

SUPPORT VECTOR MACHINES

I support vector machines (SVM) è un metodo per classificare dati in formato vettoriale, attraverso dei dati di training.

Chiameremo questi vettori **oggetti**. I dati di training hanno ovviamente una classe.

CLASSIFICAZIONE BINARIA

Abbiamo due set $A, B \subset \mathbb{R}^m$ con le label (+1 per A, -1 per B).

Abbiamo:

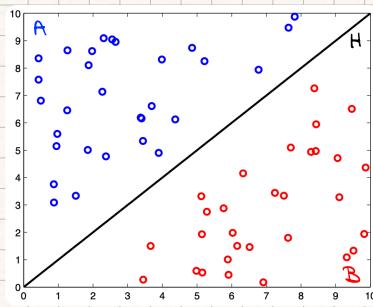
- 1) \mathbb{R}^m è l'input space
- 2) $A \cup B$ è il training set

Assumiamo che A e B siano linearmente separabili, ovvero esiste un iper piano H che divide A e B:

$$H = \{x \in \mathbb{R}^m : w^T x + b = 0\}: \begin{array}{l} w^T x_i + b > 0 \quad \forall x_i \in A \\ w^T x_j + b < 0 \quad \forall x_j \in B \end{array} \quad (\perp)$$

Funzione di decisione

$$f(x) = \text{sign}(w^T x + b) = \begin{cases} 1 & \text{se } w^T x + b > 0 \\ -1 & \text{se } w^T x + b < 0 \end{cases}$$



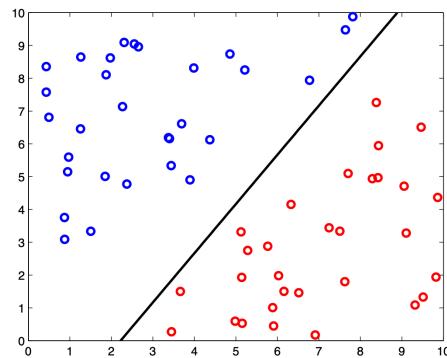
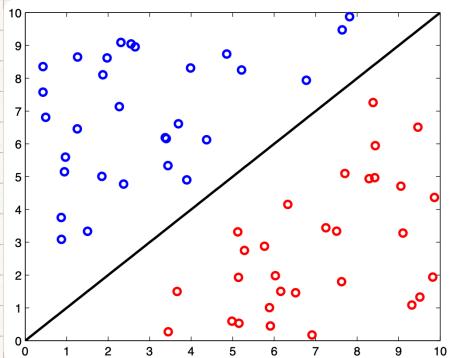
Condizione necessaria e sufficiente

La condizione necessaria e sufficiente per cui vale la (1) è che:

$$\text{Conv}(A) \cup \text{Conv}(B) = \emptyset$$

Esistono più iperpiani

Bisogna scegliere fra i possibili iperpiani, il migliore



Definizione di SEPARABILITÀ FORTE

Dati due set $A, B \subset \mathbb{R}^n$, si dicono strongly separabile se esiste $\alpha \in \mathbb{R}^n \setminus \{0\}$, $\beta \in \mathbb{R}$, $\exists \varepsilon > 0$:

$$\alpha^T x + \beta \geq \varepsilon \quad \forall x \in A$$

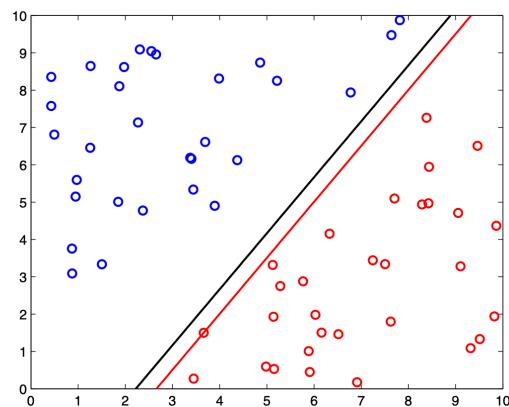
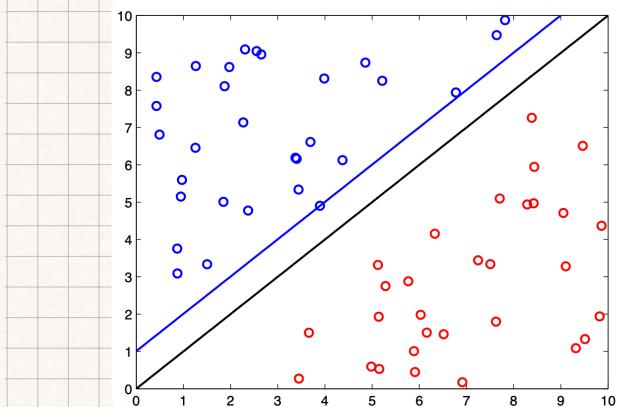
$$\alpha^T x + \beta \leq -\varepsilon \quad \forall x \in B$$

LIN̄AR SVM

Definizione

Se H è un iper piano di separazione, ALLORA il margine di separazione di H è definito come la minima distanza fra H e $A \cup B$:

$$f(H) = \min_{x \in A \cup B} \frac{|w^T x + b|}{\|w\|}$$



Obiettivo: Trovare l'iper piano che ha il massimo margine di separazione.

