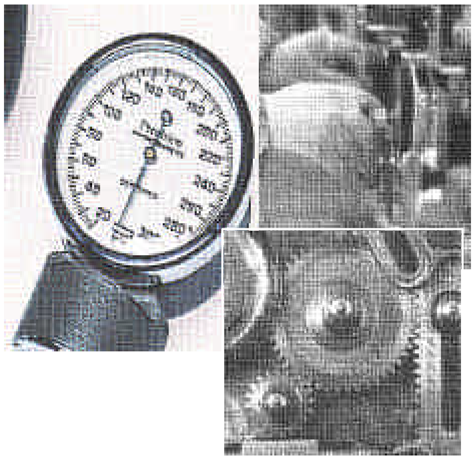


1.3 La previsione del rischio di insolvenza con modelli quantitativi

Cristina Omacini

1.3.1 La valutazione automatica del rischio: vantaggi e limiti

1.3.2 I modelli di previsione delle insolvenze basati sull'analisi discriminante: un esempio pratico

1.3.3 Limiti dell'esempio e indicazioni per un modello più realistico

✓ *Highlights*

In questo capitolo proponiamo alcune riflessioni sui limiti e sui vantaggi dei modelli di valutazione automatica del rischio. Viene inoltre proposto un esempio concreto di un possibile algoritmo quantitativo di previsione delle insolvenze, basato sull'analisi discriminante e su un set di dati semplificato. Infine, si mettono in risalto alcuni limiti di questo esempio didattico, fornendo indicazioni sulla costruzione di un modello maggiormente adatto a trattare dati reali.

1.3.1 La valutazione automatica del rischio: vantaggi e limiti

I modelli di previsione delle insolvenze basati su metodologie di carattere statistico si inquadrano nella fase quantitativa del processo di valutazione del merito creditizio. Tali modelli consentono una valutazione automatica delle aziende sottoposte ad analisi fornendo per ognuna un numero (*score*) che ne individua in misura sintetica lo stato di salute dal punto di vista dei creditori.

Come si costruisce un modello quantitativo

Al fine di poter applicare un modello quantitativo è necessario disporre di un campione composto da clienti che in passato si sono rivelati affidabili o insolventi. Successivamente, occorre individuare una combinazione di variabili che abbiano un contenuto informativo sufficiente a discriminare piuttosto nettamente tra il gruppo dei prenditori sani e quello degli insolventi. Una volta scelti il campione e le variabili, ognuna viene inserita in una funzione (per esempio, le viene assegnato un peso all'interno di una sommatoria ponderata), che contribuisce a determinare lo *score* di ogni cliente.

In base al punteggio ottenuto, l'azienda sarà considerata più o meno rischiosa. Per esempio, se le imprese affidabili presentano punteggi generalmente alti e quelle insolventi punteggi bassi, si ritiene che il cliente che abbia ottenuto uno score molto alto sarà "probabilisticamente" solvibile mentre quello con uno score basso si rivelerà con una buona probabilità insolvente.

Una volta calcolati i punteggi e le relative fasce di rischio, è possibile che la banca stabilisca un valore soglia (*cut-off score*), al di sotto del quale le richieste di finanziamento vengono respinte o sottoposte a revisione. Naturalmente la scelta del punto di *cut-off* dipende dalla propensione al rischio: un livello più elevato del valore soglia comporta minori possibilità di perdite e viceversa[1].

---

[1] Ovviamente, tuttavia, un valore soglia troppo elevato massimizza il rischio di respingere un numero consistente di prenditori "sani".

I vantaggi

I principali vantaggi presentati dai modelli di scoring rispetto alle tecniche di analisi tradizionali sono:

- ✓ una maggior oggettività dei risultati;
- ✓ una maggiore rapidità di analisi delle singole posizioni, con conseguente riduzione dei costi delle fasi di istruttoria e una più frequente verifica dell'andamento del rapporto di fido;
- ✓ la possibilità di disporre di un quadro di riferimento uniforme sulla qualità del credito, utile anche per determinare i margini di autonomia delle filiali.

Oggettività

Con riguardo al primo aspetto, è opportuno sottolineare la profonda diversità che separa i modelli di scoring dalle applicazioni automatizzate di tipo soggettivo e univariato largamente utilizzate in Italia.
Molte società di software offrono infatti modelli di monitoraggio automatico nei quali la singola banca può scegliere liberamente le variabili ritenute rilevanti e determinare la loro importanza relativa nel comporre il punteggio complessivo. La scarsa utilità a fini previsivi di questi procedimenti origina dal fatto che la scelta delle variabili e dei pesi deriva da valutazioni soggettive dell'analista e, pertanto, non generalizzabili. Allorché simili decisioni sono affidate ad un modello statistico, i coefficienti di scoring assumono invece carattere oggettivo, in quanto espressione di procedimenti quantitativi basati sui dati e non di considerazioni personali.

