SUPPORT VECTOR MACHINE

Per comprende a pieno le *Support Vector Machine (SVM)* è necessario fare un passo indietro e trattare le idee che vi stanno alla base, oltre che le origini.

Tutto ebbe inizio attorno agli anni ‘30 grazie alla grande diffusione che ebbe la statistica.

Essa si può considerare molto legata alla t*eoria della probabilità*, ma mentre in quest’ultima conoscendo il processo di generazione dei dati sperimentali (modello probabilistico), si è in grado di valutare la probabilità dei diversi possibili risultati di un esperimento, nella statistica tale processo non è noto in modo completo, anzi diventa esso stesso l'oggetto dell’indagine. Ecco che le tecniche statistiche si prefiggono di indurre le caratteristiche di tale processo sulla base dell'osservazione dei dati sperimentali da esso generati.

Come non citare in questo contesto **Andrej Nikolaevič Kolmogorov**, matematico sovietico, definito il padre della probabilità di cui ne ha stabilito assiomi. Ha inoltre formulato l’omonimo test non parametrico che verifica la forma delle distribuzioni campionarie. Adattandosi ad ogni distribuzione permette infatti di capire se i dati che si hanno a disposizione rispettano o meno una specifica distribuzione a cui si pensa si debbano adeguare *(Test di Kolmogorov Smirnov)*

O ancora personaggi come **Karl Pearson, Ronald Aylmer Fisher**, o ancora **Vladimir Naumovič Vapnik** fondamentale proprio nell’ambito degli algoritmi di SVM.

Andiamo però con ordine; Prima di tutto dobbiamo distinguere fra test parametrici e test non parametrici:

- test parametrici: si sa già che le cose si comporteranno secondo quella determinata regola o processo; quello che non si conosce invece e che si è quindi interessati a trovare, riguarda le caratteristiche, in altri termini i parametri, che contraddistinguono la regola stessa;

Basti pensare ad esempio al decadimento dell’uranio: si è già a conoscenza del fatto che tale processo presenta un’ andamento di decrescita esponenziale (in termini matematici corrispondente al %), dimezzando la sua radioattività di volta in volta al trascorrere di un tempo fissato.

- test non parametrici: caratterizzati dalla non conoscenza del tipo di distribuzione, a differenza di quelli sopra ignorano il comportamento (il processo) caratterizzante il fenomeno in esame (non implicano la stima di parametri statistici come media, deviazione standard, varianza, etc.).

Tendenzialmente i primi tendono essere preferiti in quanto a parità di potenza quelli non parametrici richiedono un numero nettamente superiore di dati; tuttavia fu proprio con Vapnik nell’ambito del SVM che vennero riscoperti.

Con l’arrivo degli anni ‘60 gli statisti si dovettero inoltre confrontare con una nuova sfida: la maledizione della dimensionalità (termine coniato da R. Bellman). Con l’avvento del computer infatti idea diffusa era che si sarebbe stati in grado di processare quantità di dati maggiori rispetto ai periodo precedenti, aumentando al contempo la precisione dei risultati ottenuti. Tuttavia le cose si mostrarono molto diverse da quanto si era immaginato: vi era certamente una capacità di elaborazione superiore ma questo non portò a guadagnarne in precisione bensì a perderne a causa dell’introduzione di un numero di errori superiore.

A cosa si dovette tale particolare fenomeno?

La risposta è alquanto semplice: l’aumento dei punti su cui si lavora (i predittori) comporta che lo spazio che si ha disposizione diventi più rarefatto. Appare evidente quindi come i punti saranno più distanziati fra loro, rendendo più complessa la loro corretta individuazione.

È proprio tale fenomeno che ha cercato di risolvere Vapnik introducendo le *Support Vector Machine*, che mostrano un aumento della loro precisione man mano che la dimensione dello spazio cresce.
Scendiamo ora più nei dettagli di ciò.

Vapnik decide di recuperare la statistica non parametrica e di concentrarsi sul principio di minimizzazione del rischio. Questo evidenzia una netta differenza rispetto alla visione impiegata fino ad ora che si articola su concetto di massima probabilità e su teoremi come il Teorema del limite centrale e la Legge dei grandi numeri.

Teorema del limite centrale

In una popolazione che non segue il modello gaussiano, le medie campionarie, se calcolate su campioni abbastanza grandi, tendono a distribuirsi secondo una legge gaussiana. In altre parole, sommando eventi che presentano una distribuzione casuale, si otterrà sempre come risultato finale una normale.

