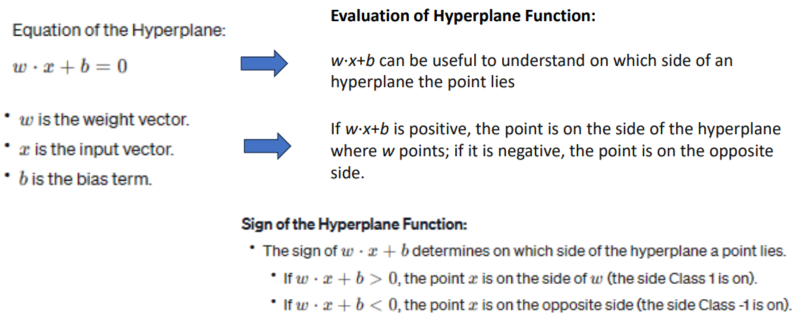
Esercitazione 6 20/11/2023

una precedente che aveva una formula

Quantità positiva: il punto è sulla parte superiore dell'iperpiano. Negativa: il punto è sulla parte inferiore.



Equation of the Hyperplane:

$w \cdot x + b = 0$

- w is the weight vector.
- x is the input vector.
- b is the bias term.

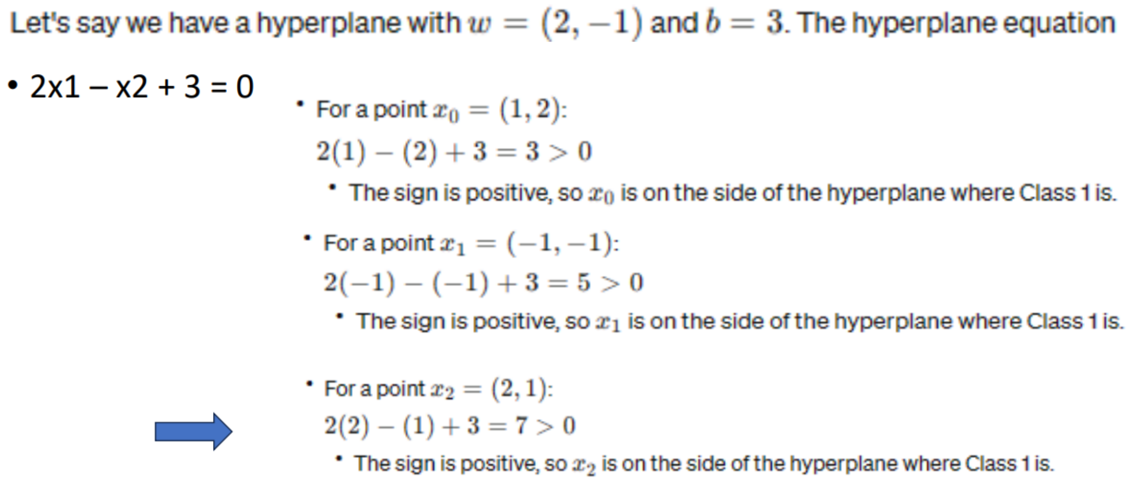
Evaluation of Hyperplane Function:

w·x+b can be useful to understand on which side of an hyperplane the point lies

If *w·x+b* is positive, the point is on the side of the hyperplane where *w* points; if it is negative, the point is on the opposite side.

Sign of the Hyperplane Function:

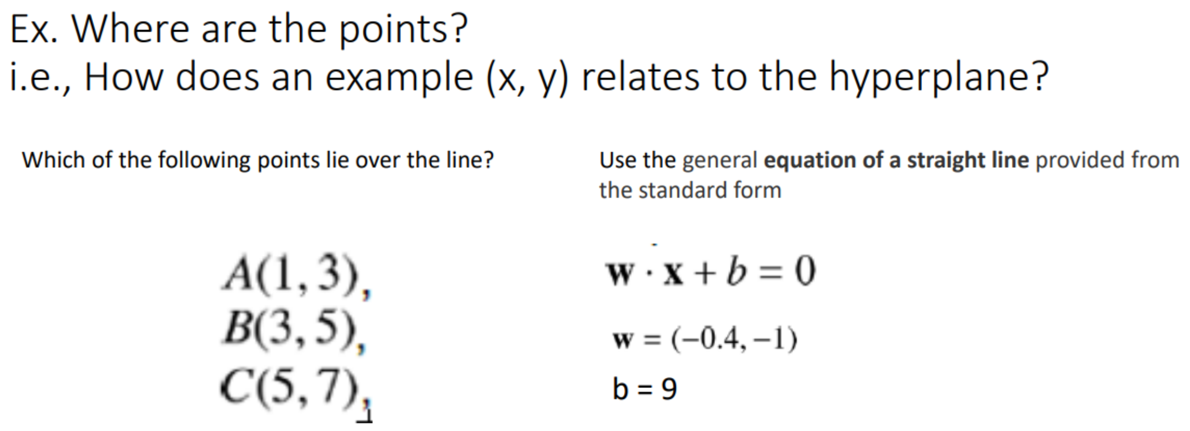
- The sign of $w \cdot x + b$ determines on which side of the hyperplane a point lies.
  - If $w \cdot x + b > 0$, the point x is on the side of w (the side Class 1 is on).
  - If $w \cdot x + b < 0$, the point x is on the opposite side (the side Class -1 is on).



Let's say we have a hyperplane with $w = (2, -1)$ and $b = 3$. The hyperplane equation

- 2x1 – x2 + 3 = 0

- For a point $x_0 = (1, 2)$:
  $2(1) - (2) + 3 = 3 > 0$
  - The sign is positive, so x_0 is on the side of the hyperplane where Class 1 is.
- For a point $x_1 = (-1, -1)$:
  $2(-1) - (-1) + 3 = 5 > 0$
  - The sign is positive, so x_1 is on the side of the hyperplane where Class 1 is.
- For a point $x_2 = (2, 1)$:
  $2(2) - (1) + 3 = 7 > 0$
  - The sign is positive, so x_2 is on the side of the hyperplane where Class 1 is.



