Metodi Matematici per il Machine Learning

Marcello Cuoghi

December 8, 2018

Abstract

Piccolo riassunto del corso Metodi Matematici per il Machine Learning interno al corso di laurea magistrale in Ingegneria Informatica presso il dipartimento di Ingegneria Enzo Ferrari, UNIMORE.

Indice

Capitolo 1

Metodo del Gradiente

1.1 Introduzione

Nel mondo del Machine Learning, si sfrutta il metodo del gradiente per arrivare ad una soluzione del problema di apprendimento che viene espresso in termini di minimizzazione di un funzionale $f : R^n \to R$:

$$\min_{x \in R^n} = f(x) \tag{1.1}$$

Questo metodo cerca di sfruttare le informazioni del primo ordine per trovare, con dei piccoli accorgimenti per velocizzarne la convergenza, la soluzione ottima del problema di minimo. Anche se le condizioni del primo ordine spesso non permettono, oltre ad una veloce convergenza, anche una precisione elevata, quest'ultima spesso è irrilevante perché nel caso del Machine Learning una volta addestrato il modello si procede alla nuova classificazione di punti sfruttando operatori quali il segno, e tale operatore risulta insensibile ad una elevata precisione.
Nel caso di ricerca del minimo di un funzionale, si deve stabilire se tale minimo esiste e, nel caso, se è globale o locale. Per questo si cerca di utilizzare dove possibile funzionali convessi, che mi garantiscono quindi l'esistenza di un minimo globale e nessun minimo locale. Spesso però questo non è possibile, e quindi si devono eseguire più prove fino al raggiungimento di una soluzione accettabile per il nostro problema. Non è detto che sia sempre possibile arrivare al globale, e spesso non è neanche necessario poiché già un minimo locale ci permette una soluzione accettabile del problema.

1.2 Convessità

Sia $f : \Omega \to R$, dove $\Omega \subseteq R^n$ è un insieme convesso, cioè la combinazione convessa di due punti qualsiasi appartenenti ad Ω appartiene sempre ad Ω.

Allora f si dice convessa se per ogni $x, y \in \Omega$ vale che:

$$f(\alpha x + (1-\alpha)y) \leqslant \alpha f(x) + (1-\alpha)f(y) \quad \forall \alpha \in [0,1] \tag{1.2}$$

Se la precedente disuguaglianza vale in senso stretto, allora si dice che f è strettamente convessa.
In poche parole, come si può notare nella Figura 1.1, tale definizione garantisce che, scelti due punti qualsiasi appartenenti al dominio, il valore della funzione calcolato nella combinazione convessa (in blu) sia minore della combinazione convessa del valori della funzione nei due punti (in rosso), per ogni α appartenente all'intervallo $[0, 1]$.

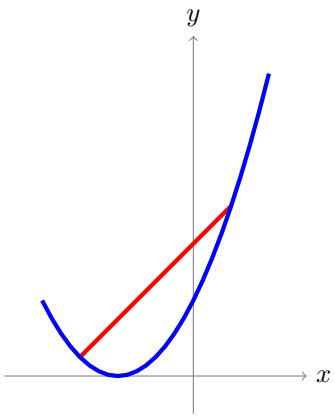

Figura 1.1: Esempio funzione convessa: $(x+1)^2$

1.3 Schemi Iterativi

Il metodo del gradiente è uno schema iterativo: a partire da un punto iniziale x^0, genera una successione di punti $x^1, x^2, ...$ che cerca di migliorare la stima della soluzione, fino al suo raggiungimento. È un metodo di tipo iterativo, sfrutta la creazione di una successione $\{x^k\}_{k=0,1,...}$ e, come tutti i metodi di tipo iterativo, deve tener contro di tre aspetti importanti:

1. regola per la costruzione delle x^k,
2. convergenza del processo iterativo $\lim_{k\to\infty} x^k = x^*$,
3. criterio di arresto, infatti il calcolatore non può gestire l'infinito.

