

Support Vector Machines

La Support Vector Machine (d'ora in poi SVM) è un algoritmo di apprendimento per classificatori lineari che, fissato un training set linearmente separabile $(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_m, y_m) \in \mathbb{R}^d \times \{-1, +1\}$, genera il classificatore lineare corrispondente all'unica soluzione $\mathbf{w}^* \in \mathbb{R}^d$ del seguente problema di ottimizzazione convessa con vincoli lineari

$$\begin{aligned} \min_{\mathbf{w} \in \mathbb{R}^d} \quad & \frac{1}{2} \|\mathbf{w}\|^2 \\ \text{s.t.} \quad & y_t \mathbf{w}^\top \mathbf{x}_t \geq 1 \quad t = 1, \dots, m. \end{aligned} \tag{1}$$

Geometricamente, \mathbf{w}^* rappresenta l'iperpiano separatore a margine massimo, come dimostrato nel seguito.

Teorema 1. Per ogni $(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_m, y_m) \in \mathbb{R}^d \times \{-1, +1\}$ linearmente separabile, il vettore \mathbf{u}^* che realizza il margine massimo

$$\gamma^* = \max_{\mathbf{u}: \|\mathbf{u}\|=1} \min_{t=1, \dots, m} y_t \mathbf{u}^\top \mathbf{x}_t$$

soddisfa $\mathbf{u}^* = \gamma^* \mathbf{w}^*$, dove \mathbf{w}^* è soluzione di (1).

DIMOSTRAZIONE. Si noti che \mathbf{u}^* è identificato dalla soluzione del seguente problema di ottimizzazione

$$\begin{aligned} \max_{\gamma > 0} \quad & \gamma^2 \\ \text{s.t.} \quad & \|\mathbf{u}\|^2 = 1 \\ & y_t \mathbf{u}^\top \mathbf{x}_t \geq \gamma \quad t = 1, \dots, m. \end{aligned}$$

Infatti, \mathbf{u} che massimizza γ è lo stesso \mathbf{u} che massimizza γ^2 dato che la funzione $f(\gamma) = \gamma^2$ è monotona crescente per $\gamma > 0$. Dividendo per $\gamma > 0$ entrambi i membri di ciascun vincolo $y_t \mathbf{u}^\top \mathbf{x}_t \geq \gamma$ otteniamo il vincolo equivalente $y_t (\mathbf{u}^\top \mathbf{x}_t) / \gamma \geq 1$. Eseguendo il cambio di variabile $\mathbf{w} = \mathbf{u} / \gamma$, e notando che $\|\mathbf{w}\|^2 = 1 / \gamma^2$ a causa del vincolo $\|\mathbf{u}\|^2 = 1$, otteniamo quindi il problema equivalente

$$\begin{aligned} \min_{\mathbf{w} \in \mathbb{R}^d} \quad & \|\mathbf{w}\|^2 \\ \text{s.t.} \quad & \gamma^2 \|\mathbf{w}\|^2 = 1 \\ & y_t \mathbf{w}^\top \mathbf{x}_t \geq 1 \quad t = 1, \dots, m. \end{aligned}$$

Si noti ora che il vincolo $\gamma^2 \|\mathbf{w}\|^2 = 1$ è superfluo in quanto, per ogni $\mathbf{w} \in \mathbb{R}^d$, posso trovare $\gamma > 0$ tale che il vincolo è soddisfatto. Quindi lo possiamo eliminare. Scalando la funzione obiettivo per la costante $\frac{1}{2}$ otteniamo

$$\begin{aligned} \min_{\mathbf{w} \in \mathbb{R}^d} \quad & \frac{1}{2} \|\mathbf{w}\|^2 \\ \text{s.t.} \quad & y_t \mathbf{w}^\top \mathbf{x}_t \geq 1 \quad t = 1, \dots, m \end{aligned}$$

che conclude la dimostrazione □

Abbiamo quindi mostrato l'equivalente fra il problema di massimizzare il margine di \mathbf{u} mantenendo la norma $\|\mathbf{u}\|$ costante ed il problema di minimizzare la norma $\|\mathbf{w}\|$ mantenendo il margine di \mathbf{w} costante.

