Intorduzione alle Support Vector Machines

Marco Sciandrone
Dipartimento di Ingegneria dell'Informazione
Università di Firenze
E-mail: marco.sciandrone@unifi.it

Generalità

Le Support Vector Machines (SVM) costituiscono una classe di "macchine di apprendimento" recentemente introdotte in letteratura. Le SVM traggono origine da concetti riguardanti la teoria statistica dell'apprendimento e presentano proprietà teoriche di generalizzazione. Approfondimenti teorici sull'argomento possono essere trovati in [1], [2], [3].

Scopo di queste note è quello di definire i concetti generali che sono alla base delle SVM e di mostrare come il problema dell'addestramento di una SVM è riconducibile ad *un problema di programmazione quadratica convessa con vincoli lineari* (in particolare, l'insieme dei vincoli è costituito da un vincolo lineare e da vincoli di box).

1 Cenni sui concetti riguardanti la teoria statistica dell'apprendimento

Si supponga di disporre di l osservazioni, relative ad un problema di classificazione, in cui ogni osservazione è costituita da una coppia: un vettore $x^i \in R^n$ e *un'etichetta* $y^i \in \{1, -1\}$, il cui valore definisce l'appartenenza del vettore ad una classe o all'altra (nel seguito, in relazione a problemi di regressione, considereremo anche il caso $y^i \in R$).

Si assuma che esista una distribuzione di probabilità $P(x, y)$ non nota, da cui sono stati estratti i dati disponibili. È supposta quindi l'esistenza di una distribuzione di y per un assegnato x (il caso più semplice è quello per cui ad ogni assegnato x corrisponde un determinato y).

Si consideri il problema della definizione di una macchina per l'apprendimento della relazione $x^i \to y^i$. Una macchina è definita mediante un insieme di possibili funzioni

$$f(\alpha) : R^n \to \{-1, 1\},$$

in cui α rappresenta *il vettore dei parametri modificabili*. La macchina è *deterministica*, cioè, assegnati il vettore di ingresso x ed il vettore dei parametri α, la macchina fornisce sempre la stessa uscita $f(x, \alpha)$. *Addestrare* la macchina significa determinare il vettore dei parametri α.

Una rete neurale multi-strato con architettura fissata rappresenta, ad esempio, una macchina di apprendimento (nel senso specificato prima), in cui il vettore α corrisponde al vettore dei pesi e delle soglie della rete.

Riepilogando:

- l'insieme di funzioni $\{f(\alpha)\}$ rappresenta una macchina di apprendimento;
- scegliere un vettore α significa selezionare una determinata funzione, cioè determinare una macchina addestrata.

Il valore atteso dell'errore di test di una macchina addestrata è

$$R(\alpha) = \int \frac{1}{2}|y - f(x, \alpha)|dP(x, y)$$

(si osservi che $P(x, y)$ è non nota). La quantità $R(\alpha)$ è denominata *rischio effettivo* (o semplicemente *rischio*). Si definisce *rischio empirico* la quantità

$$R_{emp}(\alpha) = \frac{1}{2l} \sum_{i=1}^{l} |y^i - f(x^i, \alpha)|$$

(si noti che la distribuzione di probabilità non interviene nella definizione del rischio empirico, che, una volta fissati il vettore dei parametri α e il training set $\{x^i, y^i\}_{i=1,\ldots,l}$, è un determinato numero).

Sia η un numero fissato tale che $0 \leq \eta \leq 1$. Con probabilità $1 - \eta$ vale la seguente disuguaglianza [2]

$$R(\alpha) \leq R_{emp}(\alpha) + \sqrt{\left(\frac{h(\log(2l/h) + 1) - log(\eta/4)}{l}\right)}, \tag{1}$$

in cui h è un intero non negativo denominato Vapnik Chervonenkis (VC) dimension, ed è una misura della capacità di classificazione espressa dalla macchina rappresentata dall'insieme di funzioni $\{f(\alpha)\}$. In maniera formale, la VC dimension di una macchina $\{f(\alpha)\}$ verrà definita di seguito. Il secondo termine del secondo membro della (1) viene denominato VC *confidence*. Osserviamo che la VC confidence dipende dalla macchina di apprendimento scelta (cioè dalla classe di funzioni scelta), mentre sia il rischio empirico che il rischio effettivo dipendono dalla particolare funzione determinata mediante la procedura di addestramento della macchina.

Al fine di definire la VC dimension di un insieme di funzioni, premettiamo la seguente definizione.

Definizione 1 *Un insieme di l punti in $x^i \in R^n$, $i = 1, \ldots, l$, è frammentabile da un insieme di funzioni $\{f(\alpha)\}$ se, comunque si scelgano le etichette $y^i \in \{1, -1\}$, $i = 1, \ldots, l$, esiste un elemento dell'insieme $\{f(\alpha)\}$ che classifica correttamente i punti, cioè, esiste un vettore α tale che*

$$f(x^i, \alpha) = y^i, \qquad i = 1, \ldots, l.$$

Definizione 2 *La VC dimension h di un insieme di funzioni $\{f(\alpha)\}$ è il massimo numero di punti che possono essere frammentati da $\{f(\alpha)\}$.*

Si osservi che se la VC dimension di un insieme $\{f(\alpha)\}$ è h, ciò significa che esiste almeno un insieme di h punti che possono essere frammentati, ma in generale non è vero che un qualsiasi insieme di h punti può essere frammentato da $\{f(\alpha)\}$.

Come esempio, sia $\{f(\alpha)\}$ l'insieme degli iperpiani orientati in R^n, per cui, fissato un iperpiano, tutti i punti appartenenti ad un semispazio sono etichettati con 1, e tutti i punti appartenenti all'altro semispazio sono etichettati con -1. Vale il seguente risultato.

Teorema 1 *Un insieme di punti $\{x^1, \ldots, x^m\} \subset R^n$ è frammentabile dalla famiglia degli iperpiani orientati se e solo se i vettori $x^1, \ldots, x^m$ sono affinemente indipendenti.*

Dim. Senza perdita di generalità possiamo assumere $x^1 = 0$.

(a). Supponiamo che i vettori $x^1, \ldots, x^m$ siano affinemente indipendenti. Ciò implica che i vettori $x^2 - x^1, \ldots, x^m - x^1 = x^2, \ldots, x^m$ sono linearmente indipendenti.

Occorre dimostrare che, comunque si scelgano le etichette $y^1, \ldots, y^m$, esistono un vettore dei pesi $\bar{w} \in R^n$ e una soglia $\bar{b} \in R$ tali che

$$y^i(\bar{w}^T x^i + \bar{b}) > 0 \qquad i = 1, \ldots, m. \tag{2}$$

Ponendo

$$\bar{b} = \begin{cases} 1/2 & \text{se } y^1 = +1 \\ -1/2 & \text{se } y^1 = -1 \end{cases} \tag{3}$$

si ottiene ovviamente $y^1(w^T x^1 + \bar{b}) = y^1 \bar{b} > 0$ per qualunque vettore $w \in R^n$. L'indipendenza lineare dei vettori $x^2, \ldots, x^m$ implica che il sistema lineare nel vettore incognito w

$$\begin{pmatrix} (x^2)^T \\ (x^3)^T \\ \vdots \\ (x^m)^T \end{pmatrix} w = \begin{pmatrix} y^2 \\ y^3 \\ \vdots \\ y^m \end{pmatrix}$$

ammette almeno una soluzione $\bar{w}$. Per $i = 2, \ldots, m$ risulta quindi

$$(\bar{w})^T x^i + \bar{b} = y^i + \bar{b} = 1 + \bar{b} > 0 \qquad \text{per ogni } i \;\; t.c.\; y^i = +1$$

$$(\bar{w})^T x^i + \bar{b} = y^i + \bar{b} = -1 + \bar{b} < 0 \qquad \text{per ogni } i \;\; t.c.\; y^i = -1.$$

Dalle precedenti relazioni e dalla (3) segue che la (2) è verificata, e quindi risulta dimostrato che i vettori $x^1, \ldots, x^m$ sono frammentabili.

(b) Supponiamo che i vettori $x^1, \ldots, x^m$ siano frammentabili. Dimostreremo che essi sono affinemente indipendenti.

Supponiamo per assurdo che i vettori $\{x^1, \ldots, x^m\}$ siano affinemente dipendenti e tali che, comunque si scelgano le etichette $y^1, \ldots, y^m$, è possibile determinare un vettore dei pesi $\bar{w} \in R^n$ e una soglia $\bar{b} \in R$ per i quali si ha

$$y^i(\bar{w}^T x^i + \bar{b}) > 0 \qquad i = 1, \ldots, m. \tag{4}$$

I vettori $x^2, \ldots, x^m$ sono necessariamente linearmente dipendenti, per cui possiamo scrivere

$$0 = \sum_{i \in I^+} \alpha_i x^i - \sum_{j \in J^-} |\beta_j| x^j, \tag{5}$$

dove $\alpha_i > 0$, $\beta_j < 0$, $I^+ \cup J^- \subseteq \{2, \ldots, m\}$ e $I^+ \cap J^- = \emptyset$. Si assuma $y^i = +1$ per $i \in I^+$ e $y^j = -1$ per $j \in J^-$. Di conseguenza si ha

$$\begin{array}{lll} \bar{w}^T x^i & > & -\bar{b} \qquad i \in I^+ \\ \bar{w}^T x^j & < & -\bar{b} \qquad j \in J^- \end{array}$$

da cui si ottiene

$$\begin{aligned} \sum_{i\in I^+} \alpha_i \bar{w}^T x^i &> -\bar{b} \sum_{i\in I^+} \alpha_i \\ -\sum_{j\in J^-} |\beta_j| \bar{w}^T x^j &> \bar{b} \sum_{j\in J^-} |\beta_j| \end{aligned} \tag{6}$$

Dalla (5) e dalla (6) segue

$$0 = \sum_{i\in I^+} \alpha_i \bar{w}^T x^i - \sum_{j\in J^-} |\beta_j| \bar{w}^T x^j > \bar{b} \left(\sum_{j\in J^-} |\beta_j| - \sum_{i\in I^+} \alpha_i \right), \tag{7}$$

da cui otteniamo $\sum_{j\in J^-} |\beta_j| - \sum_{i\in I^+} \alpha_i \neq 0$.