Ex. Where are the points?
i.e., How does an example (x, y) relates to the hyperplane?

Which of the following points lie over the line?

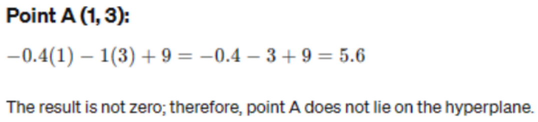
$A(1, 3)$,
$B(3, 5)$,
$C(5, 7)$,

Use the general **equation of a straight line** provided from the standard form

$$\mathbf{w} \cdot \mathbf{x} + b = 0$$

$\mathbf{w} = (-0.4, -1)$

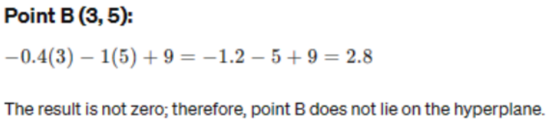
b = 9



Point A (1, 3):

$-0.4(1) - 1(3) + 9 = -0.4 - 3 + 9 = 5.6$

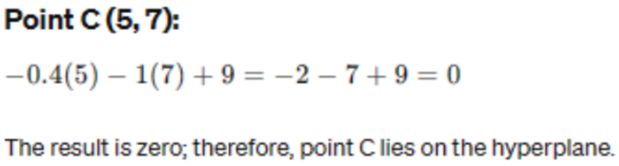
The result is not zero; therefore, point A does not lie on the hyperplane.



Point B (3, 5):

$-0.4(3) - 1(5) + 9 = -1.2 - 5 + 9 = 2.8$

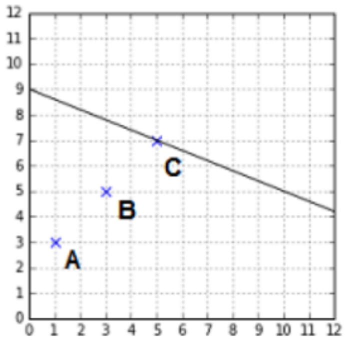
The result is not zero; therefore, point B does not lie on the hyperplane.



Point C (5, 7):

$-0.4(5) - 1(7) + 9 = -2 - 7 + 9 = 0$

The result is zero; therefore, point C lies on the hyperplane.



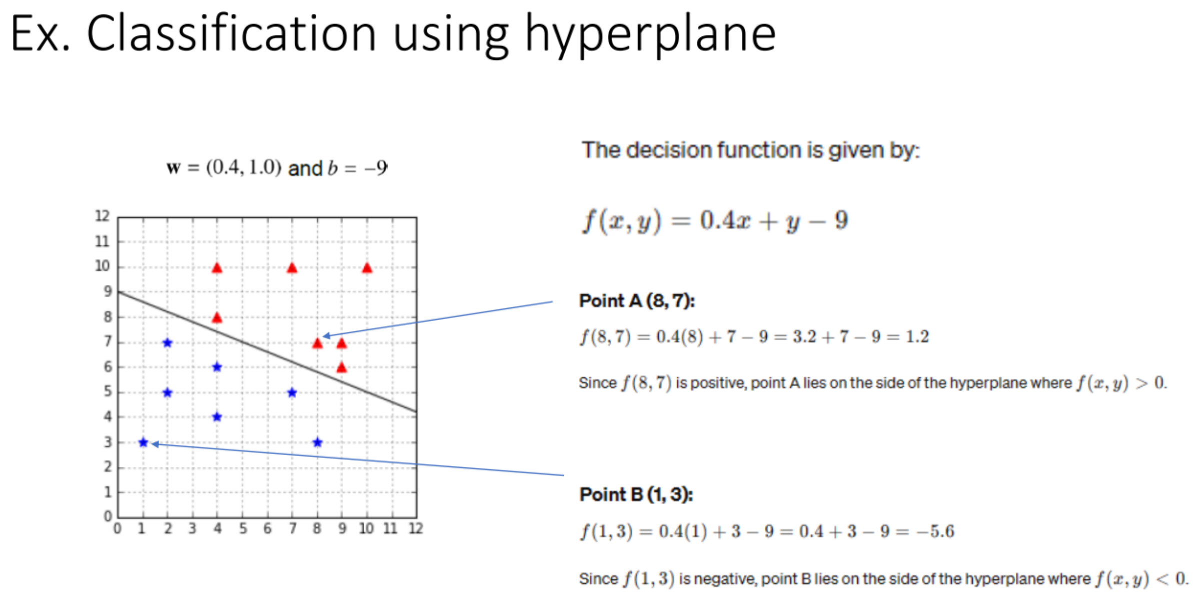
Ex. Classification using hyperplane

$$h(\mathbf{x}_i) = \begin{cases} +1 & \text{if} \quad \mathbf{w} \cdot \mathbf{x}_i + b \geq 0 \\ -1 & \text{if} \quad \mathbf{w} \cdot \mathbf{x}_i + b < 0 \end{cases}$$

$\mathbf{w} = (0.4, 1.0)$ and $b = -9$

classify the following points using the above hp

A(8,7);
B(1,3)



