UNIVERSITA' DEGLI STUDI DI ROMA TRE
FACOLTA' DI SCIENZE M.F.N.

Tesi di Laurea in Matematica

di

Jacopo Iannucci

Analisi discriminante, regressione logistica e reti per la stima delle probabilità di fallimento. Applicazioni ed estensioni del metodo Z-score

Relatore

Prof. Alessandro Ramponi

Il Candidato Il Relatore

ANNO ACCADEMICO 2003 - 2004

Luglio 2004

Classificazione AMS: 91D99, 91B99, 62J02.

Parole chiave: Rischio di credito, regressione, reti.

Indice

Introduzione

In questa tesi vengono discusse quattro diverse metodologie statistiche per risolvere problemi di classificazione. Contestualmente ne viene data un'applicazione pratica in campo finanziario nell'ambito del credit scoring, cioè nell'assegnazione di una classe di merito creditizio ad una società che riceve un prestito o emette un obbligazione. I quattro metodi illustrati sono i seguenti:

- l'analisi discriminante lineare (Cap I)
- la regressione logistica (Cap III)
- le reti neurali (Cap IV)
- gli alberi di decisione (Cap V)

Con queste metodologie affronteremo in dettaglio quasi esclusivamente problemi di classificazione binaria. Un tale problema si configura in concreto come il seguente: si hanno due classi predefinite D_1 e D_2 e vettori di osservazioni $\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_n$ relativi ad n soggetti di una popolazione; si vuole costruire una funzione, detta regola d'assegnazione, che a ciascun vettore $\underline{x}_i$ associ, sulla base di un ragionevole criterio discriminatorio, o la classe D_1 o quella D_2.
A scopo illustrativo, supponiamo di dover svolgere, in ambito finanziario, uno

studio sulla probabilità che un certo numero di individui, i quali richiedono un prestito ad una determinata banca, risanino il debito. Per ciascuno di essi rileveremo due dati "fattori" quali ad esempio reddito e beni ipotecabili, riassunti in un vettore numerico $\underline{x}$; avremo poi bisogno di un modello classificatore, magari uno dei quattro sopra citati, in base al quale far corrispondere ad $\underline{x}$ la classe $D_1 = \{\text{individui che restitueranno il prestito}\}$ o la classe alternativa D_2.

In tutti e quattro i metodi di classificazione precedendemente enunciati si parte da una base di dati costituiti da N esempi di oggetti preclassificati: il *data set*; si può decidere di dividere quest'ultimo in due sottoinsiemi corrispondentemente di n e di k elementi tali che $N = n + k$; usare il primo insieme per creare una regola di classificazione utilizzando uno dei quattro metodi in questione, il secondo per testare la validità del test vedendo come classifica degli oggetti di cui noi già conosciamo la classificazione a priori; in questo caso l'insieme di m elementi prende il nome di *training set* mentre l'altro di *validation set*.

L'analisi discriminante lineare, la regressione logistica e le reti sono metodi parametrici; metodi cioè in cui bisogna identificare, secondo una procedura appropriata, dei parametri con i quali costruire la soglia di discriminazione per poter poi effettuare la classificazione. Il quarto metodo, gli alberi decisionali non è parametrico: non sono infatti presenti pesi da identificare, ma si basa sulla ripartizione ricorsiva dello spazio dei dati osservati.

Capitolo I In questo capitolo come prima cosa viene introdotto l'obiettivo dell'analisi discriminante che è quello di riuscire a classificare un generico individuo in uno dei gruppi $D_1, D_2, \ldots, D_g$ sulla base di un vettore

di informazioni ad esso relativo $\underline{x} = (x_1, x_2, \ldots, x_p)$. La naturale applicazione di questo tipo di analisi nell'ambito bancario del rischio di credito è quello di discriminare, all'interno di una popolazione di individui che richiedono un prestito ad una determinata banca, tra quelli propensi a restituire il credito da quelli non propensi ($g = 2$) sulla base di un insieme di informazioni quali: lo stipendio,beni ipotecabili e il tipo di prestito che si richiede ($p = 3$); in questo caso $x_1 =$ *stipendio fisso*, $x_2 =$ *beni ipotecabili* $x_3 =$ *tipo di prestito richiesto*.

Il modello trattato in questo capitolo è l'analisi discriminante di Fischer applicata al caso di due gruppi ($g = 2$) e p fattori [DA01]. Egli intendeva trovare una combinazione lineare $\underline{\alpha}^t\underline{x}$ delle p variabili $\underline{\mathrm{x}}$ che separasse in modo sensato i due campioni di prova. La scelta di Fischer fu quella di determinare il vettore $\underline{\alpha}$ come quel vettore che rendesse massima la distanza fra le medie dei gruppi e al contempo minima la varianza all'interno di ciascun gruppo. Il vettore dei pesi $\underline{\alpha}$, seguendo l'idea di Fischer, è calcolato come quel vettore che massimizza il seguente rapporto

$$J(\underline{\alpha}) = Q_B(\underline{\alpha})/Q_W(\underline{\alpha}) \tag{1}$$

dove

$$Q_B(\underline{\alpha}) = \underline{\alpha}^t \widehat{S}_B \underline{\alpha}$$

e

$$Q_W(\underline{\alpha}) = \underline{\alpha}^t \widehat{S}_W \underline{\alpha}$$

sono le forme quadratiche associate rispettivamente a $\widehat{S}_B$ e $\widehat{S}_W$ con

$$\widehat{S}_B = (\underline{\widehat{x}}_2 - \underline{\widehat{x}}_1)(\underline{\widehat{x}}_2 - \underline{\widehat{x}}_1)^t \tag{2}$$

matrice delle differenze al quadrato fra le medie dei gruppi;

$$\widehat{S}_W = \frac{1}{n_1 + n_2 - 2}[n_1 \widehat{S}_{W1} + n_2 \widehat{S}_{W2}] \tag{3}$$

con

$$\widehat{S}_{W1} = \sum_{i=1}^{n1} (\underline{x}_i - \underline{\widehat{x}}_1)(\underline{x}_i - \underline{\widehat{x}}_1)^t$$

e

$$\widehat{S}_{W2} = \sum_{i=n1+1}^{n} (\underline{x}_i - \underline{\widehat{x}}_2)(\underline{x}_i - \underline{\widehat{x}}_2)^t$$

dove $\widehat{S}_{W1}$ $\widehat{S}_{W2}$ sono le matrici di covarianza all'interno del primo e del secondo gruppo e

$$\underline{\widehat{x}}_1 = (\widehat{x}_{11}, \widehat{x}_{12}, ..., \widehat{x}_{1p}) = (\frac{1}{n_1}\sum_{k=1}^{n_1} x_{k1}^{(1)}, \frac{1}{n_1}\sum_{k=1}^{n_1} x_{k2}^{(1)}, ..., \frac{1}{n_1}\sum_{k=1}^{n_1} x_{kp}^{(1)})$$

$$\underline{\widehat{x}}_2 = (\widehat{x}_{21}, \widehat{x}_{22}, ..., \widehat{x}_{2p}) = (\frac{1}{n_2}\sum_{k=1}^{n_2} x_{k1}^{(2)}, \frac{1}{n_2}\sum_{k=1}^{n_2} x_{k2}^{(2)}, ..., \frac{1}{n_2}\sum_{k=1}^{n_2} x_{kp}^{(2)})$$

vettori delle medie con $\widehat{x}_{ij}$ = *media campionaria all'interno del i-esimo gruppo del j-esimo fattore*, $n = n_1 + n_2$ dove n_1 e n_2 sono le dimensioni campionarie nei due gruppi; si può dimostrare che la (1) è massimizzata per $\underline{\alpha}$ = massimo autovalore di

$$\widehat{S}_W^{-1} \widehat{S}_B$$

da cui si ottiene come vedremo

$$\underline{\alpha} \propto \widehat{S}_W^{-1}[(\underline{\widehat{x}}_1 - \underline{\widehat{x}}_2)]$$

Una volta effettuato il calcolo dei pesi relativi ai fattori, utilizzando le informazioni contenute nel training set, un nuovo individuo per il quale è stato osservato $\underline{x} = (x_1, x_2, \ldots, x_p)$ è collocato nella classe D_1 se

$$\underline{\alpha}^t[\underline{x} - \frac{1}{2}(\underline{\widehat{x}}_1 + \underline{\widehat{x}}_2)] > 0 \tag{4}$$

dove

$$\underline{\alpha} = \widehat{S}_W^{-1}(\underline{\widehat{x}}_1 - \underline{\widehat{x}}_2) \tag{5}$$

La funzione

$$y = \underline{\alpha}^t \underline{x} \tag{6}$$

è conosciuta come: ***funzione discriminante lineare***.

L'iperpiano

$$\underline{\alpha}^t[\underline{x} - \frac{1}{2}(\underline{\widehat{x_1}} + \underline{\widehat{x_2}})] = 0$$

viene chiamato *iperpiano separatore* o *decision boundary*.

Capitolo II In questo capitolo vedremo un'applicazione al campo della finanza del metodo di classificazione esposto nel Capitolo 1. Lo Z-score di Altman è un approccio multivariato basato sulla considerazione e l'analisi di più fattori ritenuti significativi nel determinare lo stato di salute di un azienda o di un istituto finanziario [CAN98],[AS02]. Questi fattori sono pesati e combinati per produrre una misura (un punteggio del rischio di credito) che discrimini al meglio compagnie che falliscono da quelle che non lo fanno. Da un punto di vista operativo si ritiene che una tale misura esista e che sia possibile trovarla, poichè le compagnie che falliscono hanno bilanci e tendenze finanziarie molto diverse da quelle finanziariamente in attivo. Una banca che utilizzi questo metodo potrebbe accordare o rifiutare un prestito se il punteggio Z-score della compagnia in questione scende al di sotto di una determinata soglia. Altman ha basato il suo modello multivariato sui rapporti finaziari che vengono mostrati nella successiva tabella (1).

Il modello Z-score è stato costruito usando la metodologia spiegata nel capitolo precedente e utilizza cinque variabili (indici di bilancio) relative a: l'analisi della liquidità, della redditività, della leva finanziaria, della solvibilità e dell'attività. Lo Z-score è definito come:

$$Z = \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_3 + \alpha_4 X_4 + \alpha_5 X_5 \tag{7}$$

dove α_i è il peso relativo all'i-esimo fattore; Nell'esempio di Altman, il gruppo di fallimento D_2 consta di 33 aziende manufatturiere in bancarotta, mentre il gruppo del non fallimento D_1 comprende un ugual numero di compagnie manufatturiere; siamo nel caso di $n = 66$ con $n_1 = n_2 = 33$. Le aziende nel gruppo del non fallimento erano ancora in attivo al tempo dell'analisi. In questo caso (7) diviene:

$$Z = 1.2X_1 + 1.4X_2 + 3.3X_3 + 0.6X_4 + 0.999X_5 \tag{8}$$

Modello Z-score

Variabile	media gruppo fallimento	media gruppo non fallimento
$X_1 = \frac{\text{working capital}}{\text{total assets}}$	-6.1%	41.4%
$X_2 = \frac{\text{retained earning}}{\text{total assets}}$	-62.6%	35.5%
$X_3 = \frac{\text{EBIT}}{\text{total assets}}$	-31.8%	15%
$X_4 = \frac{\text{market value of equity}}{\text{book value of liabilities}}$	40.1%	247.7%
$X_5 = \frac{\text{sales}}{\text{total assets}}$	1.5 volte	1.9 volte

Tabella 1: Il campione è di 33 aziende; la funzione discriminante risulta essere $Z = 0.012X_1 + 0.014X_2 + 0.033X_3 + 0.006X_4 + 0.999X_5$

Capitolo III Affronteremo anche in questo capitolo esclusivamente il problema di classificazione in presenza di due classi ($g = 2$). Utilizzando

la tecnica della regressione logistica si modellizza direttamente la probabilità che un generico individuo si trovi nella classe D_1 o in quella D_2 dato il vettore delle osservazioni dei p fattori $\underline{x} = (x_1, x_2, \ldots, x_p)$ [CB95]. La modellizzazione è del tipo seguente: denotiamo con Y la variabile che vale 1 se l'individuo è assegnato alla classe D_1 mentre vale 0 se è assegnato a quella D_2, allora si pone

$$P(Y = 1 \mid \underline{x}) = \frac{e^{\underline{\alpha}^t \underline{x}}}{1 + e^{\underline{\alpha}^t \underline{x}}} \tag{9}$$

$$P(Y = 0 \mid \underline{x}) = \frac{1}{1 + e^{\underline{\alpha}^t \underline{x}}} \tag{10}$$

Notiamo che la (9) e la (10) possono essere riscritte, con l'uso della funzione

$$g(u) = \frac{1}{1 + e^u} \qquad con \;\; u \in \mathbb{R} \tag{11}$$

nota con il nome di "sigmoide logistica,"come:

$$P(Y = 1 \mid \underline{x}) = g(\underline{\alpha}^t \underline{x})$$

$$P(Y = 0 \mid \underline{x}) = 1 - g(\underline{\alpha}^t \underline{x})$$

Come nei capitoli precedenti, anche qui si tratta di determinare in modo appropriato i valori dei parametri liberi $\alpha_1, \alpha_2, \ldots, \alpha_p$. Uno dei possibili modi è di utilizzare la tecnica della massima verosimiglianza: denotiamo con $\mathcal{X}$ il training set e scriviamo la funzione di verosimiglianza:

$$L(\underline{\alpha}) = \prod_{i=1}^{n} P(Y = y_i \mid \underline{x}_i) \tag{12}$$

e la log-verosimiglianza :

$$l(\underline{\alpha}) = \log(L(\underline{\alpha})) = \sum_{i=1}^{n} \{y_i \underline{\alpha}^t \underline{x}_i - \log(1 + e^{\underline{\alpha}^t \underline{x}_i})\} \tag{13}$$

Data la concavità di (13), l'annullamento del gradiente della log-verosimiglianza permette di caratterizzare lo stimatore $\widehat{\alpha}$. Per il calcolo si ricorre generalmente a metodi numerici di tipo deterministico.

Una volta stimato il vettore $\underline{\hat{\alpha}}$ con il metodo di massima verosimiglianza, si calcola per ogni nuovo individuo la probabilità di trovarsi in D_1 o in D_2 mediante le formule:

$$P(Y = 1 \mid \underline{x}) = g(\underline{\hat{\alpha}}^t \underline{x}) \quad e \quad P(Y = 0 \mid \underline{x}) = 1 - g(\underline{\hat{\alpha}}^t \underline{x})$$

Capitolo IV Anche in questo capitolo affronteremo esclusivamente il problema di classificazione in presenza di 2 classi ($g = 2$); il metodo considerato presenta somiglianze, come vedremo in seguito, con la regressione logistica e con l'analisi discriminante lineare. Il metodo di discriminazione analizzato in questo capitolo fu studiato da Rosenblatt ed è noto con il nome di *perceptrone* [SH98]. Anche qui si tratta di partire da un set di dati iniziali: il training set $\mathcal{X} = \mathcal{X}_1 \cup \mathcal{X}_2$ di cardinalità n dove $\mathcal{X}_1$ e $\mathcal{X}_2$ hanno cardinalità rispettivamente n_1 e n_2 con $n_1 + n_2 = n$, inoltre $\mathcal{X}_1 =$ *training set costituito da individui appartenenti alla classe* D_1 mentre $\mathcal{X}_2 =$ *training set costituito da individui appartenenti alla classe* D_2 ; etichettiamo poi gli individui appartenenti a $\mathcal{X}_1$ e a $\mathcal{X}_2$ rispettivamente con i valori 1 e -1; per poi determinare l'equazione di un iperpiano che separi le due classi di appartenenza (decision boundary) del tipo

$$\sum_{i=1}^{p} \alpha_i x_i + b = 0 \tag{14}$$

dove b, ad esempio nel caso $p = 2$ in cui la (14) è una retta, rappresenta il punto in cui la retta in questione taglia l'asse delle ordinate. Come si

nota la (14) è simile all'equazione dell'iperpiano separatore vista precedentemente nel caso dell'analisi discriminante di Fisher e della regressione logistica; la differenza fra i due metodi di classificazione analizzati in precedenza e il metodo detto del *perceptrone*, consiste nel modo in cui si determinano i pesi $\alpha_1, \alpha_2, \ldots, \alpha_p$.