TEOREMA

Trovare l'iper piano con il massimo margine di separazione equivale a risolvere il seguente problema di programmazione quadratica convessa:

$$\begin{cases} \min_{w, b} \frac{1}{2} \|w\|^2 \\ w^T x_i + b \geq 1 \quad \forall x_i \in A \\ w^T x_j + b \leq -1 \quad \forall x_j \in B \end{cases}$$

Dimostrazione

Presi i due iperpiani:

$$1) \mathbf{w}^T \mathbf{x} + b = 1$$

$$2) \mathbf{w}^T \mathbf{x} + b = -1$$

Si può dimostrare che, la distanza fra questi due piani è pari a: $\frac{2}{\|\mathbf{w}\|}$.

Se prendiamo un punto $\hat{\mathbf{x}}$: $\mathbf{w}^T \hat{\mathbf{x}} + b = 1$, possiamo calcolare la distanza fra $\hat{\mathbf{x}}$ e l'altro iperpiano $\mathbf{w}^T \mathbf{x} + b = -1$:

$$\frac{|\mathbf{w}^T \hat{\mathbf{x}} + b + 1|}{\|\mathbf{w}\|} = \frac{2}{\|\mathbf{w}\|}$$

Che quindi, minimizzando $\|\mathbf{w}\|$, otteniamo due iperpiani che hanno distanza massima fra loro.

Note: Il nostro problema ha soluzione unica (\mathbf{w}^*, b^*) .

Assumendo $A \in \mathbb{R}^{m \times n}$, $B \in \mathbb{R}^{p \times n}$, $\mathbf{w} = (w_1 \dots w_m)^T$ il nostro problema diventa:

$$\begin{cases} \min \frac{1}{2} (\mathbf{w}, b)^T C \begin{pmatrix} \mathbf{w} \\ b \end{pmatrix} \\ D \begin{pmatrix} \mathbf{w} \\ b \end{pmatrix} \leq d \end{cases}$$

Dove con D intendiamo sia A che B :

$$-A\mathbf{w} - b \leq 1$$

$$B\mathbf{w} + b \leq -1$$

ALTRA FORMULAZIONE DEL PROBLEMA

$$\left\{ \begin{array}{l} \min_{w, b} \frac{1}{2} \|w\|^2 \\ w^T x_i + b \geq 1 \quad \forall x_i \in A \\ w^T x_j + b \leq -1 \quad \forall x_j \in B \end{array} \right. = \left\{ \begin{array}{l} \min_{w, b} \frac{1}{2} \|w\|^2 \\ 1 - y_i (w^T x_i + b) \leq 0 \quad \forall i = 1 \dots \ell \end{array} \right. \text{Linear SVM}$$

Dove $\ell = |A \cup B|$, mentre $y_i = \begin{cases} 1 & \text{se } x_i \in A \\ -1 & \text{se } x_i \in B \end{cases}$

Dato che il Linear SVM è **convesso**, è utile considerare il duale Lagrangiano:

$$\begin{aligned} L(w, b, \lambda) &= \frac{1}{2} \|w\|^2 + \sum_{i=1}^{\ell} \lambda_i (1 - y_i (w^T x_i + b)) = \\ &= \frac{1}{2} \|w\|^2 - \sum_{i=1}^{\ell} \lambda_i y_i w^T x_i - b \sum_{i=1}^{\ell} \lambda_i y_i + \sum_{i=1}^{\ell} \lambda_i \end{aligned}$$

Dove:

1) se $\sum_{i=1}^{\ell} \lambda_i y_i \neq 0$ ALLORA $\min_{w, b} L(w, b, \lambda) = -\infty$

2) se $\sum_{i=1}^{\ell} \lambda_i y_i = 0$ ALLORA

- L non dipende da b

- L è **strongly convex** rispetto a w

- $\underset{w}{\operatorname{argmin}} L(w, b, \lambda)$ è un unico punto stazionario:

$$\nabla_w L(w, b, \lambda) = w - \sum_{i=1}^{\ell} \lambda_i y_i x_i = \emptyset$$

Note: se vale quest'ultima uguaglianza e se $\sum \lambda_i y_i = 0$
allora possiamo scrivere:

$$\begin{aligned} L(w, b, \lambda) &= \frac{1}{2} \|w\|^2 - \underbrace{\sum_{i=1}^e \lambda_i y_i w^T x_i}_{} - b \underbrace{\sum_{i=1}^e \lambda_i y_i}_{=} + \sum_{i=1}^e \lambda_i \\ &= \frac{1}{2} w^T w - w^T w + \sum_{i=1}^e \lambda_i = -\frac{1}{2} w^T w + \sum_{i=1}^e \lambda_i \end{aligned}$$

Perché abbiamo $w = \sum_{i=1}^e \lambda_i y_i x_i$ da cui il passaggio in verde.