Potenza

Il secondo vantaggio consiste nella possibilità di sottoporre ad analisi preliminare un numero rilevante di imprese riducendo i tempi di istruttoria ed accrescendo la capacità competitiva dell'azienda di credito[2]. I modelli inoltre consentono potenzialmente il monitoraggio dell'intero portafoglio prestiti (o almeno del segmento per cui sono stati costruiti[3]) con cadenze ravvicinate. Ciò permette alla banca di intraprendere tempestivamente le azioni operative per il recupero funzionale del credito o per ristabilire un corretto rapporto di clientela.

Uniformità

L'ultimo importante vantaggio consiste nell'introduzione di un quadro di riferimento omogeneo per le decisioni attinenti agli affidamenti e all'autonomia decisionale delle filiali. Dall'analisi degli score medi per settori e rami di attività economica, nonché per classi dimensionali e per aree geografiche, è infatti possibile derivare utili informazioni per quanto concerne il grado di rischiosità associato a determinati aggregati settoriali e/o dimensionali. La rilevazione dei coefficienti medi per singola filiale fornisce inoltre indicazioni in merito all'ambiente in cui essa opera ed all'efficacia dimostrata nella selezione della clientela.

I limiti

I limiti dei modelli di valutazione automatica del rischio possono essere analizzati sotto differenti punti di vista. In questa sede non affronteremo le problematiche connesse alle diverse metodologie statistiche[4] che, pur se rilevanti, sono al di fuori degli obiettivi di questo capitolo. Cercheremo invece di analizzare i limiti del modello da un punto di vista più squisitamente economico-aziendale.

Instabilità

Un primo problema di fondo, che vale per qualsiasi modello statistico, riguarda la stabilità nel tempo della capacità diagnostica. Cambiamenti strutturali nel ciclo economico, mutamenti nei criteri decisionali delle banche o nella normativa possono modificare considerevolmente le relazioni tra variabili. Risulta quindi necessaria una periodica verifica delle performance dei modelli ed una ri-stima quando la loro

---

[2] Ciò vale soprattutto per talune categorie di prestiti, quali ad esempio i prestiti personali o alle imprese minori, per i quali la velocità di erogazione costituisce fattore strategico di successo.

[3] Cfr. capitolo 1.1.

[4] Si veda in questo senso Altman (1981), pp.167 e ss.

efficacia discriminante tende a ridursi. La frequenza con cui si manifestano le esigenze di revisione e la loro complessità determinano evidentemente il costo di gestione dei sistemi di valutazione automatici.

Costi di errata classificazione

Un altro costo (comune, peraltro a qualsiasi strumento di analisi del merito creditizio) è quello degli errori compiuti dal modello[5]. E' evidente che un modello poco accurato può risultare eccessivamente oneroso (in termini di perdite non evitate o di buoni clienti persi) anche se il suo utilizzo e la sua manutenzione non comportano spese rilevanti.

Qualità dei dati di input

Altri limiti possono derivare dai dati di input dell'analisi. L'uso di dati contabili suscita perplessità circa la loro effettiva capacità di rappresentare correttamente la situazione aziendale: il ritardo con cui i bilanci divengono disponibili, costringe a fondare le valutazioni su eventi accaduti parecchi mesi prima. A ciò si aggiungono i rischi derivanti da fenomeni di *window dressing* e dall'adozione di bilanci in forma abbreviata da parte di molte imprese minori.
Per ovviare parzialmente a questi limiti, i dati di bilancio vengono di norma integrati da altre informazioni reperibili in forma sistematica e standardizzata: i dati di Centrale dei rischi[6], o quelli sull'andamento del rapporto tra l'impresa e la banca. I modelli più sofisticati[7], inoltre, prevedono l'introduzione di variabili qualitative. L'uso di tali variabili tuttavia, oltre ai delicati problemi di definizione dei sistemi di codifica[8], pone numerosi problemi nell'applicazione concreta dei modelli. Ad esempio, nei modelli in cui vengano prese in considerazione variabili qualitative di natura organizzativa (connesse alle caratteristiche del management, alla composizione dei consigli di amministrazione, al prestigio dell'azienda, ecc.) sorge la difficoltà di valutare la correttezza dei segni delle variabili. In sede di prima valutazione di credito, inoltre, l'analisi ex-ante della competenza gestionale è difficilmente attuabile con l'analisi di bilancio; solo ex-post alcuni indicatori riescono ad individuare punti di debolezza dell'azienda.
I limiti dovuti all'imprecisione o alla scarsità di dati di input risultano sovente assai marcati presso le banche italiane, per effetto di carenze dei sistemi informativi e della scarsa propensione ad archiviare con regolarità i dati passati: non di rado, dunque, si ha difficoltà nel costruire una base di dati sufficientemente ampia da consentire la costruzione di un ampio dettaglio di indicatori quantitativi.

