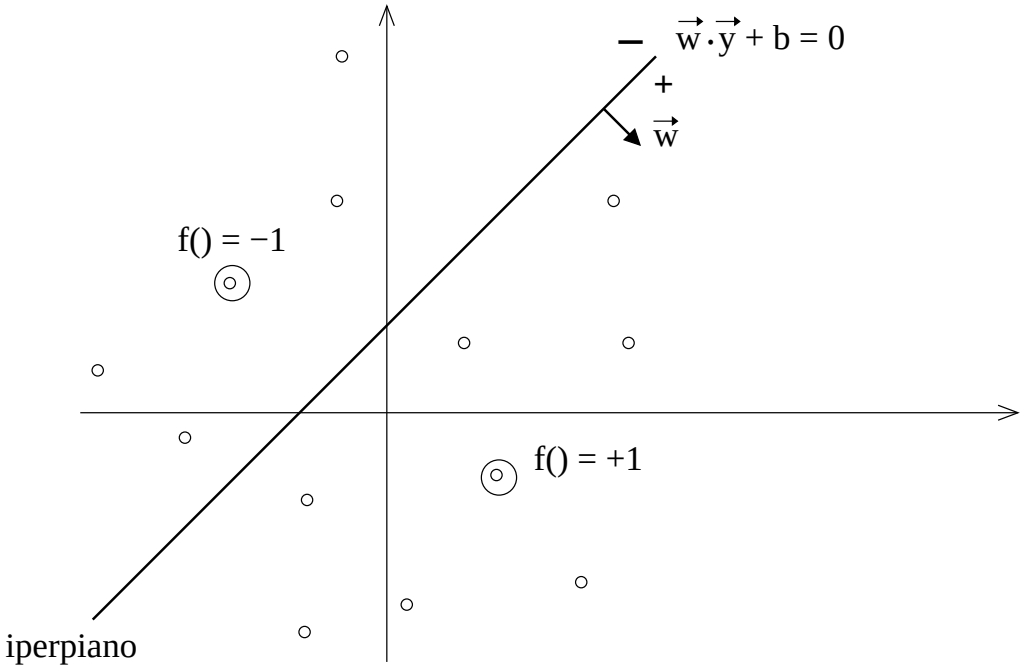
Spazio delle Ipotesi: Esempio 1

Iperpiani in $\mathbb{R}^2$

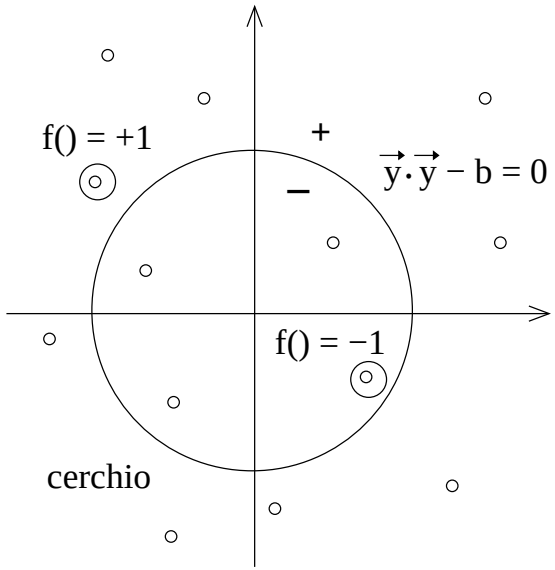
- Spazio delle Istanze $\rightarrow$ punti nel piano: $X = \{\vec{y} \in \mathbb{R}^2\}$
- Spazio delle Ipotesi $\rightarrow$ dicotomie indotte da iperpiani in $\mathbb{R}^2$:
  $\mathcal{H} = \{f_{(\vec{w},b)}(\vec{y}) | f_{(\vec{w},b)}(\vec{y}) = sign(\vec{w} \cdot \vec{y} + b), \vec{w} \in \mathbb{R}^2, b \in \mathbb{R}\}$



Spazio delle Ipotesi: Esempio 2

Dischi in $\mathbb{R}^2$

- Spazio delle Istanze $\rightarrow$ punti nel piano: $X = \{\vec{y} \in \mathbb{R}^2\}$
- Spazio delle Ipotesi $\rightarrow$ dicotomie indotte da dischi in $\mathbb{R}^2$ centrati nell'origine:
  $\mathcal{H} = \{f_b(\vec{y}) | f_b(\vec{y}) = sign(\vec{y} \cdot \vec{y} - b), b \in \mathbb{R}\}$



Spazio delle Ipotesi: Esempio 3

Congiunzione di m letterali positivi

- Spazio delle Istanze $\rightarrow$ stringhe di m bit: $X = \{s | s \in \{0,1\}^m\}$
- Spazio delle Ipotesi $\rightarrow$ tutte le sentenze logiche che riguardano i letterali positivi $l_1, \ldots, l_m$ (l_1 è vero se il primo bit vale 1, l_2 è vero se il secondo bit vale 1, etc.) e che contengono solo l'operatore $\wedge$ (**and**):

$$\mathcal{H} = \{f_{\{i_1,\ldots,i_j\}}(s) | f_{\{i_1,\ldots,i_j\}}(s) \equiv l_{i_1} \wedge l_{i_2} \wedge \cdots \wedge l_{i_j}, \{i_1, \ldots, i_j\} \subseteq \{1, \ldots, m\}\}$$

Es. $m = 3$, $X = \{0,1\}^3$

Esempi di istanze $\rightarrow s1 = 101$, $s2 = 001$, $s3 = 100$, $s4 = 111$

Esempi di ipotesi $\rightarrow h_1 \equiv l_2$, $h_2 \equiv l_1 \wedge l_2$, $h_3 \equiv true$, $h_4 \equiv l_1 \wedge l_3$, $h_5 \equiv l_1 \wedge l_2 \wedge l_3$

Notare che: h_1, h_2, e h_5 sono false per $s1$, $s2$ e $s3$ e vere per $s4$; h_3 è vera per ogni istanza; h_4 è vera per $s1$ e $s4$ ma falsa per $s2$ e $s3$

Spazio delle Ipotesi: Esempio 3

Congiunzione di m letterali positivi

- Domanda 1: quante e quali sono le ipotesi distinte nel caso $m = 3$?
- Domanda 2: quante sono le ipotesi distinte nel caso generale m ?

Spazio delle Ipotesi: Esempio 3

Congiunzione di m letterali positivi

- Domanda 1: quante e quali sono le ipotesi distinte nel caso $m = 3$?
  - Ris.(quali): $true, l_1, l_2, l_3, l_1 \wedge l_2, l_1 \wedge l_3, l_2 \wedge l_3, l_1 \wedge l_2 \wedge l_3$
  - Ris.(quante): 8
- Domanda 2: quante sono le ipotesi distinte nel caso generale m ?

Spazio delle Ipotesi: Esempio 3

Congiunzione di m letterali positivi

- Domanda 1: quante e quali sono le ipotesi distinte nel caso $m = 3$?
  - Ris.(quali): $true, l_1, l_2, l_3, l_1 \wedge l_2, l_1 \wedge l_3, l_2 \wedge l_3, l_1 \wedge l_2 \wedge l_3$
  - Ris.(quante): 8
- Domanda 2: quante sono le ipotesi distinte nel caso generale m ?
  - Ris.: 2^m, infatti per ogni possibile bit della stringa in ingresso il corrispondente letterale può apparire o meno nella formula logica, quindi:

$$\underbrace{2 \cdot 2 \cdot 2 \cdots 2}_{m \text{ volte}} = 2^m$$

Spazio delle Ipotesi: Esempio 4

Congiunzione di m letterali

- Spazio delle Istanze $\rightarrow$ stringhe di m bit: $X = \{s | s \in \{0,1\}^m\}$
- Spazio delle Ipotesi $\rightarrow$ tutte le sentenze logiche che riguardano i letterali $l_1, \ldots, l_m$ (anche in forma negata, $\neg l_i$) e che contengono solo l'operatore $\wedge$ (**and**):

$$\mathcal{H} = \{f_{\{i_1,\ldots,i_j\}}(s) | f_{\{i_1,\ldots,i_j\}}(s) \equiv L_{i_1} \wedge L_{i_2} \wedge \cdots \wedge L_{i_j},$$
$$\text{dove } L_{i_k} = l_{i_k} \text{ oppure } \neg l_{i_k}, \{i_1, \ldots, i_j\} \subseteq \{1, \ldots, 2m\}\}$$

