Newsletter Phedro – maggio 2007

I modelli MACHINE LEARNING

Abbiamo già visto la volta scorsa che i modelli di *credit assessment* possono essere classificati sostanzialmente in quattro categorie:

1. modelli euristici;
2. modelli *machine learning*;
3. modelli causali;
4. modelli ibridi.

Approfondiamo in questo articolo le tematiche riguardanti la seconda categoria di modelli:

I modelli *machine learning*

Al contrario dei metodi euristici, i quali nascono dall'esperienza soggettiva maturata da esperti di credit assessment, i modelli machine learning "apprendono" da dati contenuti in campioni opportunamente creati e contenenti indici di bilancio di aziende sane ed aziende insolventi; ciò significa che, per mezzo di procedure statistiche, tali modelli cercano di verificare ipotesi sul potenziale merito creditizio delle società. Essenzialmente tali modelli valutano, in base ai dati sui quali vengono allenati, se ci si può attendere un default quando i valori di certi indicatori siano più alti o più bassi del dovuto.
Spesso questi modelli si basano su di un processo di ottimizzazione di una c.d. funzione di verosimiglianza; tale fase permette la selezione dei fattori più rilevanti per l'analisi del merito creditizio e consente di determinare quali siano i pesi ottimali con i quali tali fattori intervengono nell'analisi. Tali pesi vengono utilizzati per la definizione di un'altra funzione, spesso denominata probability of default (PD), che, come suggerisce il nome, fornisce la probabilità che l'azienda analizzata si riveli insolvente in un certo periodo temporale.
Appare chiaro, da quanto descritto, che l'accuratezza dell'analisi condotta con i metodi machine learning dipenda tanto dalla qualità delle procedure matematiche adottate (scelta della funzione errore e della funzione PD, processo di ottimizzazione, ecc.) quanto dalla qualità delle informazioni contenute nel campione utilizzato per l'allenamento del modello.
Proprio la bontà del database a disposizione rappresenta il punto più debole di questo tipo di metodologie. Perché un metodo machine learning possa essere efficace, infatti, si dovrebbe prestare attenzione ad alcuni aspetti:
1. Il database dovrebbe contenere un numero elevato di osservazioni così da poter ritenere statisticamente attendibile ogni tipo di inferenza condotta su di esso. Essendo presenti sul mercato alcune società, come già sopra ricordato, che si occupano di

collezionare, ripulire e rivendere dati di bilancio di aziende operanti su tutto il territorio mondiale, questo della numerosità del database sembrerebbe essere un problema da poco; in realtà è difficile possedere informazioni sullo stato di default, come più volte ricordato. Le ECAI, ad esempio, possedendo le sole informazioni sul fallimento (e non su incaglio e sofferenza), se basassero le loro valutazioni sul risultato ottenuto da un metodo machine learning, otterrebbero una sottostima della reale probabilità di default delle aziende analizzate. Tale risultato sarebbe poco utile (o rischioso) per quelle banche che volessero utilizzarlo per il calcolo dei coefficienti di ponderazione.
2. Il database dovrebbe essere rappresentativo dell'intero territorio italiano e di tutte le realtà produttive esistenti in Italia. Questo vincolo è facilmente giustificabile: qualora ci si ritrovasse, infatti, a dover valutare un'azienda molto diversa per caratteristiche economico finanziare da quelle "viste" dal modello, lo score (ed il conseguente rating) che si otterrebbe non sarebbe attendibile.
3. La serie storica di dati posseduti dovrebbe essere sufficientemente estesa, così da permettere di studiare le aziende in un periodo temporale superiore ad una congiuntura economica.
Diamo di seguito, senza alcuna pretesa di completezza, una breve descrizione di alcune metodologie che afferiscono a questa categoria:

Analisi discriminante multivariata

In generale l'obiettivo dell'analisi discriminante multivariata (MDA) nel contesto del credit assessment è di distinguere tra creditori solventi ed insolventi il più accuratamente possibile usando una funzione che contenga diversi fattori indipendenti utili alla valutazione del merito creditizio. Un particolare caso di MDA è la analisi discriminante lineare: introdotta per la prima volta dal prof. Altman nel 1968 per l'analisi delle aziende americane, consiste nella somma ponderata di una combinazione di indicatori tale per cui, fissato un punto di cut-off, le due distribuzioni (quella delle aziende sane e quella delle aziende insolventi) risultino separate il più possibile.
Ipotesi che sta alla base dell'implementazione dei modelli MDA è che la relazione tra "livello di bontà" di una azienda e valori delle variabili esplicative sia monotóna.