Quindi, nel nostro caso (vedi equazione (1.1)) dato un punto iniziale x^0, per $k = 0, 1, ...$ fino al raggiungimento di un'approssimazione accettabile o a un massimo numero di iterate, calcola il nuovo punto come:

$$x^{k+1} = x^k + \alpha_k p^k \tag{1.3}$$

dove $\alpha_k > 0$ è la lunghezza del passo (o step length), e p^k è la direzione di ricerca.
I metodi di discesa generano una successione $\{f(x^k)\}_k$ monotona decrescente, ossia la successione generata $\{x^k\}_k$ deve essere tale che $f(x^{k+1}) \leqslant f(x^k)$ per ogni k, e si ha che $f(x^{k+1}) = f(x^k)$ solo se si è arrivati al minimo.

Un vettore p è di discesa per f in x^k se esiste $\overline{\alpha} > 0$ tale che:

$$f(x^k + \alpha p) < f(x^k) \quad \alpha \in (0, \overline{\alpha}] \tag{1.4}$$

Per funzioni differenziabili, la condizione per cui p è di discesa in x^k è:

$$\nabla f(x^k)^T p^k < 0 \tag{1.5}$$

Infatti, dallo sviluppo in serie di Tylor:

$$f(x_0 + \alpha s) = f(x_0) + \alpha \nabla f(x_0)^T s + \frac{1}{2}\alpha^2 s^T \nabla^2 f(x_0) s + o(\alpha^2) \tag{1.6}$$

sostituendo alla direzione s l'antigradiente in x^k e approssimando ad una funzione lineare nell'intorno di x^0, si ottiene:

$$f(x^k - \alpha \nabla f(x^k)) = f(x^k) - \alpha \nabla f(x^k)^T \nabla f(x^k) = f(x^k) - \alpha \|\nabla f(x^k)\|^2 \tag{1.7}$$

dove $\|\nabla f(x^k)\|^2$ è sempre maggiore di zero, per cui essendoci un meno, sto togliendo a $f(x^k)$ qualcosa di positivo, quindi ne riduco il valore sempre.
I più significativi metodi di discesa sono i cosiddetti metodi del gradiente, in cui $p^k = -D_k \nabla f(x^k)$, dove D_k è una matrice simmetrica definita positiva di ordine n (si ricorda che una matrice A è definita positiva se $x^T A x > 0 \ \forall x \neq 0$, nella pratica grazie ad un teorema basta verificare la positività di tutti gli autovalori di A per verificare la definita positività).
Dalla definita positività di D_k segue che:

$$\nabla f(x^k)^T p^k = -\nabla f(x^k)^T D_k \nabla f(x^k) < 0 \tag{1.8}$$

se $\nabla f(x^k) \neq 0$, cosicché p^k è sempre direzione di discesa. In base alla scelta di D_k si definiscono diversi metodi di discesa.

1.4 Steepest Descent (SD) o Discesa Ripida

Si pone $D_k = I_n$, per ogni $k \geqslant 0$. È la scelta più semplice, ma spesso implica convergenza lenta. Se considero il punto x^k, $f(x^{k+1})$ può essere approssimato come $f(x^k + p) \approx f(x^k) + \nabla f(x^k)^T p$, cerco la p che minimizza tale valore. Tra tutti i possibili versori p, quindi con $\|p\| = 1$, la derivata direzionale $\nabla f(x^k)^T p$ è minima per $\overline{p} = \frac{-\nabla f(x^k}{\|\nabla f(x^k)\|}$, infatti:

$$\nabla f(x^k)^T p \geqslant -\|\nabla f(x^k)\| \|p\| \geqslant -\|\nabla f(x^k)\| \ \forall p \ con \ \|p\| = 1 \tag{1.9}$$

da cui:

$$-\|\nabla f(x^k)\| = -\nabla f(x^k)^T \overline{p} = \nabla f(x^k)^T \frac{-\nabla f(x^k}{\|\nabla f(x^k)\|} \tag{1.10}$$

che altro non è che la direzione di più ripida discesa in x^k.