Supponiamo $\sum_{j\in J^-} |\beta_j| - \sum_{i\in I^+} \alpha_i > 0$: la (7) implica

$$\bar{b} < 0. \tag{8}$$

D'altra parte, assumendo che l'etichetta y^1 associata al vettore $x^1 = 0$ sia pari a $+1$, dalla (4) abbiamo $\bar{b} > 0$, che contraddice la (8).

Supponiamo allora che $\sum_{j\in J^-} |\beta_j| - \sum_{i\in I^+} \alpha_i < 0$: la (7) implica

$$\bar{b} > 0. \tag{9}$$

Assumendo che l'etichetta y^1 associata al vettore $x^1 = 0$ sia pari a -1, dalla (4) abbiamo $\bar{b} < 0$, che contraddice la (9). □

Il teorema precedente implica il seguente risultato.

Corollario 1 *La VC dimension dell'insieme degli iperpiani orientati in R^n è $n + 1$.*

Avendo definito la VC dimension, si consideri nuovamente la (1). Osserviamo che la VC confidence (il secondo termine del secondo membro) è una funzione monotona crescente di h. Quindi, dato un insieme di macchine di apprendimento (ognuna rappresentata da un insieme di funzioni) che presentano rischio empirico zero, dalla (1) segue che la macchina con VC dimension minima definisce il migliore upper bound del rischio effettivo. In generale, per rischio empirico maggiore di zero, si desidera selezionare la macchina che permette di minimizzare il secondo membro della (1).

Minimizzazione del rischio strutturale.

Si vuole determinare quel sottoinsieme dell'insieme dato di funzioni per cui l'upper bound del rischio effettivo è minimo.

A questo fine, si può pensare di introdurre *una struttura* nell'insieme di funzioni S definito da $\{f(\alpha)\}$, che viene decomposto in sottoinsiemi contenuti l'uno nell'altro:

$$S_1 \subset S_2 \subset \ldots S, \tag{10}$$

con $h_1 \leq h_2 \ldots \leq h$ (per ogni sottoinsieme S_k, deve essere quindi possibile o calcolare h_k oppure definire un upper bound di h_k).

In linea di principio, si può pensare di addestrare una serie di macchine, una per ogni sottoinsieme S_k, e l'obiettivo della procedura di addestramento è quello di minimizzare il rischio empirico. Tra le varie macchine addestrate si seleziona quella per cui la somma del rischio empirico e della VC confidence è minima.

1.1 Reti neurali e SVM

Per la minimizzazione del secondo membro della (1), due approcci possono essere individuati.

Nel primo approccio viene determinato un insieme di funzioni con fissata VC dimension $h^\star$. Il processo di addestramento è finalizzato in questo caso alla minimizzazione del numero di errori nel training set (minimizzazione del rischio empirico). Se $h^\star$ è troppo grande rispetto al numero l di dati disponibili (stiamo quindi considerando una macchina di "complessità" elevata), la VC confidence può risultare elevata, per cui, anche riuscendo ad ottenere rischio empirico nullo, l'errore sul test set (rischio effettivo) potrebbe essere grande. Questo fenomeno è denominato *overfitting*.

Per evitare l'overfitting si potrebbe individuare un sottoinsieme di funzioni con VC dimension $h^\star$ sufficientemente piccola (stiamo quindi considerando una macchina di "complessità" bassa), D'altra parte, se $h^\star$ ha un valore troppo piccolo, potrebbe non essere possibile (con la procedura di addestramento) ridurre sufficientemente il rischio empirico. Questo fenomeno è denominato *underfitting*.

Quindi, sarebbe auspicabile definire a priori (sulla base delle informazioni riguardanti il problema) una macchina con architettura (complessità) appropriata in modo da evitare sia l'overfitting che l'underfitting. Di conseguenza, l'addestramento della macchina è finalizzato alla minimizzazione degli errori sui dati di training.

In altre parole, la strategia del primo approccio è la seguente:

> *fissare la VC confidence (scegliendo opportunamente l'architettura della macchina) e minimizzare il rischio empirico.*

Per quanto riguarda il secondo approccio si ha la seguente strategia:

> *fissare il valore del rischio empirico e minimizzare la VC confidence.*

Macchine di apprendimento corrispondenti ai due approcci descritti sono:

(i) *Reti neurali*;

(ii) *Support Vector Machines*.

1.2 Iperpiano con gap di tolleranza

Definiamo *iperpiano con gap di tolleranza* ρ (con $\rho > 0$), un iperpiano $H(w, b)$ tale che per ogni x appartenente ad una sfera di diametro D si ha

$$y = \begin{cases} 1 & \text{se } \frac{w^T x - b}{\|w\|} \geq \rho \\ \\ -1 & \text{se } \frac{w^T x - b}{\|w\|} \leq -\rho \end{cases}$$

L'insieme degli iperpiani con gap di tolleranza ρ costituisce una famiglia di classificatori contenuta nella famiglia di classificatori rappresentata dagli iperpiani orientati. In particolare, la funzione di decisione di un iperpiano con gap di tolleranza è la seguente: punti appartenenti alla sfera di diametro D e a distanza maggiore di ρ sono etichettati con +1 oppure -1, a seconda del semispazio di appartenenza. Tutti gli altri punti sono etichettati convenzionalmente con 0. Vale il seguente risultato [2].

Teorema 2 *L'insieme degli iperpiani con gap di tolleranza ρ ha VC dimension h tale che*

$$h \leq \min \left\{ \left\lceil \frac{D^2}{\rho^2} \right\rceil, n \right\} + 1 \tag{11}$$

Dal teorema precedente segue che la VC dimension della classe di funzioni relativa agli iperpiani con gap di tolleranza può essere controllata mediante il diametro D della sfera ed il valore di gap ρ. Al variare di D e ρ si può definire una struttura "annidata" di sottoinsiemi di funzioni del tipo (10).

2 SVM lineari

Verrà considerato il problema di classificare gli elementi di due insiemi di punti di R^n mediante superfici di separazione definite da iperpiani.

2.1 Iperpiano ottimo

Si considerino due insiemi disgiunti di punti A e B in R^n. Si assuma che A e B siano *linearmente separabili*, cioè, che esista un iperpiano $H = \{x \in R^n : \ w^T x + b = 0\}$ per cui tutti i punti $x^i \in A$ appartengono ad un semispazio e quelli $x^j \in B$ appartengono all'altro. Esistono quindi un vettore $w \in R^n$ e uno scalare $b \in R$ tali che

$$\begin{array}{ll} w^T x^i + b & \geq \varepsilon, \quad \forall x^i \in A \\ w^T x^j + b & \leq -\varepsilon, \quad \forall x^j \in B \end{array} \tag{12}$$

con $\varepsilon > 0$. Senza perdita di generalità, si può pensare di riscalare la (12) dividendo per ε, e si ottiene

$$\begin{array}{ll} w^T x^i + b & \geq 1, \quad \forall x^i \in A \\ w^T x^j + b & \leq -1, \quad \forall x^j \in B \end{array} \tag{13}$$

Dato un iperpiano di separazione H, cioè, una coppia (w, b) per cui la (13) è verificata, introduciamo il concetto di margine di separazione.

Definizione 3 *Sia H un iperpiano di separazione. Si definisce margine di separazione di H la minima distanza ρ tra i punti in $A \cup B$ e l'iperpiano H, cioè*

$$\rho(w,b) = \min_{x^i \in A \cup B} \left\{ \frac{|w^T x^i + b|}{\|w\|} \right\}.$$

Definizione 4 *Si definisce iperpiano ottimo l'iperpiano di separazione $H(w^\star, b^\star)$ avente margine di separazione massimo.*

Determinare l'iperpiano ottimo equivale quindi a risolvere il seguente problema

$$\max_{w \in R^n, b \in R} \min_{x^i \in A \cup B} \left\{ \frac{|w^T x^i + b|}{\|w\|} \right\} \tag{14}$$

Dimostreremo che l'iperpiano ottimo *esiste* ed è *unico*. A questo fine proveremo l'equivalenza del problema (14) con il seguente

$$\begin{array}{lll} \min\limits_{w \in R^n, b \in R} & \|w\|^2 & \\ \text{t.c.} & w^T x^i + b \geq 1, & \forall x^i \in A \\ & w^T x^j + b \leq -1, & \forall x^j \in B \end{array} \tag{15}$$

Lemma 1 *Per ogni iperpiano di separazione, cioè per ogni coppia $(\hat{w}, \hat{b})$ tale che (13) è verificata, segue*

$$\rho(\hat{w}, \hat{b}) \geq \frac{1}{\|\hat{w}\|}.$$

Dim. Poichè

$$|\hat{w}^T x^\ell + \hat{b}| \geq 1, \quad \forall x^\ell \in A \cup B,$$

si ha

$$\rho(\hat{w}, \hat{b}) = \min_{x^\ell \in A \cup B} \left\{ \frac{|\hat{w}^T x^\ell + \hat{b}|}{\|\hat{w}\|} \right\} \geq \frac{1}{\|\hat{w}\|}. \quad \square$$