1.3.2 I modelli di previsione delle insolvenze basati sull'analisi discriminante: un esempio pratico

Il modello di previsione delle insolvenze più semplice (eppure ancora largamente usato nelle applicazioni reali) si basa sulla analisi discriminante[9] suggerita da Fisher (1936). Si tratta, in sostanza, di una tecnica di classificazione che usa informazioni tratte dai dati

---

[5] Il problema dei costi associati ai diversi tipi di errore è particolarmente rilevante nell'ambito delle concessioni e revisioni dei prestiti: per un'azienda di credito è più oneroso concedere un prestito ad un'impresa che poi diventa insolvente, che non concederlo ad una che poi si scopre essere sana. Simili "asimmetrie" nella struttura dei costi degli errori possono essere agevolmente incorporate in un modello automatico, "insegnandogli" a evitare soprattutto gli errori maggiormente onerosi. A tal fine si veda Anderson (1958).
Alcune considerazioni sulle stime dei costi di errata classificazione si possono trovare in Altman (1977).

[6] Alberici (1989), Caselli (1998).

[7] La Centrale dei Bilanci ha messo a punto un apposito modulo di analisi qualitativa che consente di completare l'analisi del bilancio delle società. Cfr. Varetto (1999), p.300.

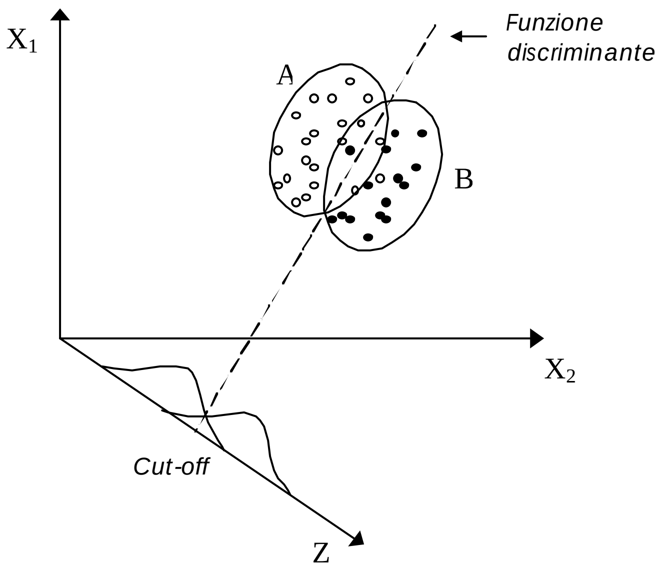
[8] Una delle più comuni codifiche è la cosiddetta *codifica disgiuntiva completa*, secondo la quale ogni variabile dà origine a tante variabili fittizie quante sono le modalità previste. Ognuna di queste, che vengono definite *variabili indicatrici*, assumerà per ciascun individuo il valore 1 o 0, a seconda che quell'individuo presenti o meno quella modalità. La tecnica che è più spesso utilizzata su matrici del genere è l'*Analisi delle Corrispondenze*. Si vedano Piccolo-Vitale (1981) e Morrison (1976).

[9] Per approfondimenti si veda Lachenbruch (1979).

del campione per tracciare un confine (per esempio, una linea retta sul piano cartesiano) capace di separare piuttosto nettamente il gruppo delle aziende affidabili da quello delle aziende insolventi. Tale separazione avviene in base ad una cosiddetta *funzione discriminante*.

La figura 1.3.1. rappresenta il modello di Fisher nel caso più semplice in cui i due gruppi di imprese affidabili (A) e insolventi (B) siano identificati da due sole variabili X_1 e X_2: sull'asse Z è rappresentato lo score generato combinando tra loro le due variabili originarie. Il meccanismo di costruzione di questo score diverrà chiaro nel seguito del paragrafo.

Figura 1.3.1
Sintesi grafica dell'analisi discriminante lineare



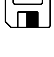
Un esempio concreto

Per meglio comprendere la natura dell'analisi discriminante, introduciamo ora un semplice esempio didattico.

Supponiamo di avere a disposizione un campione di 20 aziende, di cui 10 rivelatesi in passato insolventi e 10 affidabili. Ipotizziamo inoltre di aver calcolato per ciascuna impresa un certo numero di indicatori[10] e di avere isolato, sulla base di test statistici e di nostre valutazioni, un sottoinsieme ridotto formato da quattro variabili su cui stimare il modello.

Le variabili prescelte sono:

X_1= Roe
X_2= attivitàcorrenti/passivitàcorrenti
X_3= oneri finanziari/ricavi
X_4= debiti totali v. banche/attivo



La tavola 1.3.1 (e, in forma completa, il dischetto allegato a questo manuale) riportano i dati relativi alle aziende del nostro campione. Si noti che la variabile "stato", che rappresenta lo stato di salute dell'azienda (1= sana, 2= insolvente), è la variabile dipendente del modello.

