SVM

IDEA: ipotizziamo che i dati siano linearmente separabili, devo trovare un criterio per decidere quale fra tutte le rette, gli iperpiani, che separano l'insieme in due è **MIGLIORE,** secondo qualche punto di vista, come trovarla, come usarla per generalizzare.

La generalizzazione riguarda appunto l'apprendimento sui dati del passato per fare una buona previsione sui dati del futuro. Dire che il TS, i dati che abbiamo sono separabili in modo lineare significa che io posso sempre costruire una espressione lineare $y = w^T x + b$, la retta che abbiamo visto prima, in modo tale che tutti i dati che hanno l'etichetta $+1$ stiano dal lato positivo, cioè dal lato in cui l'espressione della retta assume valore $\geq$di un numero positivo, tutti i dati che hanno un'etichetta negativa stiano invece dall'altro lato e questo si chiama **separabilità lineare,** cioè è possibile trovare una retta, una funzione lineare specificando i parametri w e b, dove questo succede allora diciamo che i dati sono ***separabili*** in modo lineare, cioè $\exists$ un iperpiano $H(w,b)$ ossia $y = w^T x + b$, quindi $\exists \varepsilon > 0$ che è caratterizzata da:

$$w^T x_i + b \geq \varepsilon \qquad \forall i : y_i = +1$$
$$w^T x_i + b \leq -\varepsilon \qquad \forall i : y_i = -1$$

Una volta che ho trovato una retta, un iperpiano che separa i dati, io vorrei usare questa retta come previsore. Spesso, però le osservazioni (x_i, y_i) sono affette da errore, allora se la superficie di separazione è troppo vicino ai dati di un tipo e dell'altro, basterebbe una piccolissima perturbazione per classificare nel modo opposto, sarebbe una classificazione poco ***robusta***, troppo sensibile a piccole variazioni. Allora mi piacerebbe fare una classificazione che separi i dati in modo tale da mantenere un **INTERVALLO DI SEPARAZIONE** il più ampio possibile. Dividiamo tutto per $\varepsilon > 0$, in particolare $\varepsilon = 1$

$$w^T x_i + b \geq 1 \qquad \forall i : y_i = +1$$
$$w^T x_i + b \leq -1 \qquad \forall i : y_i = -1$$

Con un'unica espressione riesco a racchiudere sia il caso delle etichette positive che quelle negative

$$y_i\left(w^T x_i + b\right) \geq 1 \qquad \forall i$$

infatti, se $y_i = 1 \Rightarrow w^T x_i + b \geq 1$ ho il 1° caso,

se $y_i = -1 \Rightarrow -\left(w^T x_i + b\right) \geq 1 \Rightarrow w^T x_i + b \leq -1$ ho il 2° caso.

Dato un iperpiano di separazione H, cioè una coppia (w,b) per cui la

$$w^T x_i + b \geq 1 \qquad \forall i : y_i = +1$$
$$w^T x_i + b \leq -1 \qquad \forall i : y_i = -1$$

è verificata, introduciamo il concetto di ***margine di separazione***.

Devo definire la distanza fra le 2 rette estreme, parallele all'iperpiano $w^T x + b = 0$, ma che vanno a toccare da un lato i dati positivi, dall'altra i negativi. Questo iperpiano dipende da quanto sono distanti i punti di uno e dell'altro insieme dalla retta. La prima cosa da fare è calcolare la distanza di un punto dalla retta, distanza euclidea di norma 2 minima, la distanza minima di un punto dell'iperpiano di separazione e un punto dell'insieme.

E' un problema di minimizzazione vincolata in cui l'obiettivo è minimizzare la distanza tra un punto generico x dell'iperpiano e un punto ben preciso $\bar{x}$, uno dei punti etichettati. Il punto x lo vincolo ad appartenere all'iperpiano $w^T x_i + b = 0$. E' un problema di minimizzazione convessa, perché tutte le norme sono convesse, potrei minimizzare $\min_x \|x - \bar{x}\|$ senza al quadrato, però sarebbe convesso ma non differenziabile, siccome KKT vale per problemi differenziabili con continuità ci metto il quadrato.

Distanza di un punto da un iperpiano (KKT)

Ho un problema di minimizzazione convessa quadratica, con un vincolo lineare $w^T x + b = 0$, incognita x, w e b sono conosciuti. Posso usare KKT, valgono le C.Q. perché ho vincoli lineari.