Ex. Classification using hyperplane

$\mathbf{w} = (0.4, 1.0)$ and $b = -9$

The decision function is given by:

$$f(x, y) = 0.4x + y - 9$$

Point A (8, 7):

$$f(8, 7) = 0.4(8) + 7 - 9 = 3.2 + 7 - 9 = 1.2$$

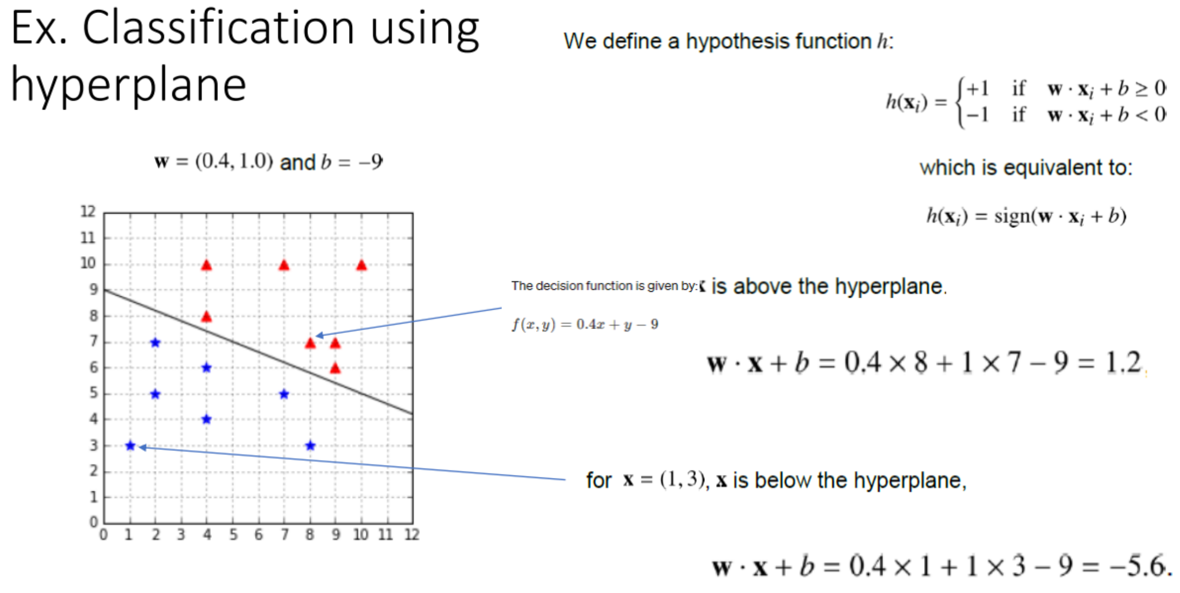
Since $f(8, 7)$ is positive, point A lies on the side of the hyperplane where $f(x, y) > 0$.

Point B (1, 3):

$$f(1, 3) = 0.4(1) + 3 - 9 = 0.4 + 3 - 9 = -5.6$$

Since $f(1, 3)$ is negative, point B lies on the side of the hyperplane where $f(x, y) < 0$.

Fisso un’ipotesi per quando il margine è positivo e per quando è negativo.



Ex. Classification using hyperplane

We define a hypothesis function h:

$$h(\mathbf{x}_i) = \begin{cases} +1 & \text{if} \quad \mathbf{w} \cdot \mathbf{x}_i + b \geq 0 \\ -1 & \text{if} \quad \mathbf{w} \cdot \mathbf{x}_i + b < 0 \end{cases}$$

which is equivalent to:

$$h(\mathbf{x}_i) = \text{sign}(\mathbf{w} \cdot \mathbf{x}_i + b)$$

$\mathbf{w} = (0.4, 1.0)$ and $b = -9$

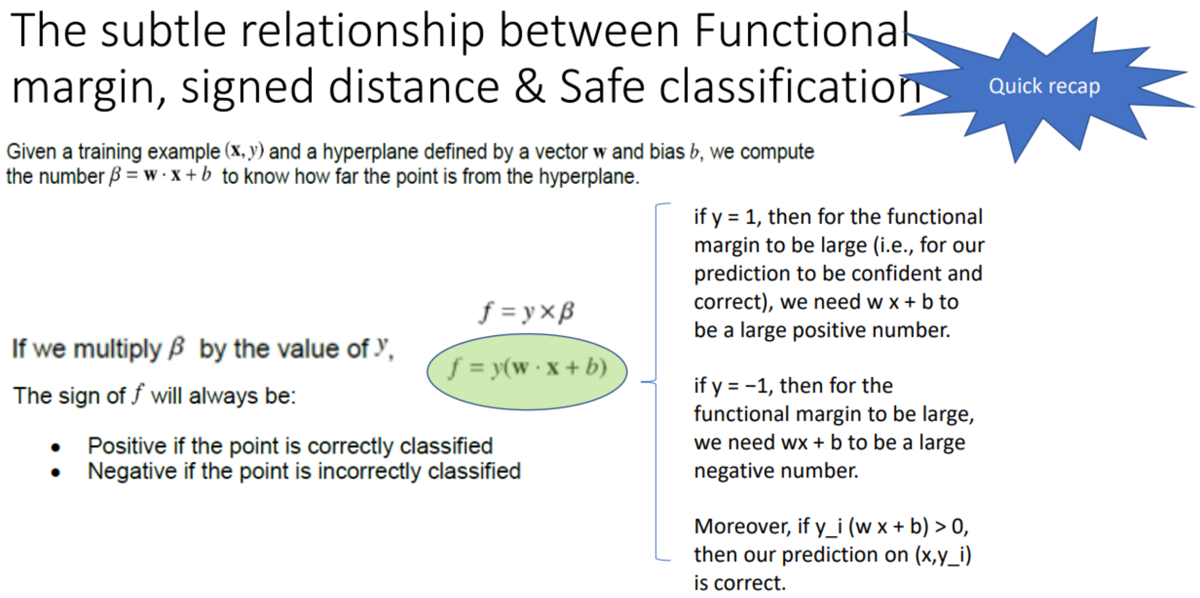
The decision function is given by:ɛ is above the hyperplane.

$f(x, y) = 0.4x + y - 9$

$$\mathbf{w} \cdot \mathbf{x} + b = 0.4 \times 8 + 1 \times 7 - 9 = 1.2$$

for $\mathbf{x} = (1, 3)$, $\mathbf{x}$ is below the hyperplane,

$$\mathbf{w} \cdot \mathbf{x} + b = 0.4 \times 1 + 1 \times 3 - 9 = -5.6.$$



The subtle relationship between Functional margin, signed distance & Safe classification

Quick recap

Given a training example $(\mathbf{x}, y)$ and a hyperplane defined by a vector $\mathbf{w}$ and bias b, we compute the number $\beta = \mathbf{w} \cdot \mathbf{x} + b$ to know how far the point is from the hyperplane.