Lemma 2 *Per ogni iperpiano di separazione $(\hat{w}, \hat{b})$, esiste un iperpiano di separazione $(\bar{w}, \bar{b})$ tale che*

$$\rho(\hat{w}, \hat{b}) \leq \rho(\bar{w}, \bar{b}) = \frac{1}{\|\bar{w}\|}. \tag{16}$$

Inoltre, esistono almeno due punti $x^+ \in A$ e $x^- \in B$ tali che

$$\begin{array}{l} \bar{w}^T x^+ + \bar{b} = 1 \\ \bar{w}^T x^- + \bar{b} = -1 \end{array} \tag{17}$$

Dim. Siano $\hat{x}^i \in A$ e $\hat{x}^j \in B$ i punti più vicini all'iperpiano $(\hat{w}, \hat{b})$, cioè i punti per i quali

$$\begin{array}{rll} \hat{d}_i & = \dfrac{|\hat{w}^T\hat{x}^i + \hat{b}|}{\|\hat{w}\|} \leq \dfrac{|\hat{w}^T x^i + \hat{b}|}{\|\hat{w}\|}, & \forall x^i \in A \\ \hat{d}_j & = \dfrac{|\hat{w}^T\hat{x}^j + \hat{b}|}{\|\hat{w}\|} \leq \dfrac{|\hat{w}^T x^j + \hat{b}|}{\|\hat{w}\|}, & \forall x^j \in B \end{array} \tag{18}$$

da cui segue

$$\rho(\hat{w}, \hat{b}) = \min\{\hat{d}_i, \hat{d}_j\} \leq \frac{1}{2}(\hat{d}_i + \hat{d}_j) = \frac{\hat{w}^T(\hat{x}^i - \hat{x}^j)}{2\|\hat{w}\|}. \tag{19}$$

Si considerino gli scalari α e β tali che

$$\begin{array}{ll} \alpha\hat{w}^T\hat{x}^i + \beta & = 1 \\ \alpha\hat{w}^T\hat{x}^j + \beta & = -1 \end{array} \tag{20}$$

cioè i valori

$$\alpha = \frac{2}{\hat{w}^T(\hat{x}^i - \hat{x}^j)}, \qquad \beta = -\frac{\hat{w}^T(\hat{x}^i + \hat{x}^j)}{\hat{w}^T(\hat{x}_i - \hat{x}_j)}.$$

Si può facilmente verificare che $0 < \alpha \leq 1$. Mostreremo che l'iperpiano $(\bar{w}, \bar{b}) \equiv (\alpha\hat{w}, \beta)$ separa gli insiemi A e B, e soddisfa la (16). Infatti, dalla (18) si ha

$$\begin{array}{ll} \hat{w}^T x^i & \geq \hat{w}^T\hat{x}^i, \quad \forall x^i \in A \\ \hat{w}^T x^j & \leq \hat{w}^T\hat{x}^j, \quad \forall x^j \in B. \end{array}$$

Poichè $\alpha > 0$, possiamo scrivere

$$\begin{array}{lll} \alpha\hat{w}^T x^i + \beta \geq \alpha\hat{w}^T\hat{x}^i + \beta = & 1, & \forall x^i \in A \\ \alpha\hat{w}^T x^j + \beta \leq \alpha\hat{w}^T\hat{x}^j + \beta = & -1, & \forall x^j \in B \end{array} \tag{21}$$

da cui segue che $\bar{w}$ e $\bar{b}$ soddisfano la (13), e quindi $(\bar{w}, \bar{b})$ è un iperpiano di separazione per gli insiemi A e B.

Inoltre, tenendo conto della (21) e del valore di α, abbiamo

$$\rho(\bar{w}, \bar{b}) = \min_{x^\ell \in A \cup B} \left\{ \frac{|\bar{w}^T x^\ell + \bar{b}|}{\|\bar{w}\|} \right\} = \frac{1}{\|\bar{w}\|} = \frac{1}{\alpha\|\hat{w}\|} = \frac{\hat{w}^T(\hat{x}^i - \hat{x}^j)}{2\|\hat{w}\|}.$$

La (16) segue dalla precedente relazione e dalla (19). Dalla (20) si ottiene la (17) con $x^+ = \hat{x}^i$ e $x^- = \hat{x}^j$. □

Proposizione 1 *Il seguente problema*

$$\begin{array}{lll} \min & \|w\|^2 & \\ \text{t.c.} & w^T x^i + b \geq 1, & \forall x^i \in A \\ & w^T x^j + b \leq -1, & \forall x^j \in B \end{array} \tag{22}$$

ammette una soluzione unica $(w^\star, b^\star)$.

Dim. Sia $\mathcal{F}$ l'insieme ammissibile

$$\mathcal{F} = \{(w,b) \in R^n \times R : \ w^T x^i + b \geq 1, \forall x^i \in A, \ w^T x^j + b \leq -1, \forall x^j \in B\},$$

e, per una assegnata coppia $(w_o, b_o) \in \mathcal{F}$, si consideri l'insieme di livello

$$\mathcal{L}_o = \{(w,b) \in \mathcal{F} : \ \|w\|^2 \leq \|w_o\|^2\}.$$

L'insieme $\mathcal{L}_o$ è ovviamente chiuso, e dimostreremo che è anche limitato. Infatti, supponiamo per assurdo che esista una sequenza illimitata $\{(w_k, b_k)\}$ appartenente a $\mathcal{L}_o$. Poichè $\|w_k\| \leq \|w_o\|, \forall k$, si ha necessariamente $|b_k| \to \infty$. Per ogni k possiamo scrivere

$$\begin{array}{ll} w_k^T x^i + b_k & \geq 1, \quad \forall x^i \in A \\ w_k^T x^j + b_k & \leq -1, \quad \forall x^j \in B \end{array}$$

e quindi, poichè $|b_k| \to \infty$, segue per k sufficientemente grande, $\|w_k\|^2 > \|w_o\|^2$, e ciò è in contraddizione con l'ipotesi che $\{(w_k, b_k)\}$ appartenga all'insieme $\mathcal{L}_o$, e quindi $\mathcal{L}_o$ è compatto. Il teorema di Weirstrass implica che la funzione $\|w\|^2$ ammette un minimo $(w^\star, b^\star)$ sull'insieme $\mathcal{L}_o$ e quindi su $\mathcal{F}$, di conseguenza $(w^\star, b^\star)$ è una soluzione del problema (22).

Per dimostrare che $(w^\star, b^\star)$ è l'unico punto di minimo globale, assumiamo per contraddizione che esista una coppia $(\bar{w}, \bar{b}) \in \mathcal{F}$, $(\bar{w}, \bar{b}) \neq (w^\star, b^\star)$, tale che $\|\bar{w}\|^2 = \|w^\star\|^2$. Supponiamo che $\bar{w} \neq w^\star$. L'insieme $\mathcal{F}$ è convesso, per cui

$$\lambda(w^\star, b^\star) + (1-\lambda)(\bar{w}, \bar{b}) \in \mathcal{F}, \quad \forall \lambda \in [0,1],$$

e poichè $\|w\|^2$ è una funzione strettamente convessa, per ogni $\lambda \in (0,1)$ segue

$$\|\lambda w^\star + (1-\lambda)\bar{w}\|^2 < \lambda \|w^\star\|^2 + (1-\lambda)\|\bar{w}\|^2.$$

Ponendo, ad esempio, $\lambda = 1/2$, cioè considerando la coppia $(\tilde{w}, \tilde{b}) \equiv \left(\frac{1}{2}w^\star + \frac{1}{2}\bar{w}, \frac{1}{2}b^\star + \frac{1}{2}\bar{b}\right)$, abbiamo $(\tilde{w}, \tilde{b}) \in \mathcal{F}$ e

$$\|\tilde{w}\|^2 < \frac{1}{2}\|w^\star\|^2 + \frac{1}{2}\|\bar{w}\|^2 = \|w^\star\|^2,$$

che contraddice il fatto che $(w^\star, b^\star)$ è un punto di minimo globale. Quindi abbiamo necessariamente $\bar{w} \equiv w^\star$.

Supponiamo allora che $b^\star > \bar{b}$ (il caso $b^\star < \bar{b}$ è analogo), e consideriamo un punto $\hat{x}^i \in A$ tale che

$$w^{\star T} \hat{x}^i + b^\star = 1$$

(L'esistenza di tale punto è assicurata dalla (17) del Lemma 2). Abbiamo

$$1 = w^{\star T} \hat{x}^i + b^\star = \bar{w}^T \hat{x}^i + b^\star > \bar{w}^T \hat{x}^i + \bar{b}$$

che contraddice il fatto che $\bar{w}^T x^i + \bar{b} \geq 1, \ \forall x^i \in A$. Di conseguenza, $\bar{b} \equiv b^\star$, e la dimostrazione è quindi completata. $\square$

Proposizione 2 *Sia* $(w^\star, b^\star)$ *la soluzione del problema* (22). *Allora,* $(w^\star, b^\star)$ *è l'unica soluzione del problema*

$$\begin{array}{ll} \max & \rho(w,b) \\ \text{t.c.} & w^T x^i + b \geq 1, \quad \forall x^i \in A \\ & w^T x^j + b \leq -1, \quad \forall x^j \in B \end{array} \tag{23}$$

.