1.5 Metodo di Newton

Si pone $D_k = (\nabla^2 f(x^k))^{-1}$, per ogni $k \geqslant 0$, assumendo $\nabla^2 f(x^k)$ definita positiva (ricordando che l'Hessiana definita positiva è condizione sufficiente affinché la funzione sia strettamente positiva).
L'idea alla base del metodo di Newton è quella di minimizzare un'approssimazione quadratica $f^k(x)$ di $f(x)$ in x^k (da sviluppo in serie di Tylor):

$$\begin{gathered} f^k(x) = f(x^k) + \nabla f(x^k)^T (x - x^k) + \frac{1}{2}(x - x^k)^T \nabla^2 f(x^k)(x - x^k) \\ dove\ x = x^k + h\ da\ cui\ h = x - x^k \end{gathered} \tag{1.11}$$

Ponendo a 0 il gradiente di $f^k(x)$, si ottiene:

$$\nabla^2 f(x^k)^T (x - x^k) + \nabla f(x^k) = 0 \tag{1.12}$$

Quindi, risolto il sistema lineare $\nabla^2 f(x^k)^T p = -\nabla f(x^k)$ in cui $p = (x - x^k)$, il punto x^{k+1} risulta:

$$x^{k+1} = x^k - \nabla^2 f(x^k)^{-1} \nabla f(x^k) \tag{1.13}$$

Questo è il metodo di Newton puro, cioè con $\alpha_k = 1$.

In generale converge molte velocemente in prossimità del minimo, ma spesso il calcolo dell'Hessiana e, a maggior ragione, dell'inversa, risulta impraticabile nei casi reali.

1.6 Scelta della Lunghezza del Passo

In base alla lunghezza del passo α, viene assicurata una sufficiente decrescita di f e una velocità di convergenza adeguata. Spesso ci si riferisce alla ricerca della lunghezza del passo con il termine line search.

1.6.1 Regola del Passo Costante

Si tiene fisso il parametro a un valore dato dall'utente. Se il passo è troppo grande si può avere divergenza, se è troppo piccolo convergenza lenta.

1.6.2 Regola di Minimizzazione Esatta

Si cerca il valore di α_k che minimizza f lungo la direzione p^k, ossia si risolve un problema di minimo unidimensionale:

$$\min_{\alpha > 0} F(\alpha) = f(x^k + \alpha p^k) \tag{1.14}$$

α_k deve essere tale che annulli la derivata del funzionale $F(\alpha)$:
$F'(\alpha) = \nabla f(x^k + \alpha p^k)^T p^k$.
Il punto x^{k+1} è il punto in cui il gradiente nel nuovo punto è ortogonale alla direzione di ricerca p^k, infatti $\nabla f(x^k + \alpha p^k)^T p^k = 0$ per costruzione.
Ciò porta a lenta convergenza, poiché ci si sposta sempre di 90° rispetto alla direzione di ricerca precedente, e quindi si ha un andamento verso il minimo che viene detto a zig-zag.

1.6.3 Regola di Minimizzazione Limitata

Si fissa uno scalare $s > 0$ che limita la ricerca di α_k:

$$\min_{\alpha \in (0,s]} F(\alpha) = f(x^k + \alpha p^k) \tag{1.15}$$

Servono algoritmi di ricerca di α sia nel caso esatto che nel caso limitato. Nel caso della ricerca esatta potrebbe essere un'operazione costosa e non necessaria, infatti spesso è meglio cambiare direzione di ricerca piuttosto che trovare la lunghezza migliore per una certa direzione. In tutto ciò però, devo sempre garantire una sufficiente decrescita della mia funzione, ed è qui che la procedura di backtracking viene in aiuto.

1.6.4 Backtracking

Dati $\alpha_0 > 0, c \in (0,1), \rho \in (0,1)$ eseguo la seguente procedura di line search:

Algorithm 1 Line search inesatta

1: **procedure** BACKTRACKING(α_0, c, ρ) ▷ Line search inesatta
2: $\alpha \leftarrow \alpha_0$
3: **while** $f(x^k + \alpha p^k) > f(x^k) + c\alpha \nabla f(x^k)^T p^k$ **do** ▷ Garantisco decrescita
4: $\alpha \leftarrow \rho\alpha$
5: $\alpha_k \leftarrow \alpha$

Lo scalare c mi serve a bilanciare il valore del gradiente rispetto a f.
La condizione di crescita deve soddisfare la regola di Armijo:

$$f(x^k + \alpha p^k) \leqslant f(x^k) + c\alpha \nabla f(x^k)^T p^k \tag{1.16}$$

$$F(\alpha) \leqslant I(\alpha) \equiv F(0) + c\alpha F'(0) \tag{1.17}$$

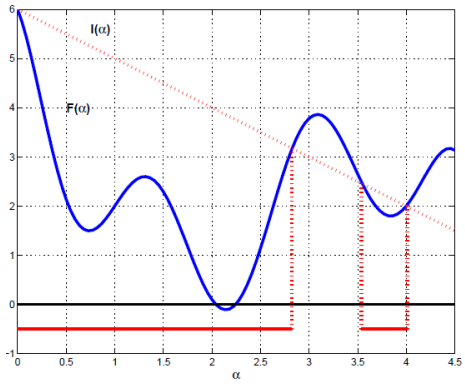

Figura 1.2: Valori di α che soddisfano la condizione (1.17) in questo esempio particolare