Basti pensare ad esempio al lancio di due dati e a quale è la probabilità che esca il valore 7. Osserverò che ci sono vari modi con qui si può ottenere tale risultato (5+2, 4+3 …), ecco che man mano che aggiungo i miei esiti la distribuzione che osservo risulta sempre più una normale.

L'immagine sotto spiega molto bene tale fenomeno.

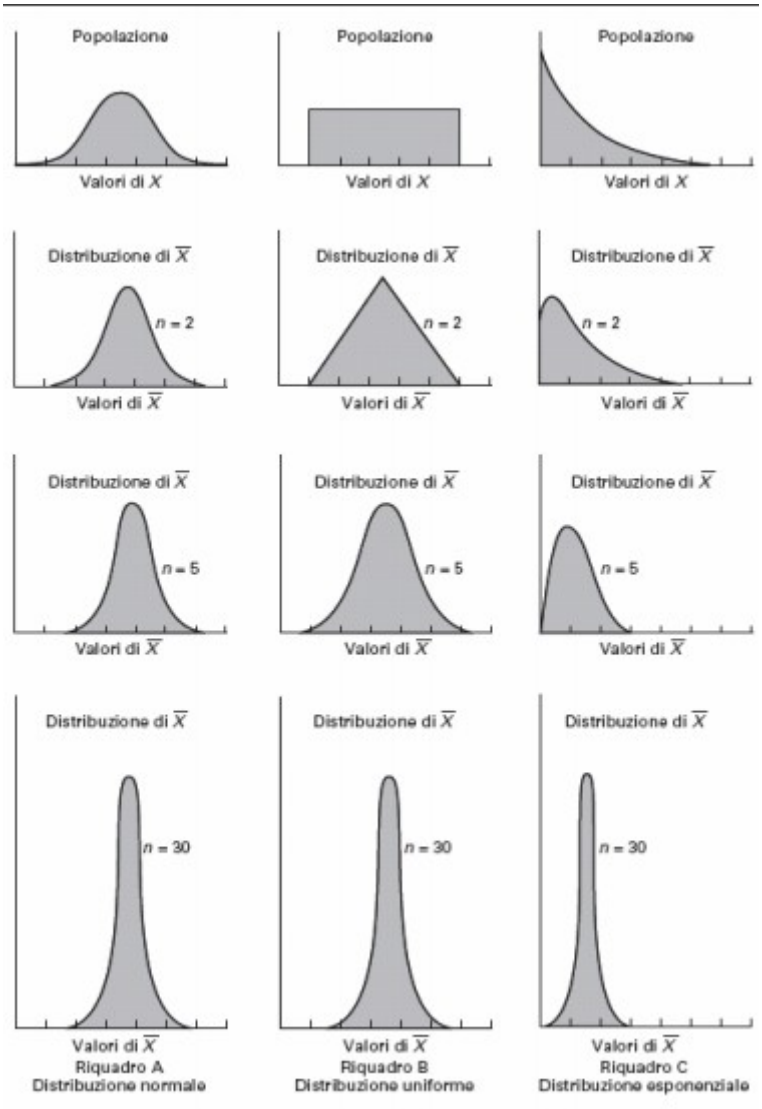

Figura 1: Distribuzione normale

Legge dei grandi numeri

Noto anche come **teorema di Bernoulli**, afferma che all'aumentare del numero delle prove fatte il valore della frequenza tende al valore teorico della probabilità.
Dunque:

- la media calcolata teoricamente è un'approssimazione di quelle sperimentali, ed aumenta la sua precisione al crescere di n;
- viceversa, si può prevedere che i risultati sperimentali mostreranno una media tanto più prossima alla media teorica, quanto più grande sarà n.

Questo teorema fornisce una possibile giustificazione alla **Legge empirica del caso** secondo la quale la frequenza relativa di un evento tende ad stabilizzarsi all'aumentare del numero delle prove.

