DATA ANALYSIS AND VISUALIZATION

INSTRUCTOR: UMME AMMARAH

A BIT OF MATH BEHIND SVM

UNDERSTANDING DOT PRODUCT

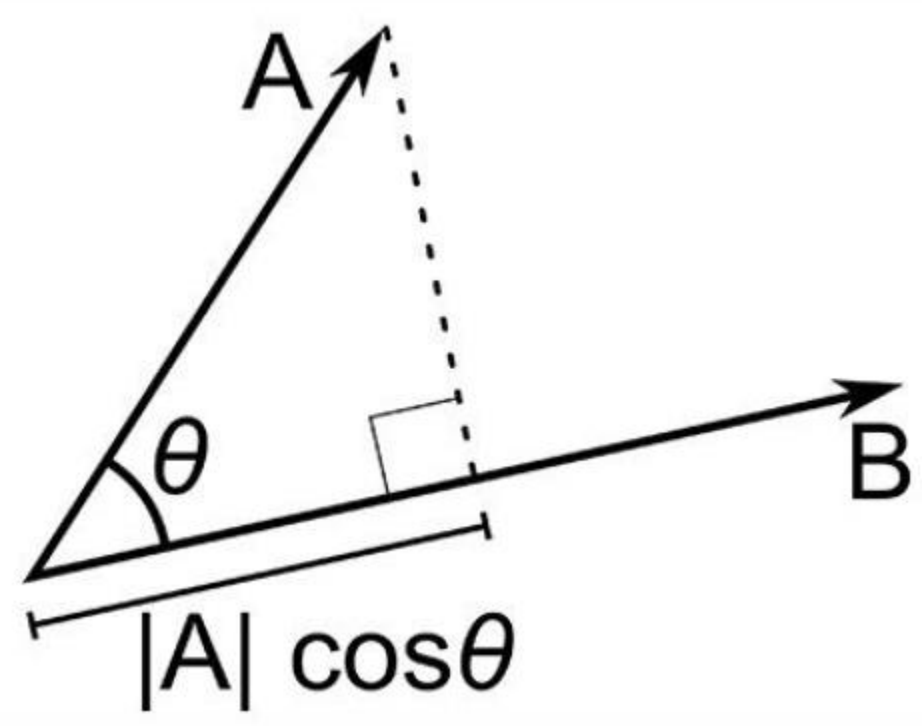
- It's the projection of one vector onto the other multiplied by the magnitude of other vector.

Mathematically it can be written as:

A . B = |A| cosθ * |B|

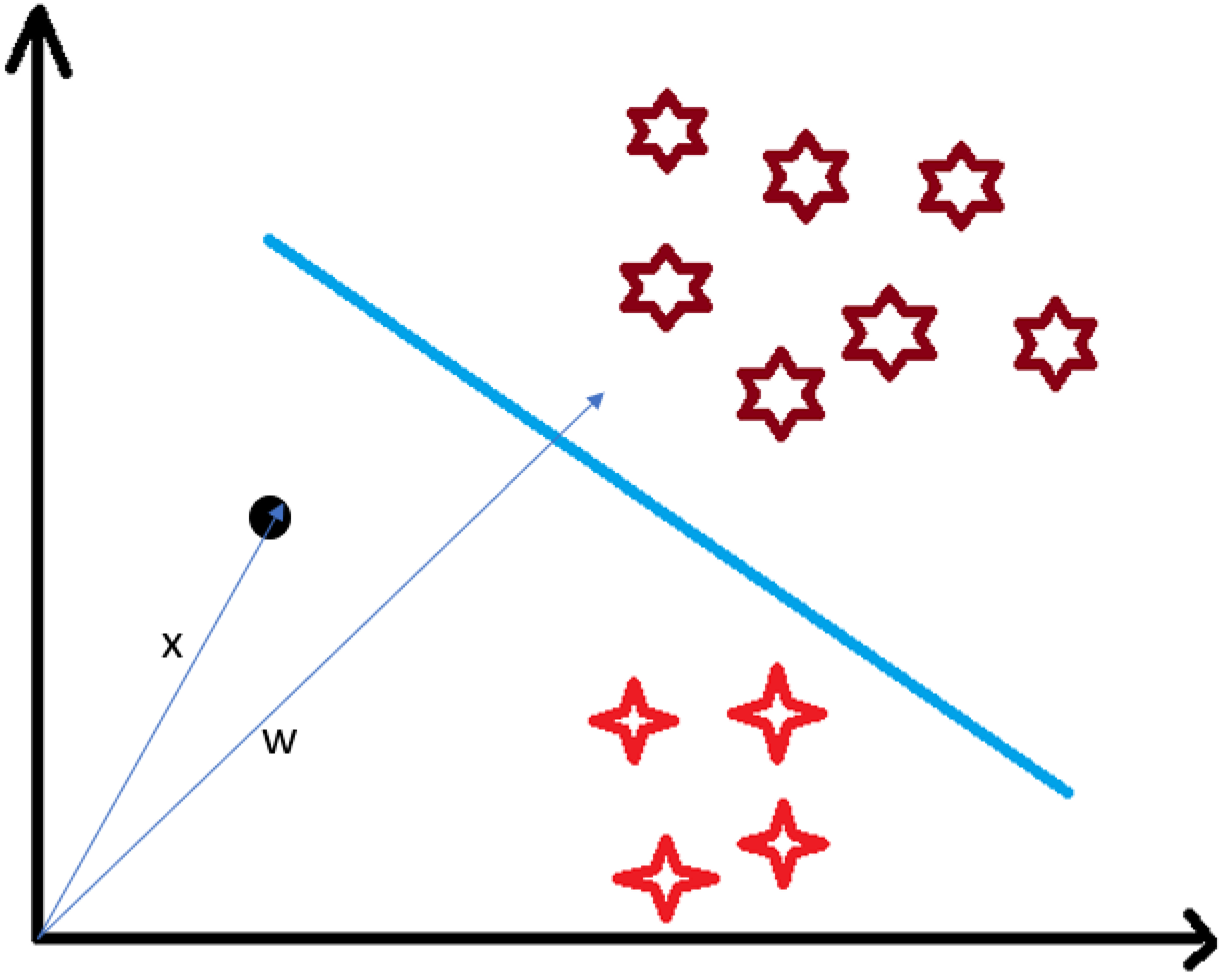
Where |A| cosθ is the projection of A on B