$I(\alpha)$ è una retta con pendenza negativa, il suo grafico per piccoli valori di α sta sopra a $F(\alpha)$, quindi la condizione per piccoli valori di α è sicuramente soddisfatta.

1.7 Risultati di Convergenza

Andiamo ora a ragionare sui risultati ottenuti attraverso l'utilizzo dei metodi del gradiente nel nostro caso. Infatti, la successione generata dai metodo del gradiente può non avere punti di accumulazione: se la funzione f non ha minimi locali, la successione $\{x^k\}$ può non essere limitata. Se però sappiamo che l'insieme di livello $\mathcal{L}_{\leqslant f(x^0)}(f) = \{x : f(x) \leqslant f(x^0)\}$ è limitato, e la scelta della lunghezza del passo forza la decrescita della funzione, allora la successione $\{x^k\}$ è limitata e dunque ha almeno un punto di accumulazione, ma non sappiamo ancora se corrisponde ad un minimo e se è unico.

Dalla sviluppo in serie di Tylor del primo ordine di f in x^k noto che la condizione che crea problemi quando ci si avvicina ad un punto stazionario (cioè con gradiente nullo) è la seguente:

$$\cos(\theta_k) = \frac{\nabla f(x^k)^T p^k}{\|\nabla f(x^k)\| \|p^k\|} \to 0 \tag{1.18}$$

in tal caso la funzione non presenta sufficiente decrescita e il metodo resta incollato al punto di non stazionarietà.

1.7.1 Direzioni Gradient Related

Una successione di direzioni $\{p^k\}$ si dice gradient related a $\{x^k\}$ se per ogni sotto-successione $\{x^j\}_{j \in J}$ convergente ad un punto di non stazionarietà, la cor-

rispondente sotto-successione $\{p^j\}_{j\in J}$ è limitata e soddisfa la condizione:

$$\lim_{j\to\infty} \sup_{j\in J} \nabla f(x^j)^T p^j < 0 \tag{1.19}$$

Se dunque p^k è gradient related, segue che se una sotto-successione $\{\nabla f(x^j)\}_{j\in J}$ tende a un vettore non nullo, la corrispondente sotto-successione$\{p^j\}_{j\in J}$ è limitata e non può tendere a essere ortogonale al gradiente, ossia p^k non diventa né troppo piccola, né troppo grande e l'angolo tra p^k e $\nabla f(x^k)$ non si avvicina a $\frac{\pi}{2}$.

Il presupposto per avere direzioni gradient related è quella di porre condizioni a priori su p^k che generalmente sono soddisfatte e non impongono grosse restrizioni. Nel caso dei metodi del gradiente, per cui $p^k = -D_k\nabla f(x^k)$, una condizione è quella di imporre gli autovalori della matrice simmetrica D_k definita positiva limitati dal basso e dall'alto da costanti positive indipendenti da k:

$$\exists\lambda_1,\lambda_2 \quad \lambda_1\|z\|^2 \leqslant z^T D_k z \leqslant \lambda_2\|z\|^2,\ \forall z\in R^n,\ k=0,1,... \tag{1.20}$$

Di conseguenza:

$$\begin{aligned} |\nabla f(x^k)^T p^k| &= |\nabla f(x^k)^T D_k \nabla f(x^k)| \geqslant \lambda_1\|\nabla f(x^k)\|^2 \\ \|p^k\|^2 &= |\nabla f(x^k)^T {D_k}^2 \nabla f(x^k)| \leqslant \lambda_2\|\nabla f(x^k)\|^2 \end{aligned} \tag{1.21}$$

Pertanto, finché $\nabla f(x^k)$ non tende a zero, $\nabla f(x^k)$ e p^k non possono diventare asintoticamente ortogonali e quindi il coseno non diventa mai nullo.