Notare che se in una formula un letterale compare sia affermato che negato, allora la formula ha sempre valore di verità $false$ (formula non soddisfacibile)

Quindi, tutte le formule che contengono almeno un letterale sia affermato che negato sono equivalenti alla funzione che vale sempre $false$

Spazio delle Ipotesi: Esempio 4

Congiunzione di m letterali

Es. $m = 3$, $X = \{0, 1\}^3$

Esempi di istanze $\rightarrow s1 = 101, s2 = 001, s3 = 100, s4 = 111, s5 = 000$

Esempi di ipotesi $\rightarrow h_1 \equiv \neg l_2, h_2 \equiv \neg l_1 \wedge l_3, h_3 \equiv true, h_4 \equiv \neg l_1 \wedge \neg l_2 \wedge \neg l_3$

Notare che:

- h_1, è falsa per $s4$, e vera per $s1$, $s2$, $s3$ e $s5$;
- h_2 è falsa per $s1$, $s3$, $s4$ e $s5$ e vera per $s2$;
- h_3 è vera per ogni istanza;
- h_4 è falsa per $s1$, $s2$, $s3$, $s4$ e vera per $s5$;

Domanda: quante sono le ipotesi distinte nel caso generale m ?

Spazio delle Ipotesi: Esempio 4

Congiunzione di m letterali

Domanda: quante sono le ipotesi distinte nel caso generale m ?

Risposta: considerando che tutte le formule non soddisfacibili sono equivalenti alla funzione sempre falsa, allora non consideriamo formule dove compare un letterale sia affermato che negato.

Quindi, per ogni possibile bit della stringa in ingresso il corrispondente letterale può non apparire nella formula logica o, se appare, è affermato o negato:

$$\underbrace{3 \cdot 3 \cdot 3 \cdots 3}_{m \text{ volte}} = 3^m$$

E considerando la funzione sempre falsa si ha $3^m + 1$

Spazio delle Ipotesi: Esempio 5

Lookup Table

- Spazio delle Istanze $\rightarrow$ stringhe di m bit: $X = \{s | s \in \{0,1\}^m\}$
- Spazio delle Ipotesi $\rightarrow$ tutte le possibili tabelle di verità che mappano istanze di ingresso ai valori $true$ e $false$: $\mathcal{H} = \{f(s) | f : X \rightarrow \{true, false\}\}$

  Es.

l_1	l_2	$\cdots$	l_m	$f(s)$
0	0	$\cdots$	0	1
0	0	$\cdots$	1	0
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
0	1	$\cdots$	0	0
0	1	$\cdots$	1	1
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
1	0	$\cdots$	0	1
1	0	$\cdots$	1	1
1	1	$\cdots$	0	0
1	1	$\cdots$	1	1
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$

Spazio delle Ipotesi: Esempio 5

Congiunzione di m letterali

Domanda: quante sono le ipotesi distinte nel caso generale m ?

Spazio delle Ipotesi: Esempio 5

Congiunzione di m letterali

Domanda: quante sono le ipotesi distinte nel caso generale m ?

Risposta: tramite una tabella è possibile realizzare una qualunque funzione dallo spazio delle istanze nei volori $true$ e $false$.

Il numero di possibili istanze è:

$$\underbrace{2 \cdot 2 \cdot 2 \cdots 2}_{m \text{ volte}} = 2^m$$

e quindi il numero di possibili funzioni realizzabili è: 2^{2^m}

Osservazioni su Esempi 3, 4 e 5

Osservare che negli esempi 3, 4 e 5 lo spazio delle istanze è sempre lo stesso.

Gli spazi delle ipotesi invece (indichiamo con $\mathcal{H}_3$ quello relativo all'esempio 3, etc.) sono diversi e per ogni m fissato vale la seguente relazione: $\mathcal{H}_3 \subset \mathcal{H}_4 \subset \mathcal{H}_5$

Per esempio, dato $m = 3$

- la funzione booleana $f(s)$ che vale vero solo per le istanze 001 e 011 è contenuta in $\mathcal{H}_4$, infatti $f(s) \equiv \neg l_1 \wedge l_3 \in \mathcal{H}_4$, e in $\mathcal{H}_5$(è facile scrivere una tabella per cui la colonna relativa all'output della funzione è 1 solo in corrispondenza delle istanze 001 e 011), ma non in $\mathcal{H}_3$ perché non esiste la possibilità di descrivere $f(s)$ usando una congiunzione di letterali positivi.
- la funzione booleana $f(s)$ che vale vero solo per le istanze 001, 011 e 100 è contenuta in $\mathcal{H}_5$ (si può procedere come sopra), ma non in $\mathcal{H}_3$ e $\mathcal{H}_4$, perché non esiste la possibilità di descrivere $f(s)$ usando una congiunzione di letterali (positivi).

In particolare $\mathcal{H}_5$ coincide con l'insieme di tutte le funzioni booleane su X.

Principali Paradigmi di Apprendimento: Richiamo

Apprendimento Supervisionato:

- dato in insieme di esempi pre-classificati, $Tr = \{(x^{(i)}, f(x^{(i)}))\}$, apprendere una descrizione generale che incapsula l'informazione contenuta negli esempi (regole valide su tutto il dominio di ingresso)
- tale descrizione deve poter essere usata in modo predittivo (dato un nuovo ingresso $\tilde{x}$ predire l'output associato $f(\tilde{x})$)
- si assume che un esperto (o maestro) ci fornisca la supervisone (cioè i valori della $f()$ per le istanze x dell'insieme di apprendimento)

Find-S è un algoritmo di apprendimento supervisionato

Dati

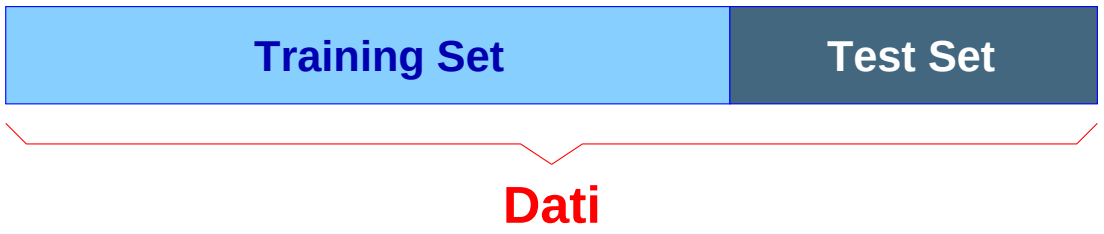
Consideriamo il paradigma di Apprendimento Supervisonato

Dati a nostra disposizione (off-line)

$$\text{Dati} = \{(x^{(1)}, f(x^{(1)})), \ldots, (x^{(N)}, f(x^{(N)}))\}$$

Suddivisione tipica ($N = N_{tr} + N_{ts}$) :

- $\text{Training Set} = \{(x^{(1)}, f(x^{(1)})), \ldots, (x^{(N_{tr})}, f(x^{(N_{tr})}))\}$
  usato direttamente dall'algoritmo di apprendimento;
- $\text{Test Set} = \{(x^{(1)}, f(x^{(1)})), \ldots, (x^{(N_{ts})}, f(x^{(N_{ts})}))\}$
  usato alla fine dell'apprendimento per stimare la bontà della soluzione.

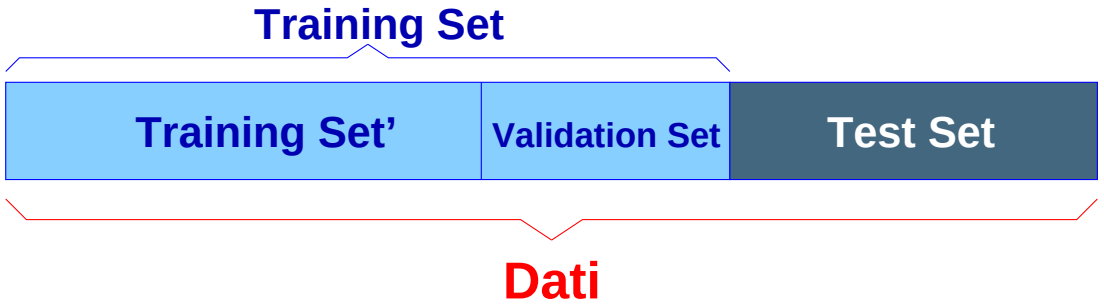


Dati (cont.)