La seguente nozione ci aiuta a calcolare la forma della soluzione ottima \mathbf{w}^* .

Lemma 2 (Condizione di ottimalità di Fritz John). *Si consideri il problema*

$$\begin{aligned} \min_{\mathbf{w} \in \mathbb{R}^d} \quad & f(\mathbf{w}) \\ \text{s.t.} \quad & g_t(\mathbf{w}) \leq 0 \quad t = 1, \dots, m \end{aligned}$$

dove le funzioni f, g_1, \dots, g_m sono differenziabili. Se \mathbf{w}_0 è una soluzione ottima, allora esiste un vettore $\boldsymbol{\alpha} \in \mathbb{R}^m$ tale che

$$\nabla f(\mathbf{w}_0) + \sum_{t \in I} \alpha_t \nabla g_t(\mathbf{w}_0) = \mathbf{0}$$

dove $I = \{1 \leq t \leq m : g_t(\mathbf{w}_0) = 0\}$.

Applicando la condizione di Fritz John alla funzione obiettivo SVM, con $f(\mathbf{w}) = \frac{1}{2} \|\mathbf{w}\|^2$ e $g_t(\mathbf{w}) = 1 - y_t \mathbf{w}^\top \mathbf{x}_t$ otteniamo che

$$\mathbf{w}^* - \sum_{t \in I} \alpha_t y_t \mathbf{x}_t = \mathbf{0}.$$

Quindi la soluzione ha forma

$$\mathbf{w}^* = \sum_{t \in I} \alpha_t y_t \mathbf{x}_t$$

dove I denota l'insieme di quegli esempi di training (\mathbf{x}_t, y_t) tali che $y_t (\mathbf{w}^*)^\top \mathbf{x}_t = 1$. Questi \mathbf{x}_t sono i cosiddetti *vettori di supporto*, ovvero quelle istanze di training sulle quali \mathbf{w}^* ha margine esattamente pari a 1. Se levassimo dal training set tutti gli esempi tranne quelli di supporto la soluzione SVM non cambierebbe.

Passiamo ora a considerare il caso di un training set non linearmente separabile. Come dobbiamo cambiare la funzione obiettivo SVM? Un modo di farlo è il seguente,

$$\begin{aligned} \min_{\mathbf{w} \in \mathbb{R}^d} \quad & \frac{\lambda}{2} \|\mathbf{w}\|^2 + \frac{1}{m} \sum_{t=1}^m \xi_t \\ \text{s.t.} \quad & y_t \mathbf{w}^\top \mathbf{x}_t \geq 1 - \xi_t \quad t = 1, \dots, m \\ & \xi_t \geq 0 \quad t = 1, \dots, m. \end{aligned}$$

Le quantità ξ_t vengono dette *variabili di slack* e misurano di quanto ciascun vincolo di margine è violato da una potenziale soluzione \mathbf{w} . La media delle violazioni viene poi aggiunta alla funzione obiettivo. Un coefficiente di regolarizzazione $\lambda > 0$ è introdotto per bilanciare i due termini della funzione obiettivo.

Consideriamo ora i vincoli che coinvolgono le ξ_t , ovvero $\xi_t \geq 1 - y_t \mathbf{w}^\top \mathbf{x}_t$ e $\xi_t \geq 0$. Per minimizzare ciascun ξ_t , possiamo porre

$$\xi_t = \begin{cases} 1 - y_t \mathbf{w}^\top \mathbf{x}_t & \text{se } y_t \mathbf{w}^\top \mathbf{x}_t < 1 \\ 0 & \text{altrimenti.} \end{cases}$$

Ovvero, se il vincolo $y_t \mathbf{w}^\top \mathbf{x}_t \geq 1$ è soddisfatto da \mathbf{w} , allora ξ_t non serve e la poniamo a zero. Altrimenti, se il vincolo non è soddisfatto da \mathbf{w} , allora scegliamo il minimo valore di ξ_t che lo soddisfa, cioè $1 - y_t \mathbf{w}^\top \mathbf{x}_t$. Riassumendo, $\xi_t = [1 - y_t \mathbf{w}^\top \mathbf{x}_t]_+$, che corrisponde alla definizione hinge loss.