Un altro risultato utile a garantire la convergenza del metodo nel nostro caso è dato dal seguente teorema.
Sia f di classe $\mathcal{C}^1$ in $\mathcal{L}_{\leqslant f(x^0)}(f)$, sia $\{x^k\}$ una successione generata da un metodo del gradiente con $\{p^k\}$ gradient related e α_k scelta con una regola qualsiasi (minimizzazione esatta, limitata o backtracking). Allora ogni punto limite di $\{x^k\}$ è punto di stazionarietà di f.

1.7.2 Teorema della Cattura

Sia f di classe $\mathcal{C}^1$ in $L_{\leqslant f(x^0)}(f)$, sia $\{x^k\}$ una successione tale che $\{f(x^k)\}$ sia monotona non crescente ed è generata da un metodo del gradiente per cui ogni punto di accumulazione è punto di stazionarietà di f.
Siano $\overline{\alpha} > 0, L > 0$ tali che $\alpha_k \leqslant \overline{\alpha}$ e $\alpha_k \leqslant \overline{\alpha}$.
Sia x^* l'unico punto di minimo locale di f in un aperto.
Allora esiste un aperto S contenente x^* tale che se $x^{\overline{k}} \in S$ per un qualche $\overline{k} > 0$, allora $x^k \in S$ per ogni $k \geqslant \overline{k}$ e $x^k \to x^*$ per $k \to +\infty$.

1.8 Velocità di Convergenza

Dopo aver studiato i teoremi che garantiscono la convergenza dei metodi del gradienti nel nostro caso, passiamo a verificare come analizzarne la velocità di

convergenza.

Si assuma che $\{x^k\}$ sia una successione convergente a un punto stazionario x^*. La velocità di convergenza viene studiata asintoticamente, usando una funzione errore $e : R^n \to R$, tale che $e(x) \geqslant 0$ per ogni $x \in R^n$ e $e(x^*) = 0$. Un esempio di funzioni di errore sono le seguenti:

$$e(x) = \|x - x^*\| \tag{1.22}$$

$$e(x) = \|f(x) - f(x^*)\| \tag{1.23}$$

Lo studio asintotico viene condotto confrontando $\{e(x^k)\}$ con qualche successione standard. Si dice che:

$\to$ la convergenza è lineare o geometrica se $\exists q > 0$ e $\beta \in (0, 1)$ tali che

$$e(x^k) \leqslant q\beta^k \quad k \geqslant 0 \tag{1.24}$$

La convergenza lineare è garantita se:

$$\lim_{k\to\infty} \sup \frac{e(x^{k+1})}{e(x^k)} \leqslant \beta \quad \beta \in (0, 1) \tag{1.25}$$

$\to$ la convergenza è super-lineare se $\forall \beta \in (0, 1)$ e $\exists q > 0$ tali che

$$e(x^k) \leqslant q\beta^k \quad k \geqslant 0 \tag{1.26}$$

La convergenza super-lineare è garantita se:

$$\lim_{k\to\infty} \sup \frac{e(x^{k+1})}{e(x^k)} = 0 \tag{1.27}$$

La convergenza super-lineare può essere quantificata confrontando $\{e(x^k)\}$ con la successione$(\beta^t)^k$, $t \geqslant 1, k \geqslant 0$; si dice che $\{e(x^k)\}$ converge superlinearmente con ordine almeno t se esiste $q > 0,\ \beta \in (0, 1),\ t > 1$ per cui:

$$e(x^k) \leqslant q(\beta^t)^k \quad k \geqslant 0 \tag{1.28}$$

per $t = 2$ si parla di velocità di convergenza quadratica. La convergenza superlineare di ordine t è garantita se:

$$\lim_{k\to\infty} \sup \frac{e(x^{k+1})}{e(x^k)^t} < \infty \tag{1.29}$$

In termini pratici l'ordine di convergenza esprime il numero di cifre decimali che il metodo guadagna, ad ogni iterazione, rispetto alla soluzione esatta: tanto più grande è l'ordine, tanto maggiore è la velocità con cui la successione $\{x_k\}_k$

converge alla soluzione esatta.