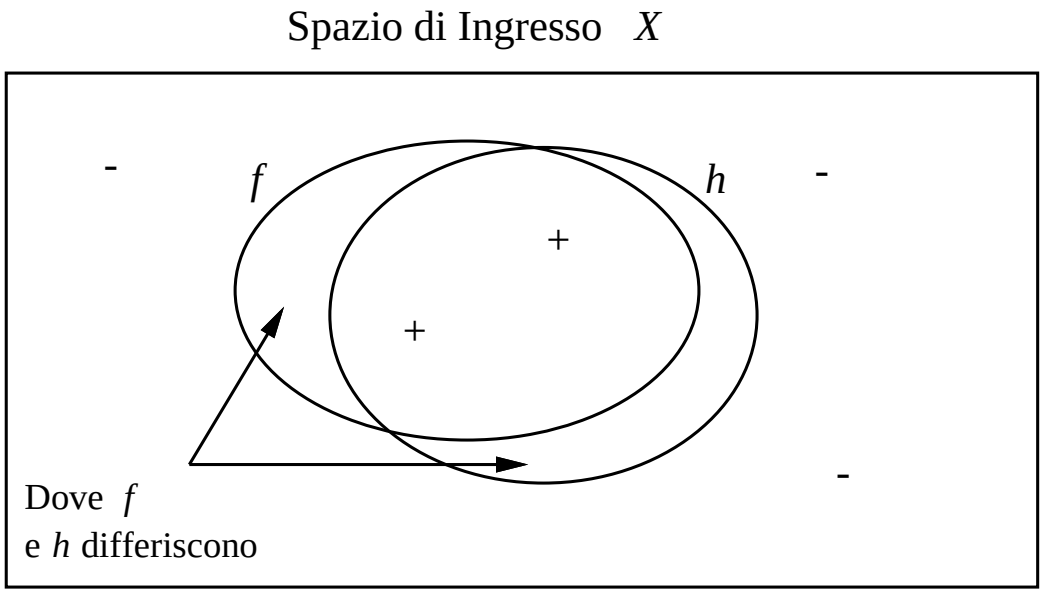
Se N abbastanza grande il Training Set è ulteriormente suddiviso in due sottoinsiemi ($N_{tr} = N_{\widehat{tr}} + N_{val}$) :

- Training Set' $= \{(x^{(1)}, f(x^{(1)})), \ldots, (x^{(N_{\widehat{tr}})}, f(x^{(N_{\widehat{tr}})}))\}$
  usato direttamente dall'algoritmo di apprendimento;
- Validation Set $= \{(x^{(1)}, f(x^{(1)})), \ldots, (x^{(N_{val})}, f(x^{(N_{val})}))\}$
  usato indirettamente dall'algoritmo di apprendimento.

Il Validation Set serve per scegliere l'ipotesi $h \in \mathcal{H}$ migliore fra quelle consistenti con il Training Set'



Errore Ideale

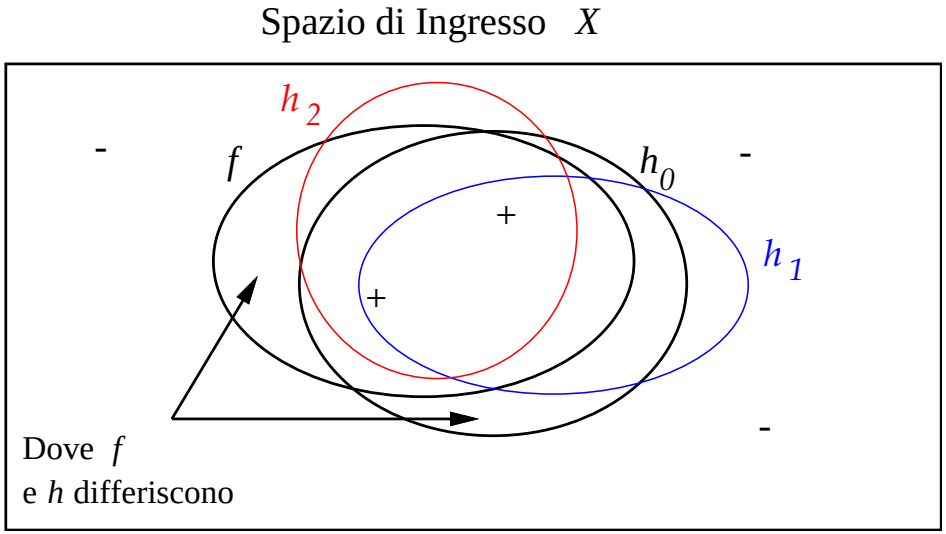


Supponiamo che la funzione f da apprendere sia una funzione booleana (concetto):

$$f : X \rightarrow \{0, 1\} \ (\{-, +\})$$

Def: L'**Errore Ideale** ($error_{\mathcal{D}}(h)$) di una ipotesi h rispetto al concetto f e la distribuzione di probabilità $\mathcal{D}$ (probabilità di osservare l'ingresso $x \in X$) è la probabilità che h classifichi erroneamente un input selezionato a caso secondo $\mathcal{D}$: $error_{\mathcal{D}}(h) \equiv \underset{x \in \mathcal{D}}{Pr}[f(x) \neq h(x)]$

Errore di Apprendimento



Dato Tr = Training Set', più ipotesi possono essere consistenti: h_0, h_1, h_2 quale scegliere ?

Def: L'**Errore Empirico** ($error_{Tr}(h)$) di una ipotesi h rispetto a Tr è il numero di esempi che h classifica erroneamente: $error_{Tr}(h) \equiv \#\{(x, f(x)) \in Tr | f(x) \neq h(x)\}$

Def: Una ipotesi $h \in \mathcal{H}$ è sovraspecializzata (**overfit**) Tr se $\exists h' \in \mathcal{H}$ tale che $error_{Tr}(h) < error_{Tr}(h')$, ma $error_{\mathcal{D}}(h) > error_{\mathcal{D}}(h')$.

Il Validation Set serve per cercare di selezionare l'ipotesi migliore (evitare **overfit**).

VC-dimension

Definizione: Frammentazione (Shattering)

Dato $S \subset X$, S è frammentato (shattered) dallo spazio delle ipotesi $\mathcal{H}$ se e solo se

$$\forall S' \subseteq S,\ \exists h \in \mathcal{H},\ \text{tale che } \forall x \in S,\ h(x) = 1 \Leftrightarrow x \in S'$$

($\mathcal{H}$ realizza tutte le possibili dicotomie di S)

Definizione: VC-dimension

La VC-dimension di uno spazio delle ipotesi $\mathcal{H}$ definito su uno spazio delle istanze X è data dalla cardinalità del sottoinsieme più grande di X che è frammentato da $\mathcal{H}$:

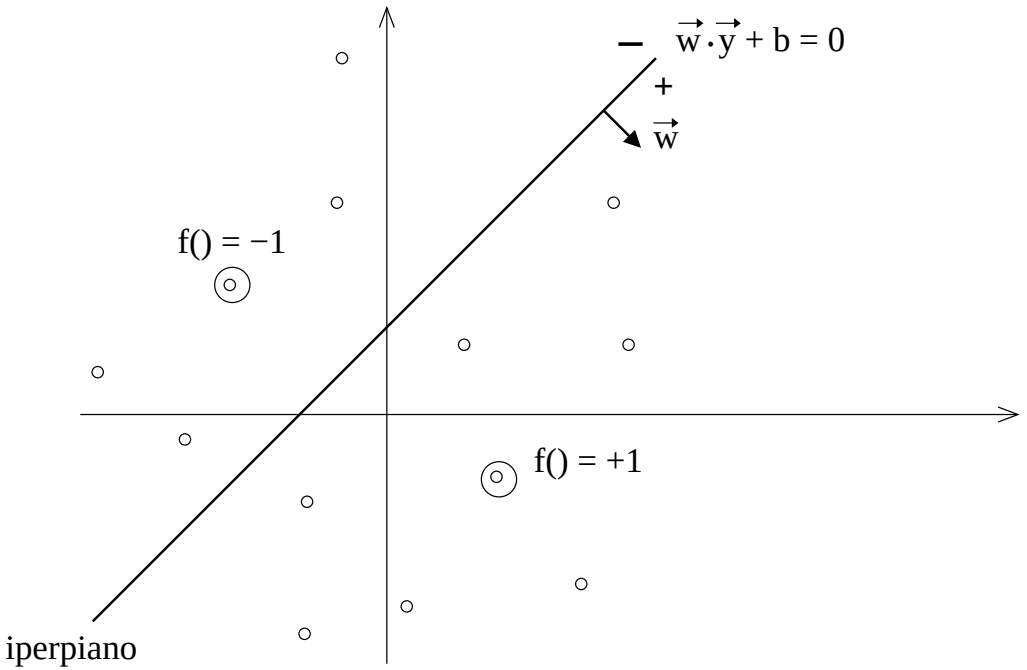
$$VC(\mathcal{H}) = \max_{S \subseteq X} |S| :\ \mathcal{H} \text{ frammenta } S$$

$VC(\mathcal{H}) = \infty$ se S non è limitato

VC-dimension: Esempio

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

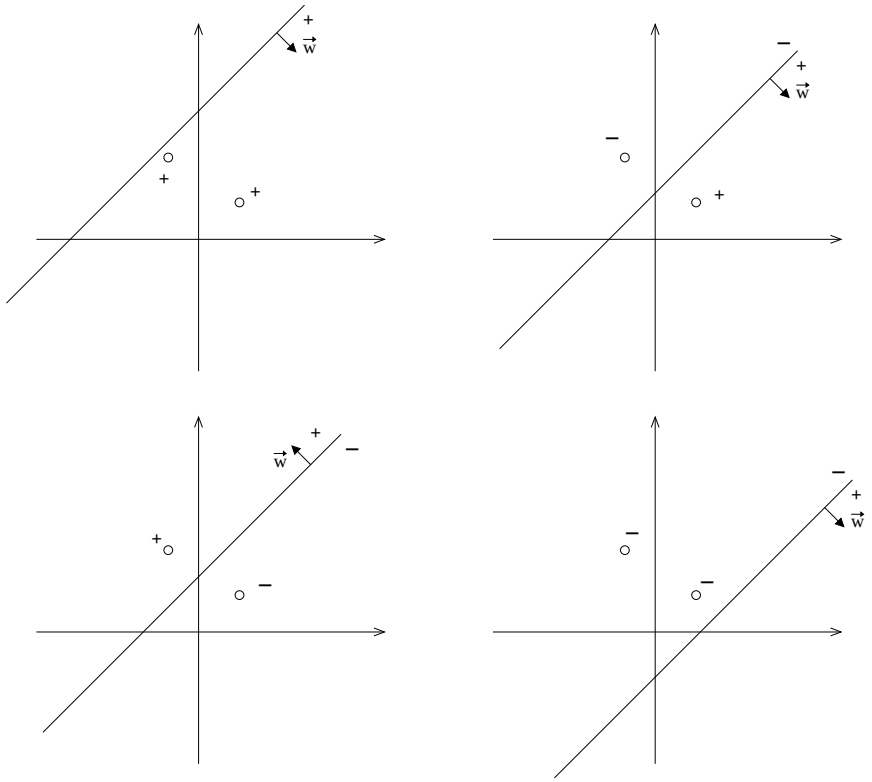
$$\mathcal{H}_1 = \{f_{(\vec{w},b)}(\vec{y}) | f_{(\vec{w},b)}(\vec{y}) = sign(\vec{w} \cdot \vec{y} + b), \vec{w} \in \mathbb{R}^2, b \in \mathbb{R}\}$$



VC-dimension: Esempio

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

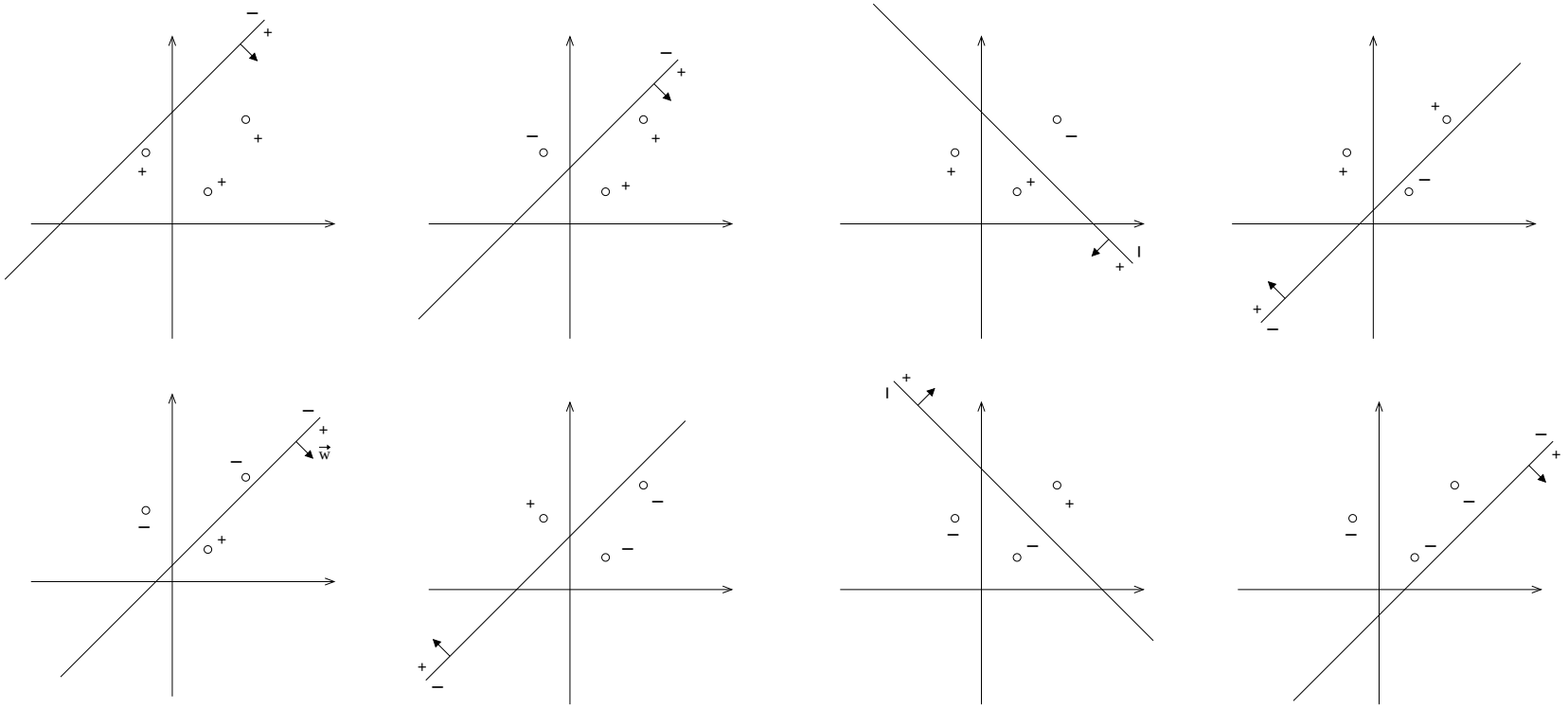
$VC(\mathcal{H}) \geq 1$ banale. Vediamo cosa succede con 2 punti:



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 $\mathcal{H}_1$?