Ponendo $h_t(\mathbf{w}) = [1 - y_t \mathbf{w}^\top \mathbf{x}_t]_+$, il problema di SVM nel caso non linearmente separabile può allora essere riscritto come $\min_{\mathbf{w} \in \mathbb{R}^d} F(\mathbf{w})$, dove

$$F(\mathbf{w}) = \frac{1}{m} \sum_{t=1}^m h_t(\mathbf{w}) + \frac{\lambda}{2} \|\mathbf{w}\|^2 .$$

Dimostriamo ora che, anche nel caso di training set non linearmente separabili, la soluzione \mathbf{w}^* appartiene al sottospazio delle combinazioni lineari di esempi del training set moltiplicati dalle loro etichette.

Teorema 3. *Il minimo di F è rappresentabile come combinazione lineare di $y_1 \mathbf{x}_1, \dots, y_m \mathbf{x}_m$.*

DIMOSTRAZIONE. Sia \mathbf{w}^* il minimo di F . Per assurdo, supponiamo che

$$\mathbf{w}^* = \sum_{t=1}^m \alpha_t y_t \mathbf{x}_t + \mathbf{u}$$

dove $\mathbf{u} \in \mathbb{R}^d$ è la componente di \mathbf{w}^* ortogonale al sottospazio descritto da $\mathbf{x}_1, \dots, \mathbf{x}_m$. Quindi, in particolare,

$$y_t \mathbf{u}^\top \mathbf{x}_t = 0 \quad t = 1, \dots, m. \quad (2)$$

Ora sia $\mathbf{v} = \mathbf{w}^* - \mathbf{u}$. Primo, $\|\mathbf{v}\|^2 \leq \|\mathbf{w}^*\|^2$ dato che abbiamo tolto a \mathbf{w}^* una componente ortogonale a \mathbf{v} , e quindi la sua lunghezza è diminuita. Secondo,

$$h_t(\mathbf{v}) = [1 - y_t \mathbf{v}^\top \mathbf{x}_t]_+ = [1 - y_t (\mathbf{w}^* - \mathbf{u})^\top \mathbf{x}_t]_+ = [1 - y_t (\mathbf{w}^*)^\top \mathbf{x}_t + y_t \mathbf{u}^\top \mathbf{x}_t]_+ = h_t(\mathbf{w}^*)$$

usando (2). Quindi $F(\mathbf{v}) \leq F(\mathbf{w}^*)$, che contraddice l'ottimalità di \mathbf{w}^* . Di conseguenza, $\mathbf{u} = \mathbf{0}$. \square

Notiamo che, come nel caso linearmente separabile, anche in questo caso più generale \mathbf{w}^* dipenderà da un sottoinsieme di vettori di supporto. Ovvero, $\alpha_t \neq 0$ solo per alcuni t . A differenza del caso lineare, questi vettori di supporto non saranno soltanto i punti del training set a margine minimo rispetto a \mathbf{w}^* , ma tutti i punti corrispondenti a variabili slack $\xi_t > 0$.

Dato che \mathbf{w}^* è determinato da un sottoinsieme dei punti del training set, possiamo limitare il rischio statistico del classificatore lineare $h^*(\mathbf{x}) = (\mathbf{w}^*)^\top \mathbf{x}$ usando le tecniche dei "Compression bounds". In particolare,

$$\text{er}(h^*) \leq \tilde{\text{er}}(h^*) + \sqrt{\frac{1}{m} \left(N + (N + 1) \ln m + \ln \frac{1}{\delta} \right)}$$

con probabilità almeno $1 - \delta$ rispetto all'estrazione del training set, dove N è il numero dei vettori di supporto di \mathbf{w}^* e $\tilde{\text{er}}(h^*)$ è la frazione degli esempi che non sono supporti e che sono classificati scorrettamente da h^* .

