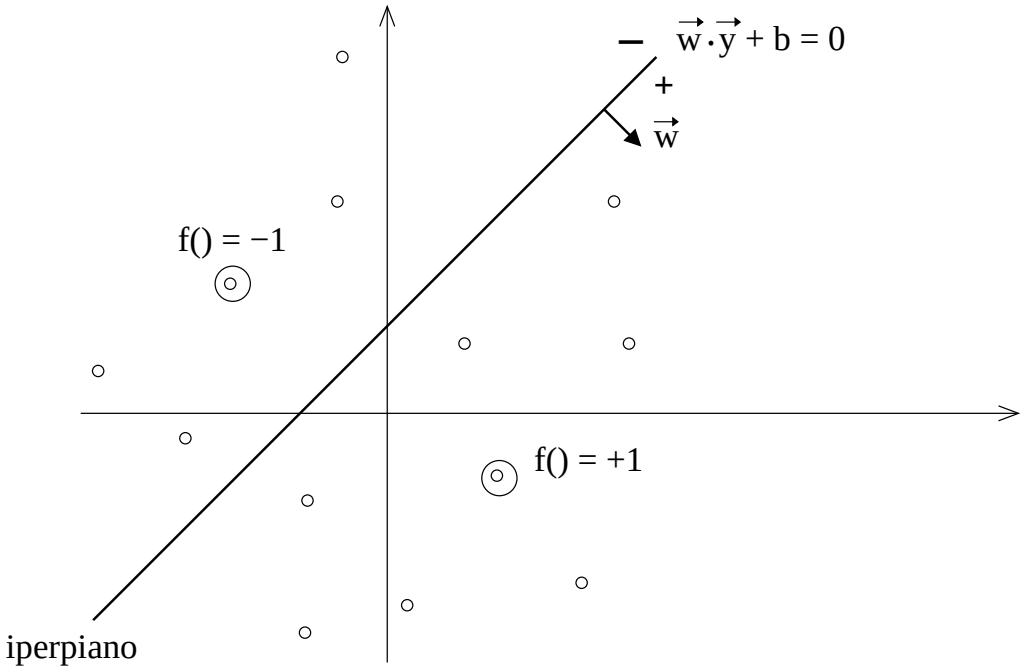
VC-dimension: Esempio

Quale è la VC-dimension di $\mathcal{H}_1$?

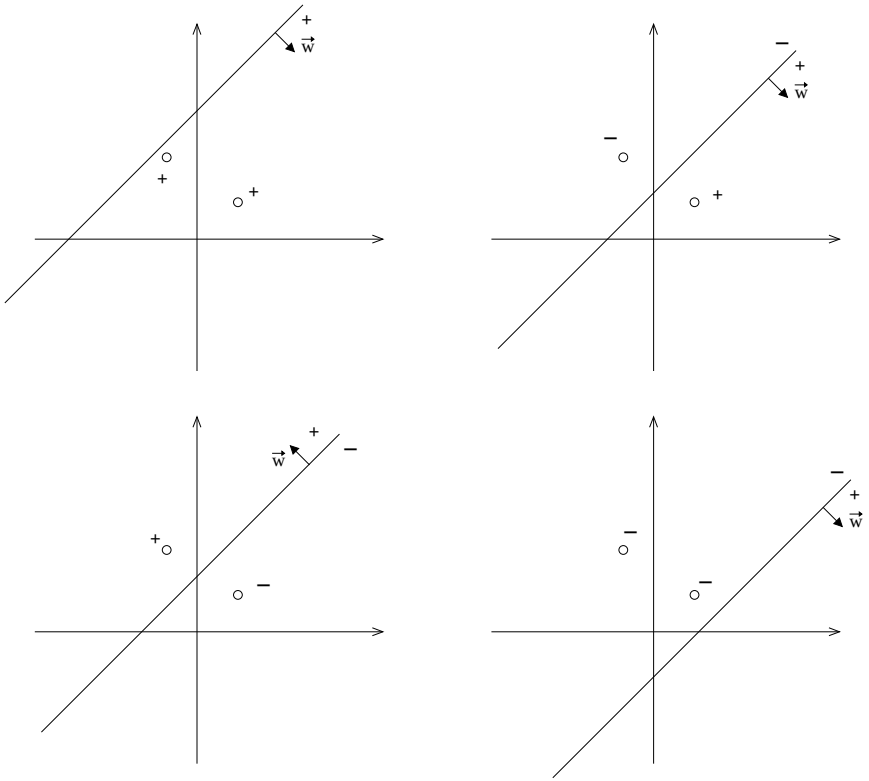
$$\mathcal{H}_1 = \{f_{(\vec{w},b)}(\vec{y}) | f_{(\vec{w},b)}(\vec{y}) = sign(\vec{w} \cdot \vec{y} + b), \vec{w} \in \mathbb{R}^2, b \in \mathbb{R}\}$$