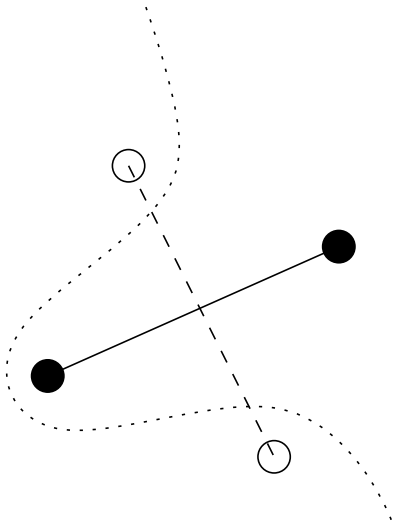
Quindi $VC(\mathcal{H}) \geq 3$. Cosa succede con 4 punti ?

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'Erorre Ideale per Classifi cazione 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 $\mathcal{H} = \{h_{\boldsymbol{w}}(\boldsymbol{x}) | \boldsymbol{w} \in I\!R^k\}$
- 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:

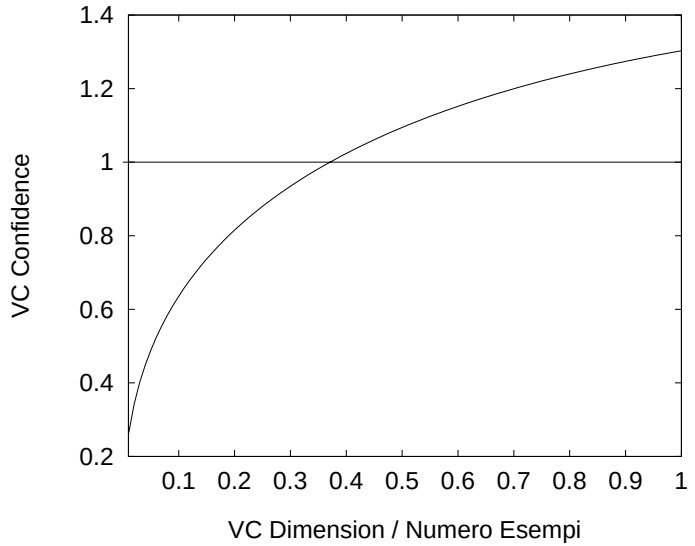
$$error_{\mathcal{D}}(h_{\boldsymbol{w}^*}(\boldsymbol{x})) \leq \underbrace{error_{Tr}(h_{\boldsymbol{w}^*}(\boldsymbol{x}))}_{A} + \underbrace{\sqrt{\frac{VC(\mathcal{H})}{N_{tr}}(\log(\frac{2N_{tr}}{VC(\mathcal{H})}) + 1) - \frac{1}{N_{tr}}\log(\delta)}}_{B}$$

Bound sull'Erorre Ideale per Classifi cazione 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

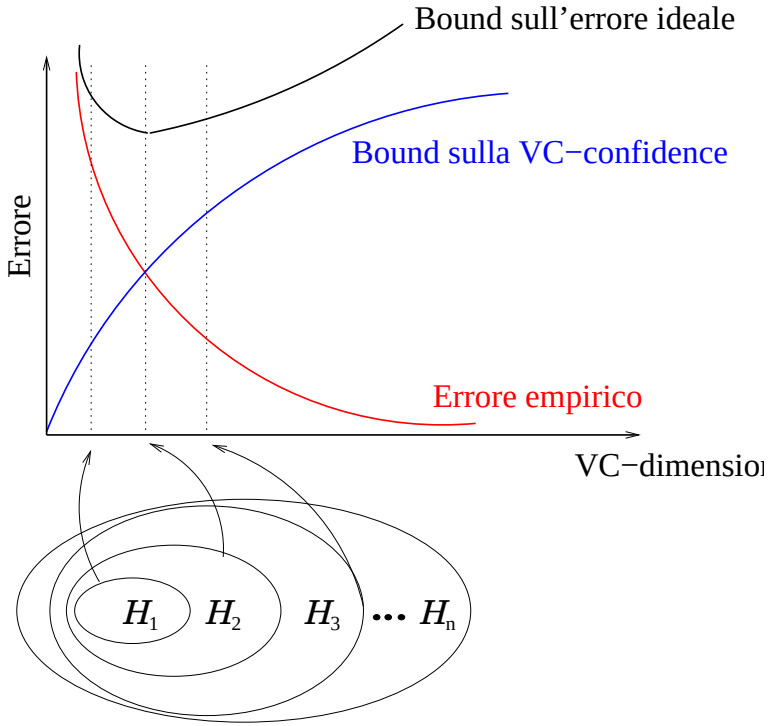
Problema: all'aumentare della VC-dimension diminuisce l'errore empirico (termine A), ma aumenta la VC confidence (termine B)!

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 ?

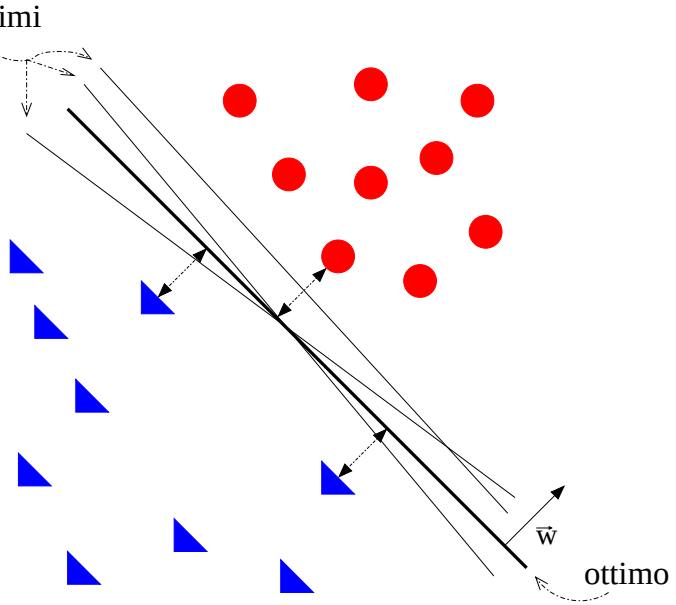
Sappiamo che un iperpiano in uno spazio a m dimensioni ha $VC = m + 1$. Come facciamo a creare una struttura di spazi delle ipotesi con VC-dimension crescente ?

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

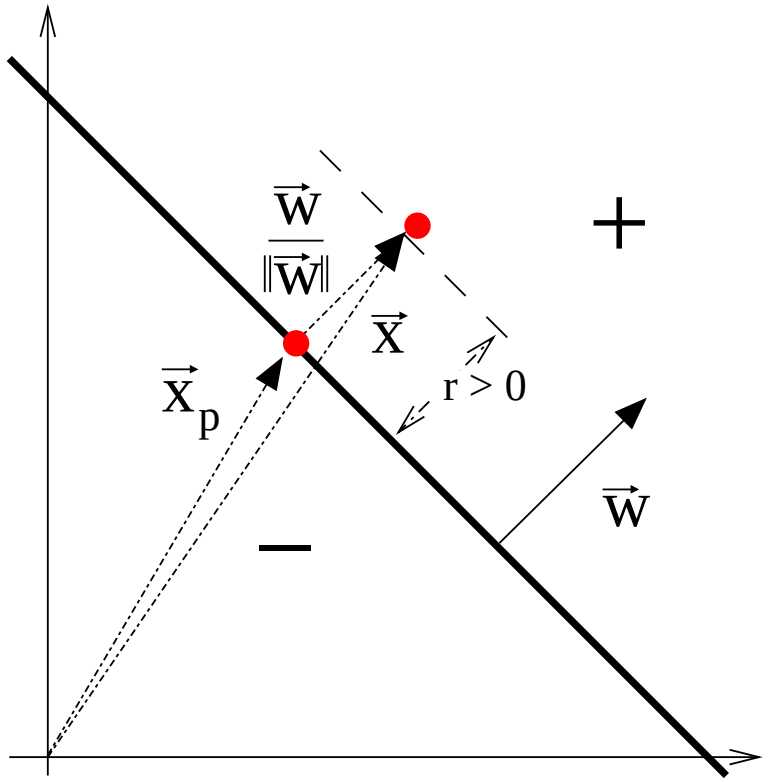
La “distanza” di un vettore da un iperpiano la possimo misurare in senso algebrico.

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

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 $||\vec{w}||$.