And |B| is the magnitude of vector B



USE OF DOT PRODUCT IN SVM

- Consider a random point X and we want to know whether it lies on the right side of the plane or the left side of the plane (positive or negative).



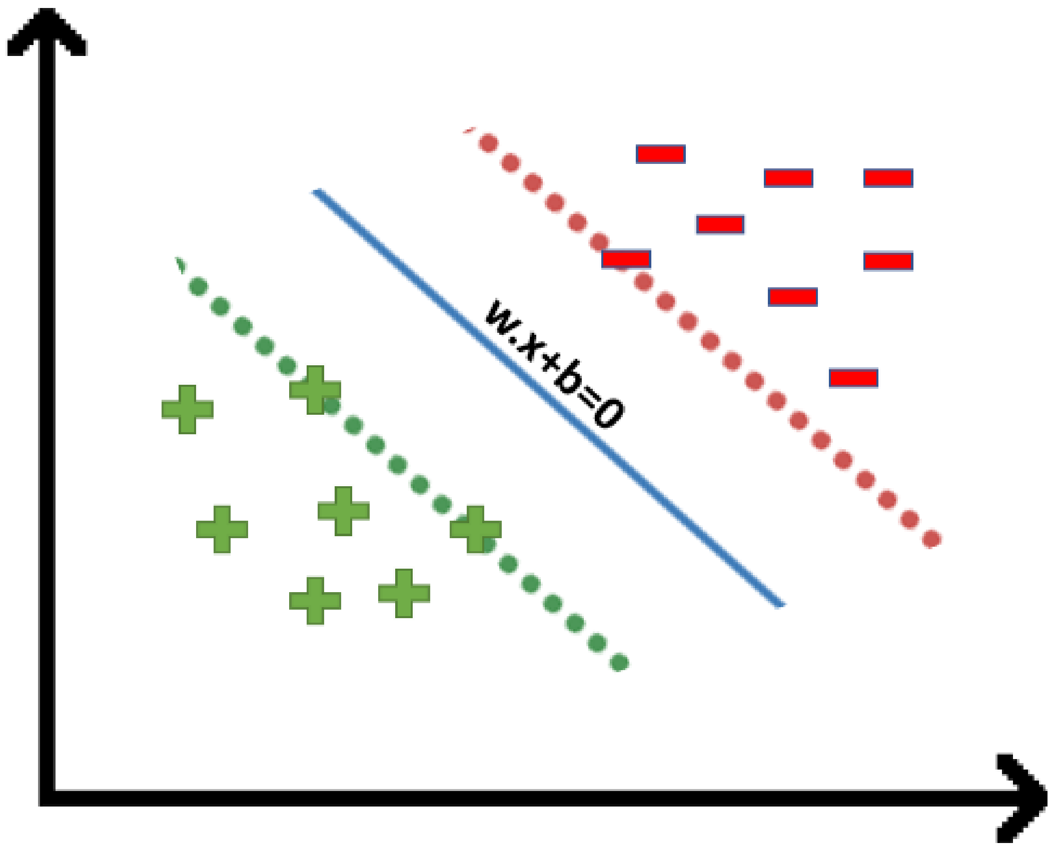
HYPERPLANE

- The equation of a hyperplane is w.x+b=0 where w is a vector normal to hyperplane and b is an offset.
- In d dimensions:

$$h(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + b = w_1x_1 + w_2x_2 + \cdots + w_dx_d + b$$