$f = y \times \beta$

If we multiply β by the value of y,

$f = y(\mathbf{w} \cdot \mathbf{x} + b)$

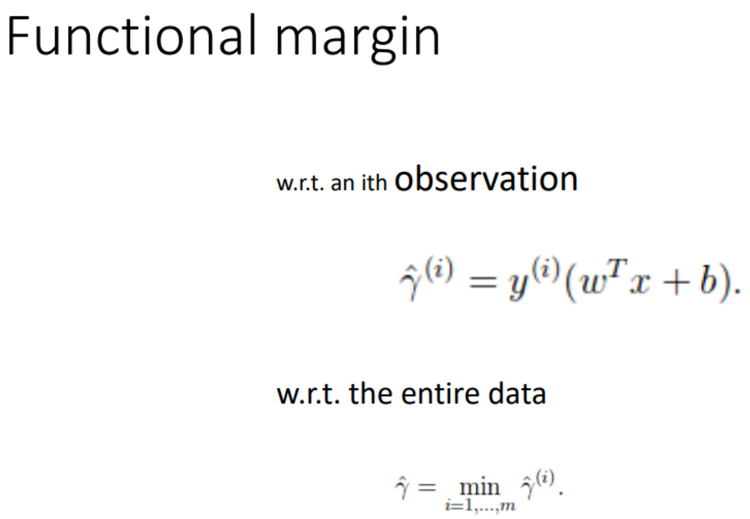
The sign of f will always be:

- Positive if the point is correctly classified
- Negative if the point is incorrectly classified

if y = 1, then for the functional margin to be large (i.e., for our prediction to be confident and correct), we need w x + b to be a large positive number.

if y = −1, then for the functional margin to be large, we need wx + b to be a large negative number.

Moreover, if y_i (w x + b) > 0, then our prediction on (x,y_i) is correct.



Functional margin

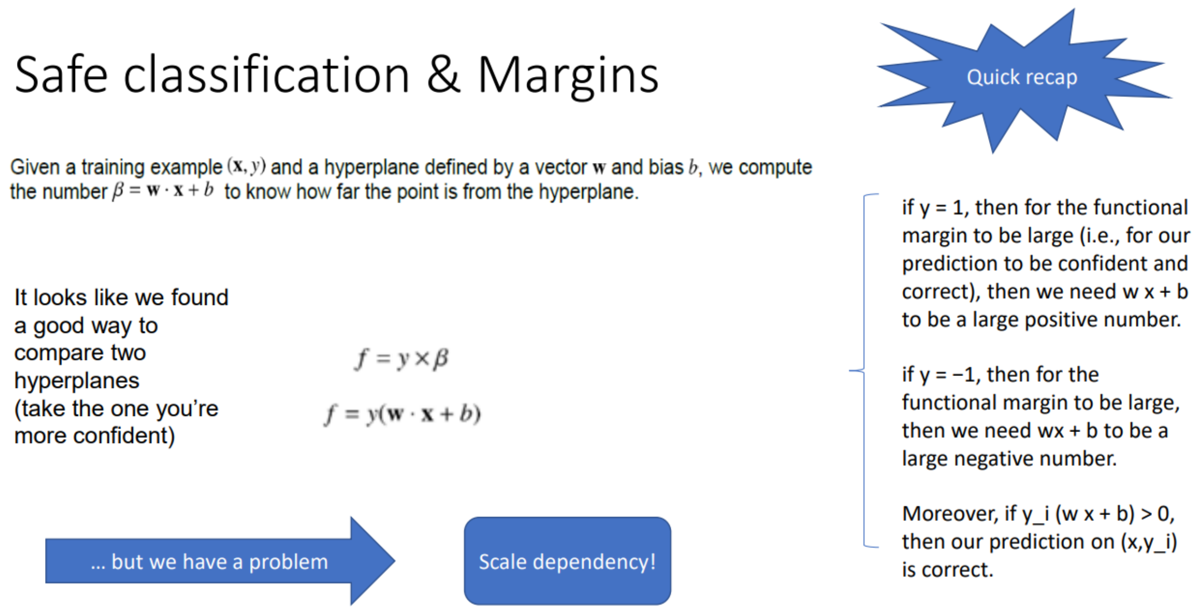
w.r.t. an ith observation

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

w.r.t. the entire data

$$\hat{\gamma} = \min_{i=1,\ldots,m} \hat{\gamma}^{(i)}.$$

Vado a calcolare tutte la quantità, e vado a prendere il minimo tra tutti i punti che sto considerando. Quindi vado a prendere il punto più vicino all’iperpiano.



Safe classification & Margins

Quick recap

Given a training example $(\mathbf{x}, y)$ and a hyperplane defined by a vector $\mathbf{w}$ and bias b, we compute the number $\beta = \mathbf{w} \cdot \mathbf{x} + b$ to know how far the point is from the hyperplane.

It looks like we found a good way to compare two hyperplanes (take the one you're more confident)

$$f = y \times \beta$$

$$f = y(\mathbf{w} \cdot \mathbf{x} + b)$$

if y = 1, then for the functional margin to be large (i.e., for our prediction to be confident and correct), then we need w x + b to be a large positive number.

if y = −1, then for the functional margin to be large, then we need wx + b to be a large negative number.