Sia $\bar{x} \in \Re^n$. Devo trovare la distanza tra $\bar{x}$ e $w^T x + b = 0$

$$\min_x \|x - \bar{x}\|^2 \qquad \text{F.O.}$$

$$w^T x + b = 0 \quad \text{vincolo lineare}$$

$$\begin{cases} 2(x - \bar{x}) + \lambda w = 0 & \lambda \text{ al posto di } \mu \\ w^T x + b = 0 \end{cases}$$

Quindi KKT riduce il problema di ottimizzazione a risolvere un problema di equazioni. Vorrei trovare λ e x in modo tale che questa funzione sia $= 0$, ma la x deve essere anche ammissibile. Dalla 1^a equazione esplicito la x

$$2x - 2\bar{x} + \lambda w = 0$$
$$2x = 2\bar{x} - \lambda w$$
$$x = \bar{x} - 0{,}5\lambda w$$

$$w^T x + b = 0 \Rightarrow w^T(\bar{x} - 0{,}5\lambda w) + b = 0 \Rightarrow w^T \bar{x} - \frac{1}{2}\lambda w^T w + b = 0$$

$$\lambda = 2\frac{w^T\bar{x}+b}{w^T w}$$

$$\|x-\bar{x}\| = \|\bar{x}-0{,}5\lambda w-\bar{x}\| = \|-0{,}5\lambda w\| = \frac{\left|w^T\bar{x}+b\right|}{w^T w}\frac{\|w\|}{1}$$

$$\|x-\bar{x}\| = \frac{\left|w^T\bar{x}+b\right|}{\|w\|}$$

Margine di separazione

Nel nostro problema in cui vogliamo trovare i margini di separazione, vado a trovare la distanza di tutti i punti, quelli pieni e quelli vuoti dall'iperpiano di separazione, si definisce margine di separazione la minima distanza dei punti osservati dall'iperpiano di separazione:

$$\rho(w,b) := \min_{x_i \in TS}\left\{\frac{\left|w^T x_i+b\right|}{\|w\|}\right\} \qquad i \in TS$$

Iperpiano ottimo (iperpiano di separazione avente margine di separazione massimo)

$$\max_{w\in\Re^n, b\in\Re}\ \min_{x_i\in TS}\left\{\frac{\left|w^T x_i+b\right|}{\|w\|}\right\}$$

$$y_i\left(w^T x_i+b\right)\geq 1 \qquad \forall i$$

In generale, se ho un insieme di k numeri, il minimo è un numero che è $\leq$ di tutti i numeri $\left(1^a \text{ richiesta}\right)$ ma uguale ad uno di loro $\left(2^a \text{ richiesta}\right)$. Qui scrivo solo la 1^a richiesta. Anziché descrivere il massimo del minimo, dico massimo di un valore che è minore o uguale a tutti:

$$\max_{w,b,\eta}\ \eta$$

$$\eta \leq \frac{\left|w^T x_i+b\right|}{\|w\|}$$

$$y_i\left(w^T x_i+b\right)\geq 1 \qquad \forall i$$

Devo dimostrare che η sarà esattamente il minimo di $\frac{\left|w^T x_i+b\right|}{\|w\|}$, quindi non solo $\leq$ ma uguale a uno di loro. Immaginiamo per assurdo che nella soluzione ottimale, ottenga un valore di η che è strettamente minore di tutti i membri di destra, quindi è una sottostima del minimo. Se al posto di η strettamente minore di tutti mettiamo $\eta+\varepsilon$, se fossero tutti strettamente minori non sarebbe ottimo, quindi per forza nell'ottimo di questo problema almeno una di queste disequazioni sarà soddisfatta con uguale e quindi nell'ottimo di questo problema η sarà proprio il massimo tra i

membri di destra e quindi effettivamente questa è una riformulazione di un problema di $\max\min$. Funziona perché è il minimo di un numero finito di oggetti. L'ottimo di questo problema ha come valore di η proprio il minimo, il margine di separazione.

Supporre che il numeratore cioè il massimo valore di η sia $\frac{1}{\|w\|}$ non è una perdita di generalità.

Immaginiamo di avere la soluzione ottimale di questo problema $\overline{\eta}, \overline{w}, \overline{b}$, cioè di aver trovato l'iperpiano di massima separazione:

$$\overline{\eta} = \min_{x_i \in TS} \frac{\left|\overline{w}^T x_i + \overline{b}\right|}{\|\overline{w}\|}$$

Facciamo vedere che da questo possiamo trovare un altro in cui il numeratore è uguale a 1, che sarà ancora di separazione, sarà ancora massimo.

Prendiamo il valore più piccolo

$$\alpha = \min_{x_i \in TS} \left|\overline{w}^T x_i + \overline{b}\right| \qquad \alpha \geq 1$$

($\alpha \geq 1$ perché se $y_i = +1$ $\overline{w}^T x_i + b \geq 1$, se $y_i = -1$ se facciamo il modulo $\overline{w}^T x_i + b \geq 1$)

è un minimo finito, esiste sempre, almeno $= 1$.

Se $\alpha > 1$, potremmo scalare tutto e ricondurci al caso $= 1$.

Sia

$$w^* = \frac{\overline{w}}{\alpha}$$

$$b^* = \frac{\overline{b}}{\alpha}$$

è sempre un iperpiano separatore, calcoliamo il margine di separazione

$$\begin{aligned}
\overline{\eta} &= \min_{x_i \in TS} \frac{\left|\overline{w}^T x_i + \overline{b}\right|}{\|\overline{w}\|} \\
&= \min_{x_i \in TS} \frac{\left|w^{*T} x_i + b^*\right|}{\|w^*\|} \\
&= \frac{1}{\|w^*\|}
\end{aligned}$$

Il problema diventa

$$\max_{w,b,\eta} \eta$$

$$\eta \leq \frac{1}{\|w\|}$$

$$y_i\left(w^T x_i + b\right) \geq 1 \qquad \forall i$$

$\eta \leq \frac{1}{\|w\|}$ è un vincolo solo, prima invece avevo tante disequazioni, una per ogni punto del dataset $\eta \leq \frac{\left|w^T x_i + b\right|}{\|w\|}$, rimetto tutto nella F.O. e il problema diventa:

$$\max_{w,b} \frac{1}{\|w\|}$$

$$y_i\left(w^T x_i + b\right) \geq 1 \qquad \forall \mathrm{i}$$

Massimizzare una frazione, in cui varia solo il denominatore, vuol dire minimizzare il denominatore e quindi posso scrivere:

$$\min_{w,b} \|w\|^2$$

$$\min_{w,b} \frac{1}{2}\|w\|^2 \text{ per comodità}$$

$$y_i\left(w^T x_i + b\right) \geq 1 \qquad \forall i$$

Un risolutore mi troverà l'iperpiano di separazione massimo.

La vera difficoltà di questo problema è $\forall i$, i può essere un numero molto grande, w è un vettore a grandissima dimensione, uno per ogni caratteristica.

Questo è un problema di ottimizzazione quadratico, convesso con vincoli lineari. E' quello che viene definito come problema di ottimizzazione **primale** in forma standard dell'SVM.

Considereremo la ***formulazione duale*** del problema. Le motivazioni sono le seguenti:

- i vincoli del problema 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.

Il problema duale del problema

$$\min_{w,b} \frac{1}{2}\|w\|^2$$

$$y_i\left(w^T x_i + b\right) \geq 1 \qquad \forall i$$

è il seguente (massimo della lagrangiana)

$$\max L(w,b,\lambda) = \frac{1}{2}\|w\|^2 - \sum_i \lambda_i \left[y_i\left(w^T x_i + b\right) - 1\right]$$

$$t.c. \qquad \nabla_w L(w,b,\lambda) = 0$$

$$\frac{\partial L(w,b,\lambda) = 0}{\partial b}$$

$$\lambda \geq 0$$

ossia,

$$\max L(w,b,\lambda) = \frac{1}{2}\|w\|^2 - \sum_i \lambda_i \left[y_i\left(w^T x_i + b\right) - 1\right]$$

$$\max L(w,b,\lambda) = \frac{1}{2} w^T w - \sum_i \lambda_i \left[y_i\left(w^T x_i + b\right) - 1\right]$$

$$\max L(w,b,\lambda) = \frac{1}{2} w^T w - \sum_i \lambda_i y_i \left(w^T x_i + b\right) + \sum_i \lambda_i$$

$$t.c. \qquad w = \sum_i \lambda_i y_i x_i$$

$$\sum_i \lambda_i y_i = 0$$

$$\lambda_i \geq 0$$

che può essere riscritto nella seguente forma

$$\max_{\lambda \geq 0} \frac{1}{2} \sum_i \sum_j \lambda_i \lambda_j y_i y_j x_i^T x_j - \sum_i \lambda_i y_i \left(\sum_j \lambda_j y_j x_j^T \right) x_i - \sum_i \lambda_i y_i b + \sum_i \lambda_i$$ [1]

[1] Se devo calcolare il prodotto scalare w con se stesso, cioè la norma al quadrato, devo fare un prodotto di 2 sommatorie, ho usato 2 indici diversi per formulare questo prodotto, attenzione che le x sono dei vettori, non posso fare $x_i \cdot x_j$, sono prodotti scalari quindi x_i è trasposto (diventa vettore riga), x_j è un vettore colonna, il risultato deve dare un numero: $(1,n) \times (n,1)$

Il problema precedente può essere riscritto nella seguente forma

$$\max_{\lambda_i \geq 0} \sum_i \lambda_i - \frac{1}{2} \sum_i \sum_j \lambda_i \lambda_j y_i y_j x_i^T x_j \; {}^2$$

$$t.c. \qquad \sum_i \lambda_i y_i = 0$$

oppure, in maniera equivalente

$$\min_{\lambda_i \geq 0} \frac{1}{2} \sum_i \sum_j \lambda_i \lambda_j y_i y_j x_i^T x_j - \sum_i \lambda_i$$

$$t.c. \qquad \sum_i \lambda_i y_i = 0$$

Osserviamo che:

- l'esistenza della soluzione ottima $\left(w^*, b^*\right)$ del problema

$$\min_{w,b} \frac{1}{2} \|w\|^2$$

$$t.c. \qquad y_i \left(w^T x_i + b\right) \geq 1 \qquad \forall i$$

e le proposizioni (Condizioni KKT)[3] e (funzioni convesse continuamente differenziabili)[4]

---

[2] Nel duale Lagrangiano b non c'è perché è moltiplicato per 0, non influisce nella soluzione del duale lagrangiano e quindi il problema duale diventa un problema in cui compare solo l'incognita λ. E' un problema quadratico perché ho $\lambda_i \cdot \lambda_j$, le incognite λ compaiono in un polinomio di grado 2 in cui i coefficienti sono i prodotti delle etichette per i prodotti scalari delle caratteristiche. Ho cambiato segno!

[3] Sia x^* un punto di minimo locale del problema

$$\min \; f(x)$$
$$g_i(x) \leq 0 \qquad \text{lineare}$$

allora x^* è un punto ammissibile ed esiste λ^* tale che

$$\nabla_x L\left(x^*, \lambda^*\right) = 0$$
$$\lambda_i^* \geq 0 \qquad \lambda_i^* g_i\left(x^*\right) = 0$$

[4] Si assuma che il problema

$$\min \; f(x)$$
$$g_i(x) \leq 0$$
$$\text{f e g convesse continuamente differenziabili}$$

ammetta almeno una soluzione ottima x^* e che esista almeno un λ^*

Sia

$$L(x, \lambda) = f(x) + \sum_i \lambda_i g_i(x)$$

Allora $\left(x^*, \lambda^*\right)$ è soluzione del seguente problema

assicurano che il problema

$$\max_{\lambda_i \geq 0} \sum_i \lambda_i - \frac{1}{2} \sum_i \sum_j \lambda_i \lambda_j y_i y_j x_i^T x_j$$ [5]

$$t.c. \qquad \sum_i \lambda_i y_i = 0$$

ammette almeno una soluzione ottima λ^*;

- dalla formulazione

$$\max L(w,b,\lambda) = \frac{1}{2}\|w\|^2 - \sum_i \lambda_i \left[y_i \left(w^T x_i + b \right) - 1 \right]$$

$$\max L(w,b,\lambda) = \frac{1}{2} w^T w - \sum_i \lambda_i \left[y_i \left(w^T x_i + b \right) - 1 \right]$$

$$\max L(w,b,\lambda) = \frac{1}{2} w^T w - \sum_i \lambda_i y_i \left(w^T x_i + b \right) + \sum_i \lambda_i$$

$$t.c. \qquad w = \sum_i \lambda_i y_i x_i$$

$$\sum_i \lambda_i y_i = 0$$

$$\lambda_i \geq 0$$

segue che il vettore $w^* =$ può essere determinato come segue

$$w^* = \sum_i \lambda_i^* y_i x_i$$

- w^* dipende esclusivamente dai vettori di training x_i (***vettori di supporto***) i cui corrispondenti moltiplicatori λ_i^* sono diversi da zero;
- in un problema quadratico, se $(\bar{x}, \bar{\lambda})$ è una soluzione del duale, allora esiste un vettore x^* (non necessariamente uguale a $\bar{x}$) tale che $(x^*, \bar{\lambda})$ è una coppia (minimo globale – vettore di moltiplicatori di Lagrange), questa proprietà assicura che (w^*, b^*, λ^*) costituisce

---

$$\max_{x,\lambda} L(x,\lambda)$$

$$\nabla_x L(x,\lambda) = 0$$

$$\lambda \geq 0$$

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

[5] Nel duale Lagrangiano b non c'è perché è moltiplicato per 0, non influisce nella soluzione del duale lagrangiano e quindi il problema duale diventa un problema in cui compare solo l'incognita λ. E' un problema quadratico perché ho $\lambda_i \cdot \lambda_j$, le incognite λ compaiono in un polinomio di grado 2 in cui i coefficienti sono i prodotti delle etichette per i prodotti scalari delle caratteristiche. Ho cambiato segno!

una coppia (soluzione ottima – vettori di moltiplicatori di Lagrange), per cui valgono le seguenti condizioni di complementarietà

$$\lambda_i^*\left[y_i\left(\left(w^*\right)^T x_i + b^*\right) - 1\right] = 0 \qquad \forall i \qquad (\#)$$

- noto w^* e considerato qualsiasi moltiplicatore $\lambda_i^* \neq 0$ lo scalare b^* può essere determinato utilizzando la corrispondente condizione definita nella (#).

Con il duale lagrangiano abbiamo guadagnato qualcosa, abbiamo perso, che succede? Intanto una cosa l'abbiamo persa, l'iperpiano, dove lo troviamo?. Qui siamo in un altro spazio, nel primale se ho 100 feature e 1 milione di esempi ho 100 variabili (i coefficienti delle feature) e un milione di vincoli; nel duale ho un solo vincolo però ha un milione di variabili. Quindi da un problema che aveva un'enormità di vincoli e relativamente poche variabili, siamo passati a un problema che ha pochissimi vincoli, 2 vincoli perché c'è anche $\lambda \geq 0$ che è speciale, è l'ortante positivo, però ha un sacco di variabili, allora a seconda dei casi può essere più facile risolvere uno o l'altro.
Se sappiamo risolvere il problema duale, dalla soluzione ottima del duale è immediato trovare l'inclinazione dell'iperpiano di supporto.
L'altra cosa importante è questa, questo è un problema di ottimalità, ***vale la condizione di complementarietà nelle KKT***, ho risolto il duale lagrangiano, ho applicato le condizioni di FJ sapendo che vale KKT perché i vincoli sono lineari, quindi vale la constraint qualification, nei vincoli $y_i\left(w^T x_i + b\right) \geq 1 \;\; \forall i$, alcuni sono soddisfatti con $=$, alcuni con $>$, nella relazione che corre tra moltiplicatori di Lagrange e vincoli c'è la condizione di complementarietà che dice che, o è uguale a 0 il moltiplicatore o è soddisfatto con l'uguaglianza il vincolo, quindi gli unici moltiplicatori che possono essere $\neq 0$, non debbono ma possono, sono quelli associati ai vincoli soddisfatti con uguale (vincoli attivi).
Quindi, se per esempio ho 4 vettori di supporto, io potrei avere 1 milione di pallini però, sono solo 4 i dati caratterizzati dal fatto che $w^T x_i + b$ è uguale a $1 \text{o} -1$, cioè sono solo 4 su 1 milione, i dati che rendono attivo il vincolo.
Questo vuol dire che nel duale, ci sono 1 milione di variabili ma solo 4 saranno diverse da zero, forse anche meno che è solo necessario. Il problema è che per risolvere il duale nessuno mi dice prima quali sono queste 4, cioè finchè non so qual è il piano non so dire quali sono i vettori di supporto, se lo sapessi sarebbe facile, per cui io non so a priori quali sono i 4 moltiplicatori di Lagrange.
I buoni algoritmi cercano di indovinare anche stocasticamente quali sono quelle poche componenti che risolvono l'ottimo del problema (LIBSVM usa metodi di decomposizione).

La complementarietà rende un po' speciale questo problema, è un problema solo in apparenza a grandissima scala, le variabili in realtà che contano sono poche, è da qui che nasce l'importanza dei vettori di supporto, questi sono soluzioni molto speciali.

Anche la forma dell'iperpiano dipende dalla combinazione dei 4 dati dei vettori di supporto, è funzione solo di 4 osservazioni.[6]

Risolviamo il duale:

identifichiamo i vettori di supporto, sono sempre associati a moltiplicatori di Lagrange $\neq 0$, e con quelli riusciamo a trovare w,b e il piano di separazione, ci interessa la funzione $w^T x + b$, cioè una funzione che dato x mi dice se $w^T x + b$ è positivo o negativo, e questa la userò per classificare i prossimi dati, fare forecasting, previsioni. Quindi mi interessa la funzione segno

$$\operatorname{sgn}\left(w^T x + b\right),$$

mi interessa capire se $w^T x + b$ è positivo o negativo per classificare, quindi se al posto di w ci metto la sua espressione

$$w^* = \sum_j \lambda_j^* y_j x_j \qquad \lambda_j^* > 0$$

estendo la somma solo agli j che hanno il moltiplicatore positivo, quindi questa è l'espressione che devo usare per verificare se la prossima osservazione, che non è nel TS, sta dal lato positivo o negativo dell'iperpiano.

Tutto quello che ci interessa per classificare una nuova osservazione è aver risolto il problema duale, trovando λ^* e aver calcolato i prodotti scalari delle osservazioni di TS con le osservazioni su cui vogliamo fare la previsione. Ma cosa succede se il problema non è separabile linearmente?

[6] Si dimostra che i vettori di supporto sono sempre 2 o più di 2, c'è ne uno positivo, uno negativo però di solito c'è ne sono altri.

Soft Margin SVM

In applicazioni reali non sempre esiste un margine, ovvero non sempre le classi sono linearmente separabili nello spazio delle *features* attraverso un iperpiano. Il concetto alla base del *Soft Margin* permette di ovviare a questo limite, introducendo una variabile ξ aggiuntiva per ogni campione, in modo da rilassare (*slack*, scarto) il vincolo sul margine

$$y_i\left(w^T x_i + b\right) \geq 1 - \xi_i \qquad \forall i$$
$$\xi_\mathrm{i} \geq 0$$

Il parametro ξ rappresenta la *slackness* associata al campione. Quando $0 < \xi \leq 1$ il campione è correttamente classificato ma è all'interno dell'area di margine. Quando $\xi > 1$ il campione entra nello spazio di decisione della classe opposta e perciò verrà classificato in maniera errata.

Per cercare ancora un iperpiano di separazione in qualche modo ottimo, la funzione costo da minimizzare deve considerare anche la distanza tra il campione e il margine:

$$\min_{w,b,\xi} \frac{1}{2}\|w\|^2 + C\sum_i \xi_i$$

soggetta ai vincoli

$$y_i\left(w^T x_i + b\right) \geq 1 - \xi_i \qquad \forall i$$
$$\xi_\mathrm{i} \geq 0$$

Il parametro C è un grado di libertà del problema per indicare quanto un campione deve pagare il violare il vincolo sul margine. Quando C è piccolo, il margine è ampio, mentre quando C è prossimo a infinito si ricade alla formulazione *Hard Margin* di SVM vista in precedenza.

Ogni campione x_i può ricadere in uno di tre possibili stati:

- può stare oltre il margine $y_i\left(w^T x_i + b\right) > 1$ e di conseguenza non contribuire alla funzione;
- può stare sul margine $y_i\left(w^T x_i + b\right) = 1$ non partecipando direttamente alla minimizzazione ma solo come *support vector*,
- può infine cadere all'interno del margine ed essere penalizzato tanto quanto si discosta dai vincoli forti.

La lagrangiana del sistema

$$\min_{w,b,\xi} \frac{1}{2}\|w\|^2 + C\sum_i \xi_i$$

$$y_i\left(w^T x_i + b\right) \geq 1 - \xi_i \qquad \forall i$$
$$\xi_\mathrm{i} \geq 0$$

con i vincoli introdotti dalle variabili ξ, è

$$L(w,b,\lambda)=\frac{1}{2}\|w\|^2+C\sum_i \xi_i-\sum_i \lambda_i\left[y_i\left(w^T x_i+b\right)-1+\xi_i\right]-\sum_i \mu_i \xi_i \qquad (°)$$

Con l'aumento del numero di vincoli, le variabili duali sono sia λ che μ.

Il risultato notevole è che, applicate le derivate, la formulazione duale di (°) diventa esattamente uguale alla duale del caso *Hard Margin*: le variabili ξ_i infatti non compaiono nella formulazione duale e l'unica differenza tra il caso *Hard Margin* e il caso *Soft Margin* è nel vincolo sui parametri λ_i, in questo caso limitati tra

$$0 \le \lambda_i \le C$$

invece che con la semplice diseguaglianza $\lambda_i \ge 0$. Il grande vantaggio di questa formulazione è proprio nella elevata semplicità dei vincoli e nel fatto che permetta di ricondurre il caso *Hard Margin* a un caso particolare $(C=\infty)$ del *Soft Margin*. La costante C è un limite superiore al valore che gli λ_i possono assumere.

Il duale lagrangiano è

$$\max_{\lambda\ge 0,\mu\ge 0}\min_{w,b,\xi}\frac{1}{2}w^T w+C\sum_i \xi_i+\sum_i \lambda_i\left[1-\xi_i-y_i\left(w^T x_i+b\right)\right]-\sum_i \mu_i \xi_i$$

Punto sella:

$$\begin{aligned}
&\nabla_w = w-\sum_i \lambda_i y_i x_i \Rightarrow \nabla_w = 0 \Rightarrow w=\sum_i \lambda_i y_i x_i\\
&\nabla_b = \sum_i \lambda_i y_i \Rightarrow \nabla_b = 0 \Rightarrow \sum_i \lambda_i y_i = 0\\
&\nabla \xi_i = -\lambda_i-\mu_i+C\\
&\lambda,\mu \ge 0
\end{aligned}$$

Il problema precedente può essere riscritto nella seguente forma

$$\begin{aligned}
&\min \frac{1}{2}\sum_i\sum_j y_i y_j x_i^T x_j \lambda_i \lambda_j-\sum_i \lambda_i\\
&\sum_i \lambda_i y_i = 0\\
&0 \le \lambda_i \le C
\end{aligned}$$

Questa è la stessa forma del duale lagrangiano che abbiamo trovato prima, non cambia niente. L'unica novità è che dobbiamo scrivere $-\lambda_i-\mu_i+C=0$. Questo vuol dire che $\lambda_i = C-\mu_i$, cioè $\lambda_i \le C$.

SVM non lineari e funzioni kernel

Per separare una superficie NON lineare, un trucco possibile è quello di trasformare lo spazio delle features x attraverso una funzione, cioè mappiamo le features in un altro spazio. Nonostante il Soft Margin, alcuni problemi sono intrinsecamente non separabili nello spazio delle feature. Tuttavia, dalla conoscenza del problema, è possibile intuire che una trasformazione non lineare $\phi : X \to H$ trasforma lo spazio delle feature di input X nello spazio delle feature H dove l'iperpiano di separazione permette di discriminare meglio le categorie. La funzione discriminante nello spazio H è

$$\text{sgn} = w^T \phi(x) + b$$

Per permettere la separazione, normalmente lo spazio H è di dimensioni maggiori dello spazio X. Questo aumento di dimensioni provoca un aumento della complessità computazionale del problema e la richiesta di risorse. I metodi Kernel risolvono questo problema.

Il vettore w è una combinazione lineare dei campioni di addestramento (i *support vector* nel caso *hard margin*):

$$w = \sum_i \lambda_i \phi(x_i)$$

La funzione discriminante assume pertanto la forma

$$\begin{aligned} \text{sgn} &= \sum_i \lambda_i \phi(x_i)^T \phi(x) + b \\ &= \sum_i \lambda_i k(x, x_i) + b \end{aligned}$$

con la valutazione della funzione kernel $k(x, x_i)$.

Al momento della valutazione della funzione discriminante pertanto è richiesto l'utilizzo dei vettori di supporto (almeno quelli con un parametro λ_i ssociato non trascurabile). Di fatto SVM con kernel individua alcuni campioni dell'insieme di addestramento come informazione utile per capire quanto vicino a loro è il campione di valutazione in esame.

Sia

$$TS = \{(x_i, y_i)\} \qquad i = 1, \ldots, l$$

$$x_i \in \Re^n, y_i = \{\pm 1\}$$

Sia

$$\phi : \Re^n \to H$$

e

$$\{\phi(x_i)\}$$

l'insieme delle features associate con i vettori di training.

Il duale lagrangiano diventa

$$\min \frac{1}{2}\sum_i \sum_j y_i y_j \langle \phi(x_i), \phi(x_J) \rangle \lambda_i \lambda_j - \sum_i \lambda_i$$
$$\sum_i \lambda_i y_i = 0$$
$$0 \le \lambda_i \le C$$

Se $k(\cdot,\cdot)$ è un Kernel

$$\min \frac{1}{2}\sum_i \sum_j y_i y_j k(x_i, x_j) \lambda_i \lambda_j - \sum_i \lambda_i$$
$$\sum_i \lambda_i y_i = 0$$
$$0 \le \lambda_i \le C$$

L'ottimo w è

$$w^* = \sum_i \lambda_i^* y_i \phi(x_i)$$

e la funzione discriminante ottima

$$\operatorname{sgn} = \left(\sum_j \lambda_j^* y_j \langle \phi(x_j), \phi(x_i) \rangle + b^* \right)$$
$$= \operatorname{sgn}\left(\sum_j \lambda_j^* y_j k(x_j, x_i) + b^* \right)$$

Qual è il vantaggio? Il vantaggio è che posso definire un problema di ottimizzazione a partire dalla funzione kernel senza bisogno di sapere qual è la trasformazione ϕ che genera questo kernel.

I kernel più diffusi, in quanto semplici da valutare, sono i kernel gaussiani nella forma

$$k(u,v) = e^{-\gamma \|u-v\|^2}$$

con γ parametro da impostare, e i kernel polinomiali di grado p nella forma

$$k(u,v) = (u^T v)^p$$

e nel caso $p = 1$ la formulazione si riconduce al caso lineare.

L'utilizzo di funzioni kernel, unita alla possibilità di precalcolare tutte le combinazioni $k(x_i x_j)$, permette di definire un'interfaccia comune tra gli addestramenti lineari e i non lineari, mantenendo di fatto lo stesso grado di prestazioni.

Anziché fare $x^T x$, valutiamo una funzione kernel nella coppia (x_i, x_j), cioè sostituiamo al posto di $u\ x_i$ e $v\ x_j$. Questo è uno dei motivi più forti per usare il DUALE perché nel duale tutto dipende dal prodotto scalare $x_i x_j$.

Programmazione quadratica per SVM lineari

Questa classe di problemi (minimizzazione con vincoli come disuguaglianze o *primal optimization problem*) si risolvono utilizzando l'approccio di Karush-Kuhn-Tucker che è il metodo dei moltiplicatori di Lagrange generalizzato a disuguaglianze. Attraverso le condizioni KKT (in questo caso necessarie e sufficienti in quanto è un problema di minimizzazione quadratica, è una norma al quadrato, quindi convessa, anzi strettamente convessa) si ottiene la funzione lagrangiana:

$$L(w,b,\lambda) = \frac{1}{2}\|w\|^2 - \sum_i \lambda_i \left[y_i \left(w^T x_i + b \right) - 1 \right]$$

da minimizzare in w e b e massimizzare in λ. I pesi $\lambda_i \geq 0$ (perché riferiti a vincoli di disuguaglianza) sono i moltiplicatori di Lagrange.

Dall'annullamento delle derivate parziali si ottiene

$$\nabla_w L(w,b,\lambda) = w - \sum_i \lambda_i y_i x_i$$

$$\nabla_w L(w,b,\lambda) = 0 \Rightarrow w = \sum_i \lambda_i y_i x_i$$

$$\frac{\partial L(w,b,\lambda)}{\partial b} = 0 \Rightarrow 0 - \sum_i \lambda_i y_i = 0 \Rightarrow \sum_i \lambda_i y_i = 0$$

Sostituendo tali risultati (le variabili primali) all'interno della lagrangiana questa diventa funzione dei soli moltiplicatori, i **dual**, da cui la forma duale di Wolfe:

$$\max_{\lambda \geq 0} \sum_i \lambda_i - \frac{1}{2} \sum_i \sum_j \lambda_i \lambda_j y_i y_j x_i^T x_j \qquad 1° \text{ vincolo } \lambda_i \geq 0$$

$$\sum_i \lambda_i y_i = 0 \qquad 2° \text{ vincolo}$$

oppure

$$\min_{\lambda \geq 0} \frac{1}{2} \sum_i \sum_j \lambda_i \lambda_j y_i y_j x_i^T x_j - \sum_i \lambda_i$$

$$\sum_i \lambda_i y_i = 0$$

Il max del duale (incognita λ) sono gli λ_i associati a ogni vettore di addestramento x_i. Tale massimo permette di trovare la soluzione del problema originale.

Su questa relazione sono valide le condizioni KKT tra le quali è di notevole importanza il vincolo, detto di *Complementary slackness*,

$$\lambda_i \left[y_i \left(w^T x_i + b \right) - 1 \right] = 0$$

Questo vincolo dice che il massimo della lagrangiana o è sul bordo del vincolo $(\lambda_i \neq 0)$ o è un minimo locale $(\lambda_i = 0)$, perché vale la condizione di complementarietà che dice che, o è uguale a 0 il moltiplicatore o è soddisfatto con l'uguaglianza il vincolo, quindi gli unici moltiplicatori che possono essere $\neq 0$, sono quelli associati ai vincoli soddisfatti con uguale (vincoli attivi).

Come conseguenza solo gli λ_i sul limite (margine) sono non nulli e contribuiscono alla soluzione: tutti gli altri campioni di addestramento sono di fatto ininfluenti. Tali vettori, associati agli $\lambda_i > 0$, sono i *Support Vectors*.

Risolvendo il problema quadratico

$$\max_{\lambda \geq 0} \sum_i \lambda_i - \frac{1}{2} \sum_i \sum_j \lambda_i \lambda_j y_i y_j x_i^T x_j \ ,$$

sotto il vincolo

$$\sum_i \lambda_i y_i = 0$$

e

$$\lambda_i \geq 0,$$

i pesi che presentano $\lambda_i \neq 0$ saranno i *Support Vectors*.

Tali pesi, inseriti nelle equazioni $w = \sum_i \lambda_i y_i x_i$ e $\lambda_i \left[y_i \left(w^T x_i + b \right) - 1 \right] = 0$, porteranno a ricavare l'iperpiano di massimo margine.

Il metodo più usato per risolvere questo problema QP è il *Sequential Minimal Optimization* (*SMO*).

Perché si chiamano vettori di supporto?

Supponiamo di aver risolto il problema

$$\min_{w,b} \|w\|^2$$

$$y_i\left(w^T x_i + b\right) \geq 1 \qquad \forall i$$

avendo trovato l'iperpiano di separazione a massimo margine succede sempre che l'iperpiano a massimo margine ha un margine che è uguale a

$$\frac{2}{\|w\|},$$

che il margine che stiamo massimizzando è $\frac{1}{\|w\|}$, questo margine è metà del margine, è la distanza tra l'iperpiano separatore e la più vicina delle osservazioni. Il margine è definito da almeno un'osservazione che sta sul bordo di questo margine, se questo "tubo" non toccasse nessuna osservazione lo potrei allargare ancora un po'. Questo vuol dire che sicuramente almeno uno di questi vincoli di ≥ 1 è attivo, deve valere per forza uguale quando troviamo la soluzione ottimale di questo problema. Se valesse il > stretto per tutti io potrei aumentare, cioè diminuire w e mantenere ancora soddisfatto il vincolo. Quindi almeno uno dei vincoli sarà soddisfatto, quindi ci sarà almeno uno dei punti che sono esattamente a distanza $\frac{1}{\|w\|}$ dall'iperpiano di separazione. Stesso ragionamento per i punti etichettati negativi.

I punti che stanno sui margini di separazione si chiamano **vettori di supporto**, perché son quelli che sostengono questo margine di separazione, sono quelli contro cui andiamo a sbattere quando cerchiamo di massimizzare il margine. Sono in realtà, i dati sperimentali, le osservazioni in un certo senso più importanti nella classificazione che abbiamo trovato. Gli altri punti non hanno nessun ruolo in questa separazione. Quindi vettori di supporto perché fungono da supporto al margine di separazione.

In definitiva se $H\left(w^*, b^*\right)$ è l'iperpiano di separazione ottimo, qualsiasi vettore caratteristica x_i tale che

$$y_i\left(\left(w^*\right)^T x_i + b^*\right) = 1$$

è detto ***vettore di supporto***.

OSSERVAZIONE: siamo sicuri che nell'ottimo di questo problema $\left|w^T x_i + b\right|$ sia uguale a 1?

Noi abbiamo preso un problema e abbiamo sfruttato la sua caratteristica per trovarne una riformulazione, però se adesso risolviamo la riformulazione chi ci garantisce che questa sia una soluzione del problema originale, cioè siamo sicuri che nella soluzione ottimale di questo problema trasformato valga che il minimo del numeratore che compare nell'espressione sia proprio 1.
La risposta è si, però non si trovano dimostrazioni di questo fatto nella stragrande maggioranza delle trattazioni