- For points that lie on the hyperplane:

$$h(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + b = 0$$



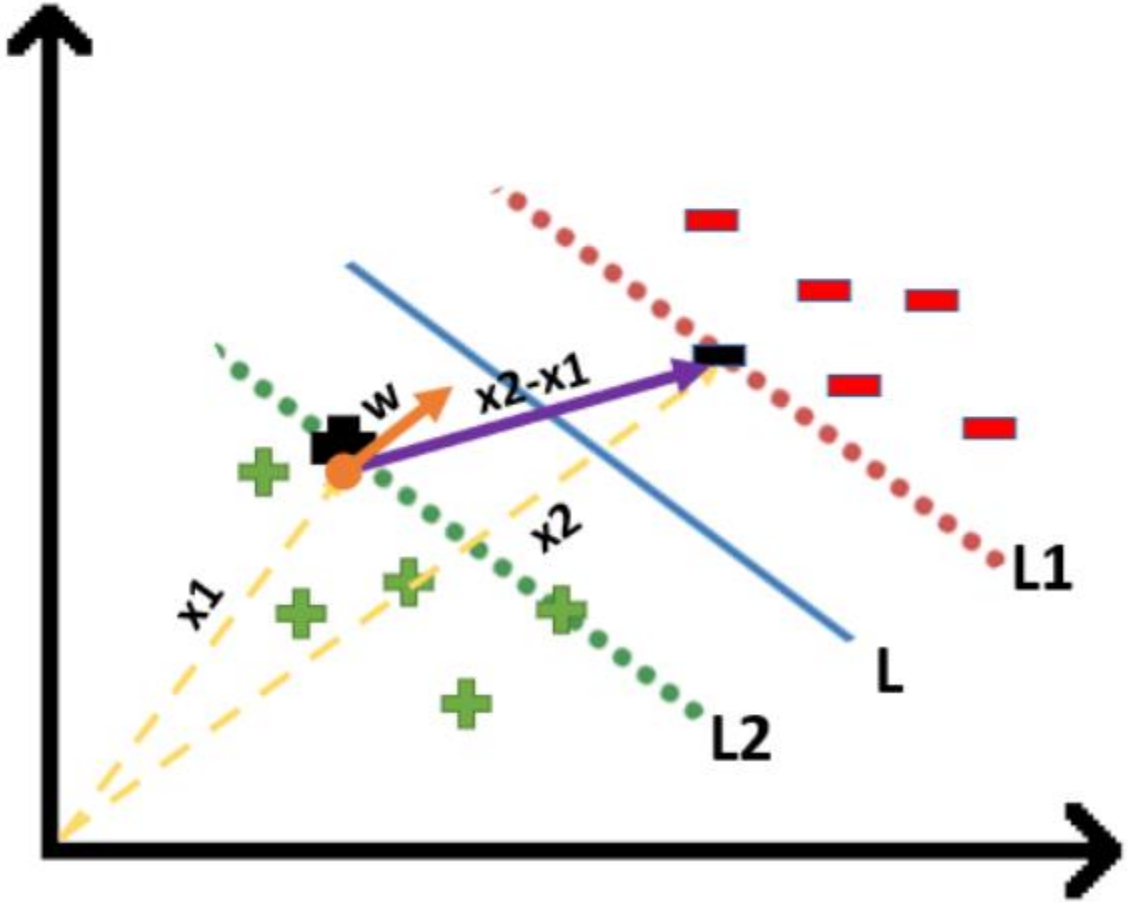
DECISION RULE

The hyperplane function $h(\boldsymbol{x})$ thus serves as a linear classifier or a linear discriminant, which predicts the class y for any given point $\boldsymbol{x}$, according to the decision rule:

$$y = \begin{cases} +1 & \text{if } h(\mathbf{x}) > 0 \\ -1 & \text{if } h(\mathbf{x}) < 0 \end{cases}$$

DISTANCE BETWEEN 2 MARGINS

$$\frac{2}{\|w\|} = d$$



DISTANCE OF A POINT TO THE HYPERPLANE

Consider a point $x \in \mathbb{R}^d$ that does not lie on the hyperplane. Let x_p be the orthogonal projection of x on the hyperplane, and let $\mathbf{r} = x - x_p$. Then we can write x as

