Support vector machines (SVMs)

Docente: Tommaso Muraca

SVM

L'algoritmo **SVM** in genere **risolve** lo stesso problema della regressione logistica (**classificazione**) e produce **prestazioni simili.**

Vale la pena conoscerlo perché l'algoritmo è di natura **geometricamente motivata**, piuttosto che essere guidato dal pensiero probabilistico.

Alcuni esempi dei problemi che le SVM può risolvere:

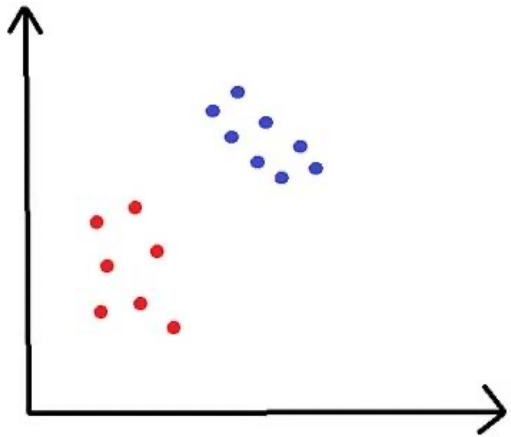
- **È l'immagine di un gatto o di un cane?**

- Questa recensione è positiva o negativa?

- I punti nel piano 2D sono rossi o blu?

Partiremo proprio da questo terzo esempio per illustrare **come funziona l'algoritmo SVM**: →

In questo esempio, abbiamo **punti in uno spazio 2D che sono rossi o blu e vorremmo separare nettamente i due.**



SVM

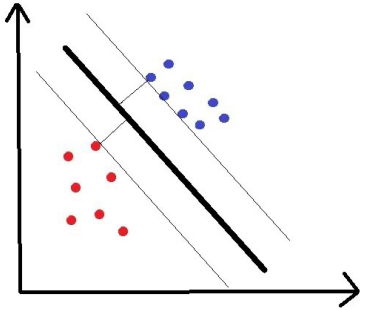
Il set di **addestramento** è tracciato nel **grafico precedente**. Vorremmo **classificare punti nuovi** e non classificati in questo piano. Per fare ciò, le SVM utilizzano una **linea di separazione** (o, in più di due dimensioni, un iperpiano multidimensionale) per **dividere lo spazio in una zona rossa e una zona blu**. Puoi già immaginare come potrebbe apparire una linea di separazione nel grafico qui sopra.

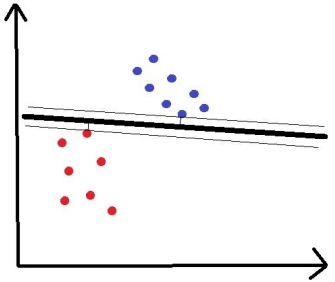
Come, nello specifico, scegliamo dove tracciare la linea?

Nei grafici a destra sono riportati 2 esempi -->

La distanza dal **punto più vicino** su entrambi i lati della linea è chiamata **margine** e **SVM** tenta di **massimizzare il margine** . Possiamo considerarlo come uno **spazio di sicurezza**: più **grande è lo spazio**, **meno è probabile** che i **punti rumorosi vengano classificati erroneamente.**

Sulla base di questa breve spiegazione, sorgono alcune grandi domande.





SVM

1.Come funziona la matematica dietro questo algoritmo?

Vogliamo trovare **l'iperpiano ottimale (una linea, nel nostro esempio 2D).**

Questo iperpiano deve

- 1) **separare i dati in modo netto**, con punti blu su un lato della linea e punti rossi sull'altro lato,
- 2) **massimizzare il margine**.

La **versione umana** per **risolvere questo problema** sarebbe quella di **prendere un righello** e continuare a provare diverse linee che separano tutti i punti **finché non si ottiene quella che massimizza il margine**.

Esiste un **metodo matematico pulito per ottenere questa massimizzazione**, i dettagli sono molto complessi quindi non li trarremo nello specifico, ma se qualcuno vuole approfondire si tratta del **metodo dei moltiplicatori di Lagrange**, una strategia per trovare i massimi e i minimi locali di una funzione soggetta a vincoli di equazione (cioè soggetta alla condizione che una o più equazioni debbano essere soddisfatte esattamente dai valori scelti delle variabili).

L'iperpiano della soluzione che si ottiene è definito in **relazione alla sua posizione rispetto a determinati x_i**, che sono chiamati **vettori di supporto** , e di solito sono quelli **più vicini all'iperpiano.**

SVM

2. Cosa succede se non riesciamo a separare i dati in modo netto?

Esistono **due metodi** per affrontare questo problema:

- **Ammorbidire la definizione di "separato".**

Permettiamo alcuni errori, nel senso che permettiamo alcuni punti blu nella zona rossa o alcuni punti rossi nella zona blu. Lo facciamo **aggiungendo un costo C** per gli esempi erroneamente classificati nella nostra funzione di perdita. Fondamentalmente, diciamo che **è accettabile ma costoso classificare erroneamente un punto.**

- **Inseriamo i dati in dimensioni più elevate.**

Possiamo creare **classificatori non lineari** aumentando il **numero di dimensioni**, ad esempio includendo x^2, x^3, anche cos(x), ecc. All'improvviso, si hanno confini che possono apparire più ondulati quando li riportiamo alla rappresentazione dimensionale inferiore.

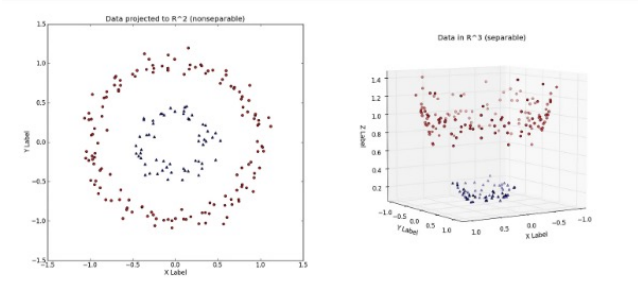
SVM

Intuitivamente, è **come avere delle biglie rosse e blu stese a terra in modo tale che non possano essere separate** nettamente da una linea - ma **se potessimo far levitare tutte le biglie rosse da terra** nel modo giusto, potremo **disegnare un piano che li separa**. Poi le **lasciamo ricadere a terra sapendo dove finiranno i blu e iniziaeranno i rossi.**

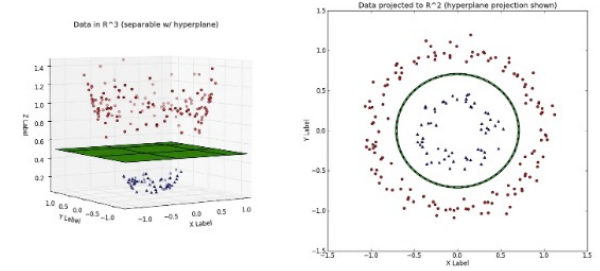
Grafici a destra -->

In **sintesi**, le **SVM** vengono **utilizzate per la classificazione**, il loro **scopo è di trovare un piano che separi nettamente le due classi**.

Quando ciò non è possibile, ammorbidiamo la **definizione di "separato"** oppure inseriamo i dati in **dimensioni più elevate** in modo da poterli separare in modo pulito.



Un set di dati non separabili in uno spazio bidimensionale R^2 e lo stesso set di dati mappato su tre dimensioni con la terza dimensione x^2+y^2 (fonte: http://www.eric-kim.net/eric-kim-net/posts/1/kernel_trucco.html)



Il confine decisionale è mostrato in verde, prima nello spazio tridimensionale (a sinistra), poi di nuovo nello spazio bidimensionale (a destra). Stessa fonte dell'immagine precedente.

Andiamo ora a vedere un po' di pratica…