ANALISI DEI DATI PER LE IMPRESE INDUSTRIALI (5)

Apprendimento supervisionato

Nell'ultimo articolo abbiamo affrontato l'apprendimento **non supervisionato**, ora iniziamo quello **supervisionato**, che per la vastità del tema, non si esaurirà in una sola trattazione.

Gli algoritmi di apprendimento supervisionato utilizzano **dati che hanno già il risultato per istruire il sistema**. Il modello creato è poi usato per predire il valore delle nuove osservazioni. Esiste un'ampia varietà di algoritmi di apprendimento supervisionato che possono essere raggruppati in due categorie principali, **la regressione e la classificazione**.

A seconda delle caratteristiche dei dati, ci può essere un'attività di regressione, per cui si ricerca un valore numerico (pubblicità e relativa vendita…), o qualitativo per la classificazione (Maschio-Femmina, Sì-No, …).

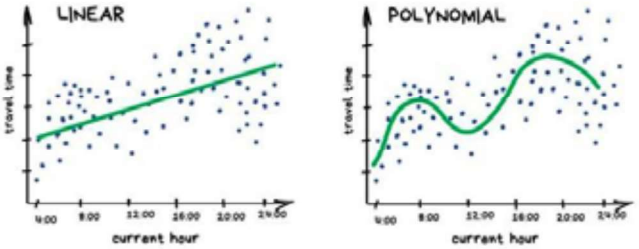
REGRESSIONE

Se vogliamo ipotizzare il prezzo di una casa, guardiamo tutte le offerte simili a quella di nostro interesse e raccogliamo le informazioni più importanti come l'anno di costruzione, il numero degli ambienti e dei bagni, …, e i prezzi. In questo modo, facendo lavorare l'algoritmo e inserendo i dati in questione dell'appartamento che ci interessa, capiremo se il prezzo ipotizzato sarà in linea con il mercato o meno.

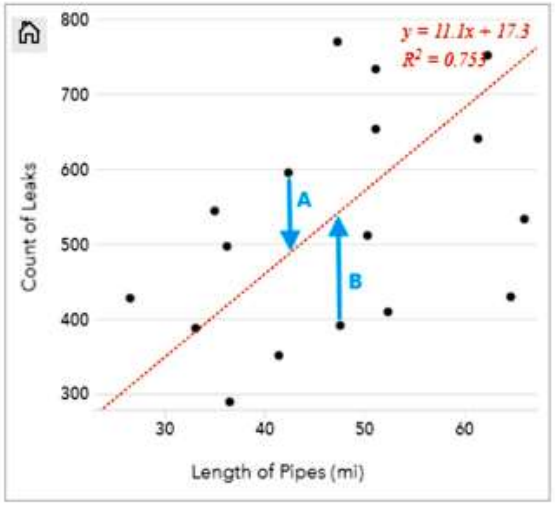
La regressione lineare **cerca la relazione tra una variabile target continua** (il prezzo) **con una o più variabili indipendenti** (anno, ambienti, bagni, …), adattando un'equazione lineare ai dati stessi.

Quando la linea che rappresenta la relazione è retta, è **una regressione lineare.** In casi più complessi, quando la linea è una curva si chiama **polinomiale**. Questi sono i due tipi principali di regressione.

Il grafico a dispersione di fianco (lineare) mostra una correlazione positiva tra una variabile indipendente (asse x, ad esempio spesa per pubblicità) e una variabile dipendente (asse y, volume vendite). All'aumentare di uno, aumenta anche l'altro.



Come si vede dalla figura 2 (sotto), **un modello di regressione lineare cerca di adattare al meglio la retta di regressione ai punti rappresentativi dei dati**. La tecnica più comune che si utilizza è quella dei minimi quadrati. Con questo metodo la migliore retta di regressione si trova minimizzando la somma dei quadrati della distanza tra i punti dati e la retta di regressione stessa. Queste distanze sono gli "errori" e si deve usare il quadrato per avere solo valori positivi.



CLASSIFICAZIONE

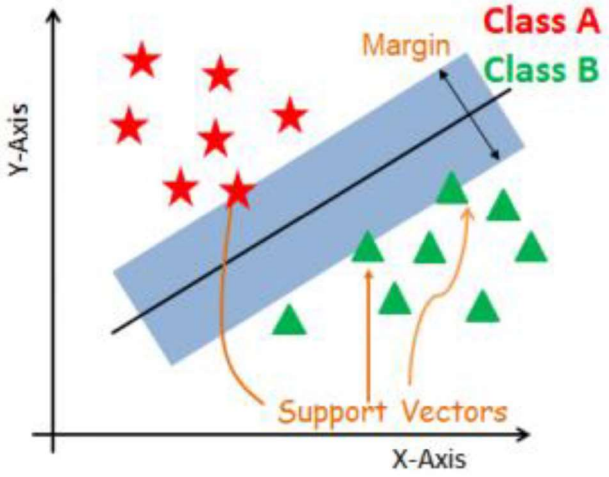
La classificazione oggi è molto utilizzata per filtri antispam, sentiment analysis, per scoprire le frodi, ad esempio con le carte di credito, e per molti altri scopi, anche in ambito scientifico e di ricerca, prevedendo **le categorie di appartenenza,** due o più.