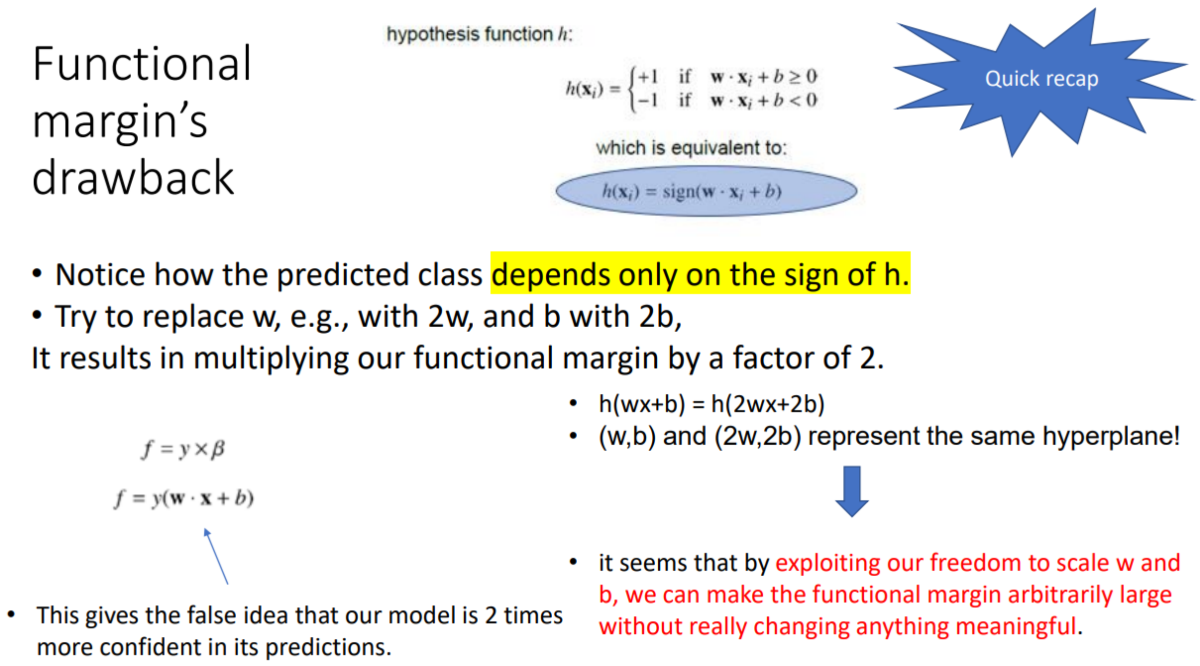
Moreover, if y_i (w x + b) > 0, then our prediction on (x,y_i) is correct.

... but we have a problem

Scale dependency!

Il nostro obiettivo è quello di andare a vedere qual è l'iperpiano migliore, quindi il margine funzionale rispetto ai punti del dataset deve essere alta.

Quindi come confrontiamo due iperpiani? Vediamo quale ha distanza maggiore. I punti sono fissi, l'iperpiano può essere mosso.



Functional margin's drawback

hypothesis function h:

$$h(\mathbf{x}_i) = \begin{cases} +1 & \text{if} \quad \mathbf{w} \cdot \mathbf{x}_i + b \geq 0 \\ -1 & \text{if} \quad \mathbf{w} \cdot \mathbf{x}_i + b < 0 \end{cases}$$

which is equivalent to:

$$h(\mathbf{x}_i) = \text{sign}(\mathbf{w} \cdot \mathbf{x}_i + b)$$

Quick recap

- Notice how the predicted class depends only on the sign of h.
- Try to replace w, e.g., with 2w, and b with 2b,

It results in multiplying our functional margin by a factor of 2.

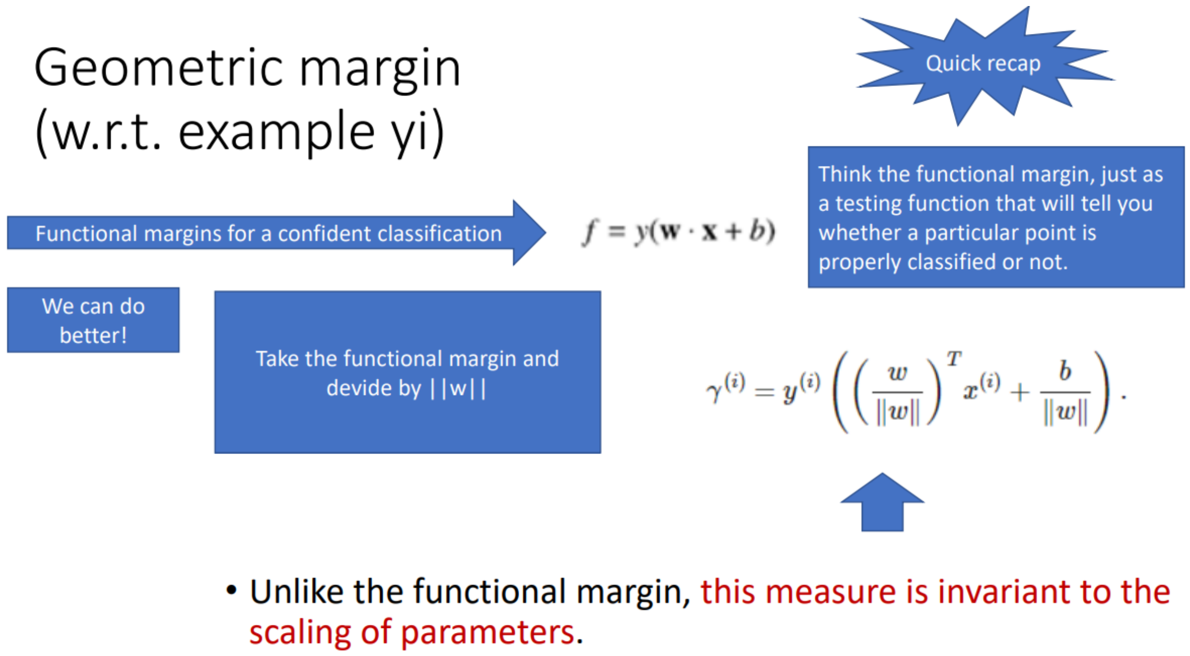
- h(wx+b) = h(2wx+2b)
- (w,b) and (2w,2b) represent the same hyperplane!

$$f = y \times \beta$$

$$f = y(\mathbf{w} \cdot \mathbf{x} + b)$$

- it seems that by exploiting our freedom to scale w and b, we can make the functional margin arbitrarily large without really changing anything meaningful.

- This gives the false idea that our model is 2 times more confident in its predictions.