[10] Naturalmente gli indicatori devono essere antecedenti alla manifestazione dell'insolvenza (si veda cap.1.2). Non è inoltre necessario considerare un elevato numero di indici. Conviene concentrarsi su un numero ridotto (per esempio, quattro o cinque) per rappresentare i principali profili della gestione aziendale (equilibrio patrimoniale-finanziario, liquidità, efficienza, redditività), in modo da raggiungere circa 20 quozienti, tra cui operare un'ulteriore selezione. Oltrepassare tale numero è inutile, perché determina problemi di collinearità tra variabili simili. Cfr. Altman, Eisenbeis (1978).

Tavola 1.3.1
Base dati

	Stato	X_1	X_2	X_3	X_4
1	1	0,25	2,15	0,02	0,01
2	1	0,05	1,20	0,01	0,02
3	1	0,12	2,36	0,03	0,00
4	1	0,20	1,50	0,01	0,13
5	1	0,10	0,80	0,05	0,10
…	…	…	…	…	…
20	2	-0,05	0,75	1,36	0,52

Una volta individuato il campione e le variabili, procediamo alla costruzione della funzione discriminante, ossia di una particolare media ponderata delle quattro variabili (il cosiddetto *valore discriminante*) in grado di classificare le singole imprese in uno dei due gruppi considerati (assegnando valori elevati a quelle affidabili, e più bassi a quelle insolventi).

La funzione discriminante assumeràla seguente forma:

$$z = a_1X_1 + a_2X_2 + a_3X_3 + a_4X_4$$

dove
z = valore discriminante calcolato per ogni singola impresa
a_i = coefficienti di discriminazione
Xi = variabili

La funzione discriminante: cosa fa...

I coefficienti a_i vengono scelti grazie ad un particolare algoritmo che garantisce che i valori discriminanti così ottenuti *massimizzeranno la differenza complessiva tra i due gruppi di imprese* (sane e insolventi).
In altri termini, gli score assegnati al gruppo delle dieci imprese sane ed alle dieci insolventi:
- ✓ renderanno massima la distanza tra le medie degli score dei due gruppi
- ✓ minimizzeranno la varianza (dispersione) degli score all'interno dei due gruppi.

In pratica, se valutate attraverso i valori discriminanti, le imprese "buone" saranno il più possibile simili tra loro e il più possibile diverse dalle "cattive".

...e come si calcola

L'algoritmo di calcolo dei coefficienti non è particolarmente complesso. Utilizzando la notazione matriciale, il vettore dei coefficienti è dato da[11]:

$$\mathbf{a} = (\overline{\mathbf{X}}_1 - \overline{\mathbf{X}}_2)'\mathbf{S}^{-1}$$

dove:
$\overline{\mathbf{X}}_1$= vettore delle medie del gruppo 1(imprese sane);
$\overline{\mathbf{X}}_2$ = vettore delle medie del gruppo 2 (imprese insolventi);
$\mathbf{S}^{-1}$= matrice inversa di varianza e covarianza del campione.

Per capire fino in fondo l'applicazione concreta della formula, torniamo ai dati della tavola 1.3.1(riportati per esteso sul dischetto). Possiamo dapprima calcolare le medie delle quattro variabili (X_1, X_2, X_3, X_4) nel gruppo 1 e nel gruppo 2 e organizzarle nei vettori $\overline{\mathbf{X}}_1$e $\overline{\mathbf{X}}_2$:

[11] La formula riportata nel testo viene derivata nell'appendice al termine di questo capitolo.

$$\overline{\mathbf{X}}_1 = \begin{bmatrix} 0{,}15 \\ 1{,}78 \\ 0{,}33 \\ 0{,}08 \end{bmatrix} \qquad \overline{\mathbf{X}}_2 = \begin{bmatrix} -0{,}14 \\ 0{,}55 \\ 1{,}22 \\ 0{,}41 \end{bmatrix}$$

Quindi, per ognuna delle quattro variabili calcoliamo la varianza e la covarianza[12] con le altre tre, e riuniamo i risultati nella matrice S[13]:

$$\mathbf{S} = \begin{bmatrix} 0{,}02 & 0{,}00 & -0{,}02 & 0{,}00 \\ 0{,}00 & 0{,}64 & -0{,}02 & -0{,}01 \\ -0{,}02 & -0{,}02 & 0{,}14 & 0{,}00 \\ 0{,}00 & -0{,}01 & 0{,}00 & 0{,}02 \end{bmatrix}$$

Si noti, ad esempio, che il terzo elemento della prima riga, che è uguale al primo valore della terza, è la covarianza tra roe (X_1) ed oneri finanziari/ricavi (X_3). Il segno negativo segnala che i due indicatori tendono a muoversi in direzione opposta, ossia che all'aumentare dell'uno corrisponde una riduzione dell'altro. Ciò rappresenta un risultato atteso, visto che un aumento del peso degli oneri finanziari porta con sé, a paritàdi altre condizioni, una riduzione della redditivitàdel patrimonio.

