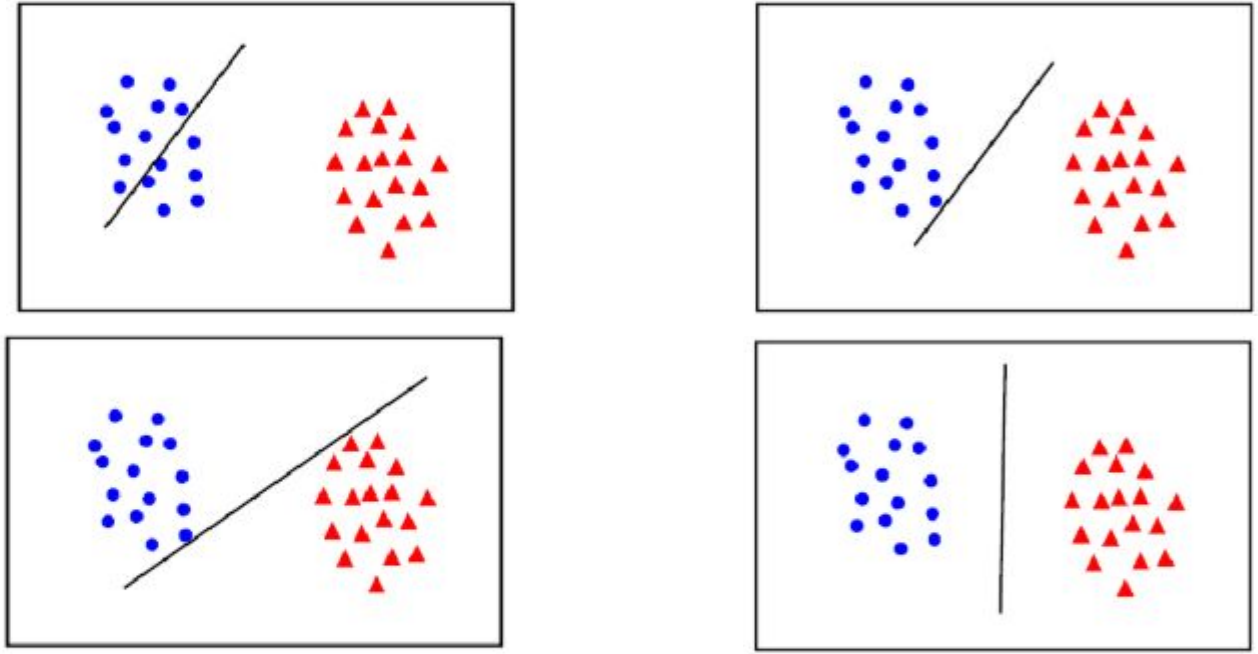
Support Vector Machines

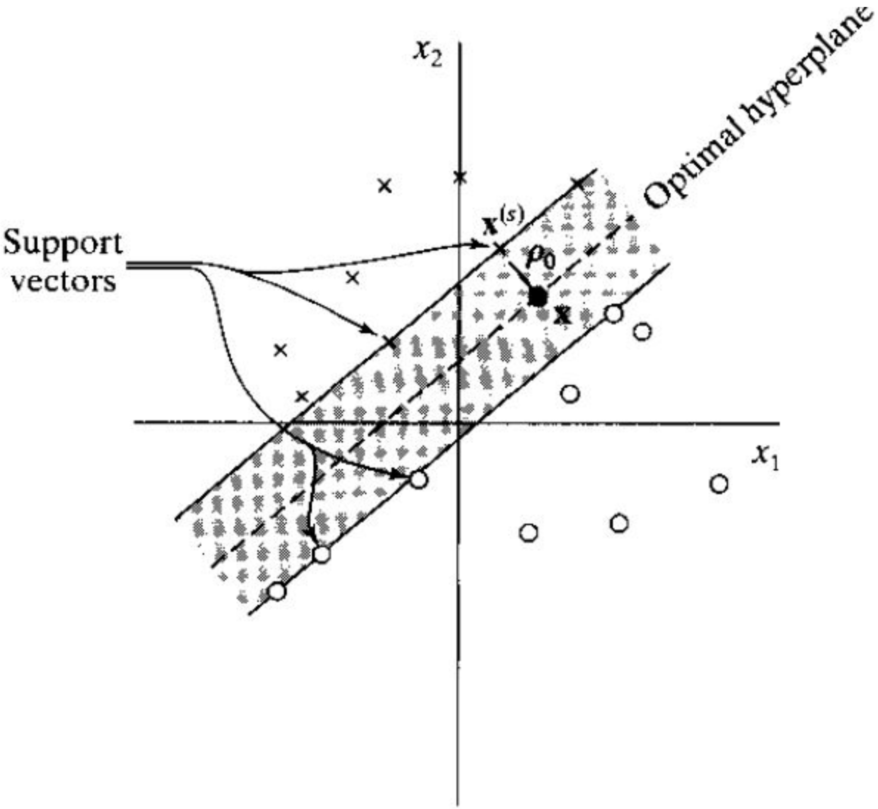
INDRAPRASTHA INSTITUTE *of*
INFORMATION TECHNOLOGY
DELHI

Optimum Separation Hyperplane



- Optimum separation hyperplane (OSH) is the linear classifier with the maximum margin for a given finite set of learning patterns.
- Better generalization!

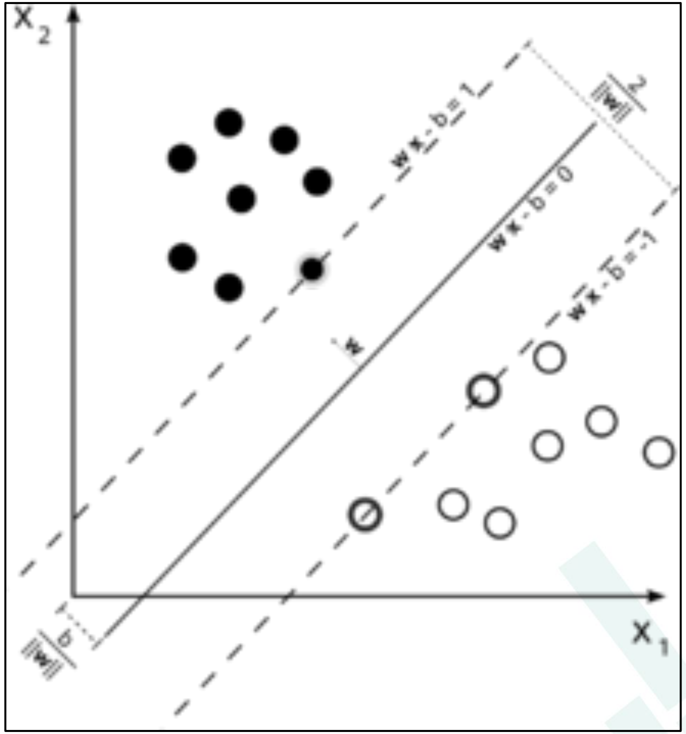
Optimum Separation Hyperplane



- *Margin of separation*: distance to the closest example
- For the optimal hyperplane
  - distance to the closest negative example = distance to the closest positive example
- The goal of SVM is to find the particular hyperplane for which the *margin of separation* is maximized.

Support Vectors

- *Support vectors* are the samples closest to the separating hyperplane
  - They are the most difficult patterns to classify.
- Optimal separation hyper-plane is completely defined by *support vectors*



Optimal Hyperplane: Problem Formulation

- Training Set: D = {(x_i,d_i); i = 1, 2, ..., n}
- Linearly Separable:
  - The Decision Boundary [Hyperplane]