I pesi del perceptrone possono essere calcolati usando una regola di correzione di errore nota come *l'algoritmo di apprendimento del perceptrone*.

Si può dimostrare un teorema per la convergenza dell'algoritmo di apprendimento del perceptrone (Rosenblatt,1962):

Teorema

Se i due training set $\mathcal{X}_1$ *e* $\mathcal{X}_2$ *sono linearmente separabili allora* $\exists\, M_0$ *indice di itarazione t.c.* $\forall$ *indice di iterazione* $m \geq N_0$ *tutti gli elementi del training set* $\mathcal{X} = \mathcal{X}_1 \cup \mathcal{X}_2$ *sono classificati correttamente.*

Una volta settati i pesi il perceptrone opera nel seguente modo: un nuovo individuo descritto da un vettore di osservazioni $x_1, \ldots, x_p$ è presentato al perceptrone, questo opera una combinazione lineare come la (14); a questo punto se la combinazione lineare risulta positiva il perceptrone darà come risposta 1, altrimenti la risposta sarà -1; l'individuo è quindi classificato in D_1 se il responso del perceptrone è 1, altrimenti è classificato in D_2 .

Infine è discussa la relazione che lega il perceptrone alla regressione logistica.

Capitolo V Un ulteriore metodo di classificazione che consideriamo è il

metodo degli *alberi decisionali* [RN98].

Un albero decisionale prende in input, come gli altri metodi analizzati fin'ora, un individuo descritto da un vettore $\underline{x} = (x_1, x_2, \ldots, x_p)$ contenente i p fattori ed emette in uscita una "decisione"del tipo sì/no. Ciascun nodo interno all'albero corrisponde ad un test sul valore di una delle proprietà e gli "archi"che partono da ciascun nodo sono etichettati con i possibili valori del relativo test. Ciascuna "foglia"specifica il valore booleano di output quando si perviene ad essa.

L'idea, per costruire un buon albero decisionale, è quella di cercarne di minimizzare la profondità cercando di selezionare l'attributo che più di ogni altro riesce a fornire una classificazione esatta degli esempi. Un attributo perfetto è quello che suddivide gli esempi in sottoinsiemi per cui è immediatamente possibile stabilirne la classificazione. Abbiamo bisogno di una misura formale di attributo *valido* e di attributo *inutile*. La misura dovrebbe avere valore massimo quando l'attributo è perfetto e minimo quando è totalmente inutile. Una misura possibile è il valore atteso della quantità di informazione fornita dall'attributo. In generale, se le possibili risposte di un attributo v_i hanno probabilità $P(v_i)$ allora il contenuto informativo I della risposta è dato da

$$I(P(v_1), P(v_2), \ldots, P(v_n)) = \sum_{i=1}^{n} -P(v_i) \log_2 P(v_i)$$

Con questo tipo di misura vedremo come sarà possibile determinare l'algoritmo che ci permetterà di costruire l'albero di decisione.

Capitolo VI In questo ultimo capitolo mostriamo l'applicazione ,tramite il paccheto statistico **R**, dell'analisi discriminante di Fischer , della re-

gressione logistica e del metodo degli alberi decisionali nel classificare in due classi D_1 e D_2 un campione di 1000 persone rivoltasi ad una determinata banca tedesca per la richiesta di un prestito [VR02],[IM03]. La stessa applicazione è stata studiata con il perceptrone utilizzando il software Matlab. Nel nostro caso D_1 è la classe delle persone che hanno ottenuto il prestito mentre D_2 è la classe di coloro che non lo hanno ottenuto. I dati sono stati scaricati dal sito http//ftp.ics.uci.edu/pub/machine-learning-databases/statlog/german Il campione di 1000 individui è caratterizzato da $p = 24$ fattori diversi indicanti il tipo di prestito che si richiede, la nazionalità, il sesso, etc... Per quanto riguarda l'analisi discriminante di Fischer e la regressione logistica l'output consiste principalmente nei pesi α_i relativi ai singoli fattori considerati; mentre nel caso dell'albero decisionale l'output consiste principalmente: nei nodi, nel numero di individui giunti nel nodo, la probabilità di appartene a D_1 o D_2 che ha un individuo giunto in quel punto. Per quanto riguarda il perceptrone con software Matlab è possibile visualizzare subito un file nel quale si possano vedere gli elementi classificati in maniera corretta od errata.

Una volta implementati i modelli si è testato il loro grado di precisione facendogli classificare una serie di individui dei quali già era nota la classificazione; i risultati di tale verifica sono presentati, poi, in una tabella nella quale sono indicati gli individui giustamente ed ingiustamente classificati dal modello in questione.

Capitolo 1

Analisi Statistica Discriminante

1.1 Elementi di base

Supponiamo di avere g gruppi $D_1, D_2, ..., D_g$: l'obiettivo dell'analisi discriminante è di collocare un individuo in uno di questi gruppi sulla base di un insieme di osservazioni, $x_1, x_2, ..., x_p$ ad esso relative. Per esempio in ambito clinico, all'interno di una popolazione di pazienti si vuole discriminare tra quelli tendenti al diabete da quelli che non lo sono ($g = 2$) sulla base delle osservazioni di glucosio nel sangue, peso corporeo ed età ($p = 3$); in questo caso $x_1 =$ *concentrazione di glucosio nel sangue*, $x_2 =$ *peso corporeo* $x_3 =$ *età*. Convenzionalmente tali variabili (x_1, x_2, x_3)rilevate per ciascun individuo vengono chiamate **fattori**. Più in generale se associato a ciascun gruppo D_j abbiamo una densità di probabilità, per le variabili osservate che descrive la probabilità che i fattori appartengono al gruppo j, della forma $f_j(\underline{x})$ dove $\underline{x} = (x_1, x_2, ..., x_p)$, una regola intuitiva per il processo di

discriminazione potrebbe essere: se

$$f_{i_0}(\underline{x}) = \max\{f_1(\underline{x}), f_2(\underline{x}), \ldots, f_g(\underline{x})\} \tag{1.1}$$

allora il soggetto per il quale si è osservato il vettore di dati $\underline{x}$ è collocato nel gruppo D_{i_0}. Per spiegare meglio questa regola facciamo dei semplici esempi:

1. Supponiamo di poter osservare una sola, $(p = 1)$, variabile, X che può assumere due valori $\{0,1\}$ e di avere due gruppi D_1 e D_2. In D_1 abbiamo

$$P(X = 0) = P(X = 1) = \frac{1}{2} \tag{1.2}$$

mentre in D_2

$$P(X = 0) = \frac{1}{4} \quad \Pr(X = 1) = \frac{3}{4} \tag{1.3}$$

La regola specificata in (1.1) associa allora un individuo con $x = 0$ a D_1 mentre uno con $x = 1$ a D_2.

2. Supponiamo di avere tre gruppi di tipo geografico spaziale: $D_N =$ *Nord Italia*, D_C *Centro Italia*, $D_S =$ *Sud Italia* e di voler stabilire se un italiano di cui ci è ignoto il luogo di nascita appartenga a D_N, D_C o D_S sulla base di un singolo fattore, ad esempio la concentrazione di colesterolo nel sangue supponendo che, in assensa di malattie epatiche, determinate abitudini alimentari abbiano radici regionali ed influenzino perciò tali concentrazioni. Supponiamo che sulla base di precedenti studi siano state stimate le seguenti probabilità discrete relative alla concentrazione di colesterolo nelle tre parti in cui abbiamo diviso l'italia, ovvero che

in D_N

$$P(X < 1) = \frac{2}{10}\ ,\ P(1 < X < 2) = \frac{3}{10}\ ,\ P(X > 2) = \frac{5}{10}$$

in D_C

$$P(X < 1) = \frac{1}{10}\ ,\ P(1 < X < 2) = \frac{7}{10}\ ,\ P(X > 2) = \frac{2}{10}$$

in D_S

$$P(X < 1) = \frac{6}{10}\ ,\ P(1 < X < 2) = \frac{3}{10}\ ,\ P(X > 2) = \frac{1}{10}$$

dove $X =$ *tasso di colesterolo*; in questo caso poichè abbiamo un solo fattore e 3 gruppi ($p = 1$ e $g = 3$). Secondo la regola (1.1) un individuo con $1 < X < 2$ verrà collocato in D_C con $X > 2$ in D_N e con $X < 1$ in D_S.

Vediamo ora come si generalizza tale tecnica di classificazione.

1.2 Il caso gaussiano univariato e multivariato

Trattiamo ora il caso gaussiano poichè per esso è possibile operare un analisi abbastanza dettagliata. Supponiamo di avere una variabile continua, X, e ancora due gruppi. In D_1 la variabile X segue una distribuzione normale con media μ_1 e varianza σ_1^2, in D_2 una distribuzione normale con media μ_2 e varianza σ_2^2 (siam $\mu_1 > \mu_2$ e $\sigma_1 > \sigma_2$); secondo la regola (1.1) un individuo per il quale si osserva x verrà assegnato a D_1 se

$$f_1(x) > f_2(x) \tag{1.4}$$

Semplici conti algebrici mostrano che si ha (1.4) se e soltanto se

$$\frac{\sigma_1}{\sigma_2} \exp\left[-\frac{1}{2}[\frac{(x-\mu_1)^2}{\sigma_1^2} - \frac{(x-\mu_2)^2}{\sigma_2^2}]\right] > 1 \tag{1.5}$$

Passando al logaritmo la (1.5) diventa

$$x^2\left[\frac{1}{\sigma_1^2}-\frac{1}{\sigma_2^2}\right]-2x\left[\frac{\mu_1}{\sigma_1^2}-\frac{\mu_2}{\sigma_2^2}\right]+\left[\frac{\mu_1^2}{\sigma_1^2}-\frac{\mu_2^2}{\sigma_2^2}\right]-2\ln\frac{\sigma_1}{\sigma_2}<0 \tag{1.6}$$

L'insieme dei valori x soluzione di

$$x^2\left[\frac{1}{\sigma_1^2}-\frac{1}{\sigma_2^2}\right]-2x\left[\frac{\mu_1}{\sigma_1^2}-\frac{\mu_2}{\sigma_2^2}\right]+\left[\frac{\mu_1^2}{\sigma_1^2}-\frac{\mu_2^2}{\sigma_2^2}\right]-2\ln\frac{\sigma_1}{\sigma_2}=0 \tag{1.7}$$

è detto *decision boundary* ed è di tipo quadratico, rappresentato da una parabola con 2 radici reali distinte e concavità rivolta verso il basso: pertanto divide l'asse $\mathbb{R}$ in 2 regioni di cui una è limitata e connessa e l'altra è illimitata e sconnessa. Se la (1.6) è verificata allora l'individuo per il quale è stato osservato il valore x in questione sarà assegnato al gruppo D_1, altrimenti a D_2

Osserviamo che nel caso in cui $\sigma_1=\sigma_2=\sigma$ la funzione discriminante è di tipo lineare: infatti la (1.7)diventa

$$-2x\left[\frac{\mu_1}{\sigma^2}-\frac{\mu_2}{\sigma^2}\right]+\left[\frac{\mu_1^2}{\sigma^2}-\frac{\mu_2^2}{\sigma^2}\right]<0 \tag{1.8}$$

Possiamo supporre senza perdita di generalità che $\mu_1>\mu_2$ e quindi dalla (1.8)otteniamo la condizione di appartenenza a D_2

$$-\frac{2x}{\sigma^2}(\mu_1-\mu_2)+\frac{1}{\sigma^2}(\mu_1^2-\mu_2^2)<0$$

$$\Leftrightarrow(\mu_1-\mu_2)(\mu_1+\mu_2)=(\mu_1^2-\mu_2^2)<2x(\mu_1-\mu_2)$$

$$\Leftrightarrow x>\frac{1}{2}(\mu_1+\mu_2) \tag{1.9}$$

Più interessante è il caso multivariato dove in un gruppo, D_1, al vettore di variabili aleatorie $\underline{x}$ si assegna una distribuzione normale multivariata

con vettore delle medie $\underline{\mu_1}$ e matrice di covarianza Σ e nell'altro , D_2 , una distribuzione normale multivariata con media $\underline{\mu_2}$ e la stessa matrice di covarianza Σ . Procedendo in maniera simile al caso univariato, utilizziamo sempre la regola (1.1) secondo la quale un individuo con vettore di osservazioni $\underline{x}$ è assegnato alla classe D_1 se

$$f_1(\underline{x}) > f_2(\underline{x})$$

che corrisponde a

$$\exp -\frac{1}{2}[(\underline{x}-\underline{\mu}_1)^t\Sigma^{-1}(\underline{x}-\underline{\mu}_1)-(\underline{x}-\underline{\mu}_2)^t\Sigma^{-1}(\underline{x}-\underline{\mu}_2)]>1 \tag{1.10}$$

Passamdo al logaritmo nella (1.10) si ottiene

$$-\frac{1}{2}[(\underline{x}-\underline{\mu}_1)^t\Sigma^{-1}(\underline{x}-\underline{\mu}_1)-(\underline{x}-\underline{\mu}_2)^t\Sigma^{-1}(\underline{x}-\underline{\mu}_2)]>0$$

$$\Leftrightarrow \frac{1}{2}\underline{x}^t(\Sigma^{-1}-\Sigma^{-1})\underline{x}+(\Sigma^{-1}\underline{\mu}_1-\Sigma^{-1}\underline{\mu}_2)^t\underline{x}+\frac{1}{2}(\underline{\mu}_2^t\Sigma^{-1}\underline{\mu}_2-\underline{\mu}_1^t\Sigma^{-1}\underline{\mu}_1)>0$$

dato che

$$\underline{x}^t(\Sigma^{-1}-\Sigma^{-1})\underline{x}=0$$

Possiamo quindi dire che la regola discriminante basata sulla (1.1) porta ad assegnare un individuo con vettore di osservazioni $\underline{x}$ a D_1 se

$$\underline{\alpha}^t(\underline{x}-\underline{\mu})>0 \tag{1.11}$$

dove

$$\underline{\alpha}=\Sigma^{-1}(\underline{\mu_1}-\underline{\mu_2}) \tag{1.12}$$

e

$$\underline{\mu}=\frac{1}{2}(\underline{\mu_1}+\underline{\mu_2}) \tag{1.13}$$

l'equazione

$$\underline{\alpha}^t(\underline{x}-\underline{\mu})=0$$

definisce un iperpiano nello spazio $\mathbb{R}^p$ dei p fattori. Individui per i quali il vettore delle osservazioni $\underline{x}$ è tale che $\underline{\alpha}^t(\underline{x}-\underline{\mu}) > 0$ vengono assegnati a D_1. Questo iperpiano prende il nome di *decision boundary*

In alcune situazioni è ragionevole assumere che membri di certi gruppi sono maggiormente osservabili che membri di altri . Ad esempio, in medicina, un raffreddore è una malattia più comune di una poliomelite ed ha perciò una maggiore probabilità a priori di essere osservata. Se i g gruppi hanno una probabilità a priori $\rho_1, \rho_2, ..., \rho_g$, la regola (1.1) cambia collocando un individuo con vettore di osservazioni, x, alla popolazione per cui

$$\rho_j f_j(\underline{x}) \tag{1.14}$$

è massimo. Nel caso di due gruppi descritti da densità multivariate normali con una comune matrice di covarianza la presa in considerazione delle distribuzioni a priori cambia la regola (1.1) in

$$\underline{\alpha}^t(\underline{x}-\underline{\mu}) > \ln\frac{\rho_2}{\rho_1} \tag{1.15}$$

1.3 Analisi Discriminante Classica

Le regole discriminanti viste nella sezione precedente per variabili distribuite in maniera normale presuppongono la conoscenza di media e matrice di covarianza. Nella pratica non è sempre possibile avere informazioni esatte su questi valori, che pertanto devono essere sostituiti con i rispettivi valori campionari. Mettiamoci nel caso precedentemente discusso di due gruppi D_1 e D_2 e di p fattori osservati descritti da densità normali multivariate con medie diverse ma stessa matrice di covarianza, allora regole come la (3.17) e la (1.15) possono essere usate semplicemente sostituendo i valori $\underline{\mu_1} = (\mu_{11}, \mu_{12}, ..., \mu_{1p})$,