Un altro risultato interessante è dato dal seguente teorema.
Sia $f(x) = \frac{1}{2}x^T Ax$, A simmetrica definita positiva. Sia $\{x^k\}$ generata dal metodo Steepest Descent, dove quindi $x^{k+1} = x^k - \alpha_k \nabla f(x^k)$, α_k è scelta con la regola di minimizzazione esatta.
Allora per ogni k si ha:

$$f(x^{k+1}) \leqslant \left(\frac{M-m}{M+m}\right)^2 f(x^k) \tag{1.30}$$

dove M e m sono, rispettivamente, il più grande e il più piccolo autovalore di A. Poiché A è definita positiva, ho il minimo in $x = 0$. Se $M \approx m$, ho forte smorzamento da un'iterazione all'altra, arrivando in fretta al minimo, infatti più M e m sono simili, più le curve di livello sono circonferenze. Nel caso opposto, quindi dove $M \gg m$, ho curve di livello molto ellittiche e quindi un andamento a zig zag e una lenta convergenza.
Stesso risultato più generale si ha con $f : R^n \to R$, $f \in \mathcal{C}^2$. Considerata la sequenza generata del metodo di discesa ripida combinata con una procedura di backtracking, se si assume che $\{x^k\}$ converga ad un punto di minimo locale forte x^* in cui l'Hessiana è definita positiva, allora

$$f(x^{k+1}) - f(x^*) \leqslant \left(\frac{k\left(\nabla^2 f(x^*)\right) - 1}{k\left(\nabla^2 f(x^*)\right) + 1}\right)^2 \left(f(x^k) - f(x^*)\right) \tag{1.31}$$

dove $k(A) = \frac{M}{m} \Rightarrow \left(\frac{M-m}{M+m}\right)^2 = \left(\frac{\frac{M}{m}-1}{\frac{M}{m}+1}\right)^2 = \left(\frac{k(A)-1}{k(A)+1}\right)^2$

1.8.1 Barzilai-Borwein

È uno dei metodi più utilizzati per accelerare notevolmente la velocità di convergenza senza dispendio in termini di risorse di calcolo. Si cercano di inglobare informazioni del secondo ordine, quindi relative all'Hessiana, nella scelta della lunghezza del passo α_k. In pratica si cercano valori di α_k tali che $\alpha_k \nabla f(x^k) \approx (\nabla^2 f(x^k))^{-1} \nabla f(x^k)$. Supposto di voler minimizzare una funzione quadratica del tipo $\frac{1}{2}x^T Ax$, dove quindi l'Hessiana è A, la precedente relazione ci porta ad ottenere:

$$\frac{1}{\alpha_k} I \approx A \tag{1.32}$$

per le proprietà di A si ha che:

$$A(x^{k+1} - x^k) = \nabla f(x^{k+1}) - \nabla f(x^k) \tag{1.33}$$

eseguendo le seguenti sostituzioni $s^k = (x^{k+1} - x^k)$ e $y^k = \nabla f(x^{k+1}) - \nabla f(x^k)$ che sono la differenza, rispettivamente, di due iterate successive e di due gradienti successivi, posso riscrivere la (1.33) come:

$$As^k = y^k \tag{1.34}$$

Procedo cercando i valori di di α che minimizzano la distanza tra i due termini dell'equazione $\frac{1}{\alpha}s^k = y^k$:

$$\arg\min_{\alpha} \frac{1}{2}\|\frac{1}{\alpha}s^k - y^k\|^2 \tag{1.35}$$

e per simmetria:

$$\arg\min_{\alpha} \|s^k - \alpha y^k\|^2 \tag{1.36}$$

Dall'equazione (1.35), ricavo gli α minimi:

$$\begin{aligned}
&\arg\min_{\alpha} \frac{1}{2}\|\frac{1}{\alpha}s^k - y^k\|^2 = \\
&\qquad = \arg\min_{\alpha} \frac{1}{2}\left(\frac{1}{\alpha}s^k - y^k\right)^T\left(\frac{1}{\alpha}s^k - y^k\right) \\
&\qquad = \arg\min_{\alpha} \frac{1}{2}\left(\left(\frac{1}{\alpha}\right)^2 {s^k}^T s^k + {y^k}^T y^k - 2\left(\frac{1}{\alpha}\right){s^k}^T y^k\right) \\
&\qquad \textit{posto } \frac{1}{\alpha} = \gamma \textit{ , } ({y^k}^T y^k) \textit{ non dipende da } \alpha, \textit{ lo elimino} \\
&\qquad = \arg\max_{\gamma} \frac{1}{2}\gamma^2 {s^k}^T s^k - \gamma {s^k}^T y^k \\
&\qquad \textit{trovo il massimo ponendo la derivata a zero} \\
&\qquad \Rightarrow \gamma {s^k}^T s^k - {s^k}^T y^k = 0 \\
&\qquad \gamma = \frac{{s^k}^T y^k}{{s^k}^T s^k}, \textit{ da cui} \\
&\qquad \alpha^{BB1} = \frac{{s^k}^T s^k}{{s^k}^T y^k} \\
&\qquad \textit{e in modo analogo :} \\
&\qquad \arg\min_{\alpha} \frac{1}{2}\|s^k - \alpha y^k\|^2 \\
&\qquad \alpha^{BB2} = \frac{{s^k}^T y^k}{{y^k}^T y^k}
\end{aligned} \tag{1.37}$$

Queste regole, prescindendo dalla conoscenza di A, possono essere inglobate a costo zero in un algoritmo per problemi non lineari (con line search inesatta) anche con vincoli semplici per cui l'Hessiana è difficilmente calcolabile.
Si osserva che $\alpha^{BB2} \leqslant \alpha^{BB1}$. Spesso si utilizzano tecniche di alternanza, per cui si eseguono alcune iterazioni con il primo parametro, altre iterazioni con il secondo parametro e così via.

Capitolo 2

Support Vector Machine

2.1 Introduzione

Il Support Vector Machine è un algoritmo di machine learning tra i più utilizzati. Deve il suo successo ai molti studi che ne hanno dimostrato l'utilità e l'efficacia in diversi contesti. L'idea base con cui è nato è quello di separare linearmente tramite un iperpiano di separazione due classi diversi, quindi un problema di classificazione binaria. In particolare cerca l'iperpiano $w^T x + b, w \in R^n \ e \ b \in R$ tale che $y_i(w^T x_i + b) \geqslant 1, i = 1, ..., N$. La classificazione di un nuovo punto avviene valutandone il segno rispetto l'iperpiano trovato:

$$F(x) = sign(w^T x + b) = \begin{cases} 1, & \text{se } w^T x_i + b > 0 \\ -1, & \text{se } w^T x_i + b < 0 \end{cases} \tag{2.1}$$

In fase di learning voglio avere quindi:

$$\begin{cases} w^T x_i + b > 0, & \text{se } y_i = 1 \\ w^T x_i + b < 0, & \text{se } y_i = -1 \end{cases} \tag{2.2}$$

scalando w e b ottengo:

$$\begin{cases} w^T x_i + b \geqslant 1, & \text{se } y_i = 1 \\ w^T x_i + b \leqslant 1, & \text{se } y_i = -1 \end{cases} = \begin{cases} y_i(w^T x_i + b) \geqslant 1, & \text{se } y_i = 1 \\ y_i(w^T x_i + b) \geqslant 1, & \text{se } y_i = -1 \end{cases} \tag{2.3}$$

da cui ricavo $y_i(w^T x_i + b) \geqslant 1 \forall i = 1, ..., N$.

2.2 Problema Primale

Passiamo ora a definire formalmente il problema di learning per poi poter successivamente andarlo a risolvere, in un primo momento nel caso di classi linearmente separabili.

Si definisce notazione canonica dell'iperpiano di separazione quella per cui la coppia (w, b) è tale che $y_i(w^T x_i + b) = 1$ per i punti più vicini all'iperpiano. L'iperpiano di separazione ottimale è quello che massimizza la distanza con i punto più vicini delle due classi. Ci sono infatti diversi studi e dimostrazioni che verificano che tra gli infiniti piani di separazione possibili tra due classi quello con le proprietà di generalizzazione migliore è quello che ne massimizza il margine. Ciò è verificabile anche a livello intuitivo geometricamente.
Si procede quindi a calcolare la distanza con segno dell'iperpiano da un punto x_i come $d_i = \frac{w^T x_i + b}{\|w\|}$, da cui, moltiplicando ambo i membri per y_i si ottiene $d_i y_i = \|w\|^{-1}$, infatti per definizione canonica di iperpiano di separazione abbiamo $w^T x_i + b = 1$. Si vuole massimizzare questa distanza, si ottiene quindi il seguente problema di minimizzazione vincolata:

$$\begin{aligned} \min_{w,b} \quad & \frac{1}{2} w^T w \\ s.t. \quad & y_i(w^T x_i + b) \geqslant 1, i = 1, ..., N \end{aligned} \tag{2.4}$$

questo nel caso di classi linearmente separabili.