$$\mathbf{x} = \mathbf{x}_p + \mathbf{r}$$

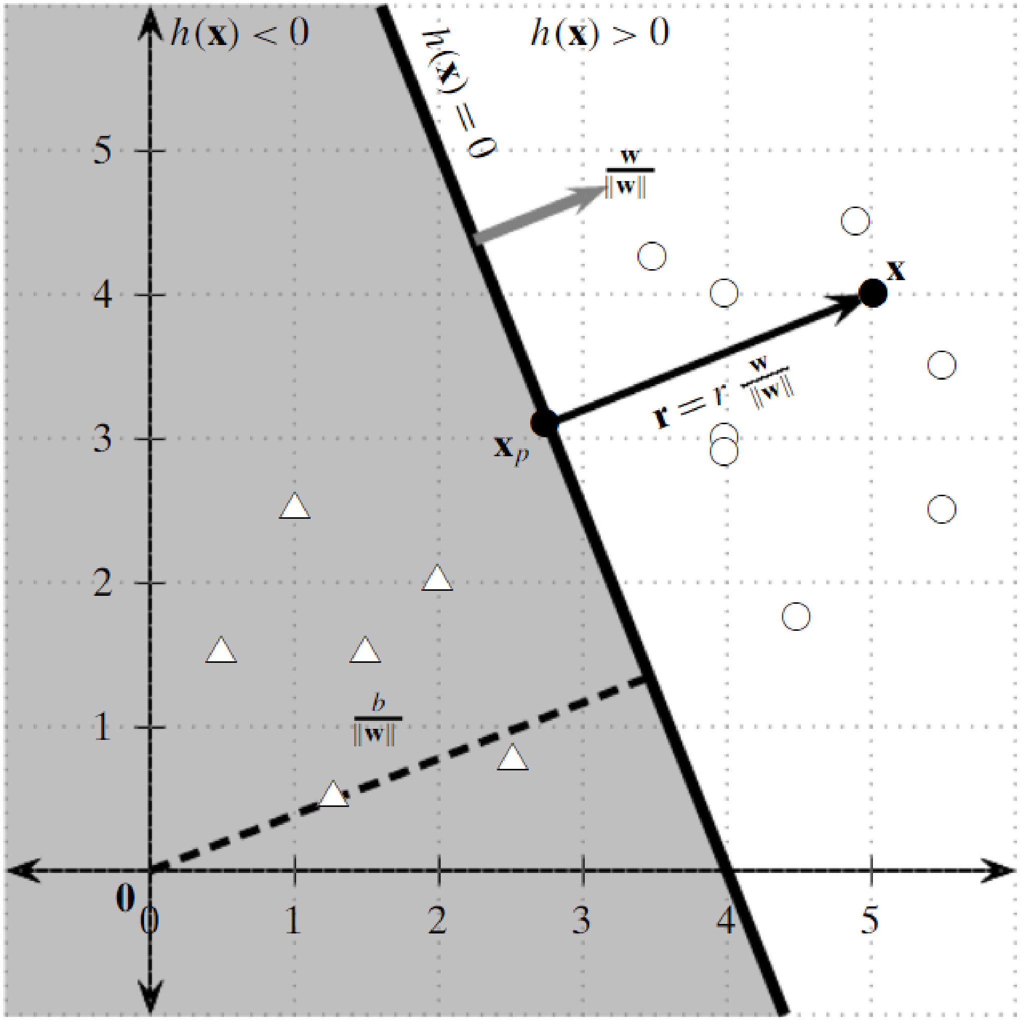
$$\mathbf{x} = \mathbf{x}_p + r\frac{\mathbf{w}}{\|\mathbf{w}\|}$$

where r is the *directed distance* of the point x from x_p.
To obtain an expression for r, consider the value $h(x)$, we have:

$$h(x) = h\left(x_p + r\frac{w}{\|w\|}\right) = w^T\left(x_p + r\frac{w}{\|w\|}\right) + b = r\|w\|$$

The directed distance r of point x to the hyperplane is thus:

$$r = \frac{h(x)}{\|w\|}$$



DISTANCE OF A POINT FROM HYPERPLANE

To obtain distance, which must be non-negative, we multiply r by the class label y_i of the point $\mathbf{x}_i$ because when $h(\mathbf{x}_i) < 0$, the class is -1, and when $h(\mathbf{x}_i) > 0$ the class is $+1$:

$$\delta_i = \frac{y_i h(\mathbf{x}_i)}{\|\mathbf{w}\|}$$

for the origin $\mathbf{x} = \mathbf{0}$, the directed distance is

$$r = \frac{h(\mathbf{0})}{\|\mathbf{w}\|} = \frac{\mathbf{w}^T \mathbf{0} + b}{\|\mathbf{w}\|} = \frac{b}{\|\mathbf{w}\|}$$

EXAMPLE

$$h(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + b = w_1 x_1 + w_2 x_2 + b = 0$$

Rearranging the terms we get

$$x_2 = -\frac{w_1}{w_2}x_1 - \frac{b}{w_2}$$

where $-\frac{w_1}{w_2}$ is the slope of the line, and $-\frac{b}{w_2}$ is the intercept along the second dimension.