Seguendo le regole per il calcolo dell'inversa di una matrice[14], ricaviamo infine

$$\mathbf{S}^{-1} = \begin{bmatrix} 56{,}80 & -0{,}08 & 8{,}57 & -15{,}07 \\ -0{,}08 & 1{,}59 & 0{,}16 & 1{,}38 \\ 8{,}57 & 0{,}16 & 8{,}56 & -3{,}56 \\ -15{,}07 & 1{,}38 & -3{,}56 & 62{,}85 \end{bmatrix}$$

Abbiamo ora tutti gli ingredienti necessari per il calcolo di $\mathbf{a} = (\overline{\mathbf{X}}_1 - \overline{\mathbf{X}}_2)'\mathbf{S}^{-1}$, che risulta pari a:

$$\mathbf{a} = \begin{bmatrix} 12{,}82 \\ 1{,}34 \\ -3{,}88 \\ -19{,}89 \end{bmatrix}$$

Interpretazione economica dei coefficienti

Soffermiamoci brevemente sull'interpretazione dei quattro coefficienti. E' importante sottolineare che il loro valore dipende dalla capacità discriminante della singola variabile, ma anche dall'unità di misura e dal campo di variazione ammissibile. Pertanto non è possibile comparare la capacitàdiscriminante per variabili aventi unità di misura o campi di variazione diversi.

Particolarmente rilevante è invece il segno, che deve risultare concorde con le aspettative a priori sulla variabile a cui si riferisce[15].

---

[12] Una covarianza elevata e positiva denota l'attitudine di due variabili a muoversi, congiuntamente, nella stessa direzione; una covarianza negativa segnala la tendenza a variare in direzioni opposte.

[13] La matrice S è ottenuta come media ponderata delle matrici di varianza e covarianza calcolate per i singoli gruppi. Per maggiori dettagli, si rimanda al dischetto allegato al manuale.

[14] Piccolo-Vitale, (1981).

[15] Tale concordanza tra aspettative a priori e risultati non è necessariamente garantita: in particolare, quando si introducono nel modello due o più variabili fortemente correlate tra loro, è possibile che i coefficienti assumano valori controintuitivi

Nel nostro esempio, sono positivi i coefficienti delle variabili X_1 e X_2, negativi quelli di X_3 e X_4. Questo indica che al crescere della redditività e della liquidità aziendale, l'impresa viene ritenuta più affidabile; al contrario, tanto maggiori sono gli indici di indebitamento, tanto peggiore è la situazione dell'impresa. Pertanto i segni dei coefficienti sono concordi con la relazione logica tra le singole variabili ed il profilo di rischio dell'impresa.

Ora che conosciamo il vettore dei pesi **a**, la funzione discriminante è semplicemente:

$$z = 10{,}95\, X_1 + 2{,}19\, X_2 - 8{,}49\, X_3 - 31{,}8\, X_4$$

Per ogni singola impresa possiamo quindi calcolare una nuova variabile z, combinazione lineare degli indici iniziali, che ci fornisce un'indicazione sul livello di rischio dell'impresa stessa.

Tavola 1.3.2
Score

	Score
1	5,80
2	1,81
3	4,58
4	1,95
5	0,17
…	…
20	-15,26

Scelta di una valore-soglia e assegnazione

A questo punto è necessario determinare un valore soglia (*cut-off point*) che ci permetta di separare le aziende insolventi da quelle sane.
Possiamo prendere come punto di *cut-off* il valore esattamente a metà strada tra le due medie dello score calcolate per i due gruppi (sane e insolventi)[16]. Pertanto, se indichiamo con $\bar{Z}_1$, $\bar{Z}_2$ e $\bar{Z}_c$ rispettivamente lo score medio del gruppo 1, lo score medio del gruppo 2 e il valore soglia, si ha

$$\bar{Z}_c = \frac{\bar{Z}_1 + \bar{Z}_2}{2} = \frac{1{,}22 - 13.90}{2} = -6{,}34$$

Poiché nel nostro esempio si ha che $\bar{Z}_1 > \bar{Z}_2$, la regola di classificazione può quindi essere espressa come: "assegna l'impresa j-esima al gruppo 1 se $z_j > \bar{Z}_c$, altrimenti assegnala al gruppo 2".