$\underline{\mu_2} = (\mu_{21}, \mu_{22}, ..., \mu_{2p})$, Σ con i corrispondenti valori campionari $\underline{\widehat{x}_1}$, $\underline{\widehat{x}_2}$ e $\widehat{S}$. Avendo a disposizione un dataset composto da $N = n + k$ individui lo si può due parti: un "training set " $\mathcal{X} = \mathcal{X}_1 \cup \mathcal{X}_2$ dove $\mathcal{X}_1$ e $\mathcal{X}_2$ hanno cardinalitàrispettivamente n_1 e n_2 con $n_1 + n_2 = n$,inoltre $\mathcal{X}_1 =$ *trainig set costituito da individui appartenenti alla classe* D_1 mentre $\mathcal{X}_2 =$ *trainig set costituito da individui appartenenti alla classe* D_2 , con il quale stimare il vettore dei pesi $\underline{\alpha}$ del modello e un "validation set "di cardinalità k con il quale verificare la validità del modello, come sarà descritto più in dettaglio nel capitolo VI. Alla luce di queste considerazioni definiamo:

$$\underline{\widehat{x}}_1 = (\widehat{x}_{11}, \widehat{x}_{12}, ..., \widehat{x}_{1p}) = (\frac{1}{n_1}\sum_{k=1}^{n_1} x_{k1}^{(1)}, \frac{1}{n_1}\sum_{k=1}^{n_1} x_{k2}^{(1)}, ..., \frac{1}{n_1}\sum_{k=1}^{n_1} x_{kp}^{(1)})$$

$$\underline{\widehat{x}}_2 = (\widehat{x}_{21}, \widehat{x}_{22}, ..., \widehat{x}_{2p}) = (\frac{1}{n_2}\sum_{k=1}^{n_2} x_{k1}^{(2)}, \frac{1}{n_2}\sum_{k=1}^{n_2} x_{k2}^{(2)}, ..., \frac{1}{n_2}\sum_{k=1}^{n_2} x_{kp}^{(2)})$$

con $\widehat{x}_{ij} =$ *media campionaria all'interno del i-esimo gruppo del j-esimo fattore* ; la matrice di covarianza $\widehat{S}_W$, nel caso di due gruppi, è data da

$$\widehat{S}_W = \frac{1}{n_1 + n_2 - 2}[n_1\widehat{S}_{W1} + n_2\widehat{S}_{W2}] \tag{1.16}$$

con

$$\widehat{S}_{W1} = \sum_{i=1}^{n1}(\underline{x}_i - \underline{\widehat{x}}_1)(\underline{x}_i - \underline{\widehat{x}}_1)^t$$

e

$$\widehat{S}_{W2} = \sum_{i=n1+1}^{n1+n2}(\underline{x}_i - \underline{\widehat{x}}_2)(\underline{x}_i - \underline{\widehat{x}}_2)^t$$

dove n_1 e n_2 sono le dimensioni campionarie nei due gruppi mentre $\widehat{S}_{W1}$ e $\widehat{S}_{W2}$ sono le matrici empiriche di covarianza del gruppo . In tale contesto la regola di classificazione data da (3.17) diventa : si colloca un individuo per il quale è stato osservato $\underline{x} = (x_1, x_2, ..., x_p)$ nel gruppo D_1 se

$$\underline{\alpha}^t[\underline{x} - \frac{1}{2}(\underline{\widehat{x_1}} + \underline{\widehat{x_2}})] > 0 \tag{1.17}$$

che equivale a

$$\underline{\alpha}^t \underline{x} > \frac{1}{2}(\underline{\widehat{x_1}} + \underline{\widehat{x_2}}) \tag{1.18}$$

dove

$$\underline{\alpha} = \widehat{S}_W^{-1}(\underline{\widehat{x_1}} - \underline{\widehat{x_2}}) \tag{1.19}$$

La funzione

$$y = \underline{\alpha}^t \underline{x} \tag{1.20}$$

è conosciuta come: ***funzione discriminante lineare***. L'iperpiano

$$\underline{\alpha}^t \underline{x} - \frac{1}{2}(\underline{\widehat{x_1}} + \underline{\widehat{x_2}}) = 0$$

viene chiamato *iperpiano separatore* o *decision boundary*. Questa idea fu suggerita per la prima volta da Fisher (1963) [DA01]. Egli voleva trovare una combinazione lineare $\underline{\alpha}^t \underline{x}$ delle p variabili $\underline{\mathrm{x}}$ che separasse in modo sensato i due campioni di prova. La scelta di Fischer fu quella di determinare il vettore $\underline{\alpha}$ come quel vettore che rendesse massima la distanza fra le medie dei gruppi e, al contempo, minima la varianza all'interno di ciascun gruppo; dimostreremo più avanti che la (1.19) è la soluzione di tale problema.

Teorema 1.3.1. *Se A è simmetrica e $Q(\underline{x}) = \underline{x}^t A \underline{x}$ è la corrispondente forma quadratica allora esiste una trasformazione $\underline{x} \to \Gamma^t \underline{x} = \underline{y}$ tale che*

$$\underline{x}^t A \underline{x} = \sum_{i=1}^{p} \lambda_i y_i^2$$

dove λ_i sono gli autovalori di A

Dimostrazione $A = \Gamma \Delta \Gamma^t$ per la decomposizione di Jordan con

$$\Lambda = diag(\lambda_1, ..., \lambda_p)$$

matrice con gli autovalori sulla diagonale e zeri altrove, e

$$\Gamma = (\gamma_1, ..., \gamma_p)$$

matrice ortogonale di autovettori di A; e $\underline{y} = \Gamma^t \underline{x}$ abbiamo che $\underline{x}^t A \underline{x} = \underline{x}^t \Gamma \Delta \Gamma^t \underline{x} = \underline{y}^t \Lambda \underline{y} = \sum_{i=1}^p \lambda_i y_i^2$ }

□

Teorema 1.3.2. *Siano A e B due matrice simmetriche con $B > 0$, allora il massimo di $\underline{x}^t A \underline{x}$ con il vincolo $\underline{x}^t B \underline{x} = 1$ è Dato dal più grande autovalre di $B^{-1}A$. Più in generale:*

$$\max_{\underline{x}:\underline{x}^t B \underline{x}=1} \underline{x}^t A \underline{x} = \lambda_1 \geq \lambda_2 \geq ... \geq_p = \min_{\underline{x}:\underline{x}^t B \underline{x}=1} \underline{x}^t A \underline{x}$$

dove $\lambda_1, ..., \lambda_p$ sono gli autovalori di $B^{-1}A$. Il vettore che massimizza (minimizza) $\underline{x}^t A \underline{x}$ sotto il vincolo $\underline{x}^t B \underline{x} = 1$ è l'autovettore di $B^{-1}A$ che corrisponde al più grande (più piccolo) autovalore di $B^{-1}A$

Dimostrazione Per definizione $B^{1/2} = \Gamma_B \Lambda_B^{1/2} \Gamma_B^T$. Sia $\underline{y} = B^{1/2} \underline{x}$ allora

$$\max_{\{\underline{x}:\underline{x}^t B \underline{x}=1\}} \underline{x}^t A \underline{x} = \max_{\{\underline{y}:\underline{y}^t \underline{y}=1\}} \underline{y}^t B^{-1/2} A B^{-1/2} \underline{y} \tag{1.21}$$

applicando la decomposizione di Jordan sia

$$B^{-1/2} A B^{-1/2} = \Gamma \Lambda \Gamma^t$$

la decomposizione spettrale di $B^{-1/2} A B^{-1/2} = \Gamma \Lambda \Gamma^t$ sia

$$\underline{z} = \Gamma^t \underline{y} \Rightarrow \underline{z}^t \underline{z} = \underline{y}^t \Gamma \Gamma^t \underline{y} = \underline{y}^t \underline{y}$$

così la (1.21) è equivalente a

$$\max_{\{\underline{z}:\underline{z}^t \underline{z}=1\}} \underline{z}^t \Lambda \underline{z} = \max_{\{\underline{z}:\underline{z}^t \underline{z}=1\}} \sum_{i=1}^p \lambda_i \underline{z}_i^2$$

ma

$$\max_{\underline{z}} \sum \lambda_i z_i^2 \leq \lambda_1 \max \sum z_i^2 = \lambda_1$$

Il massimo è così ottenuto, notando che $\max \sum z_i^2 = 1$, da $\underline{z} = (1, 0, ..., 0)^t$, cioè

$$\underline{y} = \gamma_1 \Rightarrow \underline{x} = B^{-1/2}\gamma_1$$

dato che $B^{-1}A$ e $B^{-1/2}AB^{-1/2}$ hanno gli stessi autovalori, la prova è completa

□

Come abbiamo detto in precedenza il nostro scopo è di trovare un vettore che massimizza la distanza fra le medie del gruppo ed allo stesso tempo si minimizza la varianza all'interno di ciascun gruppo. Nel caso di due gruppi ($g = 2$), la matrice di covarianza all'interno del gruppo è data da

$$\widehat{S}_W = \frac{1}{n-2}[\sum_{j=1}^{2} \sum_{\underline{x}_i \in D_j} (\underline{x}_i - \widehat{\underline{x}}_j)(\underline{x}_i - \widehat{\underline{x}}_j)^t]$$

con

$$n = n1 + n2$$

la corrispondente matrice della differenza delle medie al quadrato fra i gruppi è data da:

$$\widehat{S}_B = (\widehat{\underline{x}}_2 - \widehat{\underline{x}}_1)(\widehat{\underline{x}}_2 - \widehat{\underline{x}}_1)^t \tag{1.22}$$

Seguendo l'idea di Fischer vogliamo massimizzare il seguente rapporto

$$J(\underline{\alpha}) = Q_B(\underline{\alpha})/Q_W(\underline{\alpha}) \tag{1.23}$$

dove

$$Q_B(\underline{\alpha}) = \underline{\alpha}^t \widehat{S}_B \underline{\alpha}$$

e

$$Q_W(\underline{\alpha}) = \underline{\alpha}^t \widehat{S}_W \underline{\alpha}$$

sono le forme quadratiche associate rispettivamente a $\widehat{S}_B$ e $\widehat{S}_W$. Bisogna notare che trasformando $\underline{\alpha} \to \lambda\underline{\alpha}$ la (1.23) non cambia: infatti si raccoglierebbe un λ^2 sia al numeratore che al denominatore. Massimizzare la (1.23) equivale a trovare il seguente massimo

$$\max_{\underline{\alpha}} \{Q_B(\underline{\alpha}) \ \ t.c. \ \ Q_W(\underline{\alpha}) = 1\}$$

che esiste per il teorema di Weierstrass dato che $Q_B(\underline{\alpha})$ è una funzione continua e l'insieme $\{\alpha \ \ t.c. \ \ Q_W(\underline{\alpha}) = 1\}$ è un compatto essendo antimmagine tramite una funzione continua di un chiuso.

Per risolvere il problema del massimo di $J(\underline{\alpha})$ alla luce delle osservazioni appena fatte possiamo utilizzare i teoremi (1.3.1) e (1.3.2) o alternativamente possiamo procedere utilizzando la tecnica dei moltiplicatori di Lagrange.

In generale per calcolare il massimo di una funzione $f(x)$ vincolato a una $g(x) = C$, si ricorre ai cosiddetti moltiplicatori di Lagrange, mediante una funzione del tipo

$$L(x, \lambda) = f(x) - \lambda[g(x) - C]$$

e si eguaglia a zero la derivata di $L(x, \lambda)$ rispetto a x ovvero

$$\frac{\partial L(x, \lambda)}{\partial x} = \frac{\partial f(x)}{\partial x} - \lambda \frac{\partial g(x)}{\partial x}$$

per la nostra forma quadratica $Q_B(\underline{\alpha})$ si vede che il vettore delle derivate parziali si scrive in forma matriciale

$$\frac{\partial(\underline{\alpha}^t \widehat{S}_B \underline{\alpha})}{\partial \underline{\alpha}} = 2\widehat{S}_B \underline{\alpha}$$

parimenti per il vincolo di normalizzazione

$$\frac{\partial(\underline{\alpha}^t \widehat{S}_W \underline{\alpha})}{\partial \underline{\alpha}} = 2\widehat{S}_W \underline{\alpha}$$

La ricerca del massimo implica, come si è detto, che s'annullino le derivate parziali di:

$$L = \underline{\alpha}^t \widehat{S}_B \underline{\alpha} - \lambda(\underline{\alpha}^t \widehat{S}_W \underline{\alpha})$$

quindi l'espressione

$$\frac{\partial L}{\partial \underline{\alpha}} = 2\widehat{S}_B \underline{\alpha} - 2\lambda \widehat{S}_W \underline{\alpha} = 0$$

esprime la condizione cercata. Se ne deduce la relazione

$$\widehat{S}_B \underline{\alpha} = \lambda \widehat{S}_W \underline{\alpha} \tag{1.24}$$

da cui se $\widehat{S}_W$ è definita positiva, dunque invertibile, otteniamo

$$\widehat{S}_W^{-1} \widehat{S}_B \underline{\alpha} = \lambda \underline{\alpha} \tag{1.25}$$

moltiplicando i membri della (1.24) per $\underline{\alpha}^t$ ottengo

$$\underline{\alpha}^t \widehat{S}_B \underline{\alpha} = \lambda \underline{\alpha}^t \widehat{S}_W \underline{\alpha}$$

e tenendo conto della condizione di normalizzazione ne deriva

$$\lambda = \underline{\alpha}^t \widehat{S}_B \underline{\alpha} \tag{1.26}$$

quindi la (1.25) e la (1.26) ci dicono che la (1.23) è massimizzata per $\underline{\alpha}$ autovettore associato al massimo autovalore della matrice

$$S_W^{-1} S_B$$

Se nella (1.25) si prende $\widehat{S}_B$ come in (1.22) abbiamo

$$\widehat{S}_W^{-1}\frac{1}{g-1}[(\underline{\widehat{x}}_1-\underline{\widehat{x}}_2)(\underline{\widehat{x}}_1-\underline{\widehat{x}}_2)^t\underline{\alpha}=\lambda\underline{\alpha}$$

dato che $(\underline{\widehat{x}}_1-\underline{\widehat{x}}_2)^t\underline{\alpha}$ è uno scalare ottengo

$$\underline{\alpha}\propto S^{-1}[(\underline{\widehat{x}}_1-\underline{\widehat{x}}_2)]$$

che è esattamente la (1.19). Esponiamo ora brevemente la generalizzazione dell'analisi discriminante di Fischer nel caso in cui $g>2$ facendo poi vedere come dal caso generale con semplici calcoli ci si riconduce al caso appena esposto $g=2$

Ci limiteremo, per quanto riguarda il caso generale a dare solo la definizione della matrice di covarianza nel gruppo $\widehat{S}_W$ e la corrispondente matrice fra i gruppi $\widehat{S}_B$ dato che la procedura di risoluzione non differisce da quella riguardante il caso $g=2$ dovendo sempre massimizzare il funzionale $J(\underline{\alpha})$. Le due matrici in questione si scrivono in questo modo:

$$\widehat{S}_W=\frac{(X-GM)^t)(X-GM)}{n-g}\qquad \widehat{S}_B=\frac{(GM-1\underline{\widehat{x}})^T(GM-1\bar{x})}{g-1}$$

Dove $g=$numero delle classi; $n=$cardinalità del campione iniziale; $p=$ numero dei fattori monitorati;

$$\underline{\widehat{x}}=(\widehat{x}_1,\widehat{x}_2,...,\widehat{x}_p)=(\frac{1}{n}\sum_{i=1}^{n}x_i^{(1)},\frac{1}{n}\sum_{i=1}^{n}x_i^{(2)},...,\frac{1}{n}\sum_{i=1}^{n}x_i^{(p)})$$

è il vettore delle medie dei fattori sull'intero campione,

$$X=\begin{pmatrix}x_{11} & x_{12} & \dots & x_{14}\\ x_{21} & x_{22} & \dots & x_{24}\\ x_{31} & x_{32} & \dots & x_{34}\\ \vdots & \dots & \dots & \vdots\\ x_{n1} & x_{n2} & \dots & x_{np}\end{pmatrix}$$