Geometric margin (w.r.t. example yi)

Quick recap

Functional margins for a confident classification

$f = y(\mathbf{w} \cdot \mathbf{x} + b)$

Think the functional margin, just as a testing function that will tell you whether a particular point is properly classified or not.

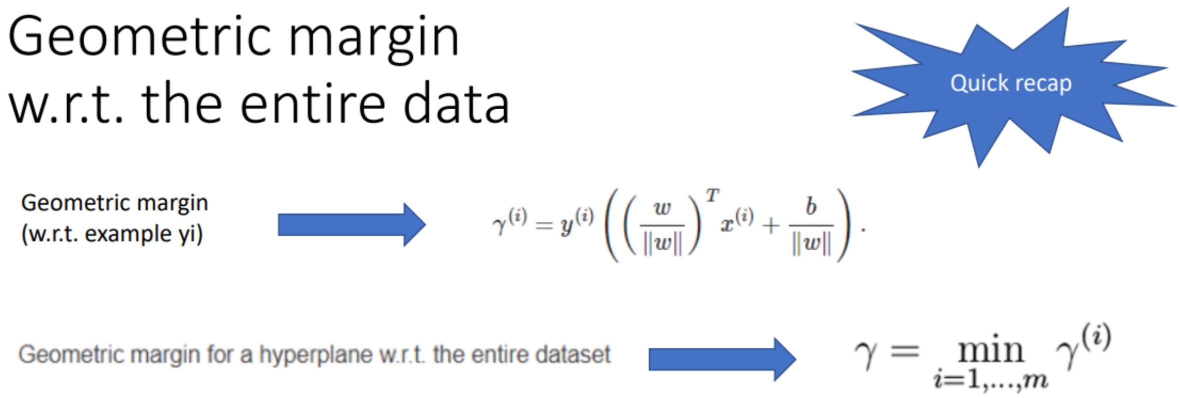
We can do better!

Take the functional margin and devide by ||w||

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

- Unlike the functional margin, this measure is invariant to the scaling of parameters.

Per riassumere: SVM cerca qualcosa meglio del percettrone, ovvero l’iperpiano che massimizzi la distanza da una parte dai punti positivi e dall’altra dai punti negativi.



Geometric margin w.r.t. the entire data

Quick recap

Geometric margin (w.r.t. example yi)

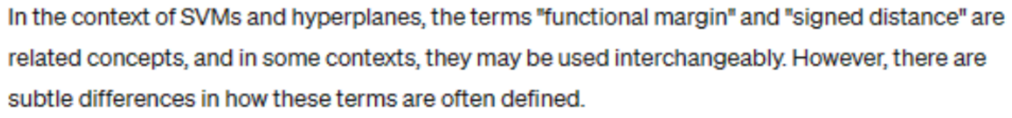
$$\gamma^{(i)} = y^{(i)} \left(\left(\frac{w}{||w||} \right)^T x^{(i)} + \frac{b}{||w||} \right).$$

Geometric margin for a hyperplane w.r.t. the entire dataset

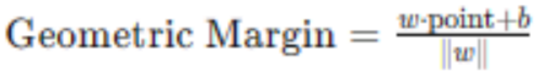
$$\gamma = \min_{i=1,\dots,m} \gamma^{(i)}$$

REMARK

- While the signed distance (functional margin) gives us a measure of how well a point is classified, it does not directly represent the distance of the point from the hyperplane.
- Additionally, the sign of the functional margin indicates the side of the hyperplane on which the point lies (positive for one side, negative for the other).

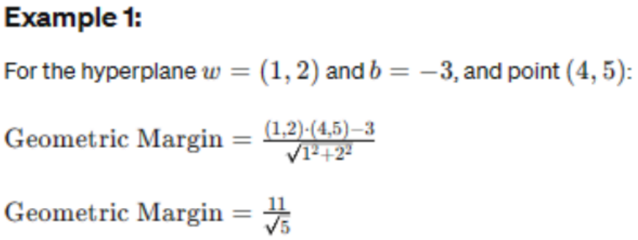


In the context of SVMs and hyperplanes, the terms "functional margin" and "signed distance" are related concepts, and in some contexts, they may be used interchangeably. However, there are subtle differences in how these terms are often defined.



$$\text{Geometric Margin} = \frac{w \cdot \text{point} + b}{\|w\|}$$

- compute the geometric margin for the points



Example 1:

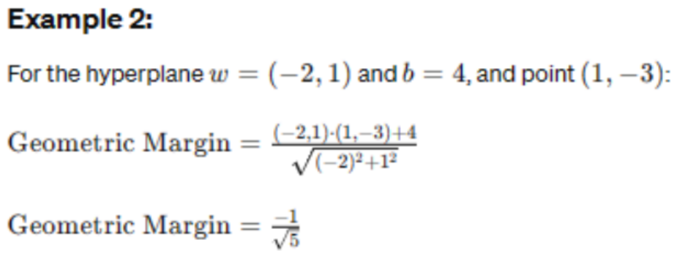
For the hyperplane $w = (1, 2)$ and $b = -3$, and point $(4, 5)$:

$$\text{Geometric Margin} = \frac{(1,2)\cdot(4,5)-3}{\sqrt{1^2+2^2}}$$

$$\text{Geometric Margin} = \frac{11}{\sqrt{5}}$$

$$\text{Geometric Margin} = \frac{w \cdot \text{point} + b}{\|w\|}$$

- compute the geometric margin for the points



Example 2:

For the hyperplane $w = (-2, 1)$ and $b = 4$, and point $(1, -3)$:

$$\text{Geometric Margin} = \frac{(-2,1)\cdot(1,-3)+4}{\sqrt{(-2)^2+1^2}}$$

$$\text{Geometric Margin} = \frac{-1}{\sqrt{5}}$$

REMARK

- the geometric margin provides a measure of the perpendicular distance from each point to the hyperplane, taking into account the scale of the weight vector *w*.
- It is a more meaningful measure in the context of Support Vector Machines, as it directly represents the margin that SVM seeks to maximize during training.