$$\sum_{i=1}^{m} w_i x_i + b = W^T x + b = 0$$

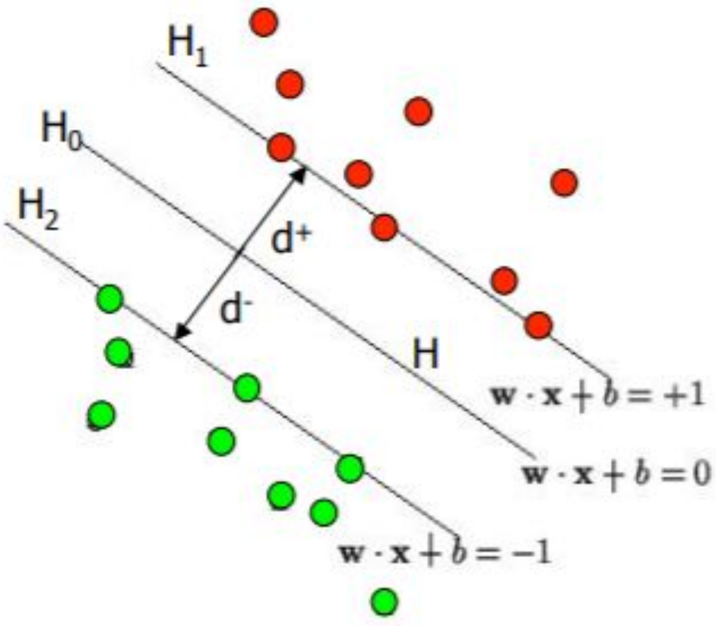
  - Correct Classification

$$w^T x_i + b \geq 0; \forall y_i = +1$$
$$w^T x_i + b < 0; \forall y_i = -1$$

- Infinitely many hyperplanes exist
  - Which is the optimal?

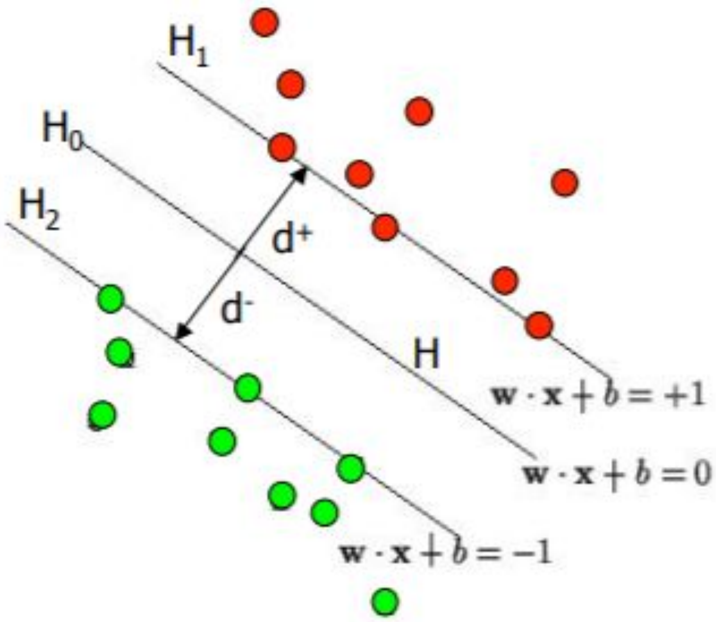
Margin of Separation

- NO training patterns exist between the two hyperplanes:
  - H_1: wx + b = 1, y = 1
  - H_2: wx + b = -1, y = -1
- The points on the planes H_1 and H_2 are the Support Vectors
- d+ = the shortest distance to the closest positive point
- d- = the shortest distance to the closest negative point
- The margin *m* of a separating hyperplane is (d+) + (d–)



Maximizing the margin

- We want a classifier (linear separator) with as big a margin as possible.
- Distance from a point(x_0,y_0) to a Line $Ax+By+c = 0$ is
  - $|Ax_0 +By_0 +c|/\mathrm{sqrt}(A^2+B^2)$
- The distance between H_0 and H_1
  - $|w\bullet x+b|/||w||=1/||w||$
- The total distance *m* between H_1 and H_2:
  - $2/||w||$



Quadratic Programming Problem

- When $// w //$ =1 then m=2
- When $// w //$ =2 then m=1
- When $// w //$ =4 then m=1/2
- The bigger the norm is, the smaller the margin become.
- *Maximize 2/ $// w //$*
  - *Minimize $// w //$ /2*
  - *= Minimize ½ $// w //^2$*
- ***Minimize f: ½$||w||^2$ s.t. g: $y_i[w \cdot x_i + b] >= 1$***
- This is a constrained optimization problem
  - It can be solved by the Lagrangian multiplier method
  - Because it is quadratic, the surface is a paraboloid, with just a single global minimum

SVM: Constrained Optimization Problem

- Given the training sample $\{(x_i, y_i)\}^n_{i=1}$, find the optimum values of the weight vector **w** and bias **b** such that they satisfy the constraints
  - $y_i*(\mathbf{w}^T\mathbf{x_i} + b) >= 1, \forall\ i=1,2, \ldots n$
    - Equality is true for support vector points and greater than condition holds true for non-support vector points.

