Classificazione Binaria e Support Vector Machines

A. Astorino*

1 Introduzione

Ci sono molti problemi, nei campi piú svariati, che possono essere formulati e risolti efficacemente come programmi matematici. Secondo Leonhard Euler, il piú grande matematico del diciottesimo secolo:

"Nulla accade nell'universo che non possa essere ricondotto ad un problema di massimo o di minimo"

L'obiettivo di questo documento é quello di analizzare alcuni approcci di programmazione matematica proposti per risolvere problemi di classificazione.

In particolare, i problemi di classificazione presi in considerazione sono quelli relativi alla determinazione di classificatori binari. In tale contesto il problema consiste principalmente nel trovare un criterio per distinguere gli elementi di due insiemi disgiunti di punti campione (o "pattern" o "training data"). Poiché i pattern sono generalmente rappresentati da punti nello spazio $\mathbb{R}^n$, il problema diventa quello di discriminare tra due insiemi finiti di punti, $\mathcal{A}$ e $\mathcal{B}$, nello spazio n-dimensionale, $\mathbb{R}^n$, attraverso un iperpiano o una superficie non lineare di separazione.

Si consideri un semplice esempio:

Esempio 1 *Dato il peso e l'altezza di una persona, si vuole trovare un criterio per determinarne il sesso.*

Si supponga, quindi, di avere a disposizione un insieme di campioni di cui si conosce peso, altezza e sesso. L'idea é quella di trovare nello spazio

*Istituto di Calcolo e Reti ad Alte Prestazioni C.N.R., c/o Dipartimento di Elettronica Informatica e Sistemistica, Università della Calabria, 87036 Rende (CS), Italia. E-mail: astorino@icar.cnr.it

$\mathbb{R}^2$ del peso e dell'altezza un iperpiano o, piú in generale, una superfice di separazione tra i campioni di sesso femminile e quelli di sesso maschile. Tale superfice verrá usata per inferire il sesso di altre persone, non appartenenti, quindi, all'insieme di campioni utilizzati, a partire dal loro peso e dalla loro altezza.

Ci sono molti settori della ricerca scientifica, quali la statistica e l'informatica attraverso le basi di dati, il "machine learning", il "pattern recognition", l'intelligenza artificiale e la visualizzazione dei dati, che si sono occupati di problemi di classificazione, proponendo per essi varie metodologie risolutive.

1.1 Machine Learning

Negi ultimi decenni vi é stato un notevole lavoro di ricerca in diverse discipline correlate, che vanno dalla psicologia all'intelligenza artificiale, con lo scopo di sviluppare modelli generali della conoscenza ed in particolare dell'apprendimento (learning). Questo sforzo ha prodotto un certo numero di tecnologie interessanti, in particolare le tecniche per la rappresentazione della conoscenza e di apprendimento hanno permesso la realizzazione di importanti applicazioni per la rappresentazione e l'analisi dei dati.

Esistono due tecniche principali di learning il *supervised* e l'*unsupervied learning*.

Nel supervised learning esiste un insegnante esterno che definisce le classi e fornisce gli esempi di ogni classe al sistema cognitivo. Il compito del sistema, quindi, consiste nello scoprire proprieá comuni negli esempi di ogni classe, cioé deve fornire una *descrizione* della classe. Questa tecnica é anche nota come *apprendimento da esempi*. Una classe, insieme alla sua descrizione, forma una *regola di classificazione* del tipo **if** *description* **then** *class* che puó essere usata per predire la classe di un nuovo oggetto.

Nell'*unsupervised learning* é il sistema che deve scoprire le classi stesse, in base a proprietá comuni degli oggetti, senza l'aiuto di un insegnante. Il sistema deve osservare gli esempi e riconoscere le relazioni tra di essi. Questa tecnica é pertanto nota come *apprendimento da osservazione e scoperta*. Il risultato di questo processo é un insieme di descrizioni di classi, una per ogni classe scoperta, che insieme coprono tutti gli oggetti dell'ambiente. Queste descrizioni formano un sommario ad alto livello degli oggetti nell'ambiente.

Il processo di costruzione di una descrizione di una classe di oggetti é un processo iterativo, in cui, fra tutte le possibili descrizioni, si cerca la migliore. Una ipotesi, cioé una descrizione iniziale, viene formulata e valutata in base a qualche funzione di qualitá. Quindi questa ipotesi é accettata, rifiutata o

migliorata finché non si trova una ipotesi corretta. Questo processo lavora su un insieme di esempi detto *insieme di addestramento (training set)* e non su tutto l'ambiente.

2 Separabilitá Lineare

2.1 Introduzione

In un problema di separabilitá lineare, l'obiettivo é quello di discriminare tra due insiemi finiti di punti, $\mathcal{A}$ e $\mathcal{B}$, nello spazio n-dimensionale, $\mathbb{R}^n$, utilizzando come superficie di separazione un iperpiano.