Quick recap

SVM & optimization

> *The SVM aims to maximize the minimum geometric margin.*

The SVM seek the hyperplane that is as far as possible from the closest member of each class.

Optimal Separating Hyperplane - Idea:
choose the hyperplane which maximize the margin from both the classes of training data

-> Maximizing the distance between the nearest points of each class (minimum of geometric margins of all the examples)

Specifically, this will result in a classifier that separates the positive and the negative training examples with a “gap”

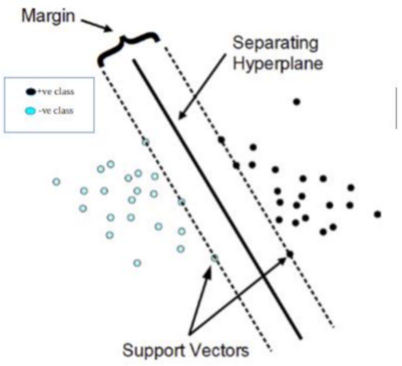
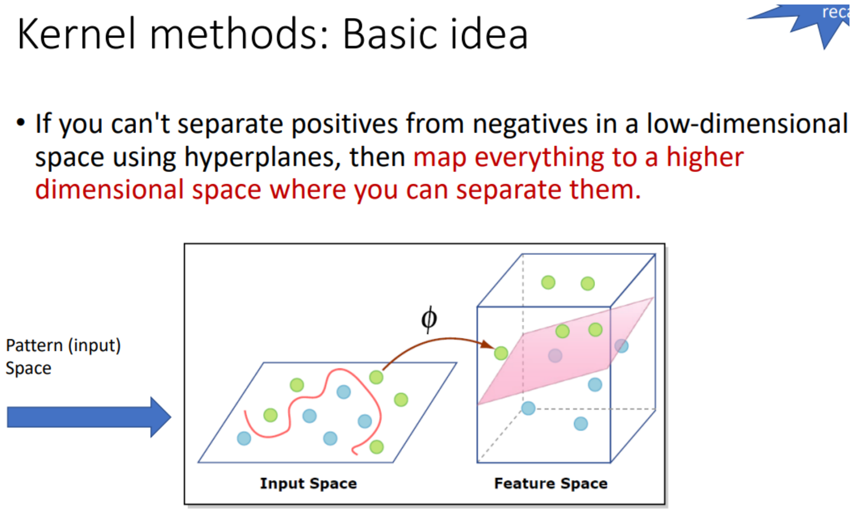


Fig 1. Diagrammatic representation of SVM for linearly separable dataset
(Source: https://www.researchgate.net/figure/Classification-of-data-by-suppo...)

Kernel methods

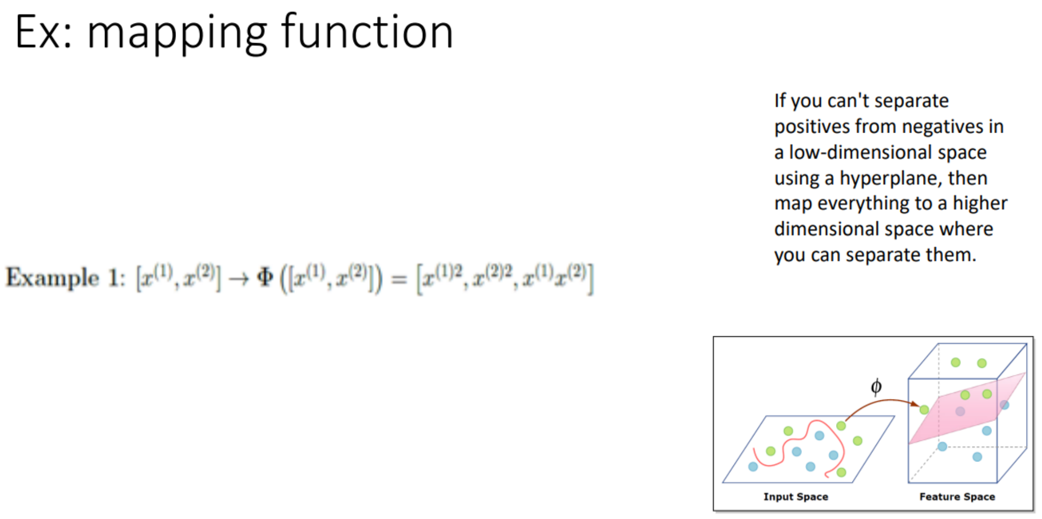
Kernel Methods operate by implicitly mapping input data into a high-dimensional feature space, where linear methods can be applied.

The kernel function plays a crucial role in this process, defining the similarity or inner product between pairs of data points.



Kernel methods: Basic idea

- If you can't separate positives from negatives in a low-dimensional space using hyperplanes, then map everything to a higher dimensional space where you can separate them.