è la matrice $n \times p$ dei dati iniziali dove x_{ij} = j-esimo fattore relativo all'n-esimo individuo

$$M = \begin{pmatrix} \widehat{x}_{11} & \widehat{x}_{12} & ... & \widehat{x}_{1p} \\ \widehat{x}_{21} & \widehat{x}_{22} & ... & \widehat{x}_{2p} \\ \widehat{x}_{31} & \widehat{x}_{32} & ... & \widehat{x}_{3p} \\ : & ... & ... & : \\ \widehat{x}_{g1} & \widehat{x}_{g2} & ... & \widehat{x}_{gp} \end{pmatrix}$$

è la matrice $g \times p$ delle medie delle classi (i.e. $\widehat{x}_{ij} =$ *media nell'i-simo gruppo del j-esimo fattore*);

$$G = \begin{pmatrix} \gamma_{11} & \gamma_{12} & \cdots & \gamma_{1g} \\ \gamma_{21} & \gamma_{22} & \cdots & \gamma_{2g} \\ \gamma_{31} & \gamma_{32} & \cdots & \gamma_{3g} \\ : & ... & ... & : \\ \gamma_{n1} & \gamma_{n2} & \cdots & \gamma_{ng} \end{pmatrix}$$

è la matrice $n \times g$ indicatrice delle classi (cioè $\gamma_{ij} = 1$ se e soltanto se l'individuo i è assegnato alla classe j). Anche nel caso generale come nel caso $g = 2$ per massimizzare il funzionale $J(\underline{\alpha})$ si procede trovando il massimo autovalore della matrice $\widehat{S}_W^{-1}\widehat{S}_B$

Quanto detto esaurisce la breve descrizione del caso generale; facciamo ora vedere il collegamento dal caso generale a quello in cui abbiamo due gruppi $g = 2$; sia $g = 2$ allora abbiamo

$$G = \begin{pmatrix} \gamma_{11} & \gamma_{12} \\ \gamma_{21} & \gamma_{22} \\ \vdots & \vdots \\ \vdots & \vdots \\ \gamma_{n1} & \gamma_{n2} \end{pmatrix}$$

matrice $n \times 2$, possiamo supporre senza perdita di generalità che i dati siano ordinati e che quindi la matrice G sia del seguente tipo

$$G = \begin{pmatrix} \gamma_{11} & \gamma_{12} \\ \vdots & \vdots \\ \gamma_{n_1 1} & \gamma_{n_1 2} \\ \gamma_{n_{1+1} 1} & \gamma_{n_{1+1} 2} \\ \vdots & \vdots \\ \gamma_{n_1 1} & \gamma_{n_2 2} \end{pmatrix}$$

con $\gamma_{i1} = 1$ per $i = 1, \ldots, n_1$, $\gamma_{i1} = 0$ per $i = n_1 + 1, \ldots, n_2$ e $\gamma_{i2} = 0$ per $i = 1, \ldots, n_1$, $\gamma_{i1} = 1$ per $i = n_1 + 1, \ldots, n_2$

$$X = \begin{pmatrix} x_{11} & x_{12} & \ldots & x_{1p} \\ x_{21} & x_{22} & \ldots & x_{2p} \\ x_{31} & x_{32} & \ldots & x_{3p} \\ \vdots & \ldots & \ldots & \vdots \\ x_{n1} & x_{n2} & \ldots & x_{np} \end{pmatrix}$$

matrice $n \times p$ con $x_i j =$ dato iniziale relativo al j-esimo fattore dell'i-esimo individuo;

$$M = \begin{pmatrix} \widehat{x}_{11} & \widehat{x}_{12} & \ldots & \widehat{x}_{1p} \\ \widehat{x}_{21} & \widehat{x}_{22} & \ldots & \widehat{x}_{2p} \end{pmatrix}$$

matrice $2 \times p$ delle medie dei fattori in ciascun gruppo.

$$GM - 1\underline{\widehat{x}} = \begin{pmatrix} \widehat{x}_{11} - \widehat{x}_1 & \widehat{x}_{12} - \widehat{x}_2 & \ldots & \widehat{x}_{1p} - \widehat{x}_p \\ \vdots & \vdots & \vdots & \vdots \\ \widehat{x}_{n_1 1} - \widehat{x}_1 & \widehat{x}_{n_1 2} - \widehat{x}_2 & \ldots & \widehat{x}_{n_1 p} - \widehat{x}_p \\ \widehat{x}_{(n_1+1)1} - \widehat{x}_1 & \widehat{x}_{(n_1+1)2} - \widehat{x}_2 & \ldots & \widehat{x}_{(n_1+1)1} \\ \vdots & \vdots & \vdots & \vdots \\ \widehat{x}_{n1} - \widehat{x}_1 & \widehat{x}_{n2} - \widehat{x}_2 & \ldots & \widehat{x}_{np} - \widehat{x}_p \end{pmatrix}$$

Quindi la matrice fra i gruppi diventa

$$\widehat{S}_B = \frac{1}{2-1}(GM - 1\underline{\widehat{x}})^t(GM - 1\underline{\widehat{x}})$$

$$= \frac{1}{2-1}[(\underline{\widehat{x}}_1-\underline{\widehat{x}})(\underline{\widehat{x}}_1-\underline{\widehat{x}})^t+(\underline{\widehat{x}}_2-\underline{\widehat{x}})(\underline{\widehat{x}}_2-\underline{\widehat{x}})^t] = \frac{1}{2-1}[(\underline{\widehat{x}}_1-\underline{\widehat{x}}_2)^t(\underline{\widehat{x}}_1-\underline{\widehat{x}}_2)] \quad (1.27)$$

dove

$$(\underline{\widehat{x}}_i - \underline{\widehat{x}}) = (x_{i1} - \widehat{x}_1, ..., m_{p1} - \widehat{x}_p)$$

Per quanto riguarda la matrice all'interno del gruppo si ha

$$X - GM = \begin{pmatrix} x_{11} - \widehat{\underline{x}}_{11} & x_{12} - \widehat{\underline{x}}_{12} & \dots & x_{1p} - \widehat{\underline{x}}_{1p} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n_1 1} - \widehat{\underline{x}}_{11} & x_{n_1 2} - \widehat{\underline{x}}_{12} & \dots & x_{n_1 p} - \widehat{\underline{x}}_{1p} \\ x_{(n_1+1)1} - \widehat{\underline{x}}_{21} & x_{(n_1+1)2} - \widehat{\underline{x}}_{22} & \dots & x_{(n_1+1)p} - \widehat{\underline{x}}_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} - \widehat{\underline{x}}_{21} & x_{n2} - \widehat{\underline{x}}_{22} & \dots & x_{np} - \widehat{\underline{x}}_{2p} \end{pmatrix}$$

dunque

$$\widehat{S}_W = \frac{1}{n-2}(X - GM)^t(X - GM)$$

$$= \frac{1}{n-2}[\sum_{j=1}^{2} \sum_{x_i \in D_j} (x_i - \widehat{\underline{x}}_j)(x_i - \widehat{\underline{x}}_j)^t]$$

Dato che vogliamo massimizzare la (1.23) tramite il teorema (1.3.2)

$$\widehat{S}_W^{-1}\widehat{S}_B\underline{\alpha} = \lambda\underline{\alpha}$$

si applica la (1.27) e si ottiene

$$\widehat{S}_W^{-1}\frac{1}{2-1}[(\underline{\widehat{x}}_1 - \underline{\widehat{x}}_2)(\underline{\widehat{x}}_1 - \underline{\widehat{x}}_2)^t\underline{\alpha} = \lambda\underline{\alpha}$$

dato che $(\underline{\widehat{x}}_1 - \underline{\widehat{x}}_2)^t\underline{\alpha}$ è uno scalare ottengo

$$\underline{\alpha} \propto S_W^{-1}[(\underline{\widehat{x}}_1 - \underline{\widehat{x}}_2)]$$

che è esattamente la (1.19)

Le probabilità a priori possono essere introdotte nella regola discriminante come suggerito in (1.15). In alcuni casi le probabilità a priori possono essere ben approssimate dalla conoscenza della dimensione delle due popolazioni. Quando non si può dire molto circa la dimensione delle popolazioni in genere si usa la stessa probabilità a priori $\rho_1 = \rho_2 = \frac{1}{2}$;

Per illustrare l'uso della funzione discriminante di Fisher useremo i dati della tabella (1.1) (dove i primi 15 pazienti sono del gruppo dei malati e dove le variabili sono le seguenti domande: $x_1 =$ "ultimamente avverti di giocare un ruolo importante nelle cose?"; $x_2 =$ "ultimamente sei soddisfatto della tua vita?"; $x_3 =$ "ultimamente ti senti capace di prendere decisioni?"; $x_4 =$ "ultimamente ti senti come se non fossi capace di iniziare qualsiasi cosa?"; $x_5 =$ "ultimamente sei spaventato dalle cose che devi fare?"; le variabili sono codificate nel seguente modo: da 1 a 3 nella direzione da "no"a "si", da 4 a 5 nella direzione da "si"a "no"). Questi dati sono stati raccolti in base ad uno studio psichiatrico su larga scala nel quale a persone sane e persone ritenute, da un punto di vista psichiatrico, malate è stata posta una serie di domande relative a sentimenti di inadeguatezza, tensione etc. L'obiettivo era di ottenere una regola di classificazione grazie alla quale, in futuro, una persona poteva essere classificata come appartenente ad uno dei due gruppi. I vettori delle medie e la relativa matrice di covarianza sono mostrate nella tabella 1.2. La funzione discriminante di Fisher è data da

$$-0.72x_1 - 0.37x_2 + 0.02x_3 - 0.34x_4 - 0.41x_5 \tag{1.28}$$

e la regola di classificazione in (1.17) diventa, collocando un individuo con

punteggi x_1, x_2, x_3, x_4, x_5 al gruppo dei *malati* se

$$-0.72x_1 - 0.37x_2 + 0.02x_3 - 0.34x_4 - 0.41x_5 + 4.20 > 0 \qquad (1.29)$$

La regola appare abbastanza intuitiva, infatti, se un individuo ha un punteggio alto verrà catalogato fra i *sani* altrimenti fra i *malati*.
Come possiamo valutare la validità della funzione discriminante in maniera più formale? Un modo ovvio è di applicarla ai dati originali e verificare quanti individui sono classificati in maniera errata; la tabella 1.3 mostra il numero dei soggetti correttamente ed incorrettamente classificati dalla (1.29). Il tasso di misclassificazione, 13%, è relativamente basso, ma questo tipo di stima, cioè applicando la funzione discriminante ai dati originali, è molto ottimistica del tasso di misclassificazione.
Una stima più realistica può essere ottenuta, oltre che con il metodo descritto in precedenza che consiste nel dividere il dataset iniziale in training set e validation set, in diversi modi: un metodo altrenativo è il cosiddetto "leaving one out method"dove la funzione discriminante è calcolata sulla base di $(n-1)$ soggetti e usata per classificare l'individuo non incluso, si ripete poi lo stesso processo per ciascun soggetto. Applicando un simile approccio all'esempio appena discusso si ottiene il cambio di valori dalla tabella 1.3 alla tabella 1.4. La stima del tasso di errore di classificazione è ora del 23% .
A parte l'assunzione di normalità, la funzione discriminante in (1.19) presuppone che la matrice di covarianza delle due popolazioni in causa sia la stessa.

Quando non è questo il caso, la procedura descritta nella sezione precedente ci conduce ad una regola di questo tipo: si assegna un individuo con

Pazienti	x_1	x_2	x_3	x_4	x_5
1	2	2	2	2	2
2	2	2	2	1	2
3	1	1	2	1	1
4	2	2	2	1	2
5	1	1	2	1	2
6	1	1	2	1	1
7	2	2	2	2	2
8	1	1	2	1	2
9	1	1	2	1	2
10	2	1	2	1	2
11	2	2	2	1	2
12	2	1	2	1	2
13	1	1	2	2	2
14	1	1	2	1	2
15	3	3	2	3	2
16	4	3	3	3	2
17	3	3	2	3	3
18	3	2	2	3	2
19	4	2	2	2	2
20	2	3	2	3	3
21	2	2	2	2	3
22	3	2	2	1	3
23	3	3	2	1	3
24	2	2	2	2	2
25	3	1	3	4	4
26	2	2	3	1	2
27	3	2	2	4	2
28	3	2	2	3	3
29	2	2	2	3	1
30	3	2	4	3	3
31	3	1	3	1	3
32	1	2	2	1	2
33	3	3	2	4	3
34	2	3	2	4	3
35	3	3	3	4	3
36	2	1	2	3	3
37	4	4	4	4	4
38	2	1	2	3	3
39	4	1	4	4	4
40	3	3	2	2	3
41	2	2	2	1	2
42	4	2	2	2	2
43	3	3	2	3	3
44	2	3	2	2	2
45	4	3	1	2	3

Tabella 1.1: dati psichiatrici

	Variabili				
	x_1	x_2	x_3	x_4	x_5
malati	1.60	1.47	2.00	1.40	1.93
sani	2.80	2.27	2.33	2.60	2.70
$S =$	0.57				
	0.19	0.55			
	0.12	0.04	0.34		
	0.23	0.15	0.16	0.90	
	0.18	0.07	0.14	0.25	0.40

Tabella 1.2: vettori di media e matrice di covarianza dei dati psichiatrici

		Disposizione secondo la regola discriminante	
		malati	sani
gruppi veri	malati	14	1
	sani	5	25

Tabella 1.3: confronto fra funzione discriminante di Fischer e dati originali

		Disposizione secondo la regola discriminante	
		malati	sani
gruppi veri	malati	14	1
	sani	9	21

Tabella 1.4: confronto fra la funzione discriminante di fischer e il metodo one out

un vettore di dati x a D_1 se

$$\underline{x}^t(\widehat{S}_{W2}^{-1} - \widehat{S}_{W1}^{-1})\underline{x} - 2\underline{x}^t(\widehat{S}_{W2}^{-1}\widehat{\underline{x}}_2 - \widehat{S}_{W1}^{-1}\widehat{\underline{x}}_1) + (\widehat{\underline{x}}_2^t\widehat{S}_{W2}^{-1}\widehat{\underline{x}}_2 - \widehat{\underline{x}}_1^t\widehat{S}_{W1}^{-1}\widehat{\underline{x}}_1) \geq$$
$$\geq \ln(|\widehat{S}_{W2}|/|\widehat{S}_{W1}|) + 2ln(\Pi_1/\Pi_2) \quad (1.30)$$

dove S_1 ed S_2 sono le matrici di covarianza stimate per ciascun gruppo. Dato che il membro di sinistra della (1.30) contiene quadrati e prodotti riga per colonna, in questo caso l'equazione è detta *funzione discriminante quadratica*. Si applica soprattutto quando le medie delle popolazioni sono uguali e quindi la funzione lineare non è di nessuna utilità.

Se ci si trova in situazioni in cui la funzione discriminante di Fischer non riesce a classificare bene il nostro dataset, vale la pena di considerare dei metodi alternativi; uno dei più usati è la *discriminazzione logistica* logistica applicheremo applicata a

1.4 Relazione fra funzione discrimante di Fischer e il metodo dei minimi quadrati

In breve il metodo dei minimi quadrati consiste nel cercare di trovare un vettore di pesi $\underline{\xi}$ che soddisfi il sistema di equazioni

$$Y\underline{\xi} = \underline{b} \tag{1.31}$$

dove Y è una matrice $n \times p$ in cui le righe sono composte dai vettori y_i^t per $i = 1 \ldots n$. E' noto che quando ci sono più equazioni che incognite il sistema è sovradeterminato e di solito non esistono soluzioni esatte. Possiamo sempre cercare, comunque, un vettore di pesi $\underline{\xi}$ che minimizzi qualche funzione di errore fra $Y\underline{\xi}$ e $\underline{b}$. Nel caso del metodo dei minimi quadrati si definisce il vettore di errore come

$$\underline{e} = Y\underline{\xi} - \underline{b}$$

e si minimizza la funzione

$$J(\underline{\xi}) = \| \underline{e} \|^2 = \sum_{i=1}^{n} (\underline{\xi}^t \underline{y}_i - \underline{b}_i)^2 \tag{1.32}$$

Ora vediamo la relazione fra il metodo appena descritto e la funzione discriminante di Fischer; supponiamo di avere due gruppi e un set di n variabili p dimensionali $\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_p$ di cui le prime n_1 appartengono al primo gruppo D_1 e sono etichettate con ω_1 mentre le rimanenti n_2 appartengono a D_2 ed etichettate con ω_2. Allora la matrice Y può essere scritta come

$$Y = \begin{pmatrix} u_1 & \mathcal{X}_1 \\ -u_2 & -\mathcal{X}_2 \end{pmatrix}$$

dove $\underline{u}_i$ è il vettore colonna degli n_i e $\mathcal{X}_i$ è la matrice $n_i \times p$ le cui righe sono gli individui etichettati con ω_i ; scegliamo i vettori $\underline{\xi}$ e $\underline{b}$ nel seguente modo

$$\underline{\xi} = \begin{pmatrix} \alpha_0 \\ \underline{\alpha} \end{pmatrix} \tag{1.33}$$

e

$$\underline{b} = \begin{pmatrix} \frac{n}{n_1}\underline{u}_1 \\ \frac{n}{n_2}\underline{u}_2 \end{pmatrix} \tag{1.34}$$

vedremo che proprio questa particolare scelta di $\underline{b}$ lega il medoto dei minimi quadrati con la funzione discriminante di Fischer.