Nel caso che i due insiemi non siano linearmente separabili, si aggiunge una variabile ausiliaria, detta variabile slack, che bilancia l'errore di classificazione di un punto. Si aggiunge tale variabile nella funzione obbiettivo volendo ridurre l'errore in fase di training. Il problema primale diventa perciò:

$$\begin{aligned} \min_{w,b,\xi} \quad & \frac{1}{2} w^T w + c \sum_{i=1}^{N} \xi_i \\ s.t. \quad & y_i(w^T x_i + b) \geqslant 1 - \xi_i \qquad i = 1, ..., N \\ & \xi_i \geqslant 0 \qquad\qquad\qquad\quad i = 1, ..., N \end{aligned} \tag{2.5}$$

Il problema è convesso poiché ho che l'Hessiana del mio problema è il seguente:

$$\begin{bmatrix} 1 & 0 & \dots & 0 & 0 & 0 \\ 0 & 1 & \dots & 0 & 0 & 0 \\ 0 & 0 & \dots & 1 & 0 & 0 \\ 0 & 0 & \dots & 0 & 0 & 0 \\ 0 & 0 & \dots & 0 & 0 & 0 \end{bmatrix}$$

infatti ho la prima parte della matrice corrispondente ai w, poi due righe e colonne tutte nulle corrispondenti a b e a ξ.
Il parametro c mi fa da trade-off tra gli errori commessi e la giusta classificazione.

Si definiscono Support Vector tutti gli esempi di training x_i tali che $y_i(w^T x_i + b) \leqslant 1$: nel caso $y_i(w^T x_i + b) < 1$ si chiamano Bound Support Vector, cioè il termine ξ_i associato a quell'esempio è maggiore di zero; se invece $y_i(w^T x_i + b) = 1$ si chiamano Support Vector non al Bound, quindi posti esattamente sul bordo del margine.

Questo modello di machine learning porta alla risoluzione di un problema di minimo vincolato complesso, quindi ora procediamo con l'introduzione di alcuni concetti utili alla creazione di un problema duale analogo ma più semplice che verrà utilizzato per la risoluzione.

2.3 Ottimizzazione Vincolata

Nel seguito si considereranno problemi del tipo $\min_{\Omega \in R^n} f(x)$, dove $f : R^n \to R$ è definita su di un aperto contenete Ω ed avente derivate continue.
Ω è detta regione ammissibile, non vuota e contenuta in R^n, chiusa e convessa.
Si definiscono vincoli semplici i vincoli la cui operazione di proiezione di un vettore sulla regione ammissibile è facile, cioè avviene a basso costo. Il costo è espresso in termini di iterazioni proporzionali alla dimensione del problema, un basso costo quindi implica un numero di iterazioni lineare rispetto alla dimensione del problema. L'operazione di proiezione consiste nel trovare il punto della regione ammissibile più vicino al punto dato: dato $z \in R^n$, $P_\Omega(z) = \arg\min_{x \in \Omega} \|x - z\|_2$. Un esempio tipico di vincoli semplici è dato dai vincoli di tipo box, in cui si limita dall'alto e dal basso i punti: $\Omega = \{x \in R^n : l \leqslant x \leqslant u\}$ $l, u \in R^n$.

Nel nostro caso la regione ammissibile è esprimibile come $\Omega = \{x \in R^n : h_i(x) = 0, i = 1, \ldots, m \wedge g_j(x) \leqslant 0, j = 1, \ldots, p\}$:

$$\begin{array}{lll} \min & f(x) & \\ s.t. & h_i(x) = 0 & i = 1, ..., m \\ & g_j(x) \leqslant 0 & j = 1, ..., p \end{array} \tag{2.6}$$

Si dice che un vincolo $g_j(x) \leqslant 0$ è attivo in x^* se $g_j(x^*) = 0$.
Si definisce insieme dei vincoli attivi in x^* l'insieme degli indici j corrispondenti ai vincoli attivi in x^*: $A(x^*) = \{j : g_j(x^*) = 0\}$.