Funzione Duale $\varphi(\lambda)$

$$\varphi(\lambda) = \begin{cases} -\infty \\ -\frac{1}{2} \sum_{i=1}^e \sum_{j=1}^e y_i y_j x_i^T x_j \lambda_i \lambda_j + \sum_{i=1}^e \lambda_i \text{ se } \sum_{i=1}^e \lambda_i y_i \neq 0 \\ \infty \text{ se } \sum_{i=1}^e \lambda_i y_i = 0 \end{cases}$$

con $\lambda \geq 0$

PROBLEMA DUALE

$$\left\{ \begin{array}{l} \max_{\lambda} -\frac{1}{2} \sum_{i=1}^e \sum_{j=1}^e y_i y_j x_i^T x_j \lambda_i \lambda_j + \sum_{i=1}^e \lambda_i \\ \sum_{i=1}^e \lambda_i y_i = 0 \\ \lambda \geq 0 \end{array} \right.$$

Oppure in forma vettoriale:

$$\left\{ \begin{array}{l} \max_{\lambda} -\frac{1}{2} \lambda^T X^T X \lambda + e^T \lambda \\ \lambda \\ e \\ \sum_{i=1}^e \lambda_i y_i = 0 \\ \lambda \geq 0 \end{array} \right.$$

Dove $X = \{y_1 x_1, y_2 x_2, \dots, y_e x_e\} \in \mathbb{R}^{n \times e}$
 e $e^T = (1 \dots 1) \in \mathbb{R}^e$

Considerazioni

- 1) $X^T X$ è sempre semidefinita positiva, quindi il duale è sempre **convesso**
- 2) Un moltiplicatore λ^* del KKT associato all'ottimo del primale (w^*, b^*) è ottimo anche per il duale.
- 3) Se $\lambda_i^* > 0$, ALLORA x_i è detto **SUPPORT VECTOR**
- 4) Se λ_i^* è ottimo per il duale ALLORA, grazie alla uguaglianza $w = \sum \lambda_i y_i x_i$, abbiamo che:

$$w^* = \sum_{i=1}^e \lambda_i^* y_i x_i$$

- 5) b^* è ottenuto usando la **condizione complementare**

$$\lambda_i^* \left(1 - y_i ((w^*)^T x_i + b^*) \right) = 0$$

da cui, per $\hat{w}_i^* > 0$:

$$b^* = \frac{1}{y_i} - (\hat{w}^*)^T x_i$$

Così da trovare

a) $H(x) = (\hat{w}^*)^T x + b = 0$

b) $f(x) = \text{Sign}(H(x))$

Cosa succede se A e B non sono linearmente separabili?

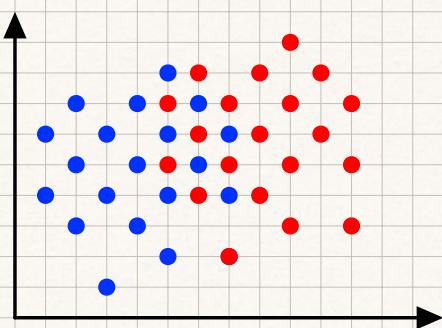
Questo vuol dire che vale la diseguazione:

$$\hat{y}_i - y_i : (\hat{w}^T x_i + b) \leq 0 \quad i = 1, \dots, \ell$$

NON HA SOLUZIONI.

Introduciamo la variabile $\xi_i \geq 0$ e consideriamo il sistema:

$$\begin{cases} \hat{y}_i - y_i : (\hat{w}^T x_i + b) \leq \xi_i & i = 1, \dots, \ell \\ \xi_i \geq 0 \end{cases}$$



È inevitabile che qualche x_i viene mal classificato, allora il ξ_i associato è $\xi_i > 1$ e $\sum_{i=1}^{\ell} \xi_i$ è un limite superiore di punti miss-classificati.