VC-dimension: Esempio

Quale è la VC-dimension di $\mathcal{H}_1$?

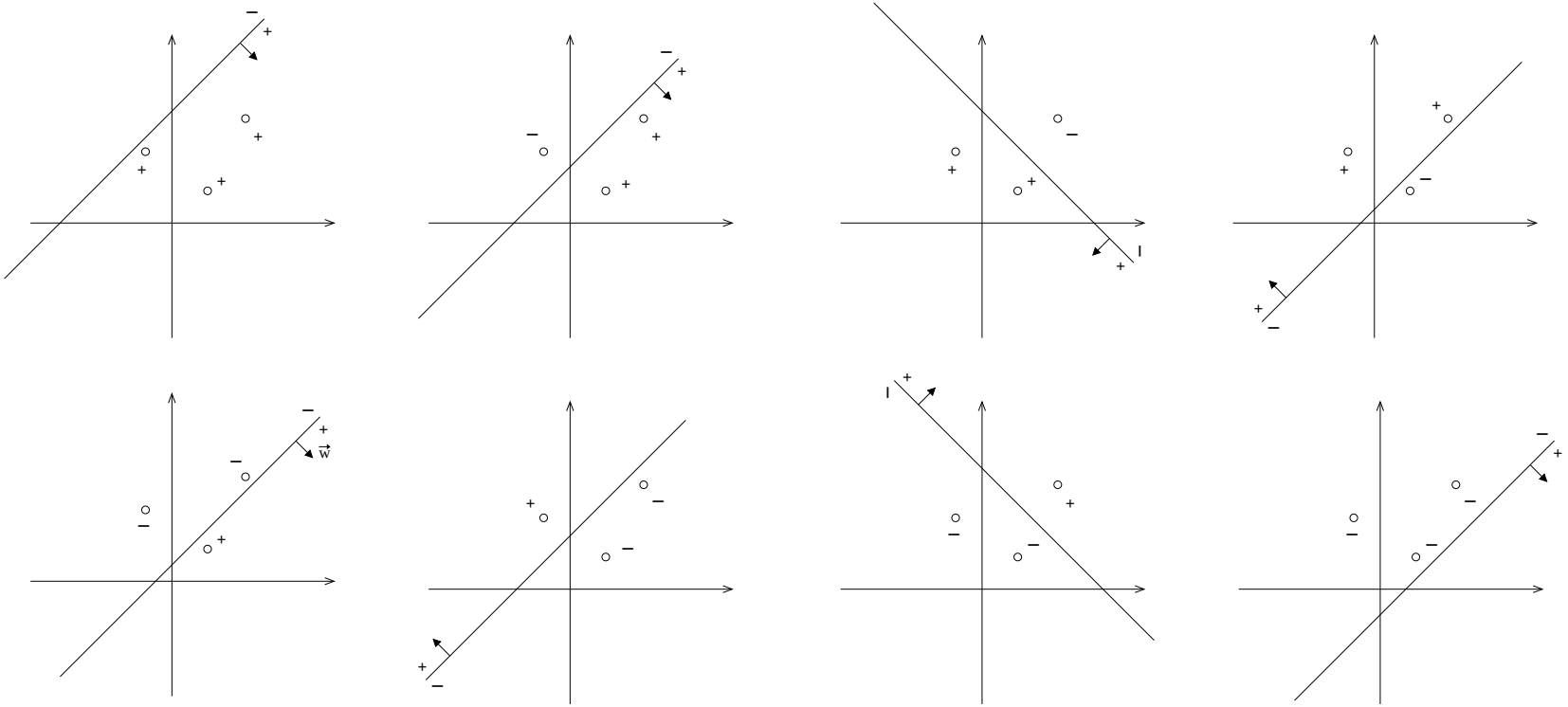
$VC(\mathcal{H}) \geq 1$ banale. Vediamo cosa succede con 2 punti:



VC-dimension: Esempio

Quale è la VC-dimension di $\mathcal{H}_1$?