Dim. Osserviamo che $(w^\star, b^\star)$ è l'unica soluzione del problema

$$\begin{array}{ll} \max & \frac{1}{\|w\|} \\ \text{t.c.} & w^T x^i + b \geq 1, \quad \forall x^i \in A \\ & w^T x^j + b \leq -1, \quad \forall x^j \in B. \end{array}$$

Il Lemma 1 e il Lemma 2 implicano che per ogni iperpiano di separazione (w, b) segue

$$\frac{1}{\|w\|} \leq \rho(w,b) \leq \frac{1}{\|w^\star\|}$$

e quindi, per la coppia $(w^\star, b^\star)$, otteniamo $\rho(w^\star, b^\star) = \dfrac{1}{\|w^\star\|}$, cioè $(w^\star, b^\star)$ è l'iperpiano ottimo. □

Osservazione. Ogni iperpiano di separazione $H(w, b)$ ha, per definizione, rischio empirico nullo, quindi si ha

$$R(w,b) \leq \sqrt{\left(\frac{h(\log(2l/h)+1) - log(\eta/4)}{l}\right)}. \tag{24}$$

Ogni iperpiano di separazione $H(w, b)$ ha un margine ρ, di conseguenza appartiene ad un sottoinsieme per cui, utilizzando il Teorema 2, è possibile definire un upper bound della VC dimension. Poichè il margine ρ è minore del margine $\rho^\star$ dell'iperpiano ottimo $H(w^\star, b^\star)$, segue che l'iperpiano ottimo appartiene ad un sottoinsieme per cui l'upper bound della VC dimension è minimo. Ricordando che il secondo membro della (24) è una funzione crescente di h, possiamo concludere che l'upper bound del rischio effettivo relativo all'iperpiano ottimo $H(w^\star, b^\star)$ è minore dell'upper bound relativo ad un qualsiasi iperpiano di separazione, e ciò giustifica l'interesse nella determinazione dell'iperpiano ottimo. □

3 Il duale di Wolfe

Il concetto di *dualità* è ampiamente utilizzato nella programmazione matematica. In relazione a determinati problemi, la teoria della dualità consente di definire formulazioni alternative che, da un lato possono essere risolte più efficientemente da un punto di vista computazionale, dall'altro possono fornire elementi teorici rilevanti.

L'idea alla base della teoria dela dualità è quella di costruire, in corrispondenza ad un problema assegnato di minimo, detto *problema primale*:

$$\min_{x \in S} f(x)$$

un problema di massimo, detto *problema duale* (definito, in genere, su uno spazio diverso)

$$\max_{u \in U} \psi(u)$$

in modo tale che valga *almeno* la condizione (detta proprietà di *dualità debole*)

$$\inf_{x \in S} f(x) \geq \sup_{u \in U} \psi(u).$$

Ove si riesca a stabilire tale corrispondenza, è possibile fornire utili caratterizzazioni delle soluzioni del primale attraverso lo studio del duale.

In particolare, dalla teoria della dualità è possibile ricavare

- *stime del valore ottimo*;
- *condizioni di ottimalità*;
- *metodi di soluzione* basati sulla considerazione del problema duale.

Per alcune classi di problemi si riesce anche a stabilire la consizione (detta proprietà di *dualità forte*)

$$\inf_{x \in S} f(x) = \sup_{u \in U} \psi(u).$$

La possibilità di costruire un duale che soddisfi una proprietà di dualità forte sussiste per una classe (ristretta ma importante) di problemi di ottimo che comprende, tipicamente, molti problemi di programmazione convessa.

Si consideri il problema

$$\begin{aligned} &\min \ f(x) \\ &g_i(x) \leq 0, \qquad i = 1, \ldots, m \\ &c_j^T x - d_j = 0, \quad j = 1, \ldots, p \end{aligned} \tag{25}$$

in cui $f : R^n \to R$, $g_i : R^n \to R$, $i = 1, \ldots, m$ sono funzioni convesse continuamente differenziabili. Sia

$$L(x, \lambda, \mu) = f(x) + \sum_{i=1}^{m} \lambda_i g_i(x) + \sum_{j=1}^{p} \mu_j (c_j^T x - d_j)$$

Proposizione 3 *Si assuma che il problema* (25) *ammetta almeno una soluzione ottima* $x^\star$*, e che esista almeno una coppia di moltiplicatori di lagrange* $(\lambda^\star, \mu^\star)$*. Allora la tripla* $(x^\star, \lambda^\star, \mu^\star)$ *è soluzione del seguente problema*

$$\begin{array}{l} \max \ L(x, \lambda, \mu) \\ x, \lambda, \mu \\ \\ \nabla_x L(x, \lambda, \mu) = 0 \\ \\ \lambda \geq 0. \end{array} \tag{26}$$

Inoltre, il gap di dualità è nullo, ossia $f(x^\star) = L(x^\star, \lambda^\star, \mu^\star)$*.*

Dim. Le condizioni KKT implicano

$$\begin{aligned} \nabla_x L(x^\star, \lambda^\star, \mu^\star) = 0, \\ (\lambda^*)^T g(x^\star) = 0, \\ \lambda^\star \geq 0. \end{aligned} \tag{27}$$

Il punto $(x^\star, \lambda^\star, \mu^\star)$ è perciò ammissibile per il problema (26), e risulta $f(x^\star) = L(x^\star, \lambda^\star, \mu^\star)$. Mostreremo che $(x^\star, \lambda^\star, \mu^\star)$ è soluzione del problema (26).

Sia (x, λ, μ) una soluzione ammissibile del problema (26), cioè tale che $\nabla_x L(x, \lambda, \mu) = 0$ e $\lambda \geq 0$. Osserviamo che, per ogni $\bar{\lambda} \geq 0$ e per ogni $\bar{\mu}$, la funzione di x

$$L(x, \bar{\lambda}, \bar{\mu}) = f(x) + \sum_{i=1}^{m} \bar{\lambda}_i g_i(x) + \sum_{j=1}^{p} \bar{\mu}_j (c_j^T x - d_j)$$

è una funzione convessa. Infatti, la funzione $f(x)$ è convessa per ipotesi, il secondo termine è una combinazione lineare, con coefficienti non negativi, di funzioni convesse, e quindi, per un noto risultato è una funzione convessa; infine, il terzo termine è una funzione affine e quindi convessa. Di conseguenza, la funzione $L(x, \bar{\lambda}, \bar{\mu})$ è espressa come somma di tre funzioni convesse, e quindi risulta essere una funzione convessa.

Tenendo conto della (27), del fatto che $g(x^\star) \leq 0$ e $\lambda \geq 0$, della convessità di L come funzione di x (per cui risulta $L(y) \geq L(x) + (y - x)^T \nabla_x L(x)$), e utilizzando la condizione $\nabla_x L = 0$, per ogni $\lambda \geq 0$ e per ogni μ, si può scrivere

$$\begin{aligned} L(x^\star, \lambda^\star, \mu^\star) = f(x^\star) \ &\geq f(x^\star) + \textstyle\sum_{i=1}^{m} \lambda_i g_i(x^\star) = L(x^\star, \lambda, \mu) \\ &\geq L(x, \lambda, \mu) + (x^\star - x)^T \nabla_x L(x, \lambda, \mu) = L(x, \lambda, \mu) \end{aligned}$$

e quindi la tesi è dimostrata. □

Il problema (26) è usualmente indicato come *problema duale di Wolfe.*

Osservazione. In generale, data una soluzione $(\bar{x}, \bar{\lambda}, \bar{\mu})$ del duale di Wolfe, non si possono trarre conclusioni sui vettori $\bar{x}$ e $(\bar{\lambda}, \bar{\mu})$. □

Programmazione quadratica

Si consideri il seguente problema quadratico

$$\begin{aligned} &\min\ f(x) = \tfrac{1}{2}x^T Q x + c^T x \\ &Ax - b \leq 0, \end{aligned} \tag{28}$$

dove $Q \in R^{n\times n}$, $A \in R^{m\times n}$, $c \in R^n$, $b \in R^m$. Posto $L(x,\lambda) = f(x) + \lambda^T(Ax-b)$, il duale di Wolfe è definito come segue

$$\begin{aligned} &\max\ L(x,\lambda) \\ &x, \lambda \\ &\nabla_x L(x,\lambda) = 0 \\ &\lambda \geq 0. \end{aligned} \tag{29}$$

Vale il seguente risultato.