Poichè esistono 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 $\hat{r}||\vec{w}|| = 1$

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.

Per passare dal primale alla sua forma duale si utilizza il teorema Kuhn-Tucker, che prescrive i seguenti due passi:

1. a partire dalla formulazione primale si costruisce un nuovo problema non vincolato utilizzando i moltiplicatori di Lagrange:

$$L(\vec{w}, b, \alpha) = \frac{1}{2}||\vec{w}||^2 - \sum_{i=1}^{n} \alpha_i (y_i(\vec{w} \cdot \vec{x}_i + b) - 1)$$

   dove le variabili $\alpha_i \geq 0$ sono i *moltiplicatori di Lagrange* (in questo caso, variabili duali). La soluzione ottima risiede nel punto di sella ottenuto minimizzando la funzione Lagrangiana $L(\vec{w}, b, \alpha)$ rispetto alle variabili primali $\vec{w}$ e b, e massimizzandola rispetto alle variabili duali α_i.

2. si utilizzano le *condizioni* di Kuhn-Tucker per esprimere le variabili primali in funzione delle variabili duali; in questo modo la funzione Lagrangiana diventa funzione esclusiva delle variabili duali e quindi deve essere massimizzata rispetto a queste variabili:

Vediamo nel dettaglio il passo 2...

Passo 2:

Il teorema Kuhn-Tucker afferma che l'ottimo si ottiene minimizzando $L(\vec{w}, b, \alpha)$ rispetto a $\vec{w}$ e b, quindi bisogna che il corrispondente gradiente della funzione sia nullo:

$$\frac{\partial L(\vec{w},b,\alpha)}{\partial \vec{w}} = 0 \quad \Leftrightarrow \quad \vec{w^*} = \sum_{i=1}^{n} y_i \alpha_i^* \vec{x}_i \text{ E1}$$
$$\frac{\partial L(\vec{w},b,\alpha)}{\partial b} = 0 \quad \Leftrightarrow \quad \sum_{i=1}^{n} y_i \alpha_i^* = 0 \text{ E2}$$

Inoltre il teorema Kuhn-Tucker afferma che all'ottimo

$$\alpha_i^* [y_i(\vec{w}^* \cdot \vec{x}_i + b^*) - 1] = 0 \;\; i = 1, \ldots, n$$

I vettori per cui $\alpha_i^* > 0$ sono detti vettori di supporto.

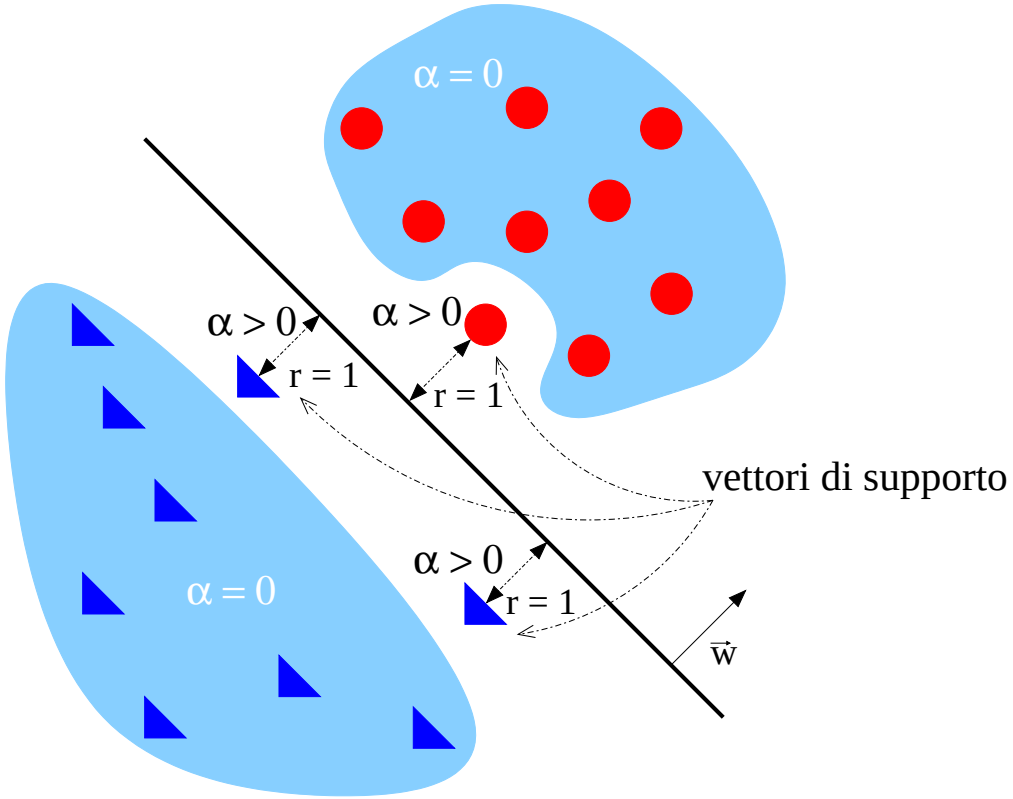
La formulazione duale si ottiene eliminando le variabili primali (equazioni E1 ed E2) dalla funzione Lagrangiana:

$$\max_\alpha \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} y_i y_j \alpha_i \alpha_j \vec{x}_i \cdot \vec{x}_j$$
soggetto a: $\forall i \in \{1, \ldots, n\} : \alpha_i \geq 0$ e $\sum_{i=1}^{n} y_i \alpha_i = 0.$

I valori ottimi delle α_i^* determinano l'iperpiano ottimo grazie ad E1, a meno del valore di b.

Il valore di b, tuttavia, si può ottenere osservando che per un qualsiasi vettore di supporto $\vec{x}_s$ deve valere $\boxed{y_s(\vec{w}^* \cdot \vec{x}_s + b^*) = 1}$ e quindi considerando un esempio positivo ($y_s = 1$)

$$b^* = 1 - \vec{w}^* \cdot \vec{x}_s$$



Caso Non Separabile

Cosa succede se gli esempi NON sono linearmente separabili ?

In questo caso si deve ammettere che alcuni dei vincoli possano essere violati. Ciò si può fare:

- introducendo le variabili *slack* $\xi_i \geq 0,\ i = 1, \ldots, n$, una per ogni vincolo:

$$y_i(\vec{w} \cdot \vec{x}_i + b) \geq 1 - \xi_i$$

- modificando la funzione costo in modo da penalizzare variabili slack che non sono a 0:

$$\frac{1}{2}||\vec{w}||^2 + C\sum_{i=1}^{n} \xi_i$$

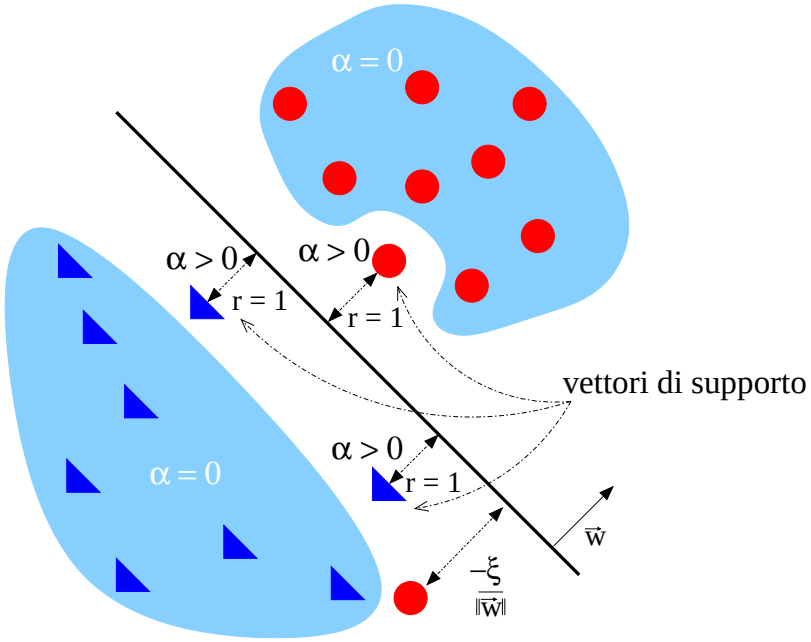
dove C (parametro di regolarizzazione) è una costante positiva che controlla il tradeoff tra la complessità dello spazio delle ipotesi e il numero di esempi non-separabili.