Siano

- l'insieme $\mathcal{A}$ costituito da m punti, $a_i \in \mathbb{R}^n$, $i = 1, ..., m$ e rappresentato dalla matrice $A \in \mathbb{R}^{m \times n}$, con $A_i = a_i^T$ i-esima riga;
- l'insieme $\mathcal{B}$ costituito da k punti, $b_l \in \mathbb{R}^n$, $l = 1, ..., k$ e rappresentato dalla matrice $B \in \mathbb{R}^{k \times n}$, con $B_l = b_l^T$ l-esima riga

con $\mathcal{A} \cap \mathcal{B} = \emptyset$. L'iperpiano

$$P \stackrel{\triangle}{=} \{x \in \mathbb{R}^n \mid \ w^T x = \xi\}$$

rappresenta un *iperpiano di separazione* tra gli insiemi $\mathcal{A}$ e $\mathcal{B}$ se

$$A_i w > \xi \quad \forall i = 1, ...m;$$

e

$$B_l w < \xi \quad \forall l = 1, ..., k.$$

Vale, quindi, la seguente definizione:

Definizione 2.1 *Gli insiemi $\mathcal{A}$ e $\mathcal{B}$ sono* linearmente separabili *se e solo se il sistema*

$$\begin{cases} Aw - e\xi & > 0 \\ -Bw + e\xi & > 0 \end{cases} \tag{1}$$

ha soluzione $w \in \mathbb{R}^n$ e $\xi \in \mathbb{R}$.

La condizione di separabilitá lineare implica che l'intersezione delle coperture convesse dei due insiemi sia vuota:

$$conv(\mathcal{A}) \cap conv(\mathcal{B}) = \emptyset.$$

Vale, a conferma di ció, il seguente teorema:

Teorema 2.1 *Gli insiemi $\mathcal{A}$ e $\mathcal{B}$ sono* linearmente inseparabili *se e solo se il sistema*

$$\begin{cases} A^T u - & B^T t & = 0 \\ e^T u & & = 1 \\ & e^T t & = 1 \\ u \geq 0 & t \geq 0 & \end{cases} \tag{2}$$

con $u \in \mathbb{R}^m$ e $t \in \mathbb{R}^k$, ha soluzione.

Dim. Dalla definizione (2.1) gli insiemi $\mathcal{A}$ e $\mathcal{B}$ sono linearmente inseparabili se e solo se il sistema

$$\begin{pmatrix} A & -e \\ -B & e \end{pmatrix} \begin{pmatrix} w \\ \xi \end{pmatrix} > 0$$

non ha soluzione. Per il teorema delle alternative di Gordan, ha, quindi, soluzione il sistema:

$$\begin{pmatrix} A^T & -B^T \\ -e^T & e^T \end{pmatrix} \begin{pmatrix} y^{(1)} \\ y^{(2)} \end{pmatrix} = 0 \tag{3}$$

con

$$\begin{pmatrix} y^{(1)} \\ y^{(2)} \end{pmatrix} \geq 0$$

e

$$\begin{pmatrix} y^{(1)} \\ y^{(2)} \end{pmatrix} \neq 0,$$

che puó essere riscritto nel seguente modo:

$$\begin{cases} A^T y^{(1)} - B^T y^{(2)} = 0 \\ -e^T y^{(1)} + e^T y^{(2)} = 0 \\ y^{(1)} \geq 0,\ y^{(2)} \geq 0 \\ \begin{pmatrix} y^{(1)} \\ y^{(2)} \end{pmatrix} \neq 0 \end{cases}$$

Poiché $e^T y^{(1)} = e^T y^{(2)}$ e $y^{(1)} \geq 0$, $y^{(2)} \geq 0$, la condizione

$$\begin{pmatrix} y^{(1)} \\ y^{(2)} \end{pmatrix} \neq 0$$

implica

$$y^{(1)} \neq 0 \quad \text{e} \quad y^{(2)} \neq 0.$$

Ponendo

$$e^T y^{(1)} = e^T y^{(2)} = M \neq 0$$

e

$$u = \frac{y^{(1)}}{M} \quad \text{e} \quad t = \frac{y^{(2)}}{M}$$

si ottiene proprio una soluzione del sistema (2). □

Poiché nella definizione (2.1) si hanno vincoli di disuguaglianza stretta e la programmazione lineare non é in grado di trattarli, questi vengono normalizzati dividendo le variabili (w, ξ) per la quantitá positiva

$$\min_{\substack{i=1,...,m \\ l=1,...,k}} \{A_i w - \xi, -B_l w + \xi\}.$$

Si ottiene, cosí, la seguente definizione equivalente a (2.1):

Definizione 2.2 *Gli insiemi $\mathcal{A}$ e $\mathcal{B}$ sono* linearmente separabili *se e solo se il sistema*

$$\begin{cases} Av \geq e(\gamma + 1) \\ Bv \leq e(\gamma - 1) \end{cases} \tag{4}$$

ha soluzione $v \in \mathbb{R}^n$ e $\gamma \in \mathbb{R}$.

Si noti che quando gli insiemi $\mathcal{A}$ e $\mathcal{B}$ sono linearmente separabili secondo la definizione (2.2), l'iperpiano

$$\{x \in \mathbb{R}^n \mid v^T x = \gamma\}$$

rappresenta un iperpiano di separazione stretta, infatti risulta:

$$\begin{cases} Av > e\gamma \\ Bv < e\gamma. \end{cases}$$

Il problema della separabilitá lineare di due insiemi di punti $\mathcal{A}$ e $\mathcal{B}$ é stato affrontato secondo diversi approcci:

- Programmazione Matematica
- Support Vector Machines (SVM)
- Reti neurali

Nella sezione successiva verr á analizzato l'approccio di programmazione matematica relativo alle SVM.