Proposizione 4 *Si assuma che la matrice Q sia semidefinita positiva. Sia $(\bar{x}, \bar{\lambda})$ una soluzione del duale di Wolfe* (29)*. Allora esiste un vettore $x^\star$ (non necessariamente uguale a $\bar{x}$) tale che*

(i) $Q(x^\star - \bar{x}) = 0$*;*

(ii) $x^\star$ *è soluzione del problema* (28)*;*

(iii) $(x^\star, \bar{\lambda})$ *è una coppia (minimo globale- vettore di moltiplicatori di lagrange).*

Dim. Si consideri il problema duale (29):

$$\begin{aligned} &\max_{x,\lambda} \tfrac{1}{2}x^T Q x + c^T x + \lambda^T(Ax-b) \\ &Qx + A^T\lambda + c = 0 \\ &\lambda \geq 0 \end{aligned}$$

Il vincolo $Qx + A^T\lambda + c = 0$ implica

$$x^T Q x + c^T x + \lambda^T A x = 0. \tag{30}$$

Utilizzando la (30), il problema duale può essere scritto nella forma

$$\begin{aligned} &\min_{x,\lambda} \tfrac{1}{2}x^T Q x + \lambda^T b \\ &Qx + A^T\lambda + c = 0 \\ &\lambda \geq 0 \end{aligned} \tag{31}$$

Sia $(\bar{x}, \bar{\lambda})$ una soluzione del problema (31). Si consideri la funzione Lagrangiana associata al problema (31)

$$W(x, \lambda, v, z) = \frac{1}{2} x^T Q x + \lambda^T b - v^T (Qx + A^T \lambda + c) - z^T \lambda.$$

Poichè $(\bar{x}, \bar{\lambda})$ è soluzione del duale di Wolfe (29), dalle condizioni KKT si ha che esistono un vettore $\bar{v} \in R^n$ e un vettore $\bar{z} \in R^r$ tali che

$$\begin{aligned} &\nabla_x W = Q\bar{x} - Q\bar{v} = 0 \\ &\nabla_\lambda W = b - A\bar{v} - \bar{z} = 0 \\ &Q\bar{x} + A^T\bar{\lambda} + c = 0 \\ &\bar{z}^T\bar{\lambda} = 0 \\ &\bar{z} \geq 0 \\ &\bar{\lambda} \geq 0 \end{aligned} \tag{32}$$

Dalla seconda e dalla quinta relazione si ricava $\bar{z} = b - A\bar{v} \geq 0$, per cui le precedenti condizioni possono essere scritte nella forma

$$\begin{aligned} &Q\bar{x} - Q\bar{v} = 0 \\ &-b + A\bar{v} \leq 0 \\ &Q\bar{x} + A^T\bar{\lambda} + c = 0 \\ &-\bar{\lambda}^T b + \bar{\lambda}^T A\bar{v} = 0 \\ &\bar{\lambda} \geq 0 \end{aligned} \tag{33}$$

Sottraendo la prima dalla terza relazione si ottiene

$$Q\bar{v} + A^T\bar{\lambda} + c = 0. \tag{34}$$

Poichè la matrice Q è semidefinita positiva, la funzione f risulta essere una funzione convessa. Dalla Proposizione 9 segue quindi che le relazioni

$$\begin{aligned} &A\bar{v} - b \leq 0 \\ &Q\bar{v} + A^T\bar{\lambda} + c = 0 \\ &\bar{\lambda}^T(A\bar{v} - b) = 0 \\ &\bar{\lambda} \geq 0 \end{aligned}$$

sono sufficienti ad assicurare che $\bar{v}$ è una soluzione del problema (28). Quindi per definizione $(\bar{v}, \bar{\lambda})$ è una coppia (minimo globale-vettore di lagrange). Posto $x^\star = \bar{v}$ si ha che $x^\star$ è soluzione del problema (28), inoltre, tenendo conto della prima relazione nella (33), risulta

$$Qx^\star = Q\bar{v} = Q\bar{x}.$$

In tal modo, le asserzioni (i)-(iii) risultano dimostrate. □

3.1 Programmazione quadratica per SVM lineari

Determinare una SVM lineare equivale a determinare l'iperpiano ottimo, cioè a risolvere il seguente problema

$$\begin{aligned} &\max\ \rho(w,b) \\ &\text{t.c.} \qquad w^T x^i + b \geq 1, \quad \forall x^i \in A \\ &\qquad\qquad w^T x^j + b \leq -1, \quad \forall x^j \in B, \end{aligned} \tag{35}$$

dove A e B sono due insiemi di punti contenuti in R^n linearmente separabili. Abbiamo visto che il problema (35) è equivalente al seguente problema di programmazione quadratica convessa

$$\begin{aligned} &\min F(w) \ = \frac{1}{2}\|w\|^2 \\ &\text{t.c.} \qquad w^T x^i + b \geq 1, \quad \forall x^i \in A \\ &\qquad\qquad w^T x^j + b \leq -1, \quad \forall x^j \in B \end{aligned} \tag{36}$$

Associando l'etichetta $y^i = +1$ ai vettori $x^i \in A$, e l'etichetta $y^j = -1$ ai vettori $x^j \in B$, il problema (37) può essere scritto nella seguente forma

$$\begin{aligned} &\min F(w) \ = \frac{1}{2}\|w\|^2 \\ &\text{t.c.} \qquad y^i \left[w^T x^i + b\right] - 1 \geq 0, \quad i = 1, \ldots, l \end{aligned} \tag{37}$$

essendo l il numero dei punti dell'insieme $A \cup B$.

Considereremo ora *la formulazione duale* del problema (37). Le motivazioni sono le seguenti:

- i vincoli del problema (37) sono sostituiti da vincoli "più semplici" sui moltiplicatori di Lagrange;
- nella formulazione duale, i vettori di training compariranno attraverso prodotti scalari tra i vettori stessi, e ciò consentirà di generalizzare la procedura al caso di insiemi che non sono linearmente separabili.

La funzione Lagrangiana del problema (37) è la seguente funzione

$$L(w,b,\lambda) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{l} \lambda_i [y^i(w^T x^i + b) - 1] \tag{38}$$

Il problema duale del problema (37) è il seguente

$$\begin{aligned} \max\ L(w,b,\lambda) = \quad & \frac{1}{2}\|w\|^2 - \sum_{i=1}^{l} \lambda_i [y^i(w^T x^i + b) - 1] \\ t.c. \qquad & \nabla_w L(w,b,\lambda) = 0 \\ & \frac{\partial L(w,b,\lambda)}{\partial b} = 0 \\ & \lambda \geq 0, \end{aligned}$$

ossia

$$\begin{aligned} \max\ L(w,b,\lambda) = \quad & \frac{1}{2}\|w\|^2 - \sum_{i=1}^{l} \lambda_i [y^i(w^T x^i + b) - 1] \\ t.c. \qquad & w = \sum_{i=1}^{l} \lambda_i y^i x^i \\ & \sum_{i=1}^{l} \lambda_i y^i = 0 \\ & \lambda_i \geq 0 \qquad i = 1, \ldots, l \end{aligned} \tag{39}$$

Il problema (39) può essere riscritto nella seguente forma

$$\begin{aligned} \max\ S(\lambda) = \quad & -\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l} y^i y^j (x^i)^T x^j \lambda_i \lambda_j + \sum_{i=1}^{l} \lambda_i \\ t.c. \qquad & \sum_{i=1}^{l} \lambda_i y^i = 0 \\ & \lambda_i \geq 0 \qquad i = 1, \ldots, l \end{aligned} \tag{40}$$

oppure, in maniera equivalente

$$\begin{aligned} \min\ \Gamma(\lambda) = \ & \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l} y^i y^j (x^i)^T x^j \lambda_i \lambda_j - \sum_{i=1}^{l} \lambda_i \\ t.c. \quad & \sum_{i=1}^{l} \lambda_i y^i = 0 \\ & \lambda_i \geq 0 \qquad i = 1, \ldots, l \end{aligned} \tag{41}$$

Osserviamo che:

- l'esistenza della soluzione ottima $(w^\star, b^\star)$ del problema (37) e le proposizioni 8 e 3 assicurano che il problema (40) ammette almeno una soluzione ottima $\lambda^\star$;
- dalla formulazione (39) segue che il vettore $w^\star$ può essere determinato come segue
$$w^\star = \sum_{i=1}^{l} \lambda_i^\star y^i x^i;$$
- $w^\star$ dipende esclusivamente dai vettori di training x^i (*vettori di supporto*) i cui corrispondenti moltiplicatori $\lambda_i^\star$ sono diversi da zero;
- l'asserzione (iii) della Proposizione 4 assicura che $(w^\star, b^\star, \lambda^\star)$ costituisce una coppia (soluzione ottima-vettore di moltiplicatori di lagrange), per cui valgono le seguenti condizioni di complementarità
$$\lambda_i^\star \left[y^i \left((w^\star)^T x^i + b^\star \right) - 1 \right] = 0 \qquad i = 1, \ldots, l \tag{42}$$
- noto $w^\star$ e considerato un qualsiasi moltiplicatore $\lambda_i^\star \neq 0$, lo scalare $b^\star$ può essere determinato utilizzando la corrispondente condizione definita nella (42).
- il problema (41) è un *problema di programmazione quadratica convessa*; infatti, posto $X = \left[y^1 x^1, \ldots, y^l x^l \right]$, $\lambda^T = \left[\lambda^1, \ldots, \lambda^l \right]$, il problema assume la forma
$$\begin{aligned} \min\ \Gamma(\lambda) = \ & \frac{1}{2}\lambda^T X^T X \lambda - e^T \lambda \\ t.c. \quad & \sum_{i=1}^{l} \lambda_i y^i = 0 \\ & \lambda_i \geq 0 \qquad i = 1, \ldots, l, \end{aligned}$$
dove $e^T = [1, \ldots, 1]$;
- la funzione di decisione risulta essere
$$f(x) = sgn\left((w^*)^T x + b^\star \right) = sgn\left(\sum_{i=1}^{l} \lambda_i^\star y^i (x^i)^T x + b^\star \right)$$

3.2 Il caso non separabile linearmente

Siano A e B due insiemi disgiunti di punti in R^n, e si assuma che A e B non siano linearmente separabili, ossia il sistema

$$\begin{array}{ll} w^T x^i + b & \geq 1, \quad \forall x^i \in A \\ w^T x^j + b & \leq -1, \quad \forall x^j \in B \end{array} \tag{43}$$

non ammette soluzione. Si introducano nel sistema (43) delle variabili positive *slack* ξ_h, con $h = 1, \ldots, l$:

$$\begin{array}{ll} w^T x^i + b & \geq 1 - \xi_i, \quad \forall x^i \in A \\ w^T x^j + b & \leq -1 + \xi_j, \quad \forall x^j \in B \\ \xi_h \geq 0, & h = 1, \ldots, l \end{array} \tag{44}$$