Quindi $VC(\mathcal{H}) \geq 2$. Vediamo cosa succede con 3 punti:



VC-dimension: Esempio

Quale è la VC-dimension di $\mathcal{H}_1$?

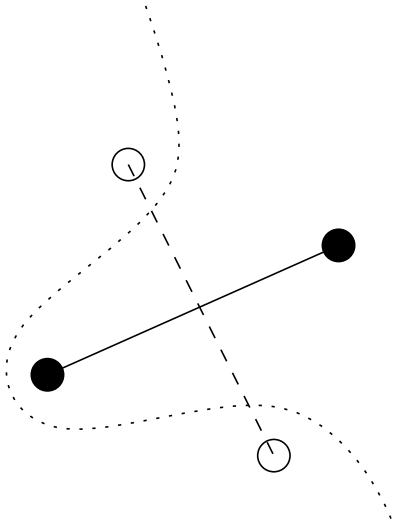
Quindi $VC(\mathcal{H}) \geq 3$. Cosa succede con 4 punti ?

VC-dimension: Esempio

Quale è la VC-dimension di $\mathcal{H}_1$?

Quindi $VC(\mathcal{H}) \geq 3$. Cosa succede con 4 punti ? Non si riesce a frammentare 4 punti!!

Infatti esisteranno sempre due coppie di punti che se unite con un segmento provocano una intersezione fra i due segmenti e quindi, ponendo ogni coppia di punti in classi diverse, per separarli non basta una retta, ma occorre una curva. Quindi $VC(\mathcal{H}) = 3$



VC-dimension

Dimostriamo che $VC(\mathcal{H}) \leq \log_2(|\mathcal{H}|)$

- Per ogni S tale che $\mathcal{H}$ frammenta S si ha $|\mathcal{H}| \geq 2^{|S|}$, infatti $\mathcal{H}$ può realizzare tutte le possibili dicotomie di S, che sono esattamente $2^{|S|}$.
- Scegliendo un S per cui vale $|S| = VC(\mathcal{H})$ si ottiene $|\mathcal{H}| \geq 2^{VC(\mathcal{H})}$

Quindi, applicando $\log_2$ ad entrambi i membri dell'ultima disuguaglianza, possiamo concludere che $\log_2(|\mathcal{H}|) \geq VC(\mathcal{H})$

Bound sull'Errore Ideale per Classificazione Binaria

Consideriamo un problema di classificazione binario (i.e., apprendimento di concetti). Dati

- Training Set $Tr = \{(\boldsymbol{x}^{(1)}, f(\boldsymbol{x}^{(1)})), \ldots, (\boldsymbol{x}^{(N_{tr})}, f(\boldsymbol{x}^{(N_{tr})}))\}$
- Spazio delle Ipotesi $\mathcal{H} = \{h_{\boldsymbol{w}}(\boldsymbol{x}) | \boldsymbol{w} \in \mathbb{R}^k\}$
- Algoritmo di Apprendimento L che restituisce l'ipotesi $h_{\boldsymbol{w}^*}(\boldsymbol{x})$, dove $\boldsymbol{w}^*$ minimizza l'errore empirico $\boldsymbol{error_{Tr}(h_w(x))}$

è possibile derivare dei bound sull'errore ideale (detto anche errore di generalizzazione), validi con probabilità $1 - \delta$, che hanno una forma del tipo

$$error_{\mathcal{D}}(h_{\boldsymbol{w}^*}(\boldsymbol{x})) \leq error_{Tr}(h_{\boldsymbol{w}^*}(\boldsymbol{x})) + \epsilon(N_{tr}, VC(\mathcal{H}), \delta)$$

Esempio:

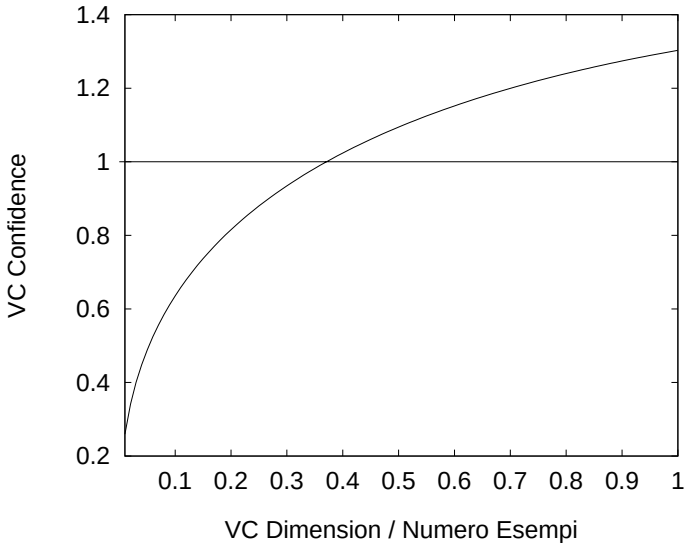
$$error_{\mathcal{D}}(h_{\boldsymbol{w}^*}(\boldsymbol{x})) \leq \underbrace{error_{Tr}(h_{\boldsymbol{w}^*}(\boldsymbol{x}))}_{A} + \underbrace{\sqrt{\frac{VC(\mathcal{H})}{N_{tr}}(\log(\frac{2N_{tr}}{VC(\mathcal{H})}) + 1) - \frac{1}{N_{tr}}\log(\delta)}}_{B}$$

Bound sull'Errore Ideale per Classificazione Binaria

Si noti che

- il termine A DIPENDE SOLO dalla ipotesi restituita dall'algoritmo di apprendimento L;
- il termine B è INDIPENDENTE dalla ipotesi restituita dall'algoritmo di apprendimento L; in particolare dipende dal rapporto fra VC-dimension dello spazio delle ipotesi $\mathcal{H}$ e il numero di esempi di apprendimento (N_{tr}), oltre ovviamente che dalla confidenza $(1 - \delta)$ con cui il bound è valido.

Il termine B è usualmente chiamato VC-confidence e risulta essere monotono rispetto al rapporto $\frac{VC(\mathcal{H})}{N_{tr}}$; fissato N_{tr} aumenta all'aumentare di $VC(\mathcal{H})$.



Structural Risk Minimization

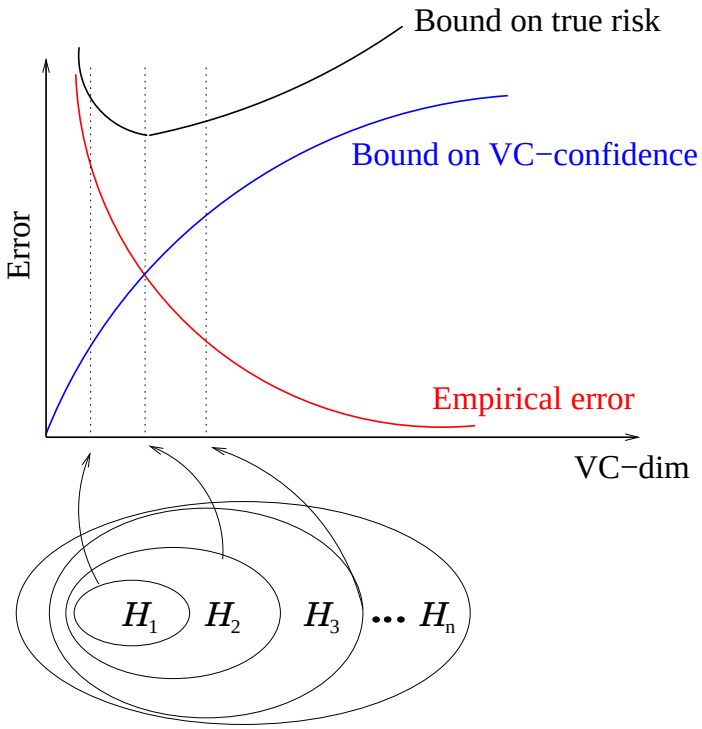
Problema: all'aumentare della VC-dimension diminuisce l'errore empirico (termine A), ma aumenta la VC confidence (termine B)!