Consider any two points on the hyperplane, say $\mathbf{p} = (p_1, p_2) = (4,0)$, and $\mathbf{q} = (q_1, q_2) = (2,5)$. The slope is given as

$$\boldsymbol{p} = (p_1, p_2) = (4,0), \ \boldsymbol{q} = (q_1, q_2) = (2,5)$$

$$-\frac{w_1}{w_2} = \frac{q_2 - p_2}{q_1 - p_1} = \frac{5-0}{2-4} = -\frac{5}{2}$$

Given $(4,0)$, the offset b is:

$$b = -5x_1 - 2x_2 = -5 \cdot 4 - 2 \cdot 0 = -20$$

Given $\boldsymbol{w} = \begin{pmatrix} 5 \\ 2 \end{pmatrix}$ and $b = -20$:

$$h(\boldsymbol{x}) = \boldsymbol{w}^T\boldsymbol{x} + b = \begin{pmatrix} 5 & 2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} - 20 = 0$$

$$\delta = y\, r = -1\, r = \frac{-b}{\|\boldsymbol{w}\|} = \frac{-(-20)}{\sqrt{29}} = 3.71$$

MARGIN AND SUPPORT VECTORS

The distance of a point $\boldsymbol{x}$ from the hyperplane $h(\boldsymbol{x}) = 0$ is thus given as

$$\boxed{\delta = y\,r = \frac{y\,h(\boldsymbol{x})}{\|\boldsymbol{w}\|}}$$

The *margin* is the minimum distance of a point from the separating hyperplane:

$$\boxed{\delta^* = \min_{\boldsymbol{x}_i} \left\{ \frac{y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b)}{\|\boldsymbol{w}\|} \right\}}$$

All the points (or vectors) that achieve the minimum distance are called *support vectors* for the hyperplane. They satisfy the condition:

$$\delta^* = \frac{y^*(\boldsymbol{w}^T\boldsymbol{x}^* + b)}{\|\boldsymbol{w}\|}$$

where y^* is the class label for $\boldsymbol{x}^*$.

CANONICAL HYPERPLANE

Multiplying the hyperplane equation on both sides by some scalar s yields an equivalent hyperplane:

$$s\,h(\mathbf{x}) = s\,\mathbf{w}^T\mathbf{x} + s\,b = (s\mathbf{w})^T\mathbf{x} + (sb) = 0$$

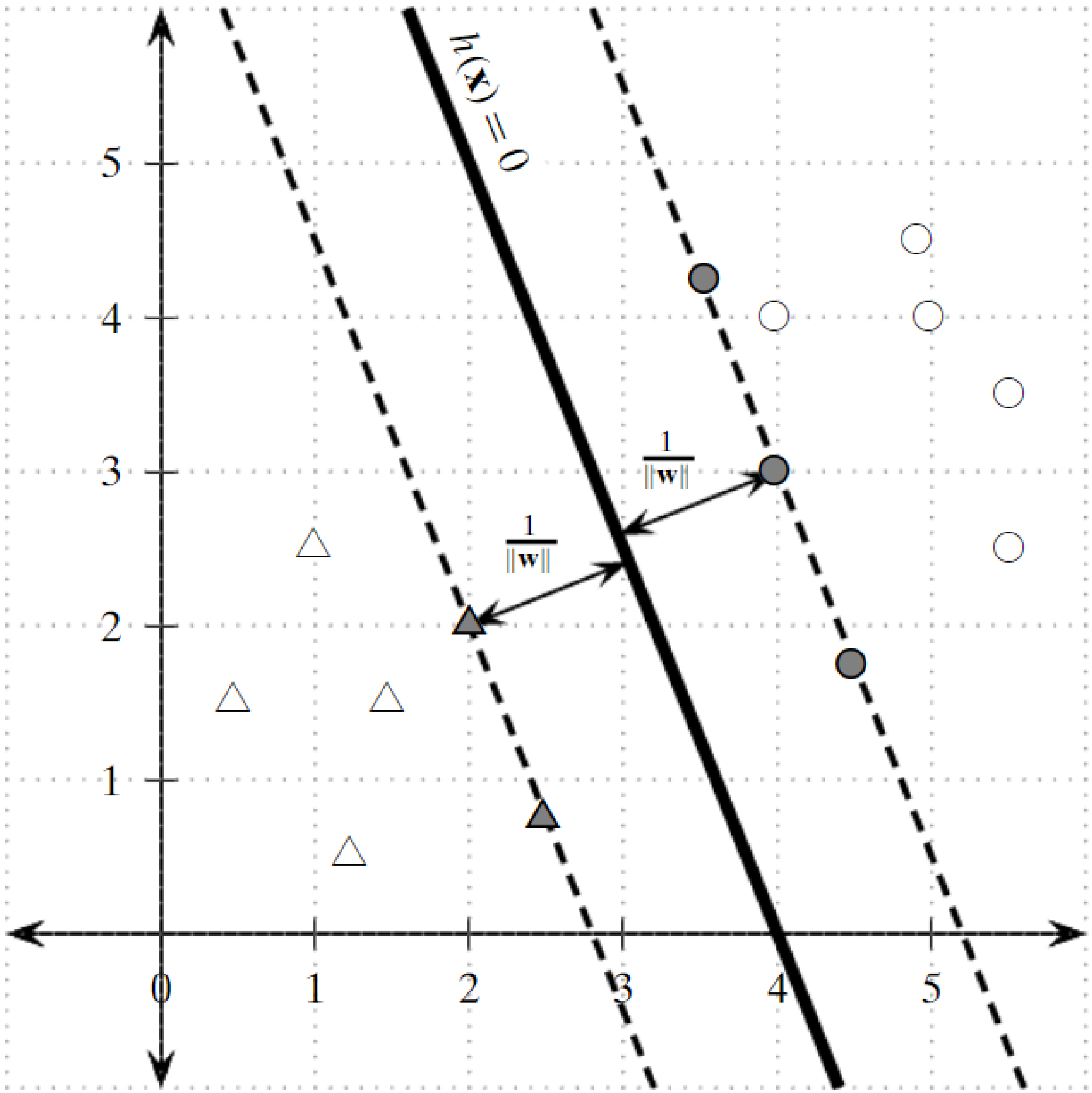
To obtain the unique or *canonical* hyperplane, we choose the scalar $s = \frac{1}{y^*(\mathbf{w}^T\mathbf{x}^*+b)}$ so that the absolute distance of a support vector from the hyperplane is 1, i.e., the margin is