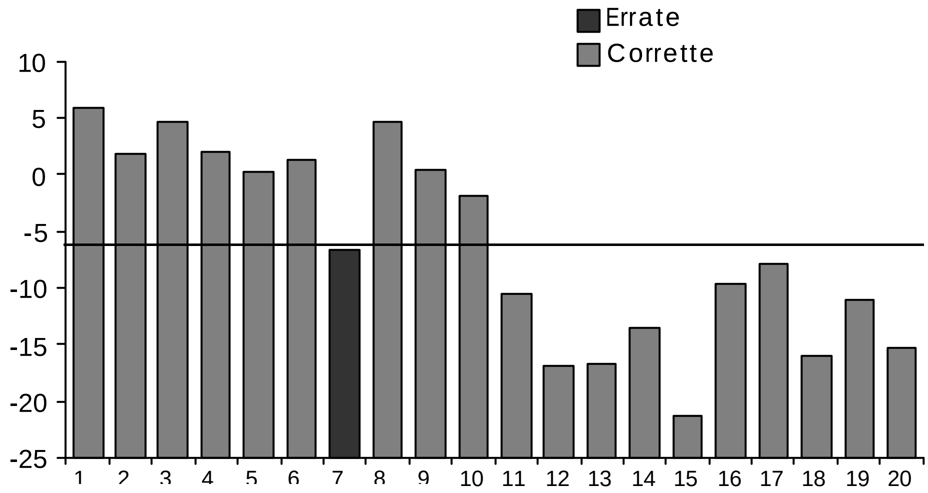
Verifica della capacità di classificazione

Confrontando gli score con il valore soglia decideremo dove allocare ogni singola azienda; potremo poi verificare la capacità di classificazione del nostro modello controllando se le diverse imprese sono state davvero assegnate al gruppo (sane o insolventi) a cui originariamente appartenevano. Nel caso del nostro campione, i risultati sono sintetizzati nella figura 1.3.2: l'altezza degli istogrammi rappresenta gli score assegnati alle singole imprese, mentre la riga orizzontale indica il valore soglia $\bar{Z}_c$: tutti e soli i debitori che non raggiungono tale soglia vengono considerati inaffidabili. Tra

[16] Solitamente il punto di *cut-off* è posizionato nella cosiddetta *grey area*, cioè un intervallo in cui per gli stessi punteggi si riscontrano sia aziende sane che insolventi. Ovviamente, tanto minore sarà quest'area, tanto maggiore sarà l'efficacia del modello poiché risulterà minima la probabilità di commettere errori.

questi, anche l'impresa n. 7 che appartiene, per la verità, al gruppo delle società sane: l'errore è evidenziato dal colore più scuro dell'istogramma[17].

Figura 1.3.2
Classificazione



Nota: errore nella classificazione = 5% (un'azienda su venti)

Dai valori discriminanti alle probabilità di insolvenza

Come anticipato in precedenza, lo score non consente soltanto di classificare ogni impresa in un gruppo, ma fornisce anche un'indicazione sul suo livello di rischio; si va da valori del punteggio molto negativi, per i quali il rischio di insolvenza è elevato, a valori molto positivi che indicano una situazione di considerevole sicurezza. Non siamo comunque ancora in grado di associare ad ogni singola azienda una probabilità di insolvenza. Tale informazione è però un ingrediente fondamentale di un buon sistema di rating, indispensabile, tra l'altro, per effettuare valutazioni di pricing del rischio di credito, di misura del capitale assorbito, di stima degli accantonamenti da iscrivere a bilancio sul portafoglio impieghi.
Attraverso l'applicazione del teorema di Bayes[18] è possibile tradurre gli score ricavati dal modello in probabilità di insolvenza che tengano conto della qualità media del portafoglio della banca. In altri termini, a parità di score, è possibile calcolare probabilità di default più alte se si ha a che fare con una banca notoriamente rischiosa (o con una filiale collocata in aree economiche disagiate), o più limitate se il campione riguarda un intermediario contraddistinto da elevata qualità degli attivi.
In questa sede non entriamo nel merito delle tecnicalità necessarie alla costruzione di queste probabilità d'insolvenza[19], ma ci limitiamo a riportare, nella tavola 1.3.3, il valore stimato per ciascuna azienda del campione nell'ipotesi che il tasso di default medio della banca sia pari a 0,01.
Osserviamo come, in base ai dati riportati nella tavola, i debitori possano essere raggruppati in classi omogenee (corrispondenti a determinati valori della probabilità di *default*) organizzate su una scala come quella della figura 1.2.5 del capitolo precedente.

---

[17] Non sarebbe tuttavia corretto, dal punto di vista logico, valutare la bontà del modello in base al tasso d'errore dimostrato sugli stessi dati utilizzati per costruirlo. E' invece preferibile verificarne le prestazioni su un campione di imprese diverso rispetto a quelle usato per la stima, e in generale continuare a monitorarne le prestazioni nel tempo.