3 SVM e problemi di Separazione Lineare

La Support Vector Machine (SVM) é una metodologia di programmazione matematica per risolvere problemi di classificazione.

Gli insiemi di punti $\{a_i\}$, $i = 1, ..., m$, di $\mathcal{A}$ e $\{b_l\}$, $l = 1, ..., k$, di $\mathcal{B}$, possono anche essere rappresentati nel seguente modo:

$$\{x_i, y_i\}, \quad i = 1, ..., m + k$$

con $x_i \in \mathbb{R}^n$ e $y_i \in \{-1, 1\}$ tali che

$$\begin{cases} x_i \in \mathcal{A} & \text{se } y_i = 1 \\ x_i \in \mathcal{B} & \text{se } y_i = -1. \end{cases}$$

Un iperpiano di separazione, $v^T x - \gamma = 0$, per tali dati soddisfa le condizioni

$$v^T x_i - \gamma \geq +1 \text{ se } y_i = 1 \tag{5}$$

e

$$v^T x_i - \gamma \leq -1 \text{ se } y_i = -1 \tag{6}$$

che possono ancora scriversi come

$$y_i[v^T x_i - \gamma] \geq 1, \quad i = 1, ..., m + k. \tag{7}$$

Tale iperpiano di separazione, tuttavia, puó, in generale, non essere "buono", nel senso che alcuni punti di $\mathcal{A}$ e/o $\mathcal{B}$ possono essere molto vicini all'iperpiano stesso. É quindi preferibile trovare un iperpiano con un largo margine di separazione o addirittura con il piú largo margine possibile. L'iperpiano con il margine massimo é noto come *iperpiano di separazione ottimale*

3.1 Insiemi linearmente separabili

Dati gli insiemi $\mathcal{A}$ e $\mathcal{B}$, linearmente separabili, rappresentati dai campioni $\{x_i, y_i\}$, $i = 1, ..., m + k$, con $x_i \in \mathbb{R}^n$ e $y_i \in \{-1, 1\}$, si supponga di poter individuare un iperpiano di separazione $v^T x - \gamma = 0$, dove $v \in \mathbb{R}^n$ é normale all'iperpiano, $|\gamma|/\|v\|$ é la distanza perpendicolare dall'iperpiano all'origine e $\|.\|$ é la norma Euclidea. Si indichi con d_+ e d_- la piú piccola distanza dall'iperpiano di separazione dei punti di $\mathcal{A}$ e $\mathcal{B}$, rispettivamente, ad esso piú vicini.

Per il caso di insiemi linearmente separabili, l'approccio SVM, cerca semplicemente l'iperpiano di separazione con il margine piú grande, dove il margine é dato dalla somma di d_+ e d_- ed é pari a $2/\|v\|$.

L'obiettivo é, quindi, quello di trovare la coppia di iperpiani che danno il massimo margine e quindi minimizzano $\|v\|^2$ sotto i vincoli (7):

$$\begin{array}{ll} \min\limits_{v,\gamma} & \frac{1}{2}\|v\|^2 \\ \text{s.v.} & y_i[v^T x_i - \gamma] - 1 \geq 0 \quad \forall i = 1, ..., m+k. \end{array} \tag{8}$$

I punti per i quali i vincoli sono soddisfatti per uguaglianza, la cui rimozione cambierebbe la soluzione trovata, sono chiamati *vettori di supporto* (support vector).

Viene ora data una formulazione lagrangiana del problema per i seguenti motivi:

1. i vincoli di (8) saranno sostituiti da vincoli sui moltiplicatori di Lagrange piú facili da trattare;

2. nella riformulazione Lagrangiana del problema i dati appariranno solo sotto forma di prodotti scalari tra vettori.

Quest'ultima é una proprietá importante che consente la generalizzazione della procedura proposta al caso non lineare.

Si introducono, quindi, i moltiplicatori di Lagrange (non negativi):

$$\alpha_i \qquad \text{per} \quad i = 1, ..., m+k.$$

Si ottiene la funzione Lagrangiana:

$$L_P \equiv \frac{1}{2}\|v\|^2 - \sum_{1=1}^{m+k} \alpha_i y_i (v^T x_i - \gamma) + \sum_{i=1}^{m+k} \alpha_i. \tag{9}$$

Essendo il problema (8) un problema di programmazione quadratica convesso, é possibile risolvere equivalentemente il seguente problema duale: *massimizzare* L_P soggetto ai vincoli che il gradiente di L_P rispetto a v e γ sia pari al vettore nullo

$$v = \sum_{i=1}^{m+k} \alpha_i y_i x_i \tag{10}$$

$$\sum_{i=1}^{m+k} \alpha_i y_i = 0 \tag{11}$$

e che i moltiplicatori α_i siano non negativi

$$\alpha_i \geq 0.$$

Andando a sostituire i vincoli di uguaglianza appena scritti nell'espressione della funzione obiettivo L_P si ottiene il seguente problema:

$$\begin{array}{ll} \max\limits_{\alpha} & L_D = \sum\limits_{i=1}^{m+k} \alpha_i - \frac{1}{2} \sum\limits_{i=1}^{m+k} \sum\limits_{j=1}^{m+k} \alpha_i \alpha_j y_i y_j x_i^T x_j \\ \text{s.v.} & \sum\limits_{i=1}^{m+k} \alpha_i y_i = 0 \\ & \alpha_i \geq 0. \end{array} \tag{12}$$