Si noti che, se un vettore di ingresso x^i non è classificato correttamente (ad esempio, $x^i \in A$ e quindi deve risultare $w^T x^i + b > 0$) la corrispondente variabile ξ^i risulta maggiore di 1. Quindi, il termine

$$\sum_{i=1}^{l} \xi_i$$

è un upper bound del numero di errori di classificazione dei vettori di training. Appare naturale perciò aggiungere alla funzione obiettivo del problema (37) il termine $C\sum_{i=1}^{l} \xi_i$, in cui $C > 0$ è un parametro che "pesa" l'errore di training. Il problema primale assume la seguente forma

$$\begin{array}{lll} \min F(w, \xi) & = \dfrac{1}{2}\|w\|^2 + C\sum_{i=1}^{l} \xi_i & \\ \text{t.c.} & y^i \left[w^T x^i + b\right] - 1 + \xi_i \geq 0 & i = 1, \ldots, l \\ & \xi_i \geq 0 & i = 1, \ldots, l \end{array} \tag{45}$$

Il duale di (45) è il problema

$$\begin{aligned}
\max\ L(w,b,\xi,\lambda,\mu) = \ & \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l}\xi_i - \sum_{i=1}^{l}\lambda_i[y^i(w^T x^i + b) - 1 + \xi_i] - \sum_{i=1}^{l}\mu_i\xi_i \\
t.c. \qquad & \nabla_w L(w,b,\xi,\lambda,\mu) = 0 \\
& \frac{\partial L(w,b,\xi,\lambda,\mu)}{\partial b} = 0 \\
& \frac{\partial L(w,b,\xi,\lambda,\mu)}{\partial \xi_i} = 0 \qquad i = 1,\ldots,l \\
& \lambda \geq 0 \\
& \mu \geq 0,
\end{aligned}$$

ossia

$$\begin{aligned}
\max\ L(w,b,\xi,\lambda,\mu) = \ & \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l}\xi_i - \sum_{i=1}^{l}\lambda_i[y^i(w^T x^i + b) - 1 + \xi_i] - \sum_{i=1}^{l}\mu_i\xi_i \\
t.c. \qquad & w = \sum_{i=1}^{l}\lambda_i y^i x^i \\
& \sum_{i=1}^{l}\lambda_i y^i = 0 \\
& C - \lambda_i - \mu_i = 0 \qquad i = 1,\ldots,l \\
& \lambda \geq 0 \\
& \mu \geq 0,
\end{aligned}$$

che può essere scritto nella forma

$$\begin{aligned}
\min\ \Gamma(\lambda) = \ & \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l} y^i y^j (x^i)^T x^j \lambda_i \lambda_j - \sum_{i=1}^{l}\lambda_i \\
t.c. \qquad & \sum_{i=1}^{l}\lambda_i y^i = 0 \\
& 0 \leq \lambda_i \leq C \qquad i = 1,\ldots,l
\end{aligned} \tag{46}$$

I vincoli $\lambda_i \leq C$, per $i = 1, \ldots, l$, derivano dalle relazioni $\lambda_i = C - \mu_i$, $\mu_i \geq 0$.

Osserviamo che:

- il vettore $w^\star$ può essere determinato come segue

$$w^\star = \sum_{i=1}^{l} \lambda_i^\star y^i x^i;$$

- in corrispondenza della soluzione ottima $(w^\star, b^\star, \xi^\star \lambda^\star, \mu^\star)$ valgono le seguenti condizioni di complementarità

$$\lambda_i^\star \left[y^i \left((w^\star)^T x^i + b^\star \right) - 1 + \xi_i^\star \right] = 0 \qquad i = 1, \ldots, l \tag{47}$$

$$\mu_i^\star \xi_i^\star = 0 \qquad i = 1, \ldots, l \tag{48}$$

- noto $w^\star$ e considerato un qualsiasi moltiplicatore $0 < \lambda_i^\star < C$, lo scalare $b^\star$ può essere determinato utilizzando la corrispondente condizione definita nella (47);
- nel caso in cui $\lambda_i^\star \in \{0, C\}$ per $i = 1, \ldots, l$, la soluzione è detta *degenere*;
- il problema (46) è un *problema di programmazione quadratica convessa*;
- la funzione di decisione risulta essere

$$f(x) = sgn\left((w^*)^T x + b^\star \right) = sgn\left(\sum_{i=1}^{l} \lambda_i^\star y^i (x^i)^T x + b^\star \right)$$

4 SVM non lineari

Verrà considerato il problema di classificare gli elementi di due insiemi di punti di R^n mediante superfici di separazione non lineari.

Funzioni kernel

Si chiama *prodotto scalare* in uno spazio lineare reale V una funzione $\langle x, y \rangle$ definita per ogni coppia di elementi $(x, y) \in V$ e soddisfacente alle seguenti condizioni:

(i) $\langle x, y \rangle = \langle y, x \rangle$;

(ii) $\langle x_1 + x_2, y \rangle = \langle x_1, y \rangle + \langle x_2, y \rangle$;

(iii) $\langle \lambda x, y \rangle = \lambda \langle x, y \rangle$;

(iv) $\langle x, x \rangle \geq 0$, inoltre $\langle x, x \rangle = 0$ soltanto se $x = 0$.

Dato un insieme $X \subseteq R^n$, una funzione

$$k : X \times X \to R$$

è un *kernel* se soddisfa la seguente proprietà

$$k(x, y) = \langle \phi(x), \phi(y) \rangle \qquad \forall x, y \in X, \tag{49}$$

dove ϕ è un'applicazione $X \to \mathcal{H}$ e $\mathcal{H}$ è uno spazio Euclideo, cioè uno spazio lineare con un prodotto scalare fissato in esso.

Osserviamo che un kernel è una funzione necessariamente simmetrica.

Proposizione 5 *Una funzione simmetrica $k : X \times X \to R$ è un kernel se e solo se, comunque si scelgano ℓ elementi $x_1, \ldots, x_\ell$ in X, la matrice*

$$K = [k(x_i, x_j)]_{i,j=1,\ldots,\ell}$$

risulta simmetrica e semidefinita positiva.

Siano $x^i \in R^n$, $i = 1, \ldots, l$ i vettori di training e $y^i \in \{-1, 1\}$ le corrispondenti etichette che individuano la classe di appartenenza di ciascun vettore.

Si consideri un'applicazione

$$\phi : R^n \to \mathcal{H}$$

in cui $\mathcal{H}$ è uno spazio Euclideo a dimensione maggiore di n (eventualmente infinita). Lo spazio $\mathcal{H}$ viene denominato *feature space* (*spazio delle caratteristiche*). Si considerino i "trasformati" $\phi(x^i)$ dei vettori di training x^i, con $i = 1, \ldots, l$ e il problema della determinazione dell'iperpiano ottimo nel feature space $\mathcal{H}$ in cui "esistono" i vettori "trasformati" $\phi(x^i)$, $i = 1, \ldots, l$.

Si può pensare di applicare la metodologia, descritta precedentemente, sostituendo x^i con $\phi(x^i)$. In particolare:

- il corrispondente del problema duale (46) è il seguente

$$\begin{aligned} \min\ \Gamma(\lambda) = \ & \tfrac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l} y^i y^j \phi(x^i)^T \phi(x^j)\lambda_i\lambda_j - \sum_{i=1}^{l}\lambda_i \\ t.c. \qquad & \sum_{i=1}^{l}\lambda_i y^i = 0 \\ & 0 \le \lambda_i \le C \qquad i = 1, \ldots, l \end{aligned} \tag{50}$$

- il vettore $w^\star$ risulta definito nel seguente modo

$$w^\star = \sum_{i=1}^{l} \lambda_i^\star y^i \phi(x^i)$$

- noto $w^\star$ e considerato un qualsiasi moltiplicatore $0 < \lambda_i^\star < C$, lo scalare $b^\star$ può essere determinato utilizzando la corrispondente condizione di complementarità

$$y^i \left(\sum_{j=1}^{l} \lambda_j^\star y^j \phi(x^j)^T \phi(x^i) + b^\star \right) - 1 = y^j \left(\sum_{j=1}^{l} \lambda_j^\star y^j k(x^j, x^i) + b^\star \right) - 1 = 0; \qquad (51)$$

- la funzione di decisione assume la seguente forma

$$f(x) = sgn\left((w^*)^T \phi(x) + b^\star \right) = sgn\left(\sum_{i=1}^{l} \lambda_i^\star y^i k(x^i, x) + b^\star \right). \qquad (52)$$

Osservazione. La (52) mostra che la superficie di separazione è:

- di tipo *lineare* nel *feature space*;

- di tipo *non lineare* nell' *input space*.