[18] Per una trattazione statistica del teorema di Bayes, si veda Piccolo-Vitale (1981), pp.144-146.

[19] Per dettagli sulla determinazione per via analitica delle probabilità di insolvenza rimandiamo a Varetto (1999), pp.290-294.

Tavola 1.3.3
Stima delle probabilità di insolvenza

	Probabilità di insolvenza (%)		Probabilità di insolvenza (%)
1	0,00	11	39,04
2	0,00	12	99,76
3	0,00	13	99,72
4	0,00	14	93,22
5	0,00	15	100,00
6	0,00	16	20,45
7	1,36	17	4,42
8	0,00	18	99,42
9	0,00	19	54,77
10	0,01	20	98,69

1.3.3 Limiti dell'esempio e indicazioni per un modello più realistico

L'esempio appena presentato aveva lo scopo di mostrare la semplicità concettuale del calcolo delle funzioni discriminanti. Ovviamente, quando si devono trattare dati reali, la costruzione della funzione diventa più complicata.
Per questo motivo, sul dischetto che accompagna questo manuale, si riporta un secondo esempio di funzione discriminante, maggiormente realistico perché riguarda un maggior numero di variabili e di aziende.
Vale comunque la pena di evidenziare, sia pur schematicamente, alcune problematiche connesse all'utilizzo pratico dell'analisi discriminante:

- ✓ Innanzitutto è necessario soffermarsi a riflettere sul criterio utilizzato per suddividere in gruppi il campione di aziende. La diversa definizione di anomalia (cfr. capitolo 1.2) influenza infatti la corretta stima della funzione discriminante e la precisione dei risultati;
- ✓ Inoltre, poiché un modello può essere stimato con tanta maggiore accuratezza quanto più alto è il numero delle imprese incluse nel campione rispetto al numero delle variabili indipendenti, è necessario disporre di un elevato numero di osservazioni. Spesso le aziende insolventi per cui è possibile disporre di dati quantitativi completi sono purtroppo quantitativamente insufficienti.
- ✓ Sarebbe opportuno scegliere campioni il più possibile omogenei in modo tale da poter costruire singoli modelli settoriali in grado di catturare le specificità delle imprese che vi operano. La limitatezza dei campioni di aziende insolventi spesso costringe invece alla costruzione di modelli generali;
- ✓ l'individuazione degli indicatori su cui stimare il modello è un compito estremamente delicato e laborioso. Generalmente il procedimento di scelta delle variabili si articola in due momenti[20]: l'individuazione soggettiva degli indici ritenuti a priori maggiormente significativi e la loro successiva riduzione attraverso un'ulteriore selezione basata su due diverse metodologie: una metodologia di natura quantitativa che identifica gli indicatori rilevanti tramite particolari procedure statistiche, ed una metodologia di natura qualitativa con la quale gli indici principali vengono selezionati sulla base dei principi delle teorie economico-finanziarie;
- ✓ al fine di aumentare il grado di realismo del modello, è infine necessario disporre di una buona stima della qualità media del portafoglio crediti (cioè della cosiddetta

[20] Si veda De Laurentis (1986), p.11.

probabilità di insolvenza a priori, che cambia da banca a banca, e che abbiamo utilizzato per produrre le stime riportate nell'ultima parte del paragrafo precedente); inoltre, per utilizzare il modello per una selezione tra clienti da accettare e controparti da rifiutare, è opportuno incorporarvi una stima dei costi di errata classificazione.

✓ L'analisi discriminante lineare non è altro che un caso particolare di un criterio di classificazione generale per il quale valgono le ipotesi di normalità multivariata delle distribuzioni delle variabili e uguaglianza delle matrici di varianza e covarianza. Se si rimuove quest'ultima ipotesi, la regola di classificazione si trasforma in una funzione discriminante quadratica[21]. Qualora non fosse soddisfatta l'ipotesi di distribuzione multinormale delle variabili, è invece possibile utilizzare, in alternativa all'analisi discriminante, una funzione logistica o probit[22].

Appendice

Dimostriamo brevemente come si deriva la formula del valore discriminante secondo Fisher. A tal fine siano:

$\mathbf{X}$ = vettore delle osservazioni (variabili);

$\overline{\mathbf{X}}_i$ = vettore delle medie del gruppo i (i=1,2);

$\mathbf{S}$ = matrice di varianza e covarianza;

$\mathbf{S}^{-1}$ = matrice inversa di varianza e covarianza;

$$\mathbf{X} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_i \end{bmatrix} \qquad \overline{\mathbf{X}}_i = \begin{bmatrix} \overline{X}_{1i} \\ \overline{X}_{2i} \\ \vdots \\ \overline{X}_{ki} \end{bmatrix} \qquad \mathbf{S} = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1k} \\ s_{21} & s_{22} & & \vdots \\ \vdots & & \ddots & \vdots \\ s_{k1} & \cdots & \cdots & s_{kk} \end{bmatrix}$$

dove:

$$s_{ij} = \mathrm{E}\,[(X_i - \overline{X}_i)(X_j - \overline{X}_j)] = \mathrm{E}\,[(X_j - \overline{X}_j)(X_i - \overline{X}_i)] = s_{ji}$$