Partendo dalla (1.31) e moltiplicando ambo i membri per Y^t ottengo

$$Y^t Y \underline{\xi} = Y^t \underline{b} \tag{1.35}$$

sostituendo nella (1.35)le (1.4),(1.33),(1.34) ottengo

$$\begin{pmatrix} \underline{u}_1^t & -\underline{u}_2^t \\ \mathcal{X}_1 & -\mathcal{X}_2 \end{pmatrix} \begin{pmatrix} \underline{u}_1 & \mathcal{X}_1 \\ -\underline{u}_2 & -\mathcal{X}_2 \end{pmatrix} \begin{pmatrix} \alpha_0 \\ \underline{\alpha} \end{pmatrix}$$

$$= \begin{pmatrix} \underline{u}_1^t & -\underline{u}_2^t \\ \mathcal{X}_1^t & -\mathcal{X}_2^t \end{pmatrix} \begin{pmatrix} \frac{n}{n_1}\underline{u}_1 \\ \frac{n}{n_2}\underline{u}_2 \end{pmatrix} \tag{1.36}$$

tenendo presente che il vettore media all'interno dell' i-esimo gruppo è dato da

$$\widehat{\underline{x}}_i = \frac{1}{n_i} \sum_{\underline{x} \in D_i} \underline{x}$$

e la matrice di varianza e covarianza all'interno del gruppo è

$$S_W = \sum_{i=1}^{2} \sum_{\underline{x} \in D_i} (\underline{x} - \widehat{\underline{x}}_i)(\underline{x} - \widehat{\underline{x}}_i)^t$$

otteniamo

$$\begin{pmatrix} n & (n_1\widehat{\underline{x}}_1 + n_2\widehat{\underline{x}}_2)^t \\ (n_1\widehat{\underline{x}}_1 + n_2\widehat{\underline{x}}_2)_{21} & S_W + n_1\widehat{\underline{x}}_1\widehat{\underline{x}}_1^t + n_2\widehat{\underline{x}}_2\widehat{\underline{x}}_2^t \end{pmatrix} \begin{pmatrix} \alpha_0 \\ \underline{\alpha} \end{pmatrix}$$

$$= \begin{pmatrix} 0 \\ n(\widehat{\underline{x}}_1 - \widehat{\underline{x}}_2) \end{pmatrix} \tag{1.37}$$

Dove $S_W + n_1\widehat{\underline{x}}_1\widehat{\underline{x}}_1^t + n_2\widehat{\underline{x}}_2\widehat{\underline{x}}_2^t$ ottenuta dalla moltiplicazione di $\mathcal{X}_1^t\ \mathcal{X}_2^t$ per $\mathcal{X}_1\ \mathcal{X}_2$; Eseguendo il calcolo si ottiene

$$\begin{pmatrix} \sum_{i=1}^n (x_{i1})^2 & \dots & \sum_{i=1}^n (x_{i1})(x_{ip}) \\ \vdots & \vdots & \vdots \\ \sum_{i=1}^n (x_{i1})(x_{ip}) & \dots & \sum_{i=1}^n (x_{ip})^2 \end{pmatrix}$$

a cui si aggiunge e si toglie la matrice $n_1\widehat{\underline{x}}_1\widehat{\underline{x}}_1^t + n_2\widehat{\underline{x}}_2\widehat{\underline{x}}_2^t$ per ottenere:

$$S_W + n_1\widehat{\underline{x}}_1\widehat{\underline{x}}_1^t + n_2\widehat{\underline{x}}_2\widehat{\underline{x}}_2^t$$

La (1.37) può essere considerata una coppia di equazioni, la prima delle quali può essere risolta in termini di $\underline{\alpha}$:

$$\alpha_0 = -\widehat{\underline{x}}^t\,\underline{\alpha} \tag{1.38}$$

dove $\widehat{\underline{x}}$ è il vettore delle medie su tutto il campione. Sostituendo quest'ultima nella seconda equazione otteniamo

$$[\frac{1}{n}S^W + \frac{n_1 n_2}{n^2}(\widehat{\underline{x}}_1 - \widehat{\underline{x}}_2)(\widehat{\underline{x}}_1 - \widehat{\underline{x}}_2)^t]\underline{\alpha} = \widehat{\underline{x}}_1 - \widehat{\underline{x}}_2 \tag{1.39}$$

dato che il vettore $(\widehat{\underline{x}}_1 - \widehat{\underline{x}}_2)(\widehat{\underline{x}}_1 - \widehat{\underline{x}}_2)^t\alpha$ è sempre nella direzione di $\widehat{\underline{x}}_1 - \widehat{\underline{x}}_2$ per ogni valore di $\underline{\alpha}$, possiamo scrivere

$$\frac{n_1 n_2}{n^2}(\widehat{\underline{x}}_1 - \widehat{\underline{x}}_2)(\widehat{\underline{x}}_1 - \widehat{\underline{x}}_2)^t\underline{\alpha} = (1-\beta)(\widehat{\underline{x}}_1 - \widehat{\underline{x}}_2)$$

dove β è qualche scalare. L'equazione (1.39) diventa

$$\underline{\alpha} = \beta n(S^W)^{-1}(\widehat{\underline{x}}_1 - \widehat{\underline{x}}_2) \tag{1.40}$$

che, tranne per un ininfluente fattore scalare, è identica alla soluzione per l'equazione discriminante di Fischer.

Capitolo 2

Lo Z-Score

In questo capitolo vedremo un'applicazione al campo della finanza del metodo di classificazione esposto nel capitolo 1. Lo Z-score di Altman è un approccio multivariato basato sulla considerazione e l'analisi di più fattori ritenuti significativi nel determinare "lo stato di salute"di un azienda o di un istituto finanziario [CAN98],[AS02]. Questi fattori sono pesati e combinati per produrre una misura (un punteggio del rischio di credito) che discrimini nel miglior modo compagnie che falliscono da quelle che non lo fanno. Da un punto di vista operativo si ritiene che una tale misura esista e che sia possibile trovarla ; in quanto le compagnie che falliscono hanno bilanci e tendenze finanziarie molto diverse da quelle finanziariamente in attivo. Una banca che utilizzi questo metodo potrebbe accordare o rifiutare un prestito se il punteggio Z-score della compagnia in questione scende al di sotto di una determinata soglia. Altman ha basato il suo modello multivariato sui rapporti finanziari che vengono mostrati nella successiva tabella 2.1. Il modello Z-score continua ad essere applicato anche a compagnie private, compagnie emergenti e industrie non manufatturiere [CAN98].

Il modello Z-score utilizza 5 variabili (indici di bilancio) relativi all'analisi della liquidità, della redditività, della leva finanziaria, della solvibilità e dell'attività per determinare lo score:

$$Z = \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_3 + \alpha_4 X_4 + \alpha_5 X_5 \tag{2.1}$$

dove α_i è il peso relativo all'e-esimo fattore; il vettore dei pesi α si calcola con le tecniche mostrate nel capitolo I. Nell'esempio di Altman, il gruppo di fallimento D_2 consiste in 33 aziende manufatturiere in bancarotta , mentre il gruppo non fallimentare D_1 comprende un ugual numero di compagnie manufatturiere; siamo nel caso di $n = 66$ con $n_1 = n_2 = 33$ Le aziende nel gruppo non fallimentare erano ancora in attivo al tempo dell'analisi. In questo caso (2.1) diviene:

$$Z = 1.2X_1 + 1.4X_2 + 3.3X_3 + 0.6X_4 + 0.999X_5 \tag{2.2}$$

Il significato delle variabili è il seguente:

Modello Z-score

Variabile	media gruppo fallimento	media gruppo non fallimento
$X_1 = \frac{\text{working capital}}{\text{total assets}}$	-6.1%	41.4%
$X_2 = \frac{\text{retained earning}}{\text{total assets}}$	-62.6%	35.5%
$X_3 = \frac{\text{EBIT}}{\text{total assets}}$	-31.8%	15%
$X_4 = \frac{\text{market value of equity}}{\text{book value of liabilities}}$	40.1%	247.7%
$X_5 = \frac{\text{sales}}{\text{total assets}}$	1.5 volte	1.9 volte

Tabella 2.1: Il campione è di 33 aziende ; la funzione discriminante risulta essere $Z = 0.012X_1 + 0.014X_2 + 0.033X_3 + 0.006X_4 + 0.999X_5$.

X_1 **WORKING CAPITAL / TOTAL ASSETS** Il rapporto working capital / total assets (WC/TA) si trova spesso negli studi di problemi societari; esprime il valore delle attività liquide dell'azienda rispetto alla capitalizzazione totale. Il Working Capital è definito come la differenza fra "current assets"e "total liabilities", rispettivamente l'attivo ed il passivo corrente. Liquidità e dimensioni aziendali sono esplicitamente considerate. Solitamente, un azienda in perdita avrà piccolo attivo corrente rispetto al totale.

X_2 **RETAINED EARNING / TOTAL ASSETS** Tale indice esprime la capacità che un'azienda ha avuto di reinvestire i propri utili. Retained earning noto anche come (earnings surplus) è l'ammontare totale dei guadagni reinvestiti e/o le perdite (in caso di guadagno negativo) di un azienda durante la sua vita totale. L'età di un azienda è presa espressamente in considerazione dato che una compagnia relativamente giovane, ad esempio , probabilmente avrà una basso RE/TA poichè non ha avuto tempo per costruire le proprie riserve. Sembrerebbe che questa analisi discrimini le aziende giovani; in realtà la presa in considerazione di tale parametro è perfettamente ragionevole dato che un compagnia neo-costituita ha un'alta probabilità di fallimento. Nel 1996, per esempio, circa il 45% delle aziende andate fallirono nei primi cinque anni di attività.

X_3 **EARNING BEFORE INTEREST AND TAXES/TOTAL ASSETS** Il rapporto "utile operativo netto/attivo totale"(EBIT/TA) misura la vera produttività delle attività di un'impresa, depurate da qualsiasi fattore di leva finanziaria o fiscale. Proprio perchè la sopravvivenza

di un'azienda è basata sul potere di profitto della sua attività, questo rapporto è particolarmente appropriato per studi che si interessano di fallimento aziendale.

X_4 **MARKET VALUE OF EQUITY/BOOK VALUE OF TOTAL LIABILITIES** Il capitale azionario è misurato dalla combinazione del valore di mercato di tutti i pacchetti azionari, ordinari e privilegiati, dove il passivo è riferito a oggetti correnti e a lungo termine. Il rapporto fra "market value of equity"e "book of total liabilities"(MVE/TL) mostra come l'attivo di un azienda possa perdere in valore prima che le perdite superino le entrate, ed essa diventi insolvente. Per esempio, una società con un patrimonio netto pari a 1,000 $ e passività pari a 500 $ può sopportare una perdita del valore di due terzi del proprio attivo prima di divenire insolvente. Invece, se la stessa azienda avesse un patrimonio netto pari a 250 $ con lo stesso ammontare di passività, diverrebbe insolvente con una riduzione di solo un terzo del proprio attivo. Questo rapporto aggiunge un parametro di valore di mercato che altri studi sul fallimento aziendale non avevano preso in considerazione. Sin dall'inizio della modellizzazione, nel 1968, Altman propose di aggiungere leasing operativi e finanziari al passivo totale di un azienda.

X_5 **SALES/TOTAL ASSETS** (S/TA) Questo parametro è un rapporto finanziario standard che evidenzia la capacità di un'azienda di generare ricavi con un determinato valore dell'attivo patrimoniale. Esso misura la capacità imprenditoriale di rapportarsi con la competitività del mercato di riferimento dell'azienda. Data la sua, unica, relazione con tutte

le altre variabili del modello il rapporto ricavo di vendita/attivo totale è il secondo per importanza nel modello.

Come mostra la tabella 2.1 le medie dei due gruppi per quattro o cinque variabili differiscono di molto. Affinchè un modello possa funzionare a dovere la deviazione standard nel gruppo deve essere relativamente bassa. L'accu-

		collocazione nel gruppo	
Gruppi veri	grandezza del campione	1	2
1 (bancarotta)	33	31(94.0%)	2(6.0%)
2 (non bancarotta)	33	1(3.0%)	32(97.0%)
Precisione totale 95.0%			

Tabella 2.2: Risultati della classificazione un anno prima della bancarotta

		collocazione nel gruppo	
Gruppi veri	grandezza del campione	1	2
1 (bancarotta)	33	23(72.0%)	9(28.0%)
2 (non bancarotta)	33	2(6.0%)	31(94.0%)
Precisione totale 82.0%			

Tabella 2.3: Risultati della classificazione due anni prima della bancarotta

ratezza del modello Z-score è mostrata nelle tabelle 2.2 e 2.3. La precisione del modello era del 95% un anno prima della bancarotta e del 82% due anni prima. L'accuratezza nella classificazione è uno dei parametri di maggior interesse utilizzati per predire se un determinato modello funzionerà bene nella pratica. Questa precisione è espressa come: accuratezza del primo tipo

(la precisione con la quale un modello classifica aziende fallite come malate) e accuratezza del secondo tipo (la precisione con la quale un modello classifica aziende in salute come tali). La precisione totale è la combinazione di queste due. Generalmente la precisione del primo tipo è ritenuta più importante rispetto a quella del secondo tipo, perchè il non riuscire a classificare una compagnia in fallimento (errore del primo tipo) costa ad un creditore molto di più della possibilità di rifiutare il prestito ad un azienda ritenuta in fallimento ma che in realtà è in salute (errore del secondo tipo).

Dato che i risultati basati sul campione potrebbero essere influenzati dalla maniera in cui si campiona, è necessario fare un secondo test. Si potrebbero stimare, ad esempio, i parametri per il modello usando solo un sottoinsieme del campione originale e poi classificare i rimanenti soggetti sulla base dei parametri stabiliti.

Altman effettuò cinque differenti replicazioni del metodo suggerito, su un sottoinsieme di 16 aziende. Le cinque prove includevano:

1. campionamento casuale
2. analisi di tutte le rimanenti aziende partendo dalla prima
3. lo stesso test ma partendo dalla seconda azienda
4. scelta delle compagnie dalla prima alla sedicesima
5. scelta delle compagnie dalla diciassettesima alla trentaduesima

Tutti i risultati hanno mostrato che la funzione discriminante era statisticamente significativa. Sono stati effettuati anche test addizionali usando campioni totalmente indipendenti; l'errore del secondo tipo, in questi altri test, ha avuto una variazione compresa fra il 15 e il 20 percento.

2.1 Z-score e Bond ratings

Uno dei primi impieghi dei modelli per il credit-scoring è quello di assegnare ad ogni score (punteggio) un appropriato bond rating. Il rating delle obbligazioni è un giudizio analitico che le società di analisi specializzate attribuiscono ad un'obbligazione e quindi alla capacità dell'emittente di rispettare gli impegni di pagamento derivanti dall'emissione obbligazionaria stessa. Questo permette all'analista di valutare la probabilità di default di un azienda che richiede un prestito osservando i dati storici di ciascun bond rating. Studi condotti da Altman mostrano che la media dello Z-score per obbligazioni AAA, nel 1995, era attorno al 5.0 fino al 1.67 per obbligazioni B. Il punteggio medio per aziende B rientra nella zona di pericolo per lo Z-score; le compagnie che emettono obbligazioni B hanno, infatti, le caratteristiche delle aziende potensialmente fallimentari.