Si noti inoltre che sia nella formulazione del problema duale (50) che nell'espressione (52) della funzione di decisione non è necessario conoscere la rappresentazione esplicita della trasformazione ϕ, ma è sufficiente utilizzare una *funzione kernel* k, cioè una funzione tale che per $x, z \in R^n$ si ha

$$k(x, z) = \langle \phi(x), \phi(z) \rangle = \phi(x)^T \phi(z). \quad \square$$

Abbiamo visto che una funzione $k(x, z)$ è un kernel solo se la matrice $l \times l$

$$\left(k(x^i, x^j) \right)_{i,j=1}^{l} = \begin{pmatrix} k(x^1, x^1) & \dots & k(x^1, x^l) \\ & \vdots & \\ k(x^l, x^1) & \dots & k(x^l, x^l) \end{pmatrix}$$

risulta semidefinita positiva per ogni insieme di vettori di training $\{x^1, \dots, x^l\}$. (Nella letteratura le funzioni kernel sono spesso denominate *kernel di tipo Mercer*). Il problema (50) può quindi essere scritto nella forma equivalente

$$\begin{aligned} \min \ \Gamma(\lambda) = \ & \tfrac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} y^i y^j k(x^i, x^j) \lambda_i \lambda_j - \sum_{i=1}^{l} \lambda_i \\ t.c. \quad & \sum_{i=1}^{l} \lambda_i y^i = 0 \\ & 0 \le \lambda_i \le C \qquad i = 1, \dots, l \end{aligned} \qquad (53)$$

Si noti che il problema (50), nel caso di kernel di tipo Mercer, risulta essere un *problema di programmazione quadratica convessa.*
Esempi di funzioni kernel sono:

$k(x,z) = (x^T z + 1)^p$ *kernel polinomiale* (p intero ≥ 1)

$k(x,z) = e^{-\|x-z\|^2/2\sigma^2}$ *kernel gaussiano* ($\sigma > 0$)

$k(x,z) = tanh(\beta x^T z + \gamma)$ *kernel di tipo tangente iperbolica* (opportuni valori di β e γ)

La funzione di decisione può essere scritta nella forma

$$f(x) = sgn\left(\sum_{i=1}^{l} \lambda_i^\star y^i k(x, x^i) + b^\star\right)$$

Osserviamo che una SVM con kernel gaussiano corrisponde ad una *rete neurale di funzioni di base radiali*, in cui il numero di funzioni e i loro centri sono determinati automaticamente dal numero di support vector e dai loro valori. Analogamente, nel caso di SVM con kernel di tipo tangente iperbolica, si ha una *rete neurale multi-layer*, con uno strato di neuroni "nascosti", in cui il numero di neuroni ed il valore dei *pesi* ad essi associati sono determinati automaticamente dal numero di support vector e dai loro valori.

5 SVM per problemi di regressione

In un problema di regressione, ogni osservazione del training set è costituita da una coppia (x^i, y^i), in cui il pattern $x^i \in R^n$ e l'etichetta $y^i \in R$. Si consideri un modello ingresso-uscita lineare, cioè una funzione $f : R^n \to R$ della forma

$$f(x; w, b) = w^T x + b.$$

Sia $\epsilon > 0$ il *grado di precisione desiderato* con il quale si vuole approssimare la funzione rappresentata dai campioni del training set, mediante il modello f. La stima del modello relativa ad un pattern x^i è considerata "corretta" se risulta

$$|y^i - w^T x^i - b| \leq \epsilon.$$

Introduciamo la *loss function*

$$|y - f(x; w, b)|_\epsilon = \max\{0, |y - f(x; w, b)| - \epsilon\} \tag{54}$$

e definiamo l'errore di training

$$E = \sum_{i=1}^{l} |y^i - f(x^i; w, b)|_\epsilon.$$

L'errore di training è nullo se e solo se il seguente sistema di disequazioni è soddisfatto

$$\begin{aligned} w^T x^i + b - y^i &\leq \epsilon \\ y^i - w^T x^i - b &\leq \epsilon \\ i &= 1, \ldots, l. \end{aligned} \tag{55}$$

Si introducano in (55) le variabili artificiali ξ^i, $\hat{\xi}^i$, con $i = 1, \ldots, l$:

$$\begin{aligned} & w^T x^i + b - y^i \leq \epsilon + \xi^i \\ & y^i - w^T x^i - b \leq \epsilon + \hat{\xi}^i \\ & \xi^i, \hat{\xi}^i \geq 0, \qquad\qquad i = 1, \ldots, l. \end{aligned}$$

Si noti che la quantità

$$\sum_{i=1}^{l} \left(\xi^i + \hat{\xi}^i \right)$$

costituisce un upper bound dell'errore di training. Si può procedere perciò come nel caso di SVM lineari per problemi di classificazione, e considerare quindi il problema

$$\begin{aligned} \min_{w,b,\xi,\hat{\xi}} \quad & \tfrac{1}{2}\|w\|^2 + C \sum_{i=1}^{l} \left(\xi^i + \hat{\xi}^i \right) \\ & w^T x^i + b - y^i \leq \epsilon + \xi^i \\ & y^i - w^T x^i - b \leq \epsilon + \hat{\xi}^i \\ & \xi^i, \hat{\xi}^i \geq 0, \qquad\qquad i = 1, \ldots l. \end{aligned} \tag{56}$$

Il duale del problema (56) è il seguente problema

$$\begin{aligned}
\max\ L(w,b,\xi,\hat{\xi},\lambda,\hat{\lambda},\mu,\hat{\mu}) = \quad & \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l}\left(\xi_i+\hat{\xi}_i\right) - \sum_{i=1}^{l}\mu_i\xi_i - \sum_{i=1}^{l}\hat{\mu}_i\hat{\xi}_i \\
& + \sum_{i=1}^{l}\lambda_i[w^T x^i + b - y^i - \epsilon - \xi_i] - \sum_{i=1}^{l}\hat{\lambda}_i[y^i - w^T x^i - b - \epsilon - \hat{\xi}_i] \\
t.c. \quad & \nabla_w L = 0 \\
& \frac{\partial L}{\partial b} = 0 \\
& \frac{\partial L}{\partial \xi_i} = 0 \qquad i = 1,\ldots,l \\
& \frac{\partial L}{\partial \hat{\xi}_i} = 0 \qquad i = 1,\ldots,l \\
& \lambda, \hat{\lambda} \geq 0 \\
& \mu, \hat{\mu} \geq 0,
\end{aligned}$$

ossia

$$\begin{aligned}
\max\ L(w,b,\xi,\hat{\xi},\lambda,\hat{\lambda},\mu,\hat{\mu}) = \quad & \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l}\left(\xi_i+\hat{\xi}_i\right) - \sum_{i=1}^{l}\mu_i\xi_i - \sum_{i=1}^{l}\hat{\mu}_i\hat{\xi}_i \\
& + \sum_{i=1}^{l}\lambda_i[w^T x^i + b - y^i - \epsilon - \xi_i] - \sum_{i=1}^{l}\hat{\lambda}_i[y^i - w^T x^i - b - \epsilon - \hat{\xi}_i] \\
t.c. \qquad & w = \sum_{i=1}^{l}\left(\lambda_i - \hat{\lambda}_i\right)x^i \\
& \sum_{i=1}^{l}\left(\lambda_i - \hat{\lambda}_i\right) = 0 \\
& C - \lambda_i - \mu_i = 0 \qquad i = 1,\ldots,l \\
& C - \hat{\lambda}_i - \hat{\mu}_i = 0 \qquad i = 1,\ldots,l \\
& \lambda, \hat{\lambda} \geq 0 \\
& \mu, \hat{\mu} \geq 0,
\end{aligned}$$

che può essere scritto nella forma

$$\begin{aligned}
\min_{\lambda,\hat{\lambda}}\ \Gamma(\lambda,\hat{\lambda}) = \quad & \tfrac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}(\hat{\lambda}_i - \lambda_i)(\hat{\lambda}_j - \lambda_j)(x^i)^T x^j - \sum_{i=1}^{l}\left(\hat{\lambda}_i - \lambda_i\right)y^i + \epsilon\sum_{i=1}^{l}\left(\hat{\lambda}_i + \lambda_i\right) \\
& \sum_{i=1}^{l}\left(\hat{\lambda}_i - \lambda_i\right) = 0 \\
& 0 \leq \lambda \leq C \qquad i = 1,\ldots,l \\
& 0 \leq \hat{\lambda} \leq C \qquad i = 1,\ldots,l.
\end{aligned}$$

La teoria della dualità e l'impiego di funzioni kernel consentono di generalizzare la trattazione al caso di modelli di regressione non lineari, in analogia con quanto fatto in problemi di classificazione. In particolare, il problema di addestramento di una SVM per la regressione non lineare risulta definito come segue

$$\min_{\lambda,\hat{\lambda}} \Gamma(\lambda,\hat{\lambda}) = \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}(\hat{\lambda}_i-\lambda_i)(\hat{\lambda}_j-\lambda_j)k(x^i,x^j) - \sum_{i=1}^{l}\left(\hat{\lambda}_i-\lambda_i\right)y^i + \epsilon\sum_{i=1}^{l}\left(\hat{\lambda}_i+\lambda_i\right)$$

$$\sum_{i=1}^{l}\left(\hat{\lambda}_i-\lambda_i\right)=0$$

$$\begin{aligned} &0\le\lambda\le C \qquad i=1,\ldots,l\\ &0\le\hat{\lambda}\le C \qquad i=1,\ldots,l, \end{aligned}$$

dove $k(x,z)$ è una funzione kernel. Il problema formulato è un problema di programmazione quadratica convessa, la cui soluzione $(\lambda^\star,\hat{\lambda}^\star)$ permette di definire la funzione di regressione nella forma

$$f(x)=\sum_{i=1}^{l}\left(\hat{\lambda}_i^\star-\lambda_i^\star\right)k(x,x^i)+b^\star,$$

in cui $b^\star$ può essere determinato utilizzando le condizioni di complementarità. Le strutture dei problemi di addestramento di SVM per la regressione e per la classificazione sono simili. Tuttavia, in pratica l'addestramento di SVM per problemi di regressione è più complesso dell'addestramento di SVM per la classificazione, a causa del fatto che occorre determinare simultaneamente i valori "ottimali" di due parametri, cioè di ϵ e di C. In genere il *tuning* di tali parametri viene effettuato con tecniche di cross-validation.