Il duale di questa nuova formulazione è molto simile al precedente:

$$\max_\alpha \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i,j=1}^n y_i y_j \alpha_i \alpha_j \vec{x}_i \cdot \vec{x}_j$$

soggetto a: $\forall i \in \{1, \ldots, n\} : 0 \leq \alpha_i \leq C$ e $\sum_{i=1}^n y_i \alpha_i = 0.$

La differenza risiede nel fatto che le variabili duali sono ora limitate superiormente da C. Per la determinazione di b^* si procede in modo simile a quanto visto in precedenza (anche se con alcune differenze...).



Caso Non Separabile

La soluzione vista in precedenza per esempi non-linearmente separabili non garantisce usualmente buone prestazioni perchè un iperpiano può solo rappresentare dicotomie dello spazio delle istanze.

Per tale motivo, quando gli esempi non sono linearmente separabili si usa la seguente strategia in due passi:

1. si mappano i dati in ingresso (input space) in uno spazio a dimensione molto superiore (feature space);
2. si calcola l'iperpiano ottimo (usando la formulazione con variabili slack) all'interno del feature space.

Caso Non Separabile

Il passo 1 si giustifica tramite il teorema di Cover sulla separabilità, il quale afferma che un problema di classificazione complesso, formulato attraverso una trasformazione non-lineare dei dati in uno spazio ad alta dimensionalià, ha maggiore probabilità di essere linearmente separabile che in uno spazio a bassa dimensionalità.

Il passo 2 è ovviamente giustificato dal fatto che l'iperpiano ottimo minimizza la VC-dimension e quindi la capacità di generalizzazione è migliorata.

Il passo 1 ci prescrive di considerare una trasforamzione $\varphi(\cdot)$ non-lineare applicata ai dati originari $\{(\vec{x}_i, y_i)\}_1^n$. In particolare, assunto che $\forall i\ \vec{x}_i \in I\!R^m$, $\varphi(\cdot)$ deve mappare tali vettori (e pi`u in generale un qualsiasi vettore a valori reali di dimensione m) in uno spazio a dimensionalità $M \gg m$ (ad esempio, $I\!R^M$).

Caso Non Separabile

Possiamo assumere che ognuna delle nuove coordinate nello spazio delle features sia generata da una funzione non-lineare $\varphi_j(\cdot)$. Quindi si considerano M funzioni $\varphi_j(\vec{x})$ con $j = 1, \ldots, M$. Un generico vettore $\vec{x}$ viene perciò mappato nel vettore M dimensionale

$$\vec{\varphi}(\vec{x}) = [\varphi_1(\vec{x}), \ldots, \varphi_M(\vec{x})]$$

Il passo 2 ci chiede di trovare un iperpiano ottimo nello spazio M dimensionale delle features. Un iperpiano in tale spazio sarà individuato dalla equazione

$$\sum_{j=1}^{M} w_j \varphi_j(\vec{x}) + b = 0$$

ovvero

$$\sum_{j=0}^{M} w_j \varphi_j(\vec{x}) = \vec{w} \cdot \vec{\varphi}(\vec{x}) = 0$$

se aggiungiamo la coordinata $\varphi_0(\vec{x}) = 1$ e $w_0 = b$.

Caso Non Separabile

Utilizzando per $\vec{w}$ la formula

$$\vec{w} = \sum_{k=1}^{n} y_k \alpha_k \vec{\varphi}(\vec{x}_k)$$

l'equazione che determina l'iperpiano diventa:

$$\sum_{k=1}^{n} y_k \alpha_k \vec{\varphi}(\vec{x}_k) \cdot \vec{\varphi}(\vec{x}) = 0$$

dove il termine $\vec{\varphi}(\vec{x}_k) \cdot \vec{\varphi}(\vec{x})$ rappresenta il prodotto interno nel feature space fra i vettori indotti dalla k-esima istanza di apprendimento e dal vettore di input $\vec{x}$.

Funzioni Kernel

Se fosse possibile definire una funzione $K(\cdot,\cdot)$ (detta kernel) tale che

$$K(\vec{x}_k,\vec{x}) = \vec{\varphi}(\vec{x}_k)\cdot\vec{\varphi}(\vec{x}) = \sum_{j=0}^{M}\varphi_j(\vec{x}_k)\varphi_j(\vec{x}) = K(\vec{x},\vec{x}_k)$$ (funzione simmetrica)

allora, si potrebbe specificare l'iperpiano nello spazio delle features SENZA calcolare esplicitamente i vettori nello spazio delle features:

$$\sum_{k=1}^{n} y_k\alpha_k K(\vec{x}_k,\vec{x})$$

Tali funzioni kernel di fatto esistono se alcune condizioni sono soddisfatte...

Funzioni Kernel

Teorema di Mercer

Sia $K(\vec{x}, \vec{x}\prime)$ un kernel continuo e simmetrico definito nell'intervallo chiuso $\vec{a} \leq \vec{x} \leq \vec{b}$ e similarmente per $\vec{x}\prime$. Il kernel $K(\vec{x}, \vec{x}\prime)$ può essere espanso nella serie

$$K(\vec{x}, \vec{x}\prime) = \sum_{i=1}^{\infty} \lambda_i \varphi_i(\vec{x}) \varphi_i(\vec{x}\prime)$$

con $\lambda_i > 0$. Affinché tale espansione sia valida e per la sua convergenza assoluta ed uniforme, è necessario e sufficiente che la condizione

$$\int_{\vec{b}}^{\vec{a}} \int_{\vec{b}}^{\vec{a}} K(\vec{x}, \vec{x}\prime) \psi(\vec{x}) \psi(\vec{x}\prime) d\vec{x} d\vec{x}\prime$$

sia vera per tutte le $\psi(\cdot)$ che soddisfano

$$\int_{\vec{b}}^{\vec{a}} \psi^2(\vec{x}) d\vec{x} < \infty$$

Funzioni Kernel

Quindi in sostanza una funzione kernel che soddisfa le condizioni del teorema di Mercer rappresenta un prodotto interno vettoriale in uno spazio delle features generato da una qualche trasformazione non-lineare.

Si noti che tale spazio delle features può essere infinito (vedi espansione) e che il fatto che $\forall i\ \lambda_i > 0$ implica che il kernel è definito positivo.

Esempi di funzioni kernel:

- kernel polinomiale di grado p, $(\vec{x} \cdot \vec{x}\prime + 1)^p$
- kernel radiale (radial-basis function), $\exp(-\frac{1}{2\sigma^2}\|\vec{x} - \vec{x}\prime\|^2)$

Formulazione con Kernel

Si noti, che l'introduzione di un kernel di fatto non modifica la formulazione del problema vincolato quadratico da risolvere per determinare l'iperpiano ottimo:

$$\max_\alpha \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i,j=1}^n y_i y_j \alpha_i \alpha_j K(\vec{x}_i, \vec{x}_j)$$

soggetto a: $\forall i \in \{1, \ldots, n\} : 0 \leq \alpha_i \leq C$ e $\sum_{i=1}^n y_i \alpha_i = 0.$

dove i valori del kernel necessari sono calcolati sulle possibili coppie di vettori di allenamento ($K(\vec{x}_i, \vec{x}_j)$, con $i, j = 1, \ldots, n$) e quindi possono essere raccolti in una matrice $\boldsymbol{K} \in I\!R^n \times I\!R^n$ (simmetrica e definita positiva) denominata matrice del kernel.