$$\delta^* = \frac{y^*(\mathbf{w}^T\mathbf{x}^* + b)}{\|\mathbf{w}\|} = \frac{1}{\|\mathbf{w}\|}$$

For the canonical hyperplane, for each support vector $\mathbf{x}_i^*$ (with label y_i^*), we have $y_i^* h(\mathbf{x}_i^*) = 1$, and for any point that is not a support vector we have $y_i h(\mathbf{x}_i) > 1$. Over all points, we have

$$\boxed{y_i\,(\mathbf{w}^T\mathbf{x}_i + b) \geq 1, \text{ for all points } \mathbf{x}_i \in \boldsymbol{D}}$$

EXAMPLE



$$h(\boldsymbol{x}) = \begin{pmatrix} 5 \\ 2 \end{pmatrix}^T \boldsymbol{x} - 20 = 0$$

Given $\boldsymbol{x}^* = (2,2)^T$, $y^* = -1$.

$$s = \frac{1}{y^* h(\boldsymbol{x}^*)} = \frac{1}{-1\left(\begin{pmatrix} 5 \\ 2 \end{pmatrix}^T \begin{pmatrix} 2 \\ 2 \end{pmatrix} - 20\right)} = \frac{1}{6}$$

$$\boldsymbol{w} = \frac{1}{6}\begin{pmatrix} 5 \\ 2 \end{pmatrix} = \begin{pmatrix} 5/6 \\ 2/6 \end{pmatrix} \qquad b = \frac{-20}{6}$$

$$h(\boldsymbol{x}) = \begin{pmatrix} 5/6 \\ 2/6 \end{pmatrix}^T \boldsymbol{x} - 20/6 = \begin{pmatrix} 0.833 \\ 0.333 \end{pmatrix}^T \boldsymbol{x} - 3.33$$

$$\delta^* = \frac{y^* \, h(\boldsymbol{x}^*)}{\|\boldsymbol{w}\|} = \frac{1}{\sqrt{\left(\frac{5}{6}\right)^2 + \left(\frac{2}{6}\right)^2}} = \frac{6}{\sqrt{29}} = 1.114$$

MAXIMUM MARGIN HYPERPLANE

The goal of SVMs is to choose the canonical hyperplane, h^*, that yields the maximum margin among all possible separating hyperplanes

$$h^* = \arg\max_{\mathbf{w},b} \left\{ \frac{1}{\|\mathbf{w}\|} \right\}$$

We can obtain an equivalent minimization formulation:

Objective Function: $\min_{\mathbf{w},b} \left\{ \frac{\|\mathbf{w}\|^2}{2} \right\}$

Linear Constraints: $y_i\left(\mathbf{w}^T\mathbf{x}_i + b\right) \geq 1, \ \forall \mathbf{x}_i \in \boldsymbol{D}$

PRIMAL SVM FORMULATION

$$\mathbf{w}^*, b^* = \arg\max_{\mathbf{w},b} \frac{1}{||\mathbf{w}||} = \arg\min_{\mathbf{w},b} ||\mathbf{w}||^2$$

which is a *quadratic programming problem*.

- Minimisation of a *quadratic function*.
- Subject to *linear constraints*.

- This is known as the *primal* SVM formulation.