6 Appendice A: richiami di ottimizzazione

Combinazioni affini di vettori

Definizione 5 *Si definisce sottospazio affine di uno spazio lineare V la traslazione di un sottospazio lineare $W \subseteq V$, ossia un insieme del tipo*

$$\begin{aligned} S &= \{x_0\} + W \\ &= \{x = x_0 + w, w \in W\}. \end{aligned}$$

Definizione 6 *Siano $x_1, \ldots, x_m$ elementi di V. L'elemento di V definito da*

$$x = \sum_{i=1}^{m} \alpha_i x_i$$

tale che

$$\sum_{i=1}^{m} \alpha_i = 1$$

si dice combinazione affine di $x_1, \ldots, x_m$.

Definizione 7 *Siano $x_1, \ldots, x_m$ elementi di V. Si dice che $x_1, \ldots, x_m$ sono affinemente dipendenti se esistono scalari $\alpha_1, \ldots, \alpha_m$ non tutti nulli tali che*

$$\sum_{i=1}^{m} \alpha_i x_i = 0 \qquad \sum_{i=1}^{m} \alpha_i = 0.$$

Vale la seguente proprietà.

I vettori $x_1, \ldots, x_m$ sono affinemente dipendenti (indipendenti) se e solo se i vettori $x_i - x_j$, $i = 1, \ldots, m$, $i \neq j$ sono linearmente dipendenti (indipendenti).

Direzioni di discesa e direzioni ammissibili

Si consideri il problema

$$\begin{aligned} &\min\ f(x) \\ &x \in S \end{aligned}$$

dove $f : R^n \to R$ e $S \subseteq R^n$.

Definizione 8 *Si dice che un vettore $d \in R^n$, $d \neq 0$ è una direzione di discesa per f in x se esiste $\bar{t} > 0$ tale che*

$$f(x + td) < f(x) \qquad \text{per ogni } t \in (0, \bar{t}].$$

Definizione 9 *Sia $x \in S$. Si dice che un vettore $d \in R^n$, $d \neq 0$ è una direzione ammissibile per S in x se esiste $\bar{t} > 0$ tale che*

$$x + td \in S \qquad \text{per ogni } t \in [0, \bar{t}].$$

Proposizione 6 *Supponiamo che f sia continuamente differenziabile nell'intorno di un punto $x \in R^n$ e sia $d \in R^n$ un vettore non nullo. Allora se risulta*

$$\nabla f(x)^T d < 0,$$

la direzione d è una direzione di discesa per f in x.

Proposizione 7 *Supponiamo che f sia una funzione convessa continuamente differenziabile nell'intorno di un punto $x \in R^n$ e sia $d \in R^n$ un vettore non nullo. Allora la direzione d è una direzione di discesa per f in x se e solo se*

$$\nabla f(x)^T d < 0.$$

Condizioni di Karush-Kuhn-Tucker

Si consideri il problema

$$\begin{array}{ll} \min\ f(x) & \\ & g_i(x) \leq 0 \qquad i = 1, \ldots, m \\ & h_j(x) = 0 \qquad j = 1, \ldots, p \end{array} \tag{57}$$

dove $f : R^n \to R$, $g : R^n \to R^m$, $h : R^n \to R^p$ sono funzioni continuamente differenziabili. Si definisce *funzione lagrangiana* $L : R^{n+m+p} \to R$ la funzione

$$L(x, \lambda, \mu) = f(x) + \sum_{i=1}^{m} \lambda_i g_i(x) + \sum_{j=1}^{p} \mu_j h_j.$$

Sia $\bar{x}$ un punto ammissibile per il problema (57). Una coppia di vettori $(\bar{\lambda}, \bar{\mu}) \in R^m \times R^p$, costituisce una coppia di *moltiplicatori di lagrange* se soddisfa le seguenti condizioni note come *condizioni di Karush-Kuhn-Tucker* (KKT)

$$\nabla_x L(\bar{x}, \bar{\lambda}, \bar{\mu}) = 0$$

$$\bar{\lambda}^T g(\bar{x}) = 0$$

$$\bar{\lambda} \geq 0.$$

Si consideri il problema

$$\begin{array}{ll} \min\ f(x) & \\ & a_i^T x - b_i \leq 0 \qquad i = 1, \ldots, m \\ & c_j^T x - d_j = 0 \qquad j = 1, \ldots, p. \end{array} \tag{58}$$

Proposizione 8 *Sia $x^\star$ un punto di minimo locale del problema* (58)*. Allora $x^\star$ è un punto ammissibile ed esiste una coppia di moltiplicatori di lagrange, cioè una coppia di vettori $(\lambda^\star, \mu^\star) \in R^m \times R^p$ tali che*

$$\begin{array}{l} \nabla_x L(x^\star, \lambda^\star, \mu^\star) = 0 \\ \\ \lambda_i^\star \geq 0 \qquad \lambda_i^\star (a_i^T x^\star - b_i) = 0 \quad i = 1, \ldots, r. \end{array} \tag{59}$$

Proposizione 9 *Si assuma che la funzione f sia convessa. Condizione necessaria e sufficiente affinchè $x^\star$ sia un punto di minimo globale del problema* (58) *è che $x^\star$ sia ammissibile e che esista una coppia di moltiplicatori di lagrange, cioè una coppia di vettori $(\lambda^\star, \mu^\star) \in R^m \times R^p$ tali che*

$$\begin{aligned} &\nabla_x L(x^\star, \lambda^\star, \mu^\star) = 0 \\ &\lambda_i^\star \geq 0 \qquad \lambda_i^\star (a_i^T x^\star - b_i) = 0 \quad i = 1, \ldots, r. \end{aligned} \tag{60}$$

Distanza punto-iperpiano

Sia $S \subseteq R^n$ un insieme convesso chiuso non vuoto, sia $x \in R^n$ un punto assegnato e sia $\|\cdot\|$ la norma euclidea. Si consideri il problema

$$\min\{\|x - y\| \quad y \in S\}, \tag{61}$$

ossia il problema di determinare un punto di S a distanza minima da x.

Proposizione 10 *Sia $S \subseteq R^n$ un insieme non vuoto, chiuso e convesso. Per ogni $x \in R^n$ esiste un unico vettore $y^\star \in S$ soluzione del problema* (61)*.*

Sia $\bar{x} \in R^n$ un punto assegnato e si consideri l'iperpiano $H = \{x \in R^n : w^T x + b = 0\}$. La distanza di $\bar{x}$ dall'insieme convesso H, cioè la distanza di $\bar{x}$ dalla sua proiezione su H, è indicata con $d(\bar{x}, H)$ e risulta

$$d(\bar{x}, H) = \frac{|w^T \bar{x} + b|}{\|w\|}.$$

Infatti, la proiezione di $\bar{x}$ su H è la soluzione del seguente problema

$$\min_y \tfrac{1}{2}\|\bar{x} - y\|^2$$

$$w^T y + b = 0$$

La funzione lagrangiana assume l'espressione

$$L(y, \lambda) = \frac{1}{2}\|\bar{x} - y\|^2 + \lambda(w^T y + b),$$

e le condizioni KKT (in questo caso necessarie e sufficienti) possone essere scritte come segue

$$\nabla_y L(y^\star, \lambda^\star) = -\bar{x} + y^\star + \lambda w = 0$$

$$w^T y^\star + b = 0,$$

dalle quali si ricava

$$\lambda^\star = -\frac{w^T \bar{x} + b}{\|w\|^2} \qquad \|\bar{x} - y^\star\| = |\lambda^\star| \|w\| = \frac{|w^T \bar{x} + b|}{\|w\|}.$$

Programmazione convessa

Proposizione 11 *Sia S un sottoinsieme convesso di R^n e sia $\bar{x}$ un qualsiasi punto di S. Allora, se $S \neq \{\bar{x}\}$, comunque si fissi $x \in S$ tale che $x \neq \bar{x}$, la direzione $d = x - \bar{x}$ è una direzione ammissibile per S in $\bar{x}$.*

Proposizione 12 *Sia S un sottoinsieme convesso di R^n e supponiamo che f sia una funzione convessa continuamente differenziabile su un insieme aperto contenente S. Allora $x^\star \in S$ è un punto di minimo globale del problema* $\min\{f(x),\ x \in S\}$ *se e solo se*

$$\nabla f(x^\star)^T d \geq 0 \qquad \text{per ogni direzione } d \text{ ammissibile in } x^\star.$$

Proposizione 13 *Sia S un sottoinsieme convesso di R^n e supponiamo che f sia una funzione convessa continuamente differenziabile su un insieme aperto contenente S. Allora $x^\star \in S$ è un punto di minimo globale del problema* $\min\{f(x),\ x \in S\}$ *se e solo se*

$$\nabla f(x^\star)^T (x - x^\star) \geq 0 \qquad \text{per ogni } x \in S.$$

Riferimenti bibliografici

[1] Burges, C.J.C. (1998), "A tutorial on support vector machines for pattern recognition," *Data Mining and Knowledge Discovery*, vol. 2, pp. 121-167.

[2] Vapnik, V.N. (1995), *The Nature of Statistical Learning Theory*, Springer-Verlag, New York.

[3] Vapnik, V.N. (1998), *Statistical Learning Theory*, Wiley, New York.