2.2 Z-score per aziende private

Altman ,nel 1993 , ritoccò il modello originario dello Z-score per poi applicarlo alle aziende private sostituendo il book value con il market value nel calcolare il rapporto X_4. Alla fine arrivò al seguente modello:

$$Z' = 0.717X_1 + 0.847X_2 + 3.107X_3 + 0.420X_4 + 0.998X_5 \qquad (2.3)$$

La tavola 2.4 mostra la precisione della classificazione, le medie del gruppo, e il modello Z′-score. La precisione del primo tipo (corretta identificazione di compagnie in bancarotta) del modello Z′-score è leggermente inferiore rispetto al modello classico (91% contro 94%); mentre l'accuratezza del secondo tipo (corretta identificazione di compagnie non in bancarotta) è la stessa:

97% . La media del gruppo delle compagnie in fallimento è più basso nel modello Z′-score che nell'altro (4.14 contro 5.02). La distribuzione dei punteggi è, quindi, più stretta con una maggiore intersezione fra i gruppi.

		collocazione nel gruppo	
Gruppi veri	grandezza del campione	1	2
1 (bancarotta)	33	30(90.9%)	3(9.1%)
2 (non bancarotta)	33	1(3.0%)	32(97.0%)
Precisione totale 95.0%			

Tabella 2.4: media del primo gruppo = 0.15; media del secondo gruppo= 4.14.

;

Capitolo 3

Discriminazione logistica

3.1 Il modello

Anche in questo capitolo affronteremo esclusivamente il problema di classificazione in presenza di due classi ($g = 2$). Nella discriminazione logistica si modellizza direttamente la probabilità che un generico individuo si trovi nella classe D_1 o in quella D_2 dato il vettore delle osservazioni dei p fattori $\underline{x} = (x_1, x_2, \ldots, x_p)$ [CB95]. La modellizzazione è del tipo seguente: denotiamo con Y la variabile che vale 1 se l'individuo è assegnato alla classe D_1 mentre vale 0 se è assegnato a quella D_2, allora si pone

$$P(Y = 1 \mid \underline{x}) = \frac{e^{\underline{\alpha}^t \underline{x}}}{1 + e^{\underline{\alpha}^t \underline{x}}} \tag{3.1}$$

$$P(Y = 0 \mid \underline{x}) = \frac{1}{1 + e^{\underline{\alpha}^t \underline{x}}} \tag{3.2}$$

Notiamo che la (3.1) e la (3.2) possono essere riscritte con l'uso della funzione

$$g(u) = \frac{1}{1 + e^u} \qquad u \in \mathbb{R} \tag{3.3}$$

nota con il nome di "sigmoide logistica "(3.2), nella forma:

$$P(Y = 1 \mid \underline{x}) = g(\underline{\alpha}^t \underline{x})$$

$$P(Y = 0 \mid \underline{x}) = 1 - g(\underline{\alpha}^t \underline{x})$$

Come nei capitoli precedenti anche qui si tratta di determinare in modo

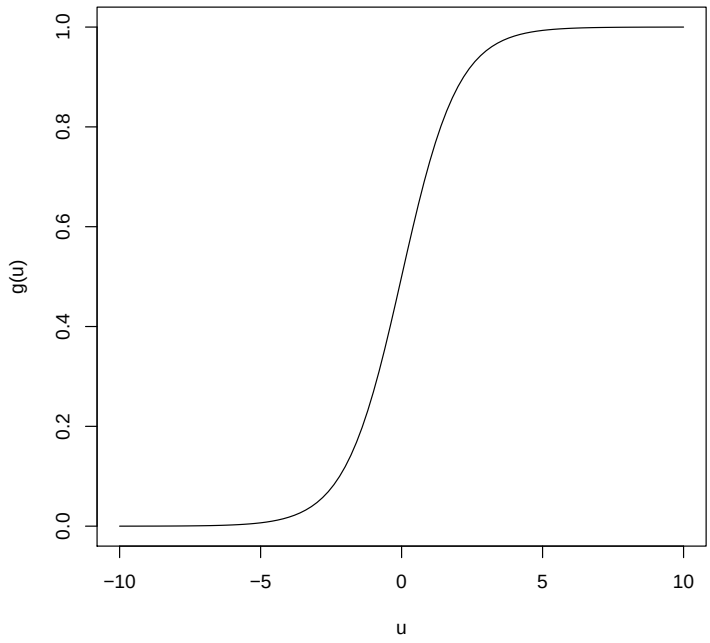

Figura 3.1: sigmoide logistica

appropriato i valori dei parametri liberi $\alpha_1, \alpha_2, \ldots, \alpha_p$. Uno dei possibili modi è di usare la tecnica della massima verosimiglianza. Supponiamo di avere un campione di $N = n + k$ individui dove $n =$ *cardinalità del trainig set* mentre $k =$ *cardinalità del validation set*, sia $\mathcal{X} = (\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_n)$ con $\underline{x}_i \in \mathbb{R}^p$ il

training set e scriviamo la funzione di verosimiglianza :

$$L(\underline{\alpha}) = \prod_{i=1}^{n} P(Y = y_i \mid \underline{x}_i) \tag{3.4}$$

e la log-verosimiglianza:

$$l(\underline{\alpha}) = \log(L(\underline{\alpha})) = \sum_{i=1}^{n} \{y_i \underline{\alpha}^t \underline{x}_i - \log(1 + e^{\underline{\alpha}^t \underline{x}_i})\} \tag{3.5}$$

Il gradiente della log-verosimiglianza è:

$$\nabla_{\underline{\alpha}} l(\underline{\alpha}) = (\frac{\partial l(\underline{\alpha})}{\partial \alpha_1}, \frac{\partial l(\underline{\alpha})}{\partial \alpha_2}, \ldots, \frac{\partial l(\underline{\alpha})}{\partial \alpha_p}) = \frac{\partial l}{\partial \underline{\alpha}} = \sum_{i=1}^{n} (y_i \underline{x}_i - \frac{e^{\underline{\alpha}^t \underline{x}_i}}{1 + e^{\underline{\alpha}^t \underline{x}_i}}) =$$

$$= \sum_{i=1}^{n} (y_i - \hat{y}_i)\underline{x}_i = (\sum_{i=1}^{n} (y_i - \hat{y}_i) x_{i1}, \sum_{i=1}^{n} (y_i - \hat{y}_i) x_{i2}, \ldots, \sum_{i=1}^{n} (y_i - \hat{y}_i) x_{ip}) \tag{3.6}$$

con

$$\widehat{y} = P(Y = 1 \mid \underline{x})$$

Calcoliamo ora la derivata seconda: per quanto detto finora abbiamo

$$\frac{\partial l(\underline{\alpha})}{\partial \alpha_r} = \sum_{i=1}^{n} (y_i - \hat{y}_i) x_{ir}$$

e quindi

$$\frac{\partial^2 l(\underline{\alpha})}{\partial \alpha_r \alpha_s} = \sum_{i=1}^{n} \frac{\partial}{\partial \alpha_s} [(y_i - \hat{y}_i) x_{ir}] = \sum_{i=1}^{n} [-\frac{\partial \hat{y}_i}{\partial \alpha_s}] x_{ir} \tag{3.7}$$

con

$$\frac{\partial \hat{y}_i}{\partial \alpha_s} = \frac{\partial}{\partial \alpha_s} (\frac{e^{\underline{\alpha}^t \underline{x}_i}}{1 + e^{\underline{\alpha}^t \underline{x}_i}}) = \frac{e^{-\underline{\alpha}^t \underline{x}_i}}{(1 + e^{-\underline{\alpha}^t \underline{x}_i})^2} x_{is} \tag{3.8}$$

In conclusione mettendo insieme la (3.13) e la (3.8) si ottiene ponendo

$$c_i(\alpha, \underline{x}_i) = \frac{e^{-\underline{\alpha}^t \underline{x}_i}}{(1 + e^{-\underline{\alpha}^t \underline{x}_i})^2} > 0$$

$$\frac{\partial^2 l(\underline{\alpha})}{\partial \alpha_r \alpha_s} = -\sum_{i=1}^{n} c_i(\alpha, \underline{x}_i) x_{ir} x_{is} = -XCX^t \tag{3.9}$$

con

$$X = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{14} \\ x_{21} & x_{22} & \dots & x_{24} \\ x_{31} & x_{32} & \dots & x_{34} \\ \vdots & \dots & \dots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix}$$

matrice $n \times p$ dei dati iniziali dove $x_{ij} =$ j-esimo fattore relativo all'n-esimo individuo;

$$C = \begin{pmatrix} c_1(\alpha, \underline{x}_1) & 0 & \dots & \dots & 0 \\ 0 & c_2(\alpha, \underline{x}_2) & 0 & \dots & 0 \\ \vdots & 0 & \ddots & \ddots & \vdots \\ \vdots & \dots & \ddots & \ddots & 0 \\ 0 & \dots & \dots & 0 & c_n(\alpha, \underline{x}_n) \end{pmatrix}$$

matrice diagonale $n \times n$. Affinchè la log-verosimiglianza abbia un massimo la (3.15) deve essere negativa e quindi la matrice $X^t C X$ deve essere definita positiva. Ricordiamo che una generica matrice A è definita positiva se per ogni vettore $\underline{v}$ si ha:

$$\underline{v}^t A \underline{v} > 0$$

La matrice C essendo diagonale ed avendo tutti gli elementi su di essa > 0 è definita positiva; dobbiamo dimostrare che lo sia anche $X^t C X$. Prendendo un generico vettore $\underline{v}$ abbiamo:

$$\underline{v}^t X^t C X \underline{v} = (X\underline{v})^t C (X\underline{v}) > 0 \quad se \;\; X\underline{v} \neq 0 \tag{3.10}$$

$X\underline{v} \neq 0$ solamente se X è iniettiva; in coordinate vogliamo che il seguente sistema abbia un unica soluzione:

$$\begin{pmatrix} x_{11} & x_{12} & \dots & x_{14} \\ x_{21} & x_{22} & \dots & x_{24} \\ x_{31} & x_{32} & \dots & x_{34} \\ \vdots & \dots & \dots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \\ v_3 \\ \vdots \\ v_n \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \tag{3.11}$$

Per il teorema di Rouche-Capelli la (3.11) ha un unica soluzione se il rango di X è uguale a p. Quindi se la matrice X ha rango p e la funzione di log-verosimiglianza è strettamente concava. Per calcolare lo stimatore di massima verosimiglianza $\widehat{\underline{\alpha}}$ data la concavità della (3.5) in $\underline{\alpha}$ si ricorre o a metodi numerici di tipo deterministico, oppure si usano algoritmi stocastici iterativi tipo "gradiente "

3.2 Uso del modello e trasformazione logit

Una volta stimato il vettore $\hat{\underline{\alpha}}$ medianante il metodo di massima verosimiglianza lo si usa per calcore i valori delle probabilità di trovarsi in D_1 o in D_2 mediante le formule:

$$P(Y = 1 \mid \underline{x}) = g(\hat{\underline{\alpha}}^t \underline{x}) \quad e \quad P(Y = 0 \mid \underline{x}) = 1 - g(\hat{\underline{\alpha}}^t \underline{x}) \tag{3.12}$$

Ad esempio supponiamo che una determinata banca non conceda mutui a privati cittadini il cui reddito annuo risulti al di sotto di 5,000 euro. Supponiamo di avere un archivio storico di dati relativi alla concessione del mutuo da parte della banca a fronte del reddito annuo percepito dal richiedente che

quindi in questo caso costituisce l'unico fattore ($p = 1$) monitorato. Per maggiore semplicità distinguiamo i possibili valori del reddito in 5 fascie definite nel seguente modo

- Fascia 1: reddito fra i 5,000 e i 10,000 euro annui.
- Fascia 2: reddito compreso fra 10,000 e 20,000 euro annui.
- Fascia 3: reddito copreso fra 20,000 e 30,000 euro annui.
- Fascia 4: reddito compreso fra 30,000 e 40,000 euro annui.
- Fascia 5: reddito superiore ai 40,000 euro annui.

I dati a nostra disposizione sono mostrati nella tabella 6.1 in cui

- *Num.richiedenti*=numero totale di persone richiedenti il mutuo per fascia di reddito.
- *Fascia*= fascia di reddito.
- *Concesso*= numero di persone relativa ad una fascia di reddito a cui è stato concesso il prestito.
- *Non concesso*=numero di persone relativa ad una fascia di reddito a cui non è stato concesso il prestito.
- *P.concesso*= probabilità, (ottenuta facendo il rapporto fra *Concesso* e *Num.richiedenti*) che il prestito venga concesso ad una persona appartenente ad una determinata fascia di reddito.

Fascia	concesso	non concesso	num. richiedenti	P.concesso
1	2	25	27	0.0741
2	17	7	24	0.7083
3	26	4	30	0.8667
4	19	8	27	0.7037
5	27	1	28	0.9643

Tabella 3.1: dati mutuo

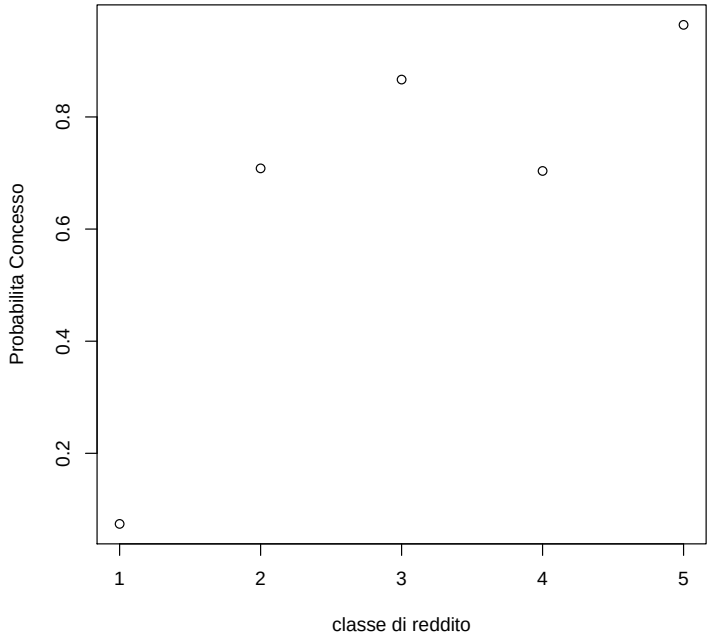

Figura 3.2: grafico probabilità stimate come nella tabella in relazione alle fascia di reddito

Utilizzando il metodo della massima verosimiglianza calcoliamo il vettore dei pesi che risulta essere

$$\underline{\hat{\alpha}} = (\hat{\alpha}_0, \hat{\alpha}_1) = (-2.122, 1.037) \tag{3.13}$$

dove α_1 è il peso relativo all'unico fattore in questione $x_1 =$ *reddito*; per la

$$P(Y = 1 \mid \underline{x})$$

abbiamo la rappresentazione

$$P(Y = 1 \mid x_1) = g(-2.122 + 1.037x_1) \tag{3.14}$$

Il grafico di tale funzione di x_1 è tracciato con linea continua nella figura(3.3). La curva graficata può ora essere usata come "decision boundary "relativamente alla erogazione o meno del prestito fatto dalla banca e di un nuovo richiedente. Per esempio dalla figura (3.3) risulta che se il richiedente ha un reddito annuo di 20,000 euro (inizio fascia 3) ha una probabilità all'incirca del 70% di ripianare completamente il debito. A questo punto se erogare o meno il prestito dipende da decisioni interne alla banca.
Esiste una trasformazione che rende il modello appena spiegato lineare; si effettua nel seguente modo:
dalle equazioni (3.1) (3.4) abbiamo

$$\frac{P(Y = 1 \mid \underline{x})}{P(Y = 0 \mid \underline{x})} = e^{\underline{\alpha}^t \underline{x}} \tag{3.15}$$

la (3.15) esprime le possibilità che si verifichi l'evento $Y = 1$ rispetto all'evento $Y = 0$; ad esempio se $P(Y = 1 \mid \underline{x}) = 0.8$ e $P(Y = 0 \mid \underline{x}) = 0.2$ la (3.15) sarà uguale a 4 quindi il primo evento sarà quattro volte più probabile del secondo.