- Computational complexity of QP for M variables is $O(M^3)$.
- For high-dimensional spaces ($M > N$), a *dual* SVM formulation exists with $O(N^3)$ complexity.
- Some QP implementations solve the dual faster than the primal.
- Derivation of the dual formulation requires a thorough understanding of Lagrange multipliers.

LAGRANGE MULTIPLIERS

We turn the constrained SVM optimization into an unconstrained one by introducing a Lagrange multiplier α_i for each constraint. The new objective function, called the *Lagrangian*, then becomes

$$\min L = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{n} \alpha_i \left(y_i(\mathbf{w}^T \mathbf{x}_i + b) - 1 \right)$$

L should be minimized w.r.t. $\mathbf{w}$ and b, and it should be maximized w.r.t. α_i.
Taking the derivative of L with respect to $\mathbf{w}$ and b, and setting those to zero, we obtain

$$\frac{\partial}{\partial \mathbf{w}} L = \mathbf{w} - \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i = 0 \quad \text{or} \quad \mathbf{w} = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i$$

$$\frac{\partial}{\partial b} L = \sum_{i=1}^{n} \alpha_i y_i = 0$$

We can see that $\mathbf{w}$ can be expressed as a linear combination of the data points $\mathbf{x}_i$, with the signed Lagrange multipliers, $\alpha_i y_i$, serving as the coefficients.

Further, the sum of the signed Lagrange multipliers, $\alpha_i y_i$, must be zero.

Once we have obtained the α_i values for $i = 1, \ldots, n$, we can solve for the weight vector $\boldsymbol{w}$ and the bias b. Each of the Lagrange multipliers α_i satisfies the conditions at the optimal solution:

$$\alpha_i \left(y_i(\boldsymbol{w}^T \boldsymbol{x}_i + b) - 1 \right) = 0$$

which gives rise to two cases:

1. $\alpha_i = 0$, or
2. $y_i(\boldsymbol{w}^T \boldsymbol{x}_i + b) - 1 = 0$, which implies $y_i(\boldsymbol{w}^T \boldsymbol{x}_i + b) = 1$

This is a very important result because if $\alpha_i > 0$, then $y_i(\boldsymbol{w}^T \boldsymbol{x}_i + b) = 1$, and thus the point $\boldsymbol{x}_i$ must be a support vector.

On the other hand, if $y_i(\boldsymbol{w}^T \boldsymbol{x}_i + b) > 1$, then $\alpha_i = 0$, that is, if a point is not a support vector, then $\alpha_i = 0$.

Once we know α_i for all points, we can compute the weight vector $\boldsymbol{w}$ by taking the summation only for the support vectors:

$$\boxed{\boldsymbol{w} = \sum_{i,\alpha_i>0} \alpha_i y_i \boldsymbol{x}_i}$$

Only the support vectors determine $\boldsymbol{w}$, since $\alpha_i = 0$ for other points.
To compute the bias b, we first compute one solution b_i, per support vector, as follows:

$$y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) = 1, \text{ which implies } b_i = \frac{1}{y_i} - \boldsymbol{w}^T\boldsymbol{x}_i = y_i - \boldsymbol{w}^T\boldsymbol{x}_i$$

The bias b is taken as the average value:

$$\boxed{b = \text{avg}_{\alpha_i>0}\{b_i\}}$$

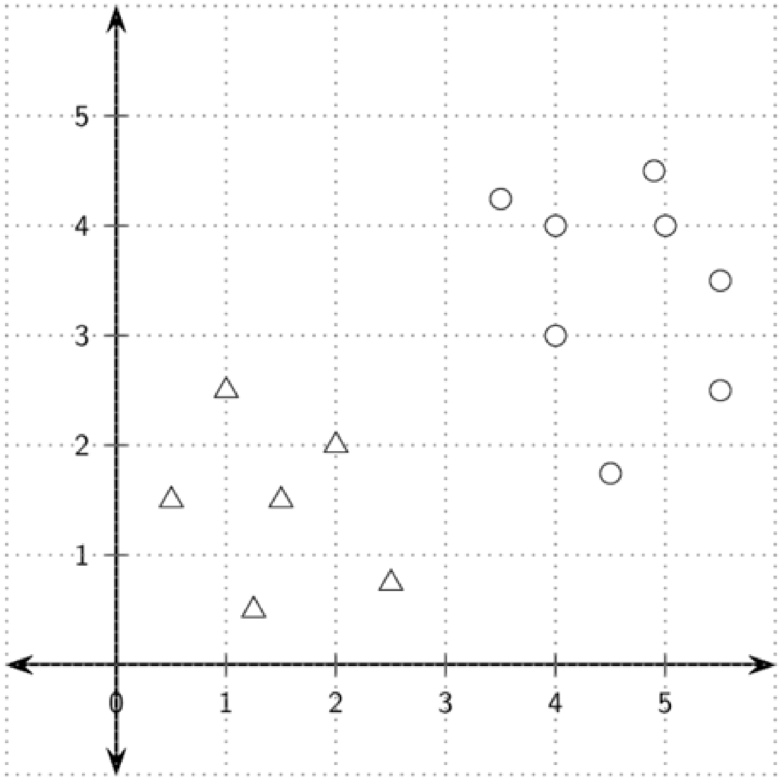
SVM CLASSIFIER

Given the optimal hyperplane function $h(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{x} + b$, for any new point $\boldsymbol{z}$, we predict its class as