L'approccio Structural Risk Minimization tenta di trovare un compromesso tra i due termini:

Si considerano $\mathcal{H}_i$ tali che

- $\mathcal{H}_1 \subseteq \mathcal{H}_2 \subseteq \cdots \subseteq \mathcal{H}_n$
- $VC(\mathcal{H}_1) \leq \cdots \leq VC(\mathcal{H}_n)$
- si seleziona l'ipotesi che ha il bound sull'errore ideale pi`u basso

Esempio: Reti neurali con un numero crescente di neuroni nascosti



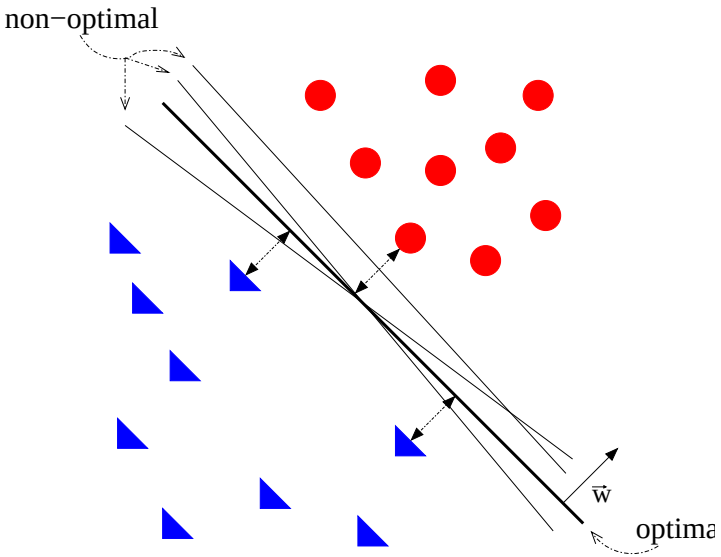
Support Vector Machines: idea base

Possiamo applicare l'approccio Structural Risk Minimization a spazi delle ipotesi costituiti da iperpiani ?

Sappiamo che un iperpiano in uno spazio a m dimensioni ha $VC = m + 1$. Come facciamo a creare una struttura di spazi delle ipotesi con VC-dimension crescente ?

Bisogna porre dei vincoli sugli iperpiani! Consideriamo iperpiani separatori con margine r

Consideriamo il caso in cui gli esempi siano linearmente separabili.
Il margine r è la "distanza" fra l'iperpiano e l'esempio più vicino.
L'iperpiano con margine maggiore è detto ottimo.



Margine

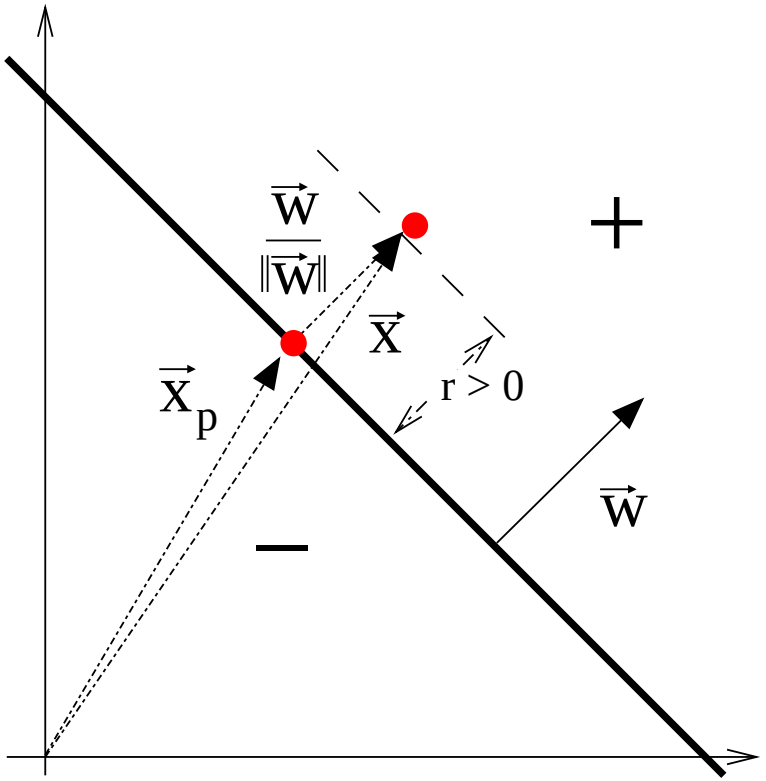
La “distanza” di un vettore da un iperpiano la possimo misurare in senso algebrico.