Modelli di regressione

Generalmente in questa categoria si inseriscono i modelli logit e probit : i tratti che caratterizzano questi modelli, sostanzialmente identici, possono essere così riassunti: regredendo una funzione non lineare degli indici di bilancio sulla variabile risposta di natura dicotomica 0/1 (non default/default), forniscono la probabilità condizionale che una determinata impresa si venga a trovare nella classe delle imprese insolventi; la probabilità è condizionata alla realizzazione delle sue variabili, ovvero degli indicatori economico-finanziari.
Se la distribuzione di probabilità è logistica la funzione si dirà logit, se viceversa sarà normale il modello è denominato probit.
Anche per i modelli di regressione si ipotizza che possa ritenersi monotóna la relazione che intercorre tra la probabilità di default e i valori delle variabili esplicative.

Reti neurali

Con tale termine si suole indicare una ampia famiglia di tecniche machine learning; il termine neural network ha origine come modello matematico di quello che in passato si riteneva essere il meccanismo di funzionamento del cervello animale. Una rete neurale è sostanzialmente uno schema di regressione non lineare a due o più stadi così costituito: vi sono strati di neuroni che, essendo collegati tra loro da ideali bottoni sinaptici, mettono in relazione le variabili di input con quelle di output. Ma cos'è un neurone? Si pensi ad esso semplicemente come ad una funzione matematica (detta funzione primitiva) delle variabili esplicative. Il processo di allenamento della rete neurale consiste nel trovare i coefficienti delle funzioni di rete (generalmente delle sigmoidi) che legano tra loro i neuroni (e quindi esprimono le relazioni che intercorrono tra le variabili di input a quelle di output) per mezzo di una minimizzazione di una funzione obiettivo espressa spesso dallo scarto quadratico medio tra il valore reale dell'output ed il valore calcolato.
Pur riuscendo a catturare le relazioni non-lineari e non-monotóne che intercorrono tra la PD e le variabili esplicative, tali modelli presenta no numerosi inconvenienti: arbitrarietà nella scelta di molti parametri, difficoltà di interpretazione dei risultati (spesso vengono indicati come black box) e, spesso, scarsa accuratezza.

Support vector machines (SVM)

Senza addentrarci troppo nei tecnicismi di quello che può essere considerato uno dei metodi di machine learning che sta riscuotendo maggior successo nella comunità scientifica, si può dire che le SVM sono delle metodologie di clustering la cui idea guida è quella di mappare i dati di allenamento in un nuovo spazio (detto feature space) nel quale sia facile individuare degli iperpiani che separino opportunamente i dati.
Tali algoritmi non solo permettono di catturare la natura non-lineare e non-monotóna della probabilità di default, ma permettono di separare quasi perfettamente le aziende insolventi da quelle sane; per contro sono di difficile implementazione ed applicabili a dataset limitati, essendo la loro complessità crescente con il numero di dati di input.

Maximum expected utility (MEU)

A differenza degli altri modelli di machine learning che, come abbiamo visto, pur usando un ottimo modello numerico per affrontare il problema in esame soffrono del fatto di non originare da considerazioni di natura finanziaria, il modello MEU (Maximum Expected Utility), sviluppato dagli analisti di Standard & Poor's, coniuga un complesso modello matematico ad una interpretazione economico- finanziaria del risultato che si vuole ottenere: l'idea base della metodologia, infatti, è cercare una misura di probabilità di default massimizzando la funzione utilità di un ipotetico investitore. Come un giocatore che dovesse scommettere su una corsa di cavalli farebbe le sue puntate in modo tale da massimizzare la vincita, così si ipotizza che un investitore dovrebbe scegliere la strategia di investimento migliore ovvero quella che massimizzi la propria utilità rispetto ad un modello in cui crede. L'approccio MEU, quindi, consiste nel ricercare asintoticamente questo risultato selezionando di volta in volta il modello che massimizza la funzione utilità sui dati non noti o sconosciuti. L'aspetto interessante della metodologia è che la massimizzazione della qualità del

modello numerico non è semplicemente mono obiettivo, bensì multi obiettivo: contemporaneamente (secondo un approccio di Pareto) si cercherà la consistenza con i dati conosciuti (normalmente denominato training set) e con la misura di probabilità a cui l'investitore crede prima di conoscere i dati. Il peso relativo tra i due obiettivi viene definito da un parametro definito dall'utente.

Nel prossimo articolo descriveremo gli algoritmi afferenti alle ultime due categorie: i modelli causali e i modelli ibridi.

www.phedro.it – info@phedro.it