Proseguiamo mostrando come minimizzare F usando OGD. Prima di tutto osserviamo che

$$F(\mathbf{w}) = \frac{1}{m} \sum_{t=1}^m \ell_t(\mathbf{w})$$

dove $\ell_t(\mathbf{w}) = h_t(\mathbf{w}) + \frac{\lambda}{2} \|\mathbf{w}\|^2$ è una funzione λ -fortemente convessa. Infatti, $\frac{\lambda}{2} \|\mathbf{w}\|^2$ è λ -fortemente convessa e h_t è convessa, il che implica che la loro somma è λ -fortemente convessa. Possiamo quindi applicare l'algoritmo OGD per funzioni fortemente convesse alle funzioni ℓ_1, \dots, ℓ_m . Questa particolare istanza di OGD prende il nome di Pegasos e può essere descritta come segue.

Parametri: numero T di cicli, coefficiente λ di regolarizzazione
 Input: Training set $(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_m, y_m) \in \mathbb{R}^d \times \{-1, +1\}$
 Inizializza $\mathbf{w}_1 = \mathbf{0}$
 Per $t = 1, \dots, T$

1. Estrai uniformemente a caso un elemento $(\mathbf{x}_{Z_t}, y_{Z_t})$ del training set
2. $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \nabla \ell_{Z_t}(\mathbf{w}_t)$

Output: $\bar{\mathbf{w}} = \frac{1}{T}(\mathbf{w}_1 + \dots + \mathbf{w}_T)$.

Procediamo quindi ad analizzare Pegasos. Sia $(\mathbf{x}_{Z_1}, y_{Z_1}), \dots, (\mathbf{x}_{Z_T}, y_{Z_T})$ la sequenza di esempi del training set che sono stati estratti nel passo 1 dell'algoritmo, e sia $\ell_{Z_1}, \dots, \ell_{Z_T}$ la corrispondente sequenza di funzioni di perdita. Cioè, $\ell_{Z_t}(\mathbf{w}) = h_{Z_t}(\mathbf{w}) + \frac{\lambda}{2} \|\mathbf{w}\|^2$ dove $h_{Z_t}(\mathbf{w}) = [1 - y_{Z_t} \mathbf{w}^\top \mathbf{x}_{Z_t}]_+$.

Sia \mathbf{w}^* la soluzione della funzione obiettivo SVM,

$$\mathbf{w}^* = \operatorname{argmin}_{\mathbf{w} \in \mathbb{R}^d} \left(\frac{1}{m} \sum_{t=1}^m h_t(\mathbf{w}) + \frac{\lambda}{2} \|\mathbf{w}\|^2 \right). \quad (3)$$

Per ogni realizzazione s_1, \dots, s_T delle variabili casuali Z_1, \dots, Z_T , l'analisi di OGD per funzioni fortemente convesse dà immediatamente il risultato

$$\frac{1}{T} \sum_{t=1}^T \ell_{s_t}(\mathbf{w}_t) \leq \frac{1}{T} \sum_{t=1}^T \ell_{s_t}(\mathbf{w}^*) + \frac{G^2}{2\lambda T} \ln(T+1) \quad (4)$$

dove $G = \max_{t=1, \dots, T} \|\nabla \ell_{s_t}(\mathbf{w}_t)\|$ è anch'essa una variabile casuale.