  and the weight vector **w** minimizes the cost function:
  - $\Phi(\mathbf{w}) = ½ // w //^2 = \mathbf{1/2w^Tw}$

- Constraints are linear
- Cost function is convex

Constrained Optimization

- *Lagrangian function: Constrained optimization can be solved through unconstrained optimization*
  - $L(x,y,\alpha)=f(x,y)-\alpha g(x,y)$
    - α: *Lagrange multipliers*
- *Solution:* $\nabla L(x,y,\alpha) = 0$
  - $\partial L(x,y,\alpha)/\partial x = 0$
  - $\partial L(x,y,\alpha)/\partial y = 0$
  - $\partial L(x,y,\alpha)/\partial a = 0$

Lagrange Multiplier: Primal Form

- $J(w,b,\alpha) = ½\, \mathbf{w}^T\mathbf{w} - \boldsymbol{\Sigma}^n{}_{i=1}\alpha_i[y_i*(\mathbf{w}^T\mathbf{x_i} + b)]$
  - Inequality constraints -> equality constraints
  - $J(w,b,\alpha) = ½\, \mathbf{w}^T\mathbf{w} - \boldsymbol{\Sigma}^n{}_{i=1}\alpha_i[y_i*(\mathbf{w}^T\mathbf{x_i} + b) - 1]$
- Karush-Kuhn-Tucker Condition
  - Multipliers that can assume non-zero values ($\alpha > 0$), must satisfy following conditions:
    - $\alpha_i[y_i*(\mathbf{w}^T\mathbf{x_i} + b) - 1] = 0$

Lagrange Multiplier

- *Solution:* $\nabla L(x,y,\alpha) = 0$
  - *Conditions of optimality*
    - $\partial J(\mathbf{w},b,\alpha)/\partial \mathbf{w} = 0$
    - $\partial J(\mathbf{w},b,\alpha)/\partial b = 0$
- $J(w,b,\alpha) = ½\, \mathbf{w}^T\mathbf{w} - \boldsymbol{\Sigma}^n_{i=1}\alpha_i[y_i*(\mathbf{w}^T\mathbf{x}_\mathbf{i} + b) - 1]$
  - $½\, \mathbf{w.w} - \boldsymbol{\Sigma}^n_{i=1}\alpha_i y_i \mathbf{w.x}_\mathbf{i} - \boldsymbol{\Sigma}^n_{i=1}\alpha_i y_i b + \boldsymbol{\Sigma}^N_{i=1}\alpha_i$
- $\mathbf{w} = \Sigma^n_{i=1}\alpha_i y_i \mathbf{x}_\mathbf{i}$ **.....(A)**
- $\Sigma^n_{i=1}\alpha_i y_i = 0$ **.....(B)**

Duality theorem (Bertsekas, 1995)

- If the primal problem has an optimal solution, the dual problem also has an optimal solution, and the corresponding optimal values are equal.
  - $\Phi(\mathbf{w}_o) = J(\mathbf{w}_o, b_o, \alpha_o) = \min J(w, b, \alpha)$
- Applying (B) and then (A) to Lagrangian equation:
  - $J(w,b,\alpha) = \frac{1}{2}\,\mathbf{w}.\mathbf{w} - \boldsymbol{\Sigma}^n_{i=1}\alpha_i y_i \mathbf{w}.\mathbf{x_i} + \boldsymbol{\Sigma}^n_{i=1}\alpha_i$
  - $J(w,b,\alpha) = \frac{1}{2}\,\Sigma^n_{i=1}\Sigma^n_{j=1}\alpha_i\alpha_j y_i y_j \mathbf{x_i}\mathbf{x_j} - \boldsymbol{\Sigma}^n_{i=1}\boldsymbol{\Sigma}^n_{j=1}\alpha_i\alpha_j y_i y_j \mathbf{x_i}\mathbf{x_j} + \boldsymbol{\Sigma}^n_{i=1}\alpha_i$
  - $Q(\alpha) = \boldsymbol{\Sigma}^N_{i=1}\alpha_i - \frac{1}{2}\,\Sigma^n_{i=1}\Sigma^n_{j=1}\alpha_i\alpha_j y_i y_j \mathbf{x_i}\mathbf{x_j}$