Per insiemi linearmente separabili, quindi, il training (addestramento) in termini di SVM consiste nel risolvere il problema (12) con soluzione v ottenuta dal vincolo di uguaglianza (10). Quei punti campione per i quali nella soluzione si ha $\alpha_i > 0$ sono chiamati vettori di supporto. A tutti gli altri punti corrispondono moltiplicatori $\alpha_i = 0$ che non danno contributo nella determinazione di v. Per queste macchine (SVM) i vettori di supporto sono gli elementi critici del training set, sono, infatti, i punti piú vicini al confine di decisione e sono quelli che determinano l'iperpiano di separazione ottimale.

3.1.1 Le condizioni di Karush-Kuhn-Tucker

Le condizioni di Karush-Kuhn-Tucker (KKT) per il problema primale (8) sono:

$$\begin{aligned} \frac{\partial}{\partial v_j} L_P = v_j - \sum_{i=1}^{m+k} \alpha_i y_i x_{ij} &= 0 \quad j = 1, ..., n \\ \frac{\partial}{\partial \gamma} L_P = \sum_{i=1}^{m+k} \alpha_i y_i &= 0 \\ y_i(v^T x_i - \gamma) - 1 &\geq 0 \quad i = 1, ..., m+k \\ \alpha_i &\geq 0 \quad i = 1, ..., m+k \\ \alpha_i[y_i(v^T x_i - \gamma) - 1] &= 0 \quad i = 1, ..., m+k. \end{aligned}$$

Tali condizioni nel caso in analisi sono necessarie e sufficienti affinché v, γ e α sia soluzione del problema SVM (8) - (12). Risolvere, quindi, il problema SVM é equivalente a trovare una soluzione alle condizioni di KKT.

Come immediata applicazione si riesce a determinare esplicitamente v dalla procedura di training (condizione (10)), mentre la soglia γ si puó calcolare facilmente dalla condizione di KKT di complementarietá scegliendo qualche punto x_i per il quale $\alpha_i > 0$ (sarebbe, comunque, numericamente piú sicuro prendere il valore medio di γ risultante da tutte queste equazioni).

3.1.2 La fase di test

Una volta addestrata la SVM, dato un nuovo punto di test x, si determina da quale parte del confine di decisione giace il punto stesso e gli si assegna l'etichetta della classe corrispondente, ossia la classe di x sará:

$$sgn(v^T x - \gamma).$$

3.2 Insiemi non linearmente separabili

L'approccio ora proposto vale per insiemi lineramente separabili. Un modo per estendere le idee appena viste al caso non separabile potrebbe essere quello di rilassare i vincoli di separabilitá (5) e (6), ma solo quando necessario, aggiungendo, quindi, un ulteriore costo nella funzione obiettivo di (8). Ció puó essere fatto introducendo delle variabili slack positive ξ_i, $i = 1, ..., m + k$, nei vincoli che allora diventano:

$$v^T x_i - \gamma \geq +1 - \xi_i \quad \text{per } y_i = +1$$

$$v^T x_i - \gamma \leq -1 + \xi_i \quad \text{per } y_i = -1$$

$$\xi_i \geq 0 \quad \forall i = 1, ..., m + k.$$

Quando si ha un errore (punto mal classificato), ξ_i risulta maggiore o uguale di uno e, cosí, $\sum_{i=1}^{m+k} \xi_i$ rappresenta un upper bound sul numero di punti mal classificati. Un modo naturale per assegnare un costo aggiuntivo agli errori commessi da un iperpiano di separazione é quello di traformare la funzione obiettivo da minimizzare da

$$\frac{1}{2}\|v\|^2$$

in

$$\frac{1}{2}\|v\|^2 + C(\sum_{i=1}^{m+k} \xi_i)^q,$$

dove C é un parametro positivo da fissare (piú é grande C piú é grande la penalitá che si assegna agli errori di classificazione). Per qualsiasi scelta di q intero positivo, il problema

$$\begin{array}{lll} \min\limits_{v,\gamma,\xi} & \frac{1}{2}\|v\|^2 + C(\sum\limits_{1}^{m+k} \xi_i)^q & \\ \text{s.v.} & y_i[v^T x_i - \gamma] - 1 + \xi_i \geq 0 & \forall i = 1, ..., m + k \\ & \xi_i \geq 0 & \forall i = 1, ..., m + k \end{array} \tag{13}$$

risulta convesso e, per $q = 2$ e $q = 1$, é anche quadratico. La scelta di $q = 1$ ha come ulteriore vantaggio il fatto che né le variabili ξ_i né i loro moltiplicatori di Lagrange appaiono nel problema duale che diventa:

$$\begin{aligned} \max_{\alpha} \quad & L_D = \sum_{i=1}^{m+k} \alpha_i - \frac{1}{2} \sum_{i=1}^{m+k} \sum_{j=1}^{m+k} \alpha_i \alpha_j y_i y_j x_i^T x_j \\ \text{s.v.} \quad & \sum_{i=1}^{m+k} \alpha_i y_i = 0 \\ & 0 \leq \alpha_i \leq C, \end{aligned} \tag{14}$$

con

$$v = \sum_{i=1}^{m+k} \alpha_i y_i x_i.$$

La sola differenza dal caso linearmente separabile risulta quindi l'upper bound C sui moltiplicatori α_i.