Dato un iperpiano determinato dalla equazione $\vec{w} \cdot \vec{x} + b = 0$, la funzione discriminante $g(\vec{x}) = \vec{w} \cdot \vec{x} + b$ restituisce la distanza algebrica di $\vec{x}$ dall’iperpiano.

Infatti, se esprimiamo $\vec{x}$ come

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

dove $\vec{x}_p$ è la proiezione normale di $\vec{x}$ sull’iperpiano ed r è la distanza algebrica desiderata ($r > 0$ se $\vec{x}$ è sul lato positivo dell’iperpiano, altrimenti $r < 0$), allora $g(\vec{x}_p) = 0$ (poiché $\vec{x}_p$ risiede sull’iperpiano).



Quindi

$$g(\vec{x}) = \vec{w} \cdot \vec{x} + b = r||\vec{w}||$$

o meglio $r = \frac{\vec{w} \cdot \vec{x} + b}{||\vec{w}||} = \frac{g(\vec{x})}{||\vec{w}||}$

Poiché esiste una infinità di soluzioni che differiscono solo per un fattore di scala su $\vec{w}$ (si noti che l'iperpiano non cambia scalando il suo vettore normale) ci si limita per convenzione a soluzioni che soddisfano l'equazione $\hat{r}||\vec{w}|| = 1$

Si noti che per l'iperpiano ottimo, la distanza assoluta da uno degli esempi positivi più vicini è uguale a quella da uno degli esempi negativi più vicini. Il margine di separazione ρ è quindi definito come il doppio del margine: $\rho = \frac{2}{||\vec{w}||}$

Inoltre, se gli esempi sono linearmente separabili con margine $\hat{r}$ da un iperpiano, allora

$$\frac{y_i g(\vec{x}_i)}{||\vec{w}||} \geq \hat{r} \quad i = 1, \ldots, n$$

dove $y_i = 1$ per esempi positivi e $y_i = -1$ per esempi negativi. Il problema di trovare l'iperpiano ottimo si riduce quindi a quello di minimizzare $||\vec{w}||$.

Margine: Legame con SRM

Theorema *Sia R il diametro della palla più piccola che contiene tutti gli esempi di apprendimento. L'insieme di iperpiani ottimi descritti dall'equazione* $\vec{w} \cdot \vec{x} + b = 0$ *possiede VC-dimension* h *limitata superiormente da*

$$h \leq \min\{\lceil \frac{R^2}{\rho^2} \rceil, m\} + 1$$

dove $\rho = \frac{2}{||\vec{w}||}$ *ed* m *è la dimensionalità dei dati di apprendimento.*

Quindi, se consideriamo gli spazi delle ipotesi

$$\mathcal{H}_k = \{\vec{w} \cdot \vec{x} + b \mid ||\vec{w}||^2 \leq c_k\} \text{ con } c_1 < c_2 < c_3 < \ldots$$

ed i dati sono linearmente separabili, allora l'errore empirico è nullo per tutti gli iperpiani e quindi per minimizzare il bound sull'errore ideale si deve selezionare l'iperpiano con VC-dimension minima, cioè quello che minimizza $||\vec{w}||^2$ (o equivalentemente massimizza il margine di separazione).

Caso Separabile: Formulazione Quadratica

Nel caso di n esempi $\{(\vec{x}_i, y_i)\}_1^n$ linearmente separabili, è possibile trovare l'iperpiano ottimo risolvendo il seguente problema vincolato di ottimizzazione quadratica:

$$\min_{\vec{w},b} \frac{1}{2}||\vec{w}||^2$$

soggetto a: $\forall i \in \{1, \ldots, n\} : y_i(\vec{w} \cdot \vec{x}_i + b) - 1 \geq 0$

Questo problema, detto problema primale, si può risolvere più facilmente passando alla sua formulazione duale.

La teoria della ottimizzazione afferma che:

1. un problema di ottimizzazione possiede una forma duale (più semplice da risolvere) se la funzione di costo e i vincoli sono strettamente convessi;
2. se le condizioni in 1 sono soddisfatte, l'ottimo per il problema duale coincide con l'ottimo del primale.

Il nostro problema primale soddisfa le condizioni in 1.