Dual Problem

- Given the training sample $\{(x_i, y_i\}^n_{i=1}$, find the Lagrange multipliers that maximize the objective function
  - $Q(\alpha) = \boldsymbol{\Sigma}^N_{i=1}\alpha_i - 1/2\Sigma^n_{i=1}\Sigma^n_{j=1}\alpha_i\alpha_j y_i y_j \mathbf{x_i x_j}$, S.T.
    - $\alpha_i >= 0, \forall\ i = 1,2, \ldots.n$
    - $\Sigma^n_{i=1}\alpha_i y_i = 0$
- Primal vs Dual
  - The dual problem is cast entirely in terms of the training data
- Objective: To find the lagrangian multipliers which maximizes the $Q(\alpha)$
  - *Some of the lagrangian multipliers will become zero*
  - *Some of the lagrangian multipliers will have high value*

Interpretation

- *Some of the lagrangian multipliers will have high value*
    - Corresponding input training sample is a support vector
- *Some of the lagrangian multipliers will become zero*
    - Corresponding input training sample is not a support vector
- *Some of the lagrangian multipliers might have very high value*
    - Corresponding input training sample is an outlier

Optimal weight and bias: Decision boundary

- *Having determined the optimum Lagrangian multipliers, the optimum weight vector may be computed*
  - $\mathbf{w}_o = \Sigma^{ns}_{i=1}\alpha_{o,i}y_i\mathbf{x_i}$
  - n_s is the number of support vectors for which the Lagrange multipliers are all non-zero
- *Having obtained w_o, the bias b_o may be computed*
  - $b_o = 1 - w_o^T\mathbf{x^{(s)}}$, $\forall\ y^{(s)} = 1$
    - $1 - \Sigma^{Ns}_{i=1}\alpha_{o,i}y_i\mathbf{x_i}\mathbf{x^{(s)}}$
- For a new sample z, calculate $\mathbf{w}_o z + b_o$
  - $\Sigma^{Ns}_{i=1}\alpha_{o,i}y_i\mathbf{x_i}\mathbf{z_j} + (1 - \Sigma^{ns}_{i=1}\alpha_{o,i}y_i\mathbf{x_i}\mathbf{x^{(s)}})$
- If the sign is positive, the sample z will belong the positive class, else the sample z will belong to the negative class.

Breakout Room Activity

i	x_i	y_i	α_i	i	x_i	y_i	α_i
1	(4,2.9)	1	0.414	6	(1.9,1.9)	-1	0
2	(4,4)	1	0	7	(3.5, 4)	1	0.018
3	(1,2.5)	-1	0	8	(0.5,1.5)	-1	0
4	(2.5,1)	-1	1.18	9	(2,2.1)	-1	0.414
5	(4.9,2.5)	1	0	10	(4.5,2.5)	1	0

Consider the training data samples and the corresponding Lagrange multipliers learned from them, as given in the following table.

From the given table above, answer the following questions?

1. What is the b for the SVM?
2. Identify the support vectors.
3. Compute w and classify the point (3,3).

References

1. https://www.svm-tutorial.com/2017/02/svms-overview-support-vector-machines/
2. https://www.syncfusion.com/ebooks/support_vector_machines_succinctly/introduction
3. https://www.youtube.com/watch?v=b-Su6aVh5yo
4. Chapter 6, Neural Networks: A Comprehensive Foundation (2nd Edition) 2nd Edition by Simon Haykin

Supplement: Scaling of weight vectors

- Distance from a point (x_o,y_o) to a Line $Ax+By+c = 0$ is
  - $|Ax_o + By_o + c|/\mathrm{sqrt}(A^2+B^2)$
- The distance between support vector point and H_o
  - $m = |w \cdot x_o + b| / ||w||$
  - The geometric margin is clearly invariant to scaling of weight parameters because it is inherently normalized by the length of $||w||$
  - This means that we can impose any scaling constraint we wish on $||w||$ without affecting the geometric margin.
- $|w \cdot x_o + b| = m^* ||w||$
- By applying a proper scaling to the weights, the factor $m^* ||w||$ can always be made $= 1$
  - $|w \cdot x_o + b| = 1$ for support vector points