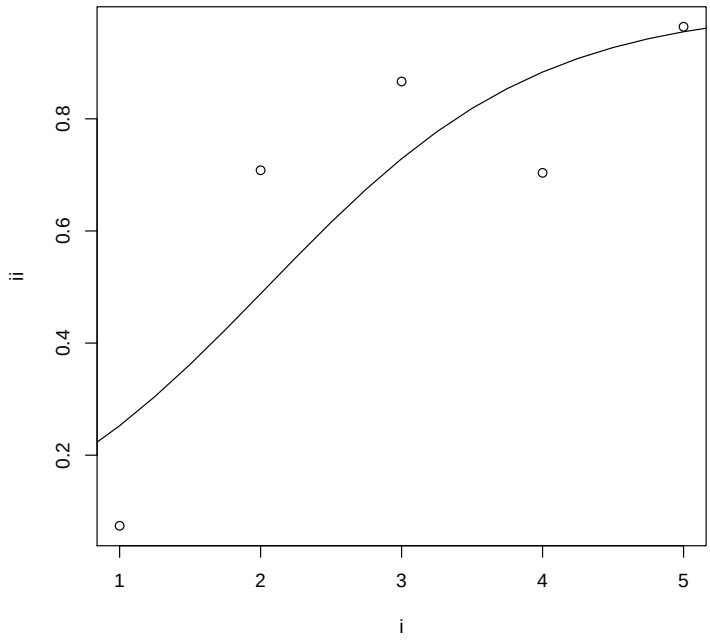

Figura 3.3: regressione logistica

La trasformazione logit consiste semplicemente nel prendere il logaritmo della (3.15).

$$logit(P(Y = 1 \mid \underline{x})) = \ln(e^{\underline{\alpha}^t \underline{x}}) = \underline{\alpha}^t \underline{x} \tag{3.16}$$

A questo punto il modello è diventato lineare possiamo quindi calcolare il vettore dei pesi $\underline{\hat{\alpha}}$ per poi ottenere tramite

$$logit(g(\underline{\hat{\alpha}})) = \hat{\alpha}_0 + \hat{\alpha}_1 \underline{x} \tag{3.17}$$

il grafico in figura (3.4)

che è equivalente a quello in figura (3.3) infatti si passa dall'uno all'altro semplicemente applicando la trasformazione logit o logit inversa.

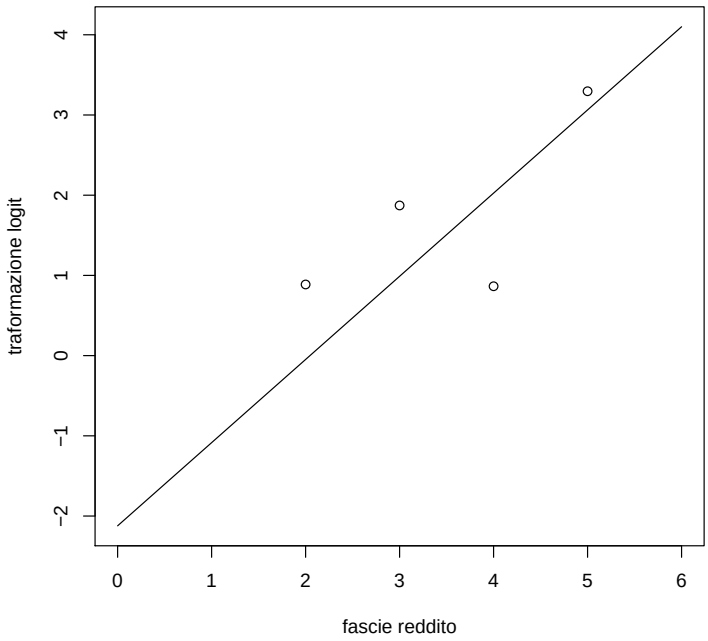

Figura 3.4: densità stimata della probabilità di concessione del prestito da parte della banca

Capitolo 4

Il Perceptrone

Anche in questo capitolo affronteremo esclusivamente il problema di classificazione in presenza di 2 classi ($g = 2$); il metodo considerato presenta somiglianze, come vedremo in seguito, con la regressione logistica e con l'analisi discriminante lineare. Anche qui si tratta di partire da un set di dati iniziali il cosiddetto training set $\mathcal{X} = \mathcal{X}_1 \cup \mathcal{X}_2$ il training set di cardinalità n dove $\mathcal{X}_1$ e $\mathcal{X}_2$ hanno cardinalità rispettivamente n_1 e n_2 con $n_1 + n_2 = n$, inoltre $\mathcal{X}_1 =$ *training set costituito da individui appartenenti alla classe* D_1 mentre $\mathcal{X}_2 =$ *training set costituito da individui appartenenti alla classe* D_2 ; etichettiamo poi gli individui appartenenti a $\mathcal{X}_1$ e a $\mathcal{X}_2$ rispettivamente con 1 e -1. e di determinare l'equazione di un iperpiano che separi le due classi di appartenenza (decision boundary) del tipo

$$\sum_{i=1}^{p} \alpha_i x_i + b = 0 \tag{4.1}$$

come si nota la (4.1) è simile all'equazione dell'iperpiano separatore vista precedentemente nel caso dell'analisi discriminante di Fisher e della regressione logistica; infatti la differenza fra i due metodi di classificazione analiz-

zati in precedenza e il metodo detto del *perceptrone* [SH98] consiste nel modo in cui si determinano i pesi $\alpha_1, \alpha_2, \ldots, \alpha_p$. Mettiamoci, ad esempio, nel caso $g = 2, p = 2$ e $n = 4$ con $\underline{x}_1, \underline{x}_2, \underline{x}_3 \in \mathcal{X}_1$ e $\underline{x}_4 \in \mathcal{X}_2$ dove $\underline{x}_i \in \mathbb{R}^p = \mathbb{R}^2$. Sia $\underline{x}_1 = (0,0), \underline{x}_2 = (1,0), \underline{x}_3 = (0,1), \underline{x}_4 = (1,1)$ in questo caso la (4.1) diventa

$$ax_1 + bx_2 + c = 0 \tag{4.2}$$

dove a, b giocano il ruolo di α_1, α_2 mentre c gioca il ruolo di b nella (4.1); diamo un valore iniziale al vettore dei pesi e a c ad esempio $\underline{\alpha}^{(0)} = (2,1)$ e $c = -1$; assumiamo, per includere nello stesso conto anche il valore $c =$ *punto in cui la retta (4.2) incontra l'asse delle ordinate*, la seguante notazione

$$\underline{\alpha} = (c, \alpha_1, \ldots, \alpha_p) \qquad \underline{x} = (1, x_1, \ldots, x_p)$$

con $\underline{x}$, $\underline{\alpha} \in \mathbb{R}^p$ pur avendo entrambi i vettori $p+1$ componenti, inoltre $\underline{\alpha}$ vettore dei pesi relativi ai p fattori e $\underline{x} =$ *vettore di p fattori relativo ad un generico individuo*; quindi la (4.2) diventa

$$\underline{\alpha}^t \underline{x} = 0 \tag{4.3}$$

A questo punto iniziamo a stimare il vettore dei pesi per classificare correttamente ogni elemento del nostro training set; come prima cosa estraiamo un elemento $\underline{x}_i$ dal training set e calcoliamo

$$sgn(\underline{\alpha}^{(0)t} \underline{y}^{(0)}) = sgn((-1,2,1)(1,0,0)) = -1 \tag{4.4}$$

dove sgn e la funzione segno e $\underline{y}^{(0)}$ è il generico vettore $\underline{x}_i$ analizzato dal perceptrone al passo indicato dall'apice, in questo caso $\underline{y}^{(0)} = (0,0,1)$; si controlla se il risultato della (4.4) equivale al valore dell'etichetta di $\underline{y}^{(0)}$ e si

pone

$$\underline{\alpha}^{(1)} = \begin{cases} \underline{\alpha}^{(0)} & se\ \underline{y}^{(0)}\ è\ correttamente\ classificato \\ \underline{\alpha}^{(0)} + \eta \underline{y}^{(0)} l(\underline{y}^{(0)}) & se\ \underline{y}^{(m)}\ è\ classificato\ in\ maniera\ errata \end{cases} \tag{4.5}$$

con $l(\underline{y}^{(0)}) = etichetta\ di\ \underline{y}^{(0)}$ e η parametro positivo minore di uno; nel nostro caso il vettore è correttamente classificato quindi il vettore dei pesi, secondo la regola espressa dalla (4.5), rimane invariato; se al successivo passo iterativo poniamo $\underline{y}^{(1)} = (1, 0, 1)$ abbiamo

$$sgn(\underline{\alpha}^{(1)t} \underline{y}^{(1)}) = sgn((-1, 2, 1)(1, 1, 0)) = 1$$

e ci accorgiamo che il vettore in questione è classificato in maniera errata quindi il vettore dei pesi viene aggiornato secondo la regola (4.5). Il procedimento appena esposto continua finche non viene individuato un vettore di pesi che permetta di costruire una retta che classifichi in maniera corretta tutti gli elementi del training set. Nel caso più generale con $g = 2$ e p generico l'algoritmo del perceptrone è riassunto nella seguente maniera:

- **Passo 1:** si inizializza il vettore dei pesi, ad esempio si pone utilizzando la convenzione precedente

$$\underline{\alpha}^{0} = (b, \underbrace{0, 0, \ldots, 0}_{pvolte}) \tag{4.6}$$

  dove con $\underline{\alpha}^{(0)}$ si vuole indicare il vettore dei pesi al passo iterativo "0".

- **Passo 2:** si prende in input, sempre conservando la notazione precedente quindi

$$\underline{x} = (1, x_1, \ldots, x_p)$$

un vettore $\underline{y}^{(m)} \in \mathcal{X}$, dove $\underline{y}^{(m)}$ è il generico vettore di osservazioni $\underline{x} \in \mathcal{X}$ dato come input al perceptrone al passo "m", si calcola il valore di

$$sgn(\underline{\alpha}^{(m)t}\underline{y}^{(m)}) \tag{4.7}$$

dove

$$sgn(x) = \begin{cases} 1 & se\ x > 0 \\ -1 & se\ x \leq 0 \end{cases}$$

- **Passo 3:** si confronta il valore della (4.7) con quello dell'etichetta $\{-1, 1\}$ del vettore $\underline{y}^{(m)}$; se sono uguali, quindi il vettore $\underline{x}_i$ è stato classificato correttamente, allora il vettore dei pesi si lascia invariato; mentre se $\underline{x}_i$ è stato classificato in maniera errata, cioè il valore della (4.7) è diverso dal valore dell'etichetta allora $\underline{\alpha}^{(m)}$ viene aggiornato. La procedura di aggiornamento appena descritta è riassunta in maniera seguente:

$$\underline{\alpha}^{(m+1)} = \begin{cases} \underline{\alpha}^{(m)} & se\ \underline{y}^{(m)}\ è\ correttamente\ classificato \\ \underline{\alpha}^{(m)} + \eta\underline{y}^{(m)}l(\underline{y}^{(m)}) & se\ \underline{y}^{(m)}\ è\ classificato\ in\ maniera\ errata \end{cases} \tag{4.8}$$

dove η si prende uguale a 1 e

$$l(\underline{y}^{(m)}) = \begin{cases} +1, & \text{se } \underline{y}^{(m)} \text{ appartiene al training set } \mathcal{X}_1 \\ -1, & \text{se } \underline{y}^{(m)} \text{ appartiene al training set } \mathcal{X}_2 \end{cases}$$

è il valore dell'etichetta del vettore $\underline{y}^{(m)}$. Si ritorna al passo 2.

4.1 Teorema di convergenza del perceptrone

Analizzando la procedura appena descritta ci si accorge che non c'è la garanzia di trovare un vettore di pesi $\underline{\alpha}$ che sia in grado di classificare correttamente ogni vettore $\underline{x} \in \mathcal{X}$, il procedimento potrebbe essere iterato all'infinito senza mai trovare il vettore $\underline{\alpha}$ in questione; Rosenblatt nel 1962 dimostrò che se i dati presenti in $\mathcal{X}$ sono linearmente separabili allora è possibile trovare, in un numero finito di passi, il vettore di pesi $\underline{\alpha}$ che classifichi correttamente ogni $\underline{x} \in \mathcal{X}$.

Teorema 4.1.1. *Se i due training set $\mathcal{X}_1$ e $\mathcal{X}_2$ sono linearmente separabili allora $\exists\, M_0$ indice di itarazione t.c. $\forall$ indice di iterazione $m \geq N_0$ tutti gli elementi del training set $\mathcal{X} = \mathcal{X}_1 \cup \mathcal{X}_2$ sono classificati correttamente.*

Dimostrazione

sia $M_0 >> n$ dove, come detto in precedenza, $n = n_1 + n_2$ è la cardinalità di $\mathcal{X}$ mentre $n_i =$ cardinalità di $\mathcal{X}_i$; sia

$T_i^M = T^M(\underline{x}_i) =$ *numero di volte che* $\underline{x}_i$ *è classificato in modo errato in* M *iterazioni*

chiaramente

$$T_M = T_1^M + T_2^M + \ldots + T_n^M \leq M$$

con T_M numero totale di classificazioni errate in M passi. Affinchè il teorema risulti provato deve esistere $C > 0$ t.c.

$$T_M < C \qquad \forall M \tag{4.9}$$

supponiamo che

$$\underline{\alpha}^{(0)} = (\underbrace{0, 0, \ldots, 0}_{p \text{ volte}}, b)$$

dall'ipotesi sappiamo che esite $\widehat{\underline{\alpha}}$ t.c. l'iperpiano

$$\widehat{\underline{\alpha}}^t \underline{x} = 0$$

separa linearmente i due training set $\mathcal{X}_1$ e $\mathcal{X}_2$ in $\mathbb{R}^p$ quindi abbiamo

$$\widehat{\underline{\alpha}}^t \underline{x}_i l(\underline{x}_i) > 0 \qquad i = 1, \ldots, n$$

da cui

$$\widehat{\underline{\alpha}}^t \underline{y}^{(m)} l(\underline{y}^{(m)}) > 0 \qquad \forall m \tag{4.10}$$

seguendo la procedura descritta in precedenza il vettore dei pesi all' M-esima iterazione è:

$$\widehat{\underline{\alpha}}^{(M)} = \sum_{m=1}^{M} T^M(\underline{y}^m) \underline{y}^m l(\underline{y}^m) \tag{4.11}$$

moltiplicando entrambi i membri della (4.11) per $\widehat{\underline{\alpha}}$ si ottiene:

$$\widehat{\underline{\alpha}}^t \underline{\alpha}^{(M)} = \sum_{m=1}^{M} T^M(\underline{y}^m) \widehat{\underline{\alpha}}^t \underline{y}^m l(\underline{y}^m) \tag{4.12}$$

utilizzando la (4.10) definiamo la quantità

$$r = \min_{\{m=1,\ldots,M\}} \{\widehat{\underline{\alpha}}^t \underline{y}^{(m)} l(\underline{y}^{(m)})\}$$

quindi ottengo

$$\widehat{\underline{\alpha}}^t \underline{\alpha}^{(M)} \leq r \sum_{m=1}^{M} T^M(\underline{y}^m) = r T_M \leq rM \tag{4.13}$$

La disugualianza di Cauchy-Schwarz ci dice che

$$\|\widehat{\underline{\alpha}}\|^2 \|\underline{\alpha}^{(M)}\|^2 \geq [\widehat{\underline{\alpha}}^t \underline{\alpha}^{(M)}]^2$$

utilizzando la (4.13) notiamo che

$$[\widehat{\underline{\alpha}}^t \underline{\alpha}^{(M)}]^2 \geq r^2 M^2$$

quindi ottieniamo:

$$\|\underline{\alpha}^{(M)}\|^2 \geq \frac{r^2 M^2}{\|\widehat{\underline{\alpha}}\|^2} \tag{4.14}$$