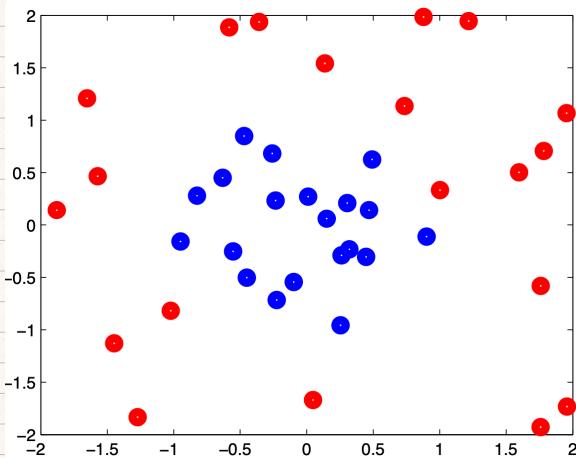
Aggiungiamo alla funzione obiettivo $C \sum_{i=1}^e \xi_i$ con $C > 0$.

LIN̄AR SVM CON SOFT MARGIN

$$\begin{cases} \min_{w, b} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^e \xi_i \\ 1 - y_i (w^\top x_i + b) \leq \xi_i \quad \forall i = 1 \dots e \\ \xi_i \geq 0 \quad \forall i = 1 \dots e \end{cases}$$

SVM NON LINEARI

Consideriamo due set A e B che non sono linearmente separabili in \mathbb{R}^n .



Ma sono linearmente separabili in uno spazio diverso?

Usiamo una mappa $\Phi: \mathbb{R}^n \rightarrow H$, dove H è uno spazio con molte dimensioni (Anche infinite), e H è definito **feature space**.

L'idea è separare linearmente i punti dei due set A e B nello spazio delle caratteristiche $\Phi(x_i)$.

PRIMALI SVM NON-LINEARI

$$\begin{cases} \min_{w, b} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{\ell} \xi_i \\ 1 - y_i (w^\top \Phi(x_i) + b) \leq \xi_i \quad \forall i=1 \dots \ell \\ \xi_i \geq 0 \quad \forall i=1 \dots \ell \end{cases}$$

DUALS SVM NON-LINEARE

$$\left\{ \begin{array}{l} \max_{\lambda} -\frac{1}{2} \sum_{i=1}^e \sum_{j=1}^e y_i y_j \Phi(x_i^\top) \Phi(x_j) \lambda_i \lambda_j + \sum_{i=1}^e \lambda_i \\ \sum_{i=1}^e \lambda_i y_i = 0 \\ 0 \leq \lambda_i \leq C \quad i = 1, \dots, e \end{array} \right.$$

Sia λ^* la soluzione del duale allora :

$$w^* = \sum_{i=1}^e \lambda_i^* y_i \Phi(x_i)$$

Fissando $0 < \lambda_i^* < C$ possiamo trovare b^* con

$$y_i \left[\sum_{i=1}^e \lambda_i^* y_i \Phi(x_i)^\top \Phi(x_i) + b^* \right] - 1 = 0$$

Decision function

$$f(x) = \text{sign}\left((w^*)^\top \Phi(x_i) + b^* \right)$$

$$= \text{sign}\left(\sum_{i=1}^e \lambda_i^* y_i \Phi(x_i)^\top \Phi(x) + b^* \right)$$

Dove $f(x)$ dipende da: λ^* , $\Phi(x_i)^\top \Phi(x)$, b^*

Note: non bisogna conoscere $\Phi(x_i)$ ma solo $\Phi(x_i)^\top \Phi(x_j)$

FUNZIONI KERNEL

Definiamo $K: \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}$ e chiamata Kernel se
 $\exists \Phi: \mathbb{R}^n \rightarrow H:$

$$K(x, y) = \langle \Phi(x), \Phi(y) \rangle$$

Prodotto scalare nel feature space.

TEOREMA

Se $K: \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}$ è una funzione Kernel, e $x_1 \dots x_e \in \mathbb{R}^n$,
allora la matrice K , così definita:

$$K_{ij} = K(x_i, x_j)$$

è semidefinita positiva.

Possiamo ridefinire il doppio D :

$$\left\{ \begin{array}{l} \max_{\lambda} -\frac{1}{2} \sum_{i=1}^e \sum_{j=1}^e y_i y_j K(x_i^\top, x_j) \lambda_i \lambda_j + \sum_{i=1}^e \lambda_i \\ \sum_{i=1}^e \lambda_i y_i = 0 \\ 0 \leq \lambda_i \leq C \quad i = 1, \dots, e \end{array} \right.$$

IN PRATICA

- 1) Scegliere una funzione Kernel
- 2) Trovare una soluzione ottimale del duale: λ^*
- 3) Scegliere $i: \emptyset < \lambda_i^* < C$ e trovare b^*

$$b^* = \frac{1}{y_i} - \sum_{j=1}^e \lambda_j^* y_j K(x_i, x_j)$$

- 4) Calcolare la decision function:

$$f(x) = \text{sign} \left(\sum_{j=1}^e \lambda_j^* y_j K(x_j, x) + b^* \right)$$

Note: la superficie di separazione $f(x) = 0$ è:

- 1) Lineare nello spazio delle caratteristiche
- 2) Non Lineare nello spazio degli ingressi