Sia z una funzione discriminante (cioè una combinazione lineare delle variabili presenti nel vettore **X,** pesate per i coefficienti del vettore **a**)

$$z = \mathbf{a}'\mathbf{X}$$

Allora il valore medio di z nel gruppo 1 può scriversi $\mathrm{E}(\mathbf{a}'\mathbf{X}) = \mathbf{a}'\overline{\mathbf{X}}_1$, e analogamente nel gruppo 2 si ha $\mathrm{E}(\mathbf{a}'\mathbf{X}) = \mathbf{a}'\overline{\mathbf{X}}_2$. La varianza di z (calcolata a partire dalla matrice S comune a entrambi i gruppi) sarà $\mathrm{Var}\,(\mathbf{a}'\mathbf{X}) = \mathbf{a}'\mathbf{S}\mathbf{a}$.

Costruire la "migliore" funzione discriminante equivale a scegliere **a** in modo tale che sia massima la distanza tra le medie dei due gruppi (presa in valore assoluto o, come nell'espressione qui di seguito, al quadrato) pesata per la varianza, cioè:

$$d = \frac{(\mathbf{a}'\overline{\mathbf{X}}_1 - \mathbf{a}'\overline{\mathbf{X}}_2)^2}{\mathbf{a}'\mathbf{S}\mathbf{a}}$$

---

[21] Per una panoramica dei principali approcci metodologici, si vedano Altman et al.(1981), De Laurentis (1986).

[22] L'ipotesi chiave dei modelli logistico e probit riguarda la forma della distribuzione cumulata delle probabilità di insolvenza: nel primo, si assume che la forma della distribuzione sia la logistica cumulata, mentre nel secondo si ipotizza che essa sia la normale standardizzata cumulata. Sull'argomento si veda Altman et al. (1981).

Calcolando la derivata prima di questa espressione e uguagliandola a zero si ottiene:

$$\frac{d\,\mathrm{d}}{d\,\mathrm{a}} = \frac{2(\mathbf{a}'\overline{\mathbf{X}}_1 - \mathbf{a}'\overline{\mathbf{X}}_2)(\overline{\mathbf{X}}_1 - \overline{\mathbf{X}}_2)\mathbf{a}'\mathbf{S}\mathbf{a} - 2\mathbf{S}\mathbf{a}(\mathbf{a}'\overline{\mathbf{X}}_1 - \mathbf{a}'\overline{\mathbf{X}}_2)^2}{(\mathbf{a}'\mathbf{S}\mathbf{a})^2} = 0 \quad (\overline{\mathbf{X}}_1 - \overline{\mathbf{X}}_2)\mathbf{a}'\mathbf{S}\mathbf{a} - \mathbf{S}\mathbf{a}(\mathbf{a}'\overline{\mathbf{X}}_1 - \mathbf{a}'\overline{\mathbf{X}}_2) = 0$$

$$(\overline{\mathbf{X}}_1 - \overline{\mathbf{X}}_2) = \mathbf{S}\mathbf{a}\underbrace{\left(\frac{(\mathbf{a}'\overline{\mathbf{X}}_1 - \mathbf{a}'\overline{\mathbf{X}}_2)}{\mathbf{a}'\mathbf{S}\mathbf{a}}\right)}_{\text{costante}}$$

Poiché la scala di **a** è assolutamente arbitraria (è solo un punteggio utilizzato solo per assegnare ogni azienda a un gruppo), possiamo moltiplicarlo per qualsiasi costante a piacere.
Di conseguenza si ottiene:

$$\mathbf{a} = (\overline{\mathbf{X}}_1 - \overline{\mathbf{X}}_2)'\mathbf{S}^{-1}$$

Il punto di *cut-off* è:

$$\overline{Z}_c = \frac{(\overline{\mathbf{X}}_1 - \overline{\mathbf{X}}_2)'\mathbf{S}^{-1}\overline{\mathbf{X}}_1 + (\overline{\mathbf{X}}_1 - \overline{\mathbf{X}}_2)'\mathbf{S}^{-1}\overline{\mathbf{X}}_2}{2} = \frac{\overline{Z}_1 + \overline{Z}_2}{2}$$

La regola decisionale diventa:
assegna un'impresa al gruppo 1 se $Z > \overline{Z}_c$, altrimenti assegnala al gruppo 2 (supponendo che $\overline{Z}_1 > \overline{Z}_2$)