Gli algoritmi più popolari sono: **Support Vector Machine, Naive Bayes, Decision Tree & Random Forest, Logistic Regression, K-NN**,…. Oltre a quelli citati, che sono la base storica, vi sono tante evoluzioni, con specialità mirate e con prestazioni vicine a quelle delle reti neurali per dati di non enormi dimensioni. Tutti questi algoritmi erano noti da molti anni, ma solo i recenti sviluppi nei processori e nelle memorie hanno permesso il loro uso in maniera estesa.

SUPPORT VECTOR MACHINE

Le Support Vector Machine o SVM (in italiano Macchine a Vettori di Supporto) sono modelli di classificazione il cui obiettivo è quello di trovare **la linea di separazione delle classi** che massimizza il margine tra le classi stesse, dove con margine si intende la distanza minima dalla linea ai punti delle due classi. Con questo sistema si usa un numero di punti ridotto rispetto a tutti i dati perché lavoriamo solo con quelli di frontiera, che sono quelli più

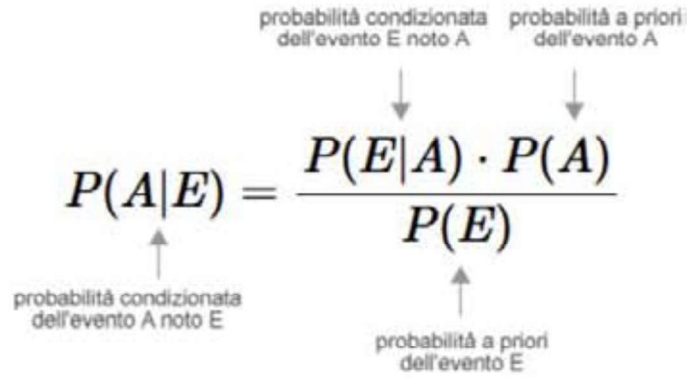
difficilmente separabili. Si ripresenta il problema del numero delle dimensioni dei dati, che rendono il calcolo complesso. Per esempio, se usiamo solamente "lunghezza" e "larghezza" per la classificazione, saremo in uno **spazio bidimensionale** e il confine per la decisione è **una linea**. Se utilizziamo 3 parametri**,** il confine diventa **un piano nello spazio tridimensionale.**
Se usiamo più di 3 caratteristiche, il confine decisionale diventa **un iperpiano**, che è davvero difficile da visualizzare. Succede in realtà che la linea di separazione non sia così ben definita come nell'immagine sootostante e occorre utilizzare delle tecniche, definite **kernel**, che utilizzano uno stratagemma matematico (aumentano il numero delle dimensioni per definire un iperpiano di livello superiore) per trovare una soluzione. Le SVM sono uno dei pochi algoritmi che riesce a fornire risultati attendibili se il numero delle osservazioni è minore del numero delle caratteristiche. Ricordando il foglio excel dei precedenti articoli, se il numero delle righe è inferiore al numero delle colonne, si possono utilizzare. Le SVM sono utilizzabili in



qualche caso per regressioni.

NAIVE BAYES

Il teorema di Bayes, enunciato da Thomas Bayes (1702-1761), utilizza due caposaldi della teoria della probabilità: **il teorema della probabilità composta e quello della probabilità assoluta.** Per esempio, si può calcolare la probabilità di un errore. Una persona che ha eseguito un controllo medico, può in realtà essere malata anche se il risultato è negativo, oppure viceversa. Ciò si può fare conoscendo la frequenza con cui si presenta la malattia e la percentuale di efficacia del test diagnostico. **L'importanza del teorema di Bayes per la statistica è tale che la divisione tra le due scuole principali** (statistica bayesiana e statistica frequentista) **nasce dall'interpretazione che si dà al teorema stesso.**
Naive Bayes presuppone che tutte **le caratteristiche siano indipendenti l'una dall'altra** e che non vi sia alcuna correlazione tra le caratteristiche stesse. Nella realtà le caratteristiche sono spesso connesse e questo è il motivo per cui questo algoritmo è chiamato **"naive" (ingenuo).**
Un'automobile può essere classificata come "sportiva" se è bassa, ha due posti e supera i 200 km/h. Sono tre caratteristiche differenti. Naive Bayes valuta ogni caratteristica come un contributo indipendente alla probabilità complessiva. Classifica sportiva l'auto e non prende in considerazione le possibili correlazioni tra le caratteristiche. Malgrado queste ipotesi semplificative, l'algoritmo funziona molto bene.
Se abbiamo due eventi A e E, la formula enunciata da Bayes è:



- P(A) è la probabilità a priori di A, cioè la probabilità dell'evento A senza considerare l'evento E. È anche detta probabilità marginale di A.
- P(E) è la probabilità a priori di E, cioè la probabilità dell'evento E senza considerare l'evento A. È anche detta probabilità marginale di E.
- P(A|E) è la probabilità condizionata dell'evento A, considerando le informazioni sull'evento E. È anche detta probabilità a posteriori dell'evento A perché dipende dal valore di E.
- P(E|A) è la probabilità condizionata dell'evento E, considerando le informazioni sull'evento A. È anche detta probabilità a posteriori dell'evento E perché dipende dal valore di A.

P(A|E) si legge "Probabilità di A dato E"; P(E) normalmente non si considera ma non può essere zero.

Nel prossimo articolo proseguiremo l'analisi degli algoritmi di classificazione.

Fonti e figure:
http://www.developersmaggioli.it/blog/support-vector-machine/
http://www.eage.it/machine-learning/naive-bayes
http://doc.arcgis.com
Wikipedia.