Il Lagrangiano del problema primale risulta essere

$$L_P = \frac{1}{2}\|v\|^2 + C \sum_{1=1}^{m+k} \xi_i - \sum_{1=1}^{m+k} \alpha_i [y_i(v^T x_i - \gamma) - 1 + \xi_i] - \sum_{i=1}^{m+k} \mu_i \xi_i$$

dove i μ_i sono i moltiplicatori di Lagrange introdotti per rafforzare la positivitá delle variabili ξ_i. Le condizioni di KKT per il problema primale sono:

$$\begin{aligned} \frac{\partial}{\partial v_j} L_P = v_j - \sum_{i=1}^{m+k} \alpha_i y_i x_{ij} = & \quad 0 \quad j = 1, ..., n \\ \frac{\partial}{\partial \gamma} L_P = \sum_{i=1}^{m+k} \alpha_i y_i = & \quad 0 \\ \frac{\partial}{\partial \xi_i} L_P = C - \alpha_i - \mu_i = & \quad 0 \quad i = 1, ..., m+k \\ y_i(v^T x_i - \gamma) - 1 + \xi_i \geq & \quad 0 \quad i = 1, ..., m+k \\ \xi_i \geq & \quad 0 \quad i = 1, ..., m+k \\ \alpha_i \geq & \quad 0 \quad i = 1, ..., m+k \\ \mu_i \geq & \quad 0 \quad i = 1, ..., m+k \\ \alpha_i [y_i(v^T x_i - \gamma) - 1 + \xi_i] = & \quad 0 \quad i = 1, ..., m+k \\ \mu_i \xi_i = & \quad 0 \quad i = 1, ..., m+k. \end{aligned}$$

Come per il caso separabile si possono usare le condizioni di complementarietá per determinare la soglia γ. Dalle condizioni di KKT appena scritte, infatti, si ha

$$\xi_i = 0 \quad \text{quando } \alpha_i < C.$$

Segue, quindi, che il calcolo di γ si puó ottenere dalla relazione di complementarietá relativa ad un qualsiasi punto campione per il quale risulta $0 < \alpha_i < C$ (come nel caso precedente sarebbe numericamente piú sicuro prendere il valore medio di γ risultante da tutte queste equazioni).

4 SVM e problemi di Separazione NonLineare

I metodi di separazione lineare basati sulla teoria delle SVM, che hanno come obiettivo quello di determinare un iperpiano di separazione ottimale, possono essere generalizzati al caso di separazione non lineare nel seguente modo: si effettua una trasformazione dello spazio degli input in uno spazio dimensionalmente maggiore, attraverso una funzione non lineare, quindi, si costruiscono iperpiani di separazione ottimali nello spazio aumentato, che corrispondono a superfici non lineari nello spazio originario degli input.

Ritornando al caso lineare, é possibile notare come nel problema di training i dati entrano in gioco solo sotto forma di prodotti scalari, $x_i^T x_j$:

$$\begin{aligned} \max_{\alpha} \quad & \sum_{i=1}^{m+k} \alpha_i - \frac{1}{2} \sum_{i=1}^{m+k} \sum_{j=1}^{m+k} \alpha_i \alpha_j y_i y_j x_i^T x_j \\ \text{s.v.} \quad & \sum_{i=1}^{m+k} \alpha_i y_i = 0 \\ & 0 \leq \alpha_i \leq C. \end{aligned}$$

Supponendo, ora, di traformare lo spazio degli input $\mathbb{R}^n$ in un generico spazio Euclideo $\mathcal{H}$, attraverso la trasformazione Φ:

$$\Phi \ : \ \mathbb{R}^n \mapsto \mathcal{H},$$

l'algoritmo di training dipenderá solo dai prodotti scalari dei dati in $\mathcal{H}$ (i trasformati dei punti campione), ossia da funzioni della forma $\Phi(x_i)^T \Phi(x_j)$. Se esiste una funzione "kernel" K, tale che

$$K(x_i, x_j) = \Phi(x_i)^T \Phi(x_j),$$

l'algoritmo di training fará uso solo di K senza avere mai la necessitá di conoscere esplicitamente Φ. Naturalmente tutte le considerazioni fatte per il caso lineare continuano a valere, in quanto si continua ancora a fare una

separazione lineare, ma in un nuovo spazio:

$$\begin{aligned} \max_{\alpha} \quad & \sum_{i=1}^{m+k} \alpha_i - \frac{1}{2} \sum_{i=1}^{m+k} \sum_{j=1}^{m+k} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \\ \text{s.v.} \quad & \sum_{i=1}^{m+k} \alpha_i y_i = 0 \\ & 0 \leq \alpha_i \leq C. \end{aligned} \tag{15}$$