Ad esempio, se si usa un kernel polinomiale di grado $p = 3$ si ha $\boldsymbol{K}_{i,j} = (\vec{x}_i \cdot \vec{x}_j + 1)^3$ e una nuova istanza $\vec{x}$ è classificata dalla funzione

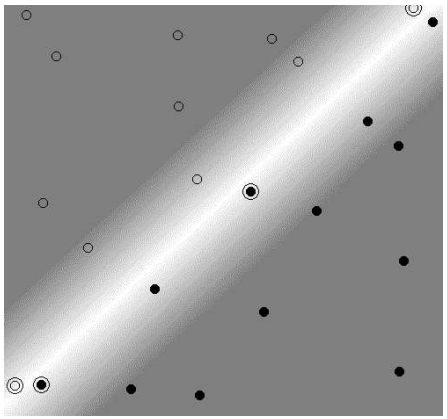
$$sign(\sum_{\vec{x}_k \in SV} y_k \alpha_k^* K(\vec{x}_k, \vec{x})) = sign(\sum_{\vec{x}_k \in SV} y_k \alpha_k^* (\vec{x}_k \cdot \vec{x} + 1)^3)$$

dove SV è l'insieme dei vettori di supporto all'ottimo e α_k^* sono i valori ottimi per i vettori di supporto (gli altri sono a 0 e quindi non contribuiscono alla sommatoria).

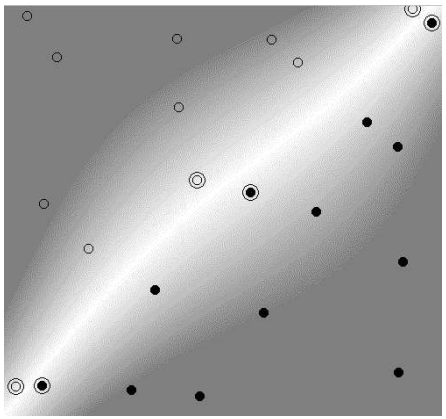
Formulazione con Kernel

Si noti come ciò permetta di effettuare la trasformazione non-lineare $\vec{\varphi}(\cdot)$ in modo IMPLICITO, in quanto quello che importa non sono i vettori nello spazio delle features, ma il prodotto interno fra di loro, che si può calcolare direttamente tramite la funzione kernel senza passare attraverso lo spazio delle features.

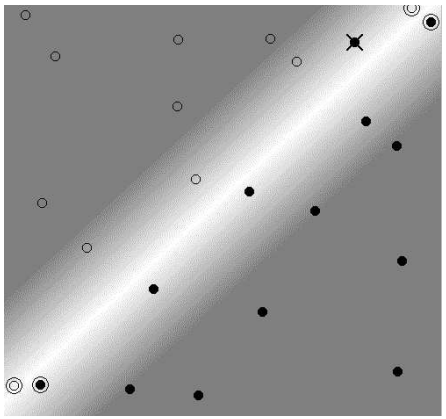
Esempi a confronto delle superfici di decisione generate NELLO SPAZIO DELLE ISTANZE senza e con kernel (polinomiale di grado 3) sia nel caso separabile che non-separabile:



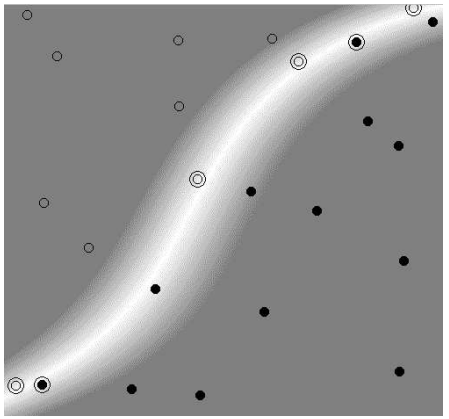
separable



degree 3 poly



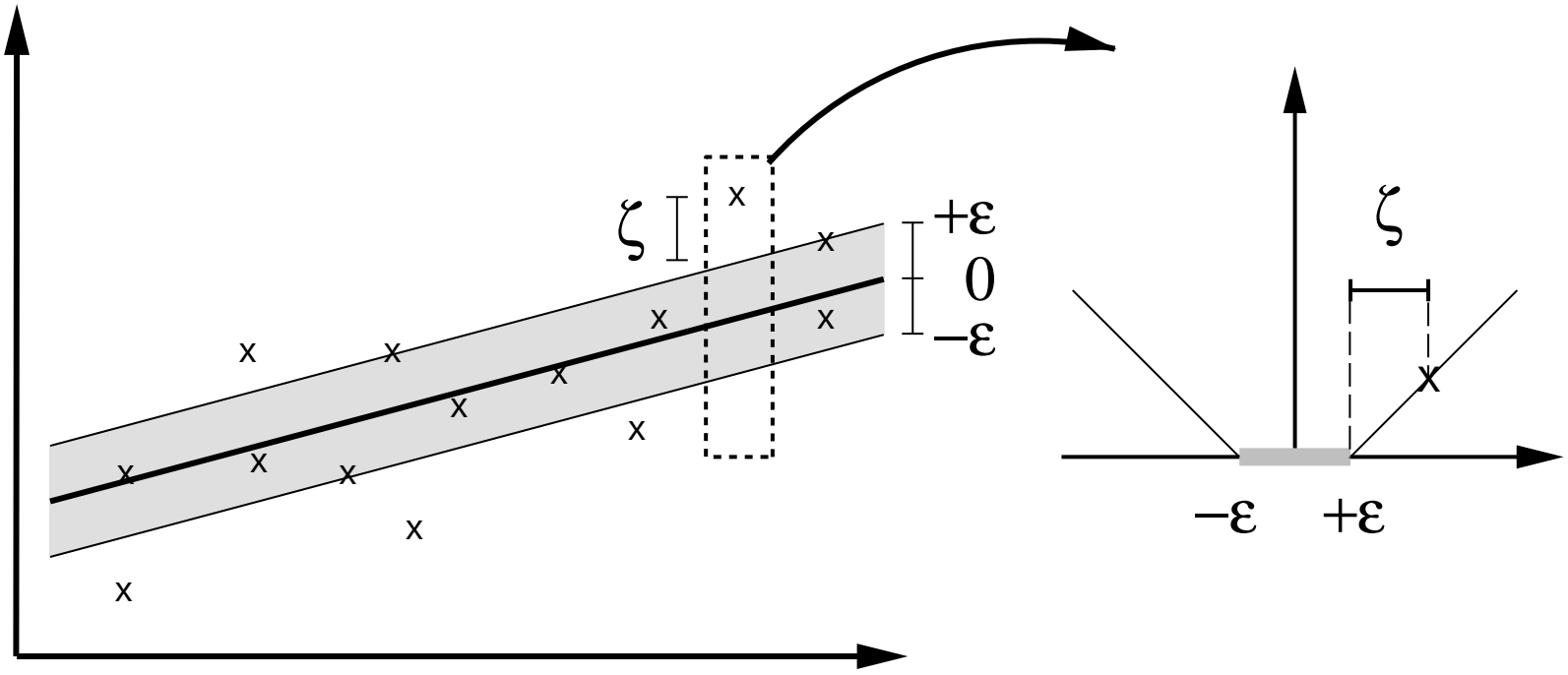
not separable



degree 3 poly

Regressione: Idea Base

Quando si considera il problema di approssimazione di funzioni a valori reali (regressione) si utilizza l'ϵ-tubo: output che differiscono dal valore di target per più di ϵ in valore assoluto vengono penalizzati linearmente, altrimenti non vengono considerati errori.



Regressione: Forma Primale

Questa idea da' origine alla seguente formulazione primale

$$\min_{\vec{w},b,\xi,\xi^*} \frac{1}{2}||\vec{w}||^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*)$$

soggetto a:

$$\forall i \in \{1,\ldots,n\}$$

$$y_i - \vec{w}\cdot\vec{x}_i - b \leq \epsilon + \xi_i$$

$$\vec{w}\cdot\vec{x}_i + b - y_i \leq \epsilon + \xi_i^*$$

$$\xi_i, \xi_i^* \geq 0$$

la cui forma duale ...

Regressione: Forma Duale

... è la seguente

$$\max_{\alpha,\alpha^*} -\epsilon \sum_{i=1}^{n}(\alpha_i + \alpha_i^*) + \sum_{i=1}^{n} y_i(\alpha_i - \alpha_i^*) +$$
$$-\frac{1}{2}\sum_{i,j=1}^{n}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)K(\vec{x}_i, \cdot\vec{x}_j)$$

soggetto a:

$$\sum_{i=1}^{n}(\alpha_i - \alpha_i^*) = 0$$

$$\alpha_i, \alpha_i^* \in [0, C]$$