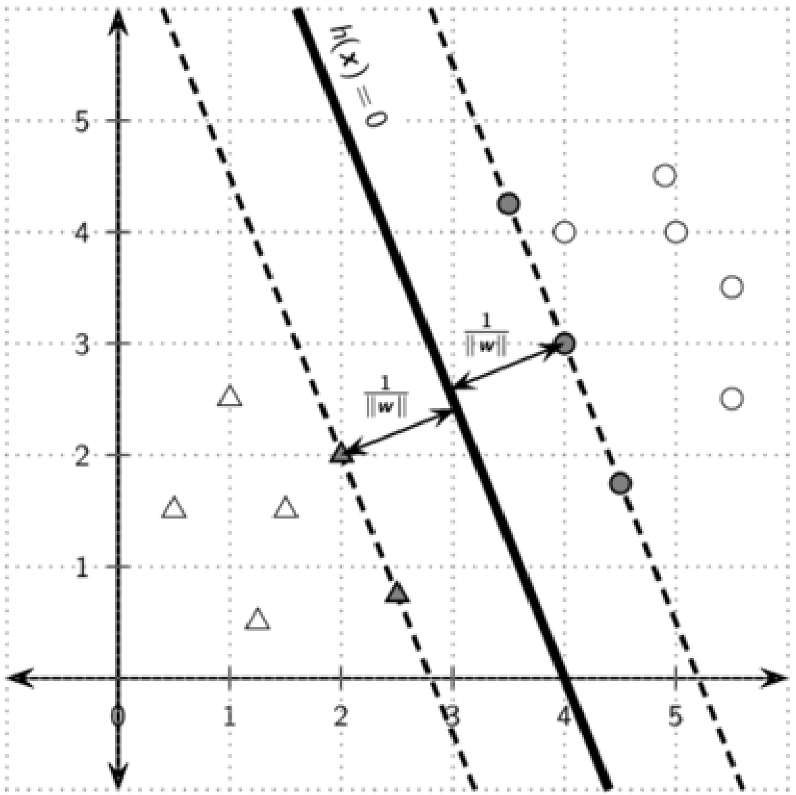
$$\boxed{\hat{y} = \text{sign}(h(\boldsymbol{z})) = \text{sign}(\boldsymbol{w}^T \boldsymbol{z} + b)}$$

where the $\text{sign}(\cdot)$ function returns $+1$ if its argument is positive, and -1 if its argument is negative.

EXAMPLE



$\mathbf{x}_i$	x_{i1}	x_{i2}	y_i
$\mathbf{x}_1$	3.5	4.25	+1
$\mathbf{x}_2$	4	3	+1
$\mathbf{x}_3$	4	4	+1
$\mathbf{x}_4$	4.5	1.75	+1
$\mathbf{x}_5$	4.9	4.5	+1
$\mathbf{x}_6$	5	4	+1
$\mathbf{x}_7$	5.5	2.5	+1
$\mathbf{x}_8$	5.5	3.5	+1
$\mathbf{x}_9$	0.5	1.5	−1
$\mathbf{x}_{10}$	1	2.5	−1
$\mathbf{x}_{11}$	1.25	0.5	−1
$\mathbf{x}_{12}$	1.5	1.5	−1
$\mathbf{x}_{13}$	2	2	−1
$\mathbf{x}_{14}$	2.5	0.75	−1



Solving the L_{dual} quadratic program yields

$\mathbf{x}_i$	x_{i1}	x_{i2}	y_i	α_i
$\mathbf{x}_1$	3.5	4.25	$+1$	0.0437
$\mathbf{x}_2$	4	3	$+1$	0.2162
$\mathbf{x}_4$	4.5	1.75	$+1$	0.1427
$\mathbf{x}_{13}$	2	2	-1	0.3589
$\mathbf{x}_{14}$	2.5	0.75	-1	0.0437

The weight vector and bias are:

$$\mathbf{w} = \sum_{i,\alpha_i>0} \alpha_i y_i \mathbf{x}_i = \begin{pmatrix} 0.833 \\ 0.334 \end{pmatrix}$$

$$b = \text{avg}\{b_i\} = -3.332$$

The optimal hyperplane is given as follows:

$$h(\mathbf{x}) = \begin{pmatrix} 0.833 \\ 0.334 \end{pmatrix}^T \mathbf{x} - 3.332 = 0$$