Per quanto riguarda il vettore v, che caratterizza l'iperpiano di separazione, esso apparterrá allo spazio $\mathcal{H}$ e per il suo calcolo si deve conoscere esplicitamente Φ:

$$v = \sum_{i=1}^{N_S} \alpha_i y_i \Phi(x_i).$$

Si puó ,peró, osservare che nella fase di test é necessario solo calcolare i prodotti scalari di un dato punto di test x con v, o piú specificatamente calcolare il segno di $(v^T \Phi(x) - \gamma)$, si ottiene cosí

$$sgn\left(\sum_{i=1}^{m+k} \alpha_i y_i \Phi(x_i)^T \Phi(x) - \gamma\right) = sgn\left(\sum_{i=1}^{N_S} \alpha_i y_i K(x_i, x) - \gamma\right),$$

evitando, di nuovo, il calcolo di $\Phi(x)$ ed usando, invece, la funzione kernel.

Le prime funzioni kernel utilizzate per problemi di pattern recognition sono state le seguenti:

- $K(x, y) = \left(x^T y + 1\right)^p$
- $K(x, y) = \exp\left\{-\frac{\|x - y\|^2}{\sigma^2}\right\}$
- $K(x, y) = \tanh\left(b(x^T y) - c\right)$

4.1 SVM per problemi a larga scala

Per addestrare una SVM si deve, quindi, risolvere un problema di ottimizzazione quadratico (15) con vincoli bound ed un vincolo lineare di uguaglianza.

Tale problema messo sotto forma di minimizzazione diventa:

$$\begin{aligned} \min_{\alpha} \quad & -\sum_{i=1}^{m+k} \alpha_i + \frac{1}{2} \sum_{i=1}^{m+k} \sum_{j=1}^{m+k} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \\ \text{s.v.} \quad & \sum_{i=1}^{m+k} \alpha_i y_i = 0 \\ & 0 \leq \alpha_i \leq C \quad \forall i = 1, ..., m+k \end{aligned} \tag{16}$$

dove si ricorda che con $m+k$ si é indicato il numero di punti campione, che servono per l'addestramento e con α un vettore a $m+k$ componenti, una per ciascun punto (x_i, y_i).

Definendo la matrice Q come

$$(Q)_{ij} = y_i y_j K(x_i, x_j)$$

il problema (16) puó, equivalentemente, scriversi come

$$\begin{aligned} \min_{\alpha} \quad & -e^T \alpha + \frac{1}{2} \alpha^T Q \alpha \\ \text{s.v.} \quad & \alpha^T y = 0 \\ & 0 \leq \alpha \leq Ce. \end{aligned} \tag{17}$$

La dimensione di questo problema di ottimizzazione dipende dal numero di campioni, $m+k$. Se tale numero é molto elevato, poiché la dimensione della matrice Q é pari a $(m+k)^2$, sará impossibile mantenere Q in memoria. Molte implementazioni di risolutori di problemi quadratici richiedono esplicitamente l'immagazzinamento di Q, il che rende proibitiva la loro applicazione. Una alternativa per evitare tale problema, potrebbe essere quella di ricalcolare Q ogni volta che serve, ma anche questo sarebbe proibitivamente costoso qualora la matrice Q fosse necessaria molte volte.

Un approccio per rendere trattabile l'addestramento di una SVM per problemi a larga scala é quello di decomporlo in una serie di problemi piú piccoli.

4.1.1 Algoritmo "Chunking"

L'algoritmo noto come "chunking" sfrutta ancora il fatto che il valore della funzione obiettivo del problema quadratico (17) é lo stesso se si rimuovono le righe e le colonne della matrice Q che corrispondono a moltiplicatori di Lagrange con valore nullo. Il problema quadratico di partenza, quindi, viene decomposto in una serie di sotto-problemi quadratici piú piccoli che hanno

come ultimo obiettivo quello di identificare tutti i moltiplicatori di Lagrange non nulli e di scartare gli altri.

Ad ogni passo l'algoritmo risolve un problema di programmazione quadratica relativo ai punti che al passo precedente hanno avuto associati moltiplicatori di Lagrange con valore non nullo piú gli M punti "peggiori" che hanno violato le condizioni di ottimalitá di KKT, per qualche M fissato. Ciascun sottoproblema quadratico viene inizializzato con i risultati del sottoproblema precedente. All'ultimo passo l'intero insieme di moltiplicatori di Lagrange non nulli viene identificato e, quindi, si risolve il problema quadratico di partenza. Si puó notare come l'algoritmo chunking rispetta le condizioni di convergenza relative alla decomposizione di Osuna.

4.1.2 Sequential Minimal Optimization

Mentre l'algoritmo di decomposizione di Osuna puó in alcuni casi convergere molto lentamente verso la soluzione ottima, l'uso dell'algoritmo chunking, che pur riduce la dimensione della matrice Q dal quadrato del numero di punti campione al quadrato del numero di punti cui corrispondono moltiplicatori di Lagrange non nulli, puó ancora risultare proibitivo per problemi a larga scala. Viene, quindi, proposto un ulteriore algoritmo di decomposizione noto come Sequential Minimal Optimization (SMO) che continua a rispettare le condizioni di convergenza di Osuna.

A differenza degli altri metodi SMO sceglie di risolvere ad ogni passo il problema di ottimizzazione piú piccolo, ossia quello corrispondente a soli due moltiplicatori di Lagrange.