Per mostrare come questo risultato possa essere usato per migliorare $F(\bar{\mathbf{w}})$ useremo il fatto seguente

$$\mathbb{E}[\ell_{Z_t}(\mathbf{w}_t) \mid Z_1, \dots, Z_{t-1}] = \frac{1}{m} \sum_{s=1}^m \ell_s(\mathbf{w}_t) = F(\mathbf{w}_t). \quad (5)$$

Ovvero, condizionato sulle prime $t-1$ estrazioni (le quali determinano \mathbf{w}_t), il valore atteso di $\ell_{Z_t}(\mathbf{w}_t)$ è pari a $F(\mathbf{w}_t)$. L'altro fatto che useremo è che per ogni coppia di variabili casuali X, Y

vale $\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X | Y]]$. Quindi possiamo scrivere

$$\begin{aligned}
\mathbb{E}[F(\bar{\mathbf{w}})] &= \mathbb{E}\left[F\left(\frac{1}{T}\sum_{t=1}^T \mathbf{w}_t\right)\right] \\
&\leq \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^T F(\mathbf{w}_t)\right] \quad \text{usando la dis. di Jensen, dato che } F \text{ è convessa} \\
&= \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^T \mathbb{E}[\ell_{Z_t}(\mathbf{w}_t) | Z_1, \dots, Z_{t-1}]\right] \quad \text{usando (5)} \\
&= \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^T \ell_{Z_t}(\mathbf{w}_t)\right] \quad \text{usando } \mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X | Y]] \\
&\leq \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^T \ell_{Z_t}(\mathbf{w}^*)\right] + \frac{\mathbb{E}[G^2]}{2\lambda T}(\ln T + 1) \quad \text{usando (4)} \\
&= \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^T \mathbb{E}[\ell_{Z_t}(\mathbf{w}^*) | Z_1, \dots, Z_{t-1}]\right] + \frac{\mathbb{E}[G^2]}{2\lambda T}(\ln T + 1) \quad \text{usando } \mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X | Y]] \\
&= F(\mathbf{w}^*) + \frac{\mathbb{E}[G^2]}{2\lambda T} \ln(T + 1) \quad \text{usando (5)}.
\end{aligned}$$

Abbiamo così ottenuto

$$\mathbb{E}[F(\bar{\mathbf{w}})] \leq F(\mathbf{w}^*) + \frac{\mathbb{E}[G^2]}{2\lambda T}(\ln T + 1). \quad (6)$$

Quindi, se $\mathbb{E}[G^2]$ è limitato da una costante, la media $\bar{\mathbf{w}}$ dei vettori generati da OGD converge (in valore atteso rispetto all'estrazione a caso dei T esempi dal training set) a \mathbf{w}^* con tasso $\frac{\ln T}{T}$. Con un po' di fatica in più è possibile dimostrare (ma non lo facciamo qui) che $\bar{\mathbf{w}}$ converge a \mathbf{w}^* non solo in valore atteso ma anche in probabilità.

Ora maggioriamo il valore di G per ogni realizzazione s_1, \dots, s_T delle variabili casuali Z_1, \dots, Z_T . Abbiamo $\nabla \ell_{s_t}(\mathbf{w}_t) = -y_{s_t} \mathbf{x}_{s_t} \mathbb{I}\{h_{s_t}(\mathbf{w}_t) > 0\} + \lambda \mathbf{w}_t$. Sia $\mathbf{v}_t = y_{s_t} \mathbf{x}_{s_t} \mathbb{I}\{h_{s_t}(\mathbf{w}_t) > 0\}$. Dato che $\eta_t = 1/(\lambda t)$, notiamo che l'aggiornamento di \mathbf{w}_t ha una forma particolarmente semplice,

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \nabla \ell_t(\mathbf{w}_t) = \mathbf{w}_t + \eta_t \mathbf{v}_t - \eta_t \lambda \mathbf{w}_t = \left(1 - \frac{1}{t}\right) \mathbf{w}_t + \frac{1}{\lambda t} \mathbf{v}_t.$$