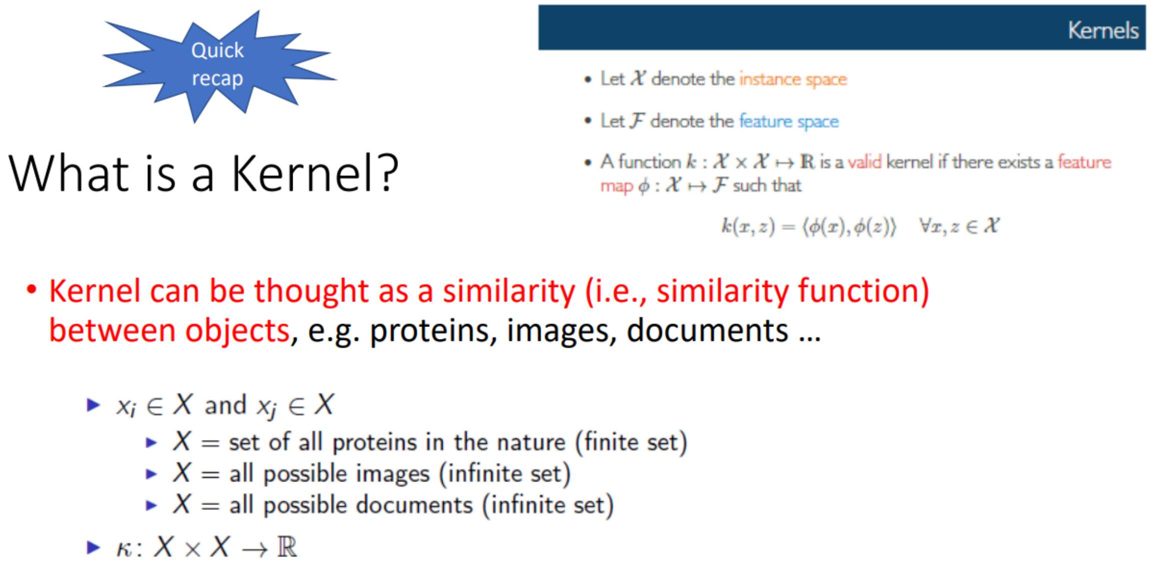
Ex: mapping function

If you can't separate positives from negatives in a low-dimensional space using a hyperplane, then map everything to a higher dimensional space where you can separate them.

Example 1: $[x^{(1)}, x^{(2)}] \rightarrow \Phi\left([x^{(1)}, x^{(2)}]\right) = [x^{(1)2}, x^{(2)2}, x^{(1)}x^{(2)}]$

x e z sono gli argomenti della funzione Kernel, sono gli input originali del problema.

Una funzione K per essere un Kernel deve calcolare il prodotto scalare tra due oggetti trasformati nello spazio delle feature.



Quick recap

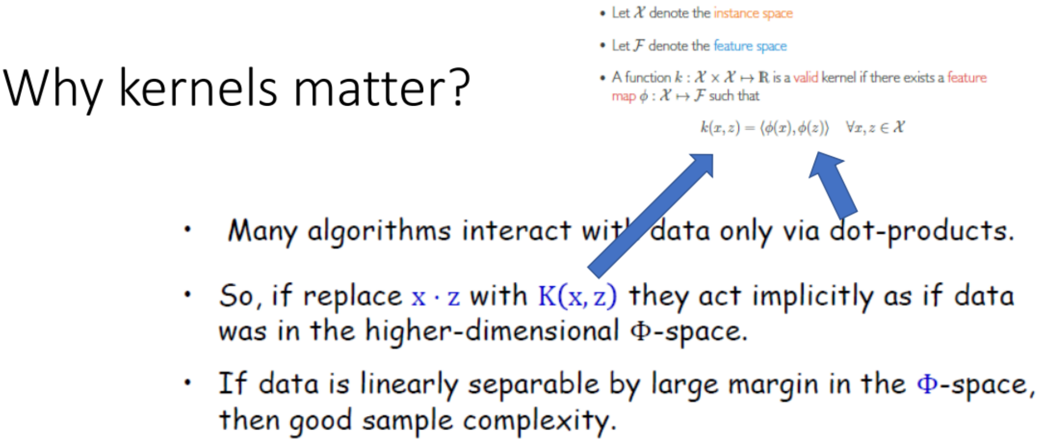
What is a Kernel?

Kernels

- Let $\mathcal{X}$ denote the instance space
- Let $\mathcal{F}$ denote the feature space
- A function $k : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ is a valid kernel if there exists a feature map $\phi : \mathcal{X} \mapsto \mathcal{F}$ such that

$$k(x, z) = \langle \phi(x), \phi(z) \rangle \quad \forall x, z \in \mathcal{X}$$

- Kernel can be thought as a similarity (i.e., similarity function) between objects, e.g. proteins, images, documents …
  - $x_i \in X$ and $x_j \in X$
    - $X =$ set of all proteins in the nature (finite set)
    - $X =$ all possible images (infinite set)
    - $X =$ all possible documents (infinite set)
  - $\kappa : X \times X \to \mathbb{R}$



Why kernels matter?

- Let $\mathcal{X}$ denote the instance space
- Let $\mathcal{F}$ denote the feature space
- A function $k : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ is a valid kernel if there exists a feature map $\phi : \mathcal{X} \mapsto \mathcal{F}$ such that

$$k(x, z) = \langle \phi(x), \phi(z) \rangle \quad \forall x, z \in \mathcal{X}$$

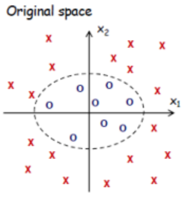
- Many algorithms interact with data only via dot-products.
- So, if replace $\mathrm{x} \cdot \mathrm{z}$ with $\mathrm{K}(\mathrm{x}, \mathrm{z})$ they act implicitly as if data was in the higher-dimensional Φ-space.
- If data is linearly separable by large margin in the Φ-space, then good sample complexity.

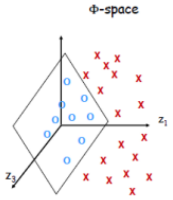
Quindi con il kernel rimpiazzo il prodotto con il calcolo del kernel. Questo trasforma i punti.

Ex:
Polinomial kernel

$$k(\mathbf{x}, \mathbf{z}) = (\mathbf{x}'\mathbf{z})^d$$

Let x = (x1,x2) and z = (z1,z2), take d=2





$$\begin{aligned}
(\mathbf{x} \cdot \mathbf{z})^2 &= (x_1 z_1 + x_2 z_2)^2 \\
&= (x_1^2 z_1^2 + x_2^2 z_2^2 + 2 x_1 z_1 x_2 z_2) \\
&= < (x_1^2, x_2^2, \sqrt{2} x_1 x_2), (z_1^2, z_2^2, \sqrt{2} z_1 z_2) > \\
&= < \Phi(\mathbf{x}), \Phi(\mathbf{z}) >
\end{aligned}$$

In this case we have $\Phi(x_1, x_2) = (x_1^2, x_2^2, \sqrt{2} x_1 x_2).$