Per rispettare le condizioni di convergenza di Osuna, tale algoritmo, a ciascun passo, ottimizza e cambia i due moltiplicatori in modo tale che almeno uno dei due, prima dell'ottimizzazione, violi le condizioni di ottimalitá di KKT. In questo modo si garantisce il miglioramento progressivo della funzione obiettivo di partenza.

Per rendere veloce la convergenza verso la soluzione ottima, SMO usa appropriate euristiche per la scelta dei due moltiplicatori.

5 Possibili Applicazioni Mediche

Le applicazioni dell'ottimizzazione nel settore medico rappresentano un'area di ricerca di recente e rapido sviluppo sia nel campo della diagnosi che in quello della terapia.

In ambito diagnostico, le tecniche di ottimizzazione si sono rilevate particolarmente utili per ricavare informazioni circa la presenza o assenza di

una data patologia a partire da dati clinici non facilmente interpretabili dal medico. Sono state inoltre utilizzate per prevedere l'evoluzione di pazienti verso uno stato patologico in processi di "screening" di massa su popolazioni estese.

In ambito terapeutico le tecniche di ottimizzazione numerica sono alla base di sistemi di supporto alle decisioni del medico specialmente nella fase di realizzazione di diete speciali, perfettamente bilanciate, per patologie particolari.

Ai fini diagnostici stanno riscuotendo sempre maggiore rilevanza tutte quelle metodologie che permettono di elaborare grandi moli di dati originari e ricavare informazioni non ovvie e di cruciale importanza nell'individuazione della giusta diagnosi.

Una delle funzioni piú comuni realizzata da tali metodi ed algoritmi é la classificazione. Si ricorda che si ha una Classificazione quando vengono identificati schemi o insiemi di caratteristiche che definiscono il gruppo cui appartiene un dato elemento.

Il problema di classificazione considerato é quello di discriminare tra due insiemi finiti di punti, $\mathcal{A}$ e $\mathcal{B}$, nello spazio n-dimensionale delle caratteristiche attraverso uno o piú iperpiani di separazione che utilizzano solo alcune delle caratteristiche possibili.

5.1 Wisconsin Breast Cancer Database

L'applicazione presentata é relativa allo studio di pazienti con cancro al seno. In particolare il problema studiato é quello di stabilire la malignitá o benignitá del tumore attraverso l'analisi di immagini di cellule ed utilizzando, nello spazio dei parametri (dati clinici), tecniche di classificazione.

Generalmente tali tipi di tumori sono scoperti dai pazienti come un nodulo al seno. La maggior parte dei noduli sono benigni, pertanto, é responsabilitá del medico, diagnosticare il male, ossia distinguere tra i tanti noduli quello maligno.

Ci sono tre metodi disponibili per diagnosticare il male: mammografia, FNA con interpretazione visuale e biopsia chirurgica. L'abilitá di diagnosticare il male quando esso é presente varia dal 68% al 79% per la mammografia, dal 65% al 98% per FNA e vicino al 100% per la biopsia chirurgica. Tuttavia, nonostante la biopsia sia il metodo piú accurato, essa é invasiva e costosa sia in termini temporali che di risorse mediche. Il giusto compromesso tra la mammografia e la biopsia é, quindi, l'esame FNA, anche noto come ago aspirato sottile.

I campioni del WBCD (Wisconsin Breast Cancer Database) sono stati

raccolti periodicamente dal Dr. William H. Wolberg dell'ospedale dell'Universitá del Wisconsin, Madison, e si riferiscono a casi clinici in osservazione presso il suo ospedale. Tali campioni consistono in caratteristiche delle cellule valutate visivamente e tratte dai noduli anomali del seno del paziente mediante asportazione da essi di una piccola quantitá di fluido per mezzo dell'esame FNA.

Ciascun paziente é stato cosí associato ad un vettore a 11 dimensioni: l'attributo 1 é semplicemente il codice di identificazione del paziente, mentre l'attributo 2 indica se il campione in questione é benigno o maligno; in tal senso, la benignitá/malignitá delle cellule é diagnosticata rilevando un campione di tessuto dal seno del paziente ed effettuando su di esso una biopsia. I rimanenti attributi (dal 3 al 11) si riferiscono per l'appunto ai parametri del nucleo soggetti all'interpretazione visiva da parte del dottor Wolberg: il loro valore é sempre nell'intervallo 1-10, con il valore 1 corrispondente allo stato normale e il 10 al massimo stato anormale:

1. Numero di codice
2. Classe: 2 per benignitá, 4 per malignitá
3. Spessore della massa
4. Regolaritá della dimensione della cellula
5. Regolaritá della forma della cellula
6. Aderenza marginale
7. Dimensione della singola cellula epiteliale
8. Nuclei vuoti
9. Cromatina
10. Nucleoli normali
11. Mitosi

Il database contiene 682 punti suddivisi in 8 gruppi, di diversa entitá numerica, relativi a differenti periodi di osservazione:

- Numero di parametri = 9
- Numero di campioni = 682
- $\mathcal{A}$ = Tumore maligno
- $\mathcal{B}$ = Tumore benigno

5.2 Wisconsin Breast Cancer Prognosis Database

L'applicazione presentata é relativa allo studio del decorso post-operatorio di pazienti con cancro al seno. In particolare il problema studiato é quello di stabilire la ricorrenza o non ricorrenza del tumore entro 24 mesi dall'operazione.