Sia $X = \max_{s=1, \dots, m} \|\mathbf{x}_s\|$. Dato che $\|\nabla \ell_{s_t}(\mathbf{w}_t)\| \leq \|\mathbf{v}_t\| + \lambda \|\mathbf{w}_t\| \leq X + \lambda \|\mathbf{w}_t\|$, dobbiamo calcolare un maggiorante di $\|\mathbf{w}_t\|$. Per far ciò, esaminiamo la ricorrenza

$$\mathbf{w}_{t+1} = \left(1 - \frac{1}{t}\right) \mathbf{w}_t + \frac{1}{\lambda t} \mathbf{v}_t.$$

Come è facilmente dimostrabile per induzione, \mathbf{w}_{t+1} è esprimibile come una combinazione lineare di \mathbf{v}_s per $s = 1, \dots, t$ ma con quali coefficienti? Fissiamo un $s \leq t$ e notiamo che quando \mathbf{v}_s è aggiunto a questa somma esso ha coefficiente $1/(\lambda s)$. Quando viene calcolato \mathbf{w}_{t+1} , \mathbf{v}_s avrà coefficiente pari a

$$\frac{1}{\lambda s} \prod_{r=s+1}^t \left(1 - \frac{1}{r}\right) = \frac{1}{\lambda s} \prod_{r=s+1}^t \frac{r-1}{r} = \frac{1}{\lambda t}.$$

Quindi otteniamo una semplice espressione per \mathbf{w}_{t+1} ,

$$\mathbf{w}_{t+1} = \frac{1}{\lambda t} \sum_{s=1}^t \mathbf{v}_s . \quad (7)$$

Dato che \mathbf{w}_{t+1} è una media dei \mathbf{v}_s divisi per λ , abbiamo infine $\|\mathbf{w}_{t+1}\| \leq \frac{1}{\lambda} \max_s \|\mathbf{v}_s\| \leq \frac{1}{\lambda} X$. Questo ci fa concludere che

$$\|\nabla \ell_t(\mathbf{w}_t)\| \leq X + \lambda \|\mathbf{w}_t\| \leq 2X .$$

Riportando questo maggiorante di G in (6) otteniamo

$$\mathbb{E}[F(\bar{\mathbf{w}})] \leq F(\mathbf{w}^*) + \frac{2X^2}{\lambda T} \ln(T+1) .$$

Il Teorema 3 stabilisce che la soluzione \mathbf{w}^* del problema SVM è rappresentabile come

$$\mathbf{w}^* = \sum_{s \in S} y_s \alpha_s \mathbf{x}_s$$

dove $\alpha_s > 0$ e $S \equiv \{t = 1, \dots, m : h_t(\mathbf{w}^*) > 0\}$. Una importante conseguenza di questo risultato è che possiamo risolvere il problema (3) in uno spazio di kernel \mathcal{H}_K , dove la funzione obiettivo F diventa

$$F_K(g) = \frac{1}{m} \sum_{t=1}^m h_t(g) + \frac{\lambda}{2} \|g\|_K^2 \quad g \in \mathcal{H}_K$$

con $h_t(g) = [1 - y_t g(\mathbf{x}_t)]_+$. Nello spazio \mathcal{H}_K , la soluzione SVM diventa rappresentabile come

$$\sum_{s \in S} y_s \alpha_s K(\mathbf{x}_s, \cdot)$$

che è chiaramente un elemento dello spazio kernel

$$\mathcal{H}_K \equiv \left\{ \sum_{i=1}^N \alpha_i K(\mathbf{x}_i, \cdot) : \mathbf{x}_1, \dots, \mathbf{x}_N \in \mathbb{R}^d, \alpha_1, \dots, \alpha_N \in \mathbb{R}, N \in \mathbb{N} \right\}$$

Così come avevamo fatto per il Perceptrone, possiamo implementare Pegasos nello spazio kernel. In \mathcal{H}_K il predittore Pegasos (7) diventa

$$g_{T+1} = \frac{1}{\lambda T} \sum_{t=1}^T f_t$$

con $f_t = y_{s_t} K(\mathbf{x}_{s_t}, \cdot) \mathbb{I}\{h_{s_t}(g_t) > 0\}$.