Supponiamo adesso che il perceptrone classifichi incorrettamente un vettore $\underline{y}^{(m)}$; in base a quanto detto finora si ha

$$\underline{\alpha}^{(m)} = \underline{\alpha}^{(m-1)} + \underline{y}^{(m-1)} l(\underline{y}^{(m-1)})$$

passando alle norme al quadrato si ottiene:

$$\|\underline{\alpha}^{(m)}\|^2 = \|\underline{\alpha}^{(m-1)}\|^2 + \|\underline{y}^{(m-1)}\|^2 + 2\underline{\alpha}^{(m-1)t}\underline{y}^{(m-1)} l(\underline{y}^{(m-1)}) \tag{4.15}$$

dato che $\underline{\alpha}^{(m-1)t}\underline{y}^{(m-1)} l(\underline{y}^{(m-1)}) < 0$ perchè il vettore $\underline{y}^{(m-1)}$ è classificato incorrettamente e $\|l(\underline{y}^{(m-1)}\| = 1$ la (4.15)diventa:

$$\|\underline{\alpha}^{(m)}\|^2 - \|\underline{\alpha}^{(m-1)}\|^2 \leq \|\underline{y}^{(m-1)}\|^2 \tag{4.16}$$

sommando la disuguaglianza (4.16) per $m = 1, \ldots, M$ e prendendo il vettore dei pesi iniziale come in (4.6) si ottiene:

$$\|\underline{\alpha}^{(M)}\|^2 \leq \sum_{m=1}^{M} \|\underline{\alpha}^{(m)}\| \leq Ms \tag{4.17}$$

dove

$$s = \max_{\{\underline{x} \in \mathcal{X}\}} \|\underline{x}\|^2$$

La (4.17) e la (4.14) sono in contraddizione fra loro per grandi valori di M; possiamo però trovare un valore massimo di M_{max} oltre il quale le due disuguaglianze in questione non sono più verificate contemporaneamente; il valore in questione è trovato uguagliando la (4.17) e la (4.14) e risolvendo per M

$$\frac{M_{max}^2 r^2}{\|\widehat{\underline{\alpha}}\|^2} = M_{max} s \Leftrightarrow M_{max} = \frac{s\|\widehat{\underline{\alpha}}\|^2}{r^2}$$

Variabili e parametri :

$\underline{x}$ = vettore dei dati iniziali a $p+1$ componenti

$= [+1, x_1, x_2, \ldots, x_p]^t$

$\underline{\alpha}^{(m)}$ = vettore dei pesi, a $p+1$ componenti, al passo "m"

$= [b, \alpha_1^{(m)}, \alpha_2^{(m)}, \ldots, \alpha_p^{(m)}]^t$

$\underline{y}^{(m)}$ = generico vettore $\underline{x} \in \mathcal{X}$ dato come input al perceptrone al passo "m"

$l(\underline{y}^{(m)})$ = risposta sperata

η = parametro di apprendimento positivo minore o uguale a 1

Primo passo : si pone $\underline{\alpha}^{(0)} = \underline{0}$ poi si procede come segue

per $m = 1, 2, \ldots$

Secondo passo : al passo m-esimo il perceptrone si attiva

applicando ad esso il vettore di dati iniziali $\underline{y}^{(m)}$ e la risposta

desiderata $l(\underline{y}^{(m)})$

Terzo passo : si calcola la risposta $sgn[\underline{\alpha}^{(m)t}\underline{y}^{(m)}]$

Quarto passo : adattamento del vettore dei pesi

$\underline{\alpha}^{(m+1)} = \underline{\alpha}^{(m)} + \eta\underline{y}^{(m)}l(\underline{y}^{(m)})$ *se* $\underline{y}^{(m)}$ *è classificato in maniera errata*

$\underline{\alpha}^{(m+1)} = \underline{\alpha}^{(m)}$ *se* $\underline{y}^{(m)}$ *è correttamente classificato*

Quinto passo : incrementare di 1 il passo e ritornare al Secondo passo

Tabella 4.1: Algoritmo di apprendimento del perceptrone

Abbiamo così dimostrato che adottando la procedura descritta in precedenza dopo al massimo M_{max} passi il perceptrone riesce a classificare correttamente tutti gli elementi del training set $\mathcal{X}$.

□

L'algoritmo di settaggio dei pesi del perceptrone è riportato in tabella (4.1)

4.2 Relazione fra il perceptrone e la regressione logistica

Mostriamo ora sempre nel caso di due gruppi $g = 2$ la relazione fra la regressione logistica e il metodo di classificazione, appena descritto, del perceptrone. Nell'affrontare un problema di classificazione con $g = 2$ nel caso in cui i dati provengono da due distribuzioni di tipo gaussiano la regressione logistica si riduce ad una combinazione lineare molto simile a quella del perceptrone. Per illustrare analogie e differenze fra i due metodi di discriminazione mettiamoci nel caso gaussiano in cui i vettori media nei due gruppi siano differenti

$$\underline{\mu}_1 \neq \underline{\mu}_2$$

mentre le matrici di covarianza sono uguali

$$\Sigma = \Sigma_1 = \Sigma_2$$

La densità condizionate alla classe di appartenenza è, quindi, data da

$$p(\underline{x} \mid D_k) = \frac{1}{(2\pi)^{n/2} \mid \Sigma \mid^{1/2}} \exp\{-\frac{1}{2}(\underline{x} - \underline{\mu}_k)^T \Sigma^{-1} (\underline{x} - \underline{\mu}_k)\} \qquad (4.19)$$

dove n è la dimenzione del vettore dei fattori ovvero $p = n$ Si assume anche che

$$p(D_1) = p(D_2) = \frac{1}{2}$$

cioè la probabilità che un vettore di osservazione $\underline{x}$ appartenga al sottospazio $\mathcal{X}_i$ con $i = 1, 2$ sia uguale, cioè le due classi sono equiprobabili[CB95]. Usando il teorema di Bayes otteniamo:

$$P(Y = 1 \mid \underline{x}) = P(D_1 \mid \underline{x}) = \frac{P(\underline{x} \mid D_1)P(D_1)}{P(\underline{x} \mid D_1)P(D_1) + P(\underline{x} \mid D_2)P(D_2)}$$

$$\frac{1}{1 + \frac{P(\underline{x}|D_2)P(D_2)}{P(\underline{x}|D_1)P(D_1)}} \tag{4.20}$$

Ponendo

$$a = \ln \frac{P(\underline{x} \mid D_1)P(D_1)}{P(\underline{x} \mid D_2)P(D_2)}$$

si ottiene che la (4.20) è uguale a

$$\frac{1}{1 + g(-a)} = g(a) \tag{4.21}$$

con $g(a)$ =sigmoide logistica. Tenendo conto delle assunzioni precedenti a diventa il logaritmo del rapporto delle funzioni di verosimiglianza nei due gruppi

$$a = \ln \frac{P(\underline{x} \mid D_1)}{P(\underline{x} \mid D_2)}$$

svolgendo i conti esattamente come nel capito I nell'analisi del caso gaussiano otteniamo:

$$y = \underline{\alpha}^t \underline{x} + \alpha_0 \tag{4.22}$$

dove volendo fare il collegamento con il metodo del percettere α_0 gioca il ruolo di "b"e

$$y = log \frac{p(\underline{x} \mid D_1)}{p(\underline{x} \mid D_2)} \tag{4.23}$$

dove Il vettore dei pesi è dato da:

$$\underline{\alpha} = S_w^{-1}(\underline{\mu}_1 - \underline{\mu}_2) \tag{4.24}$$

con

$$\alpha_0 = \frac{1}{2}(\underline{\mu}_2^t S_w^{-1} \underline{\mu}_2 - \underline{\mu}_1^t S_w^{-1} \underline{\mu}_1) \tag{4.25}$$

la regola discriminante in questo caso è : assegnare un vettore di osservazioni $\underline{x}$ alla classe D_1 se $y > 0$ e alla classe D_2 altrimenti.
Le similitudini e le differenze, fra questi due procedimenti , che vale la pena di sottolineare sono:

- La forma dell'equazione (4.22) è praticamente identica alla (1.20)solo che il metodo per calcolare il vettore dei pesi $\underline{\alpha}$ è totalmente differente poichè il perceptrone si basa su una metodologia di correzione di errore iterativa partendo da una condizione iniziale mentre la regressione logistica o anche il metodo di massima verosimiglianza, il calcolo si effettua a partire dalle distribuzioni di probabilità all'interno delle due classi.

- Il metodo del perceptrone non dipende dal tipo di distribuzione dei dati ed è semplice da implementare infatti come input richiede una distorsione, un vettore di dati e un vettore di pesi a differenza della regressione che richiede, come fatto notare anche prima, l'assegnazione di una qualche densità di probabilità all'interno delle classi.

- Il perceptrone converge solo se le classi sono linearmente separabili. Le distribuzioni gaussiane prese in considerazione hanno una zona di sovraposizione e quindi le due classi non sono sicuramente separabili. In

questo caso quindi la convergenza del metodo del perceptrone potrebbe creare problemi.

Capitolo 5

Alberi decisionali

Esaminiamo ora un altro metodo di classificazione detto degli *alberi decisionali* [RN98].

Un albero decisionale prende in input come gli altri metodi analizzati fin'ora, un individuo descritto da un vettore $\underline{x} = (x_1, x_2, \ldots, x_p)$ contenente i p fattori ed emette in uscita una "decisione "del tipo sì/no. Ciascun nodo interno all'albero corrisponde ad un test sul valore di una delle proprietà e gli archi che partono da ciascun nodo sono etichettati con i possibili valori del relativo test. Ciascuna foglia specifica il valore booleano di output se si perviene a tale foglia.

Supponiamo, ad esempio, di trovarci nel caso in cui una banca debba decidere se concedere o meno un prestito ad un privato cittadino utilizzando un albero decisionale. Lo scopo, in questo caso, è quello di trovare una *definizione* espressa sotto forma di albero di decisione per il nostro problema: cioè concedere o meno il prestito. Come prima cosa bisogna decidere quali proprietà sono le più adatte per descrivere il problema. Supponiamo di aver deciso la lista dei seguenti p attributi:

1. *Single*: se è single o no.
2. *Storia dei prestiti*: se, fin'ora, ha restituito tutti i prestiti in tempo.
3. *Cittadinanza*: se il richiedente è extracomunitario oppure no.
4. *Reddito*: se ha un reddito fisso o no.
5. *Conto bancario*: ammontare del conto del richiedente; classificato con 1, 2 o 3 a seconda se il conto sia in rosso o uguale a 0; compreso fra 0 e 2000 euro; oppure superiore a 2000 euro.
6. *Prestito*: quanti soldi vengono richiesti in prestito; le categorie sono catalogate dalla richiesta minore a quella maggiore ($, $$, $$$).
7. *Crediti*: se il richiedente ha ancora dei debiti da saldare.
8. *Casa*: se il richiedente è proprietario di un immobile.
9. *Tipo*: come il richiedente ha intenzione di utilizzare il prestito (mutuo, macchina, educazione, spese per la casa).
10. *Durata del prestito*: il tempo in cui il richiedente si impegna a restituire il prestito espresso in mesi (entro 12 mesi, 12-18, 18-24, >24)

5.1 Induzione di alberi decisionali a partire da esempi

Un *esempio* è descritto dai valori dei p attributi in questione e dal valore dell'esito (nel nostro caso decisione). L'esito ci permette di classificare l'esempio; se l'esito è positivo per un determinato esempio definiamo tale esempio

positivo; in caso contrario diciamo che è negativo. Un insieme di esempi descritti dai vettori

$$\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_n$$

con

$$n = n_1 + n_2$$

per il problema del prestito è mostrato in tabella (1.4). Gli esempi positivi sono quelli per cui l'esito, *Concedo il prestito*, è positivo

$$(\underline{x}_1, \underline{x}_3, \ldots)$$

in totale n_1 elementi; mentre gli esempi negativi sono i rimanenti $n_2 = n - n_1$ elementi

$$(\underline{x}_2, \underline{x}_5, \ldots)$$

L'insieme completo è chiamato *insieme di addestramento.*

Ora dobbiamo trovare un albero decisionale che sia consistente con l'insieme di addestramento: potremmo semplicemente costruire un albero di decisione con un cammino completo fino a una foglia per ciascun esempio, in cui in tale cammino si verifica che ciascun attributo possieda lo stesso valore assunto nell'esempio e la foglia abbia la stessa classificazione dell'esempio. Qualora lo stesso esempio, o un esempio con la stessa descrizione, venga presentato nuovamente all'albero di decisione esso sarà in grado di fornire la classificazione corretta. Sfortunatamente non sarà molto utile negli altri casi!

Il problema di questo albero costruito in modo banale è che si limita a memorizzare le osservazioni senza estrarre alcuno schema dagli esempi, di conseguenza, non ci si può attendere che sia in grado di classificare esempi che

Esempio	attributi										decisione
	Sin	St.pr	Cit	Red	Con.ban	Pre	Cre	Ca	Ti	Du.pre	
$\underline{x}_1$	sì	no	no	sì	3	$$$	no	sì	ma	$0-12$	sì
$\underline{x}_2$	sì	no	no	sì	1	$	no	no	edu	$18-24$	no
$\underline{x}_3$	no	sì	no	no	3	$	no	no	ca	$0-12$	sì
$\underline{x}_4$	sì	no	sì	sì	1	$	no	no	edu	$12-18$	sì
$\underline{x}_5$	sì	no	sì	no	1	$$$	no	si	ma	> 24	no
$\underline{x}_6$	no	sì	no	sì	3	$$	sì	sì	mu	$0-12$	sì
$\underline{x}_7$	no	sì	no	no	2	$	sì	no	ca	$0-12$	no
$\underline{x}_8$	no	no	no	sì	3	$$	sì	sì	edu	$0-12$	sì
$\underline{x}_9$	no	sì	sì	no	1	$	si	no	ca	> 24	no
$\underline{x}_{10}$	sì	sì	sì	sì	1	$$$	no	sì	mu	$12-18$	no
$\underline{x}_{11}$	no	no	no	no	2	$	no	no	edu	$0-12$	no
$\underline{x}_{12}$	sì	sì	sì	sì	1	$	no	no	ca	$18-24$	sì

Tabella 5.1: esempi per il problema del prestito

non ha mai visto.

Estrarre uno schema significa essere in grado di descrivere un gran numero di casi in maniera concisa. L'idea fondamentale alla base dell'algoritmo Apprendimento-Alberi-Decisione consiste nel cercare di verificare per primi gli attributi più importanti, cioè quelli che fanno più differenza nella classificazione di un esempio. In questo modo possiamo sperare di arrivare alla classificazione corretta con un piccolo numero di test, il che significa che i cammini dell'albero saranno brevi e l'albero nel suo complesso risulterà piccolo.

La figura (5.1) mostra come l'algoritmo comincia ad operare. Abbiamo

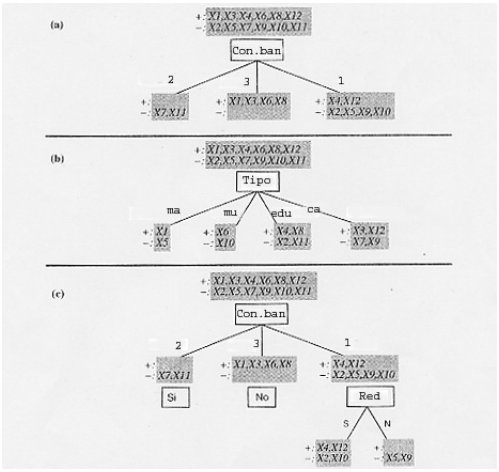

Figura 5.1: Suddivisione degli esempi a seguito di un test sugli attributi. La figura è tratta da [RN98], in cui Xi equivale al nostro $\underline{x}_i$

dodici esempi di addestramento che classifichiamo in un insieme positivo e uno negativo. A questo punto decidiamo quale attributo usare come primo test in un albero. La figura (5.1a) mostra che *Conto bancario* è un attributo importante perchè se il suo valore è 2 oppure 3 ci ritroviamo con insiemi di

esempi per cui possiamo dare una risposta sicura (no e sì rispettivamente). Se il valore è 1 allora abbiamo bisogno di test aggiuntivi. In figura (5.1b) vediamo che *Tipo* è un attributo poco utile, perchè ci lascia con quattro risultati possibili, ciascuno dei quali ha lo stesso numero di risposte positive e negative. Valutiamo tutti i possibili attributi in questo modo e scegliamo quello più importante come test associato alla radice. Vedremo in seguito come viene misurata l'importanza. Per il momento assumiamo che l'attributo più importante sia *Conto bancario*.