I campioni del WBCPD (Wisconsin Breast Cancer Prognosis Database), raccolti sempre dal Dr. William H. Wolberg dell'ospedale dell'Universitá del Wisconsin, Madison, includono i soli casi clinici che al momento della diagnosi non presentavano evidenza di metastasi del tumore.

A ciascun paziente é stato associato un vettore a 34 dimensioni: l'attributo 1 é semplicemente il codice di identificazione del paziente, mentre l'attributo 2 indica se vi é stata ricorrenza o meno del tumore; gli attributi dal 4 al 33 si riferiscono ai parametri calcolati da un'immagine digitalizzata dei nuclei delle cellule tumorali asportate dai noduli anomali del seno del paziente mediante l'esame FNA; gli utlimi due attributi si riferiscono alla dimensione del tumore ed allo stato dei linfonodi:

1. Numero di codice

2. Classe: R per la ricorrenza del tumore, N per la non ricorrenza

3. Tempo di ricorrenza se l'attributo 2 é pari a R, tempo in assenza di malattia se l'attributo 2 é pari a N.

4.-33. attributi a valori reali caratterizzanti i nuclei delle cellule tumorali

34. Dimensione del tumore (dimensione del tumore asportato in centimetri)

35. Stato dei linfonodi

Il database contiene 198 campioni, di cui quattro senza un valore per l'attributo 35, quindi:

- Numero di parametri = 32
- Numero di campioni = 194
- $\mathcal{A}$ = Ricorrenza del tumore
- $\mathcal{B}$ = Non ricorrenza del tumore

5.3 Cleveland Heart Disease Database

L'applicazione presentata é relativa all'analisi di pazienti con disfunzioni cardiache creato da Robert Detrano, V.A. Medical Center, Long Beach and Cleveland Clinic Foundation.

Il quattoridicesimo attributo, in particolare, rappresenta la classe di appartenenza del paziente ed assume i valori interi da 0 (per l'assenza di disfunzioni) a 4.

Tale database contiene 297 punti:

- Numero di parametri = 13
- Numero di campioni = 297
- $\mathcal{A}$ = Assenza di disfunzioni cardiache
- $\mathcal{B}$ = Presenza di disfunzioni cardiache

5.4 BUPA Liver Disorders Database

Rimanendo nell'ambito della diagnostica medica, si considera ora un database di malati di fegato creato dal DR. Richard S. Forsyth.

A ciascun paziente sono stati associati sette attributi: i primi cinque si riferiscono tutti a prove del sangue che si ritiene siano strettamente correlate ai disturbi epatici, in particolare alla cirrosi provocata da un eccessivo consumo di bevande alcoliche; con il sesto attributo si definisce il consumo medio giornaliero di tali bevande; l'ultimo attributo, associato al singolo paziente maschio, rappresenta un codice che ne specifica la classe di appartenenza (1 per i sani e 2 per i malati):

1. Volume globulare medio
2. Fosfatasi alcalina
3. Amino-transferasi alanina
4. Aspartato amino-transferasi
5. Gamma-glutamil-transferasi
6. Consumo giornaliero di bevande alcoliche (in pinte)
7. Classe: 1 per i sani, 2 per i malati

Tale database contiene 345 punti:

- Numero di parametri = 6
- Numero di campioni = 345
- $\mathcal{A}$ = Assenza di disturbi epatici
- $\mathcal{B}$ = Presenza di disturbi epatici

5.5 Decorso post-operatorio di trapiantati renali

L'applicazione ora presentata si riferisce allo studio del decorso post operatorio di pazienti sottoposti a trapianto renale. In questo caso il paziente puó evolvere verso diversi stati. Il problema é quello di prevedere tale evoluzione sulla base dei dati clinici ottenuti dopo il trapianto.

Le diverse possibili evoluzioni sono:

- Pieno Successo
- Forme di Intossicazione piú o meno serie
- Rigetto

Il problema affrontato é quello di ottenere la classificazione dei pazienti sulla base di quattro parametri:

1. Citrato
2. T. M. A. O.
3. Creatinina
4. Hyppurato

misurati nei periodi immediatamente successivi all'operazione.

I dati clinici riguardanti tali parametri sono stati estratti dal KTD (Kidney Transplant Database), nato dalla collaborazione del Dipartimento di Chimica dell'Universitá degli Studi della Calabria con il Reparto di Nefrologia dell'Ospedale Civile dell'Annunziata di Cosenza. I campioni del KTD riguardano misurazioni effettuate mediante uno Spettrometro a Risonanza Magnetica Nucleare (NMR) per alcuni metaboliti presenti nelle urine dei pazienti esaminati. In particolare nell'applicazione presentata si sono considerati quattro indici di funzionalit'ênale rilevandone la concentrazione rispetto ad una quntitá nota di urina liofilizzata.

Il database contiene 71 campioni, 21 per il rigetto e 50 per le altre evoluzioni:

- Numero di parametri = 4
- Numero di campioni = 71
- $\mathcal{A}$ = Rigetto
- $\mathcal{B}$ = Tutti gli altri