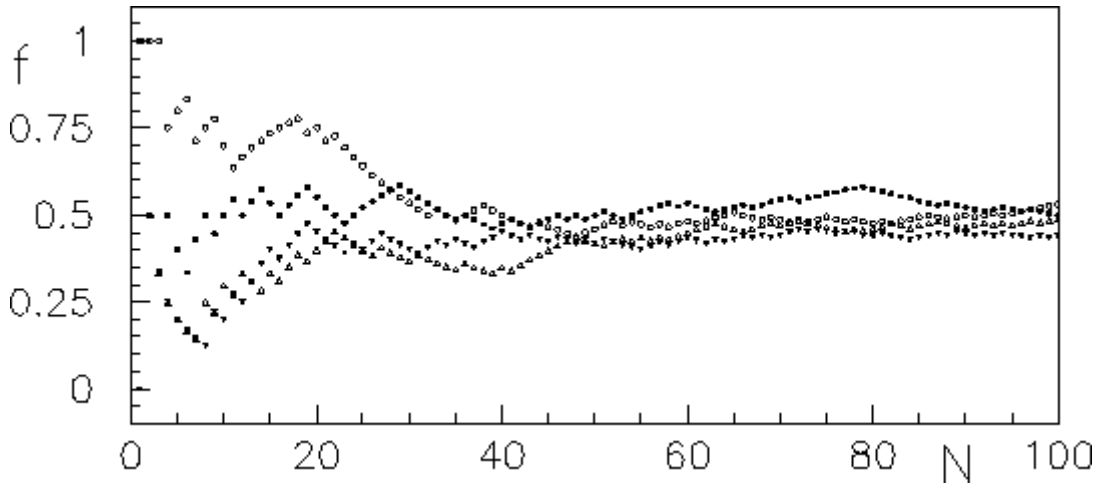

Figura 2: Legge empirica del caso

È evidente come il Teorema del limite centrale e la Legge dei grandi numeri siano fortemente connesse fra loro sancendo le basi della statistica classica.

Criterio di minimizzazione del rischio

Vi è il passaggio dalla probabilità al concetto di rischio: *"Cosa rischio facendo questa previsione?"*. Vapnik infatti introduce l'idea che si hanno a disposizione i predittori con le loro specifiche caratteristiche, ed i dati di test. Quest'ultimi grazie ai risultati che mi forniscono (R2, varianza ecc.) rappresentano l'errore noto, permettendomi poi di risalire all'errore che non conosco.

L'obiettivo in questo contesto è quello di individuare il modello ottimo, esso è tale proprio perché minimizza il rischio di commettere errori, gli fornisce quindi una buona capacità predittiva non andando però a sovraccaricare eccessivamente di predittori il modello (compromesso fra bias e varianza). Questo è proprio ciò che sta alla base del **Teorema di Vapnik Chervonenkis** e delle **SVM.**

Per capirlo meglio vediamo cosa sono le *Support Vector Machine*.

Idealmente dobbiamo figurarci in uno spazio dove vi sono delle osservazioni. L'obiettivo in tutto ciò consiste nell'individuare l'iperpiano che taglia meglio questo spazio e che permette di classificare i vari punti in due classi distinte. Uno spazio qualsiasi può infatti venire attraversato da un infinito numero di iperpiani, noi siamo interessati a quello ottimo, inoltre la potenza delle SVM sta proprio nel fatto che non temono la dimensionalità anzi, più il grado del polinomio è elevato più lavorano meglio, in quanto puntano ad

individuare solo i punti più importanti: **i vettori di supporto**. Sono proprio quest'ultimi che permettono di individuare il l'iperpiano in questione.

A seguire l'immagine mostra meglio quanto detto.

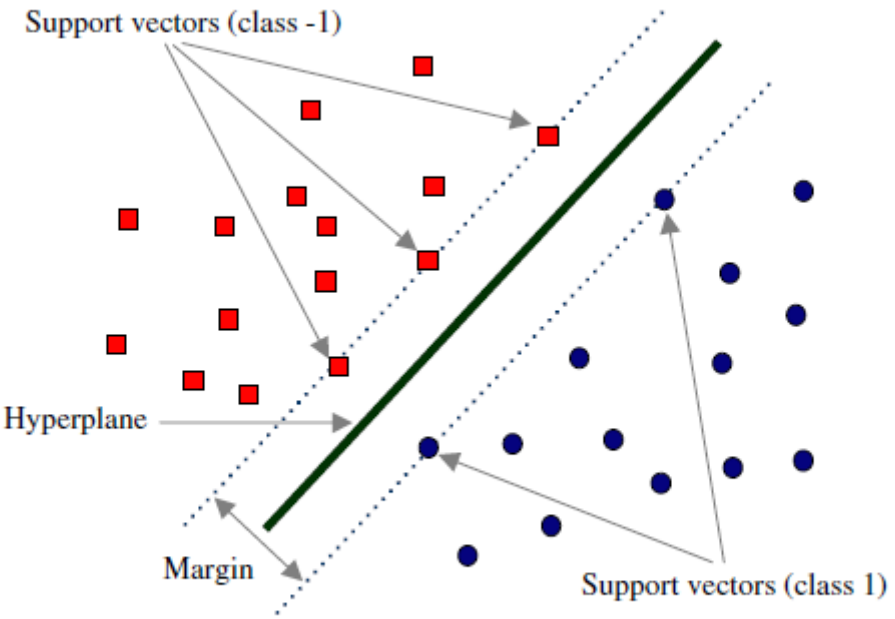

Figura 3: Support Vector Machine

Si può notare come il *margine*, non sia altro che la distanza tra i vettori di supporto delle due classi, alla cui metà si posiziona l'iperpiano (o retta nel caso si stia lavorando a due dimensioni).

Come si seleziona l'iperpiano migliore?

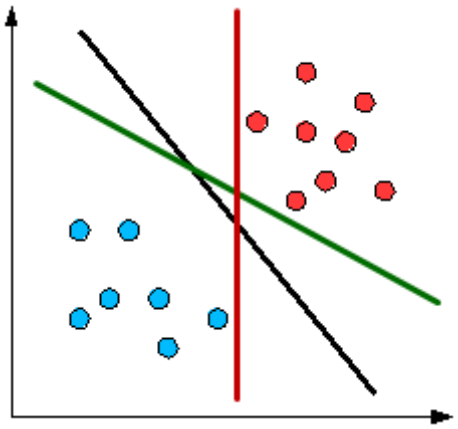

Figura 4: Iperpiani SVM

Come mostra l'immagine a fianco, si seleziona l'iperpiano che si trova alla **massima distanza** dai vettori di supporto delle varie classi; maggiore è infatti la distanza fra di lui e i punti, e più cresce la fiducia che la classificazione sia stata svolta correttamente.

È proprio in questo ragionamento che si posiziona il *Teorema di Vapnik Chervonenkis:*

"La macchina con la capacità più piccola è la migliore."

Principio che si muove in simbiosi con il rasoio di Occam:

"A parità di fattori la spiegazione più semplice è da preferire"

Da notare che qui non va confuso il concetto di "capacità minima" con quello di *"semplicità" nel senso di un ridotto numero di parametri, ma come* ***dimensione Vapnik Chervonenkis*** *che deve quindi puntare ad essere la più* piccola possibile (*infatti vi possono essere modelli con tanti parametri che hanno una bassa VC e viceversa).*

A questo punto appare spontanea una domanda:

Cos'è la dimensione Vapnik Chervonenkis?

Essa rappresenta la cardinalità dell'insieme più grande frantumabile.

In generale per uno spazio a k dimensioni corrisponde a:

$$VC(H) = k+1$$

La dimensione VC per un classificatore lineare è almeno 3 perché non si riesce a frammentare 4 punti.
Ma mostriamo alcuni esempi che spiegano tale cosa.

Nel caso di 2 punti avremo una VC(H) >= 1

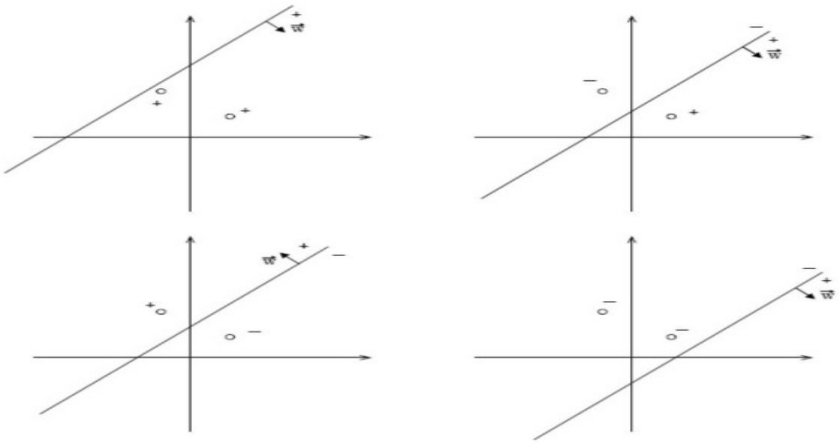

Figura 5: VC-dimension 2 punti

Con 3 punti avremo una VC(H) >= 2

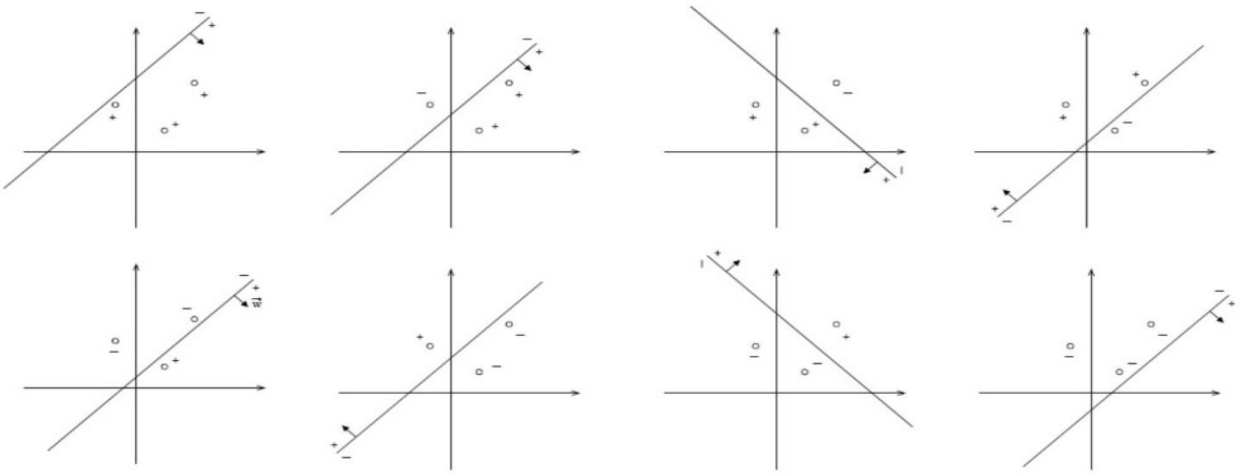

Figura 6: VC-dimension 3 punti

Con 4 punti

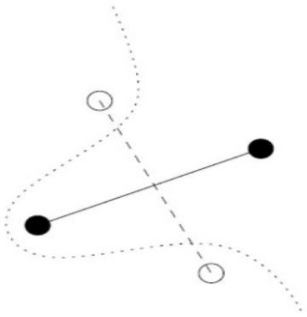

Figura 7: VC-dimension 4 punti

Dall'immagine possiamo notare La dimensione VC per un classificatore lineare è almeno 3 perché non si riescono a frammentare 4 punti!!

$$VC(H) = 3$$

Ecco che la dimensione VC viene usata per classificare i diversi tipi di algoritmi in base alla loro complessità.

In particolare con l'aumentare della complessità di un modello, si passa dal sottoadattamento a quello eccessivo (da underfitting a overfitting); ciò rappresenta il passaggio chiave che permette di comprendere il sopra citato Teorema di Vapnik Chervonenkis. L'aggiungere complessità infatti non fornisce una certezza sicura che il modello sia migliore di quello meno complesso. Nello specifico tale aggiunta si mostra valida fino a un certo punto, dopo di che avremo una discesa causata da un sovraccarico dei dati di addestramento.

Un altro modo di pensare la cosa, ma che racchiude il medesimo significato, è attraverso i concetti di *Bias* e *Varianza* e di come sia necessario trovare il giusto trade-off fra questi due indici. Un modello a bassa complessità infatti se da un lato a causa del suo ridotto potere espressivo non permette di fidarsi completamente dei valori ottenuti, dall'atro recupera in semplicità, cosa che porta prestazioni molto prevedibili e bassa varianza. Al contrario, un modello più complesso avrà un bias inferiore proprio per la sua maggiore espressività, ma presenterà una varianza più elevata a causa della presenza di più parametri.

Appare quindi evidente come ad un certo livello di complessità del modello esisterà un equilibrio ideale tra distorsione e varianza, in corrispondenza della quale non si è né insufficienti né adatti ai propri dati. Ma questo non ci riporta altro che al Teorema di Vapnik Chervonenkis: si dovrebbe mirare a scegliere un modello con un livello di complessità (una bassa VC) appena sufficiente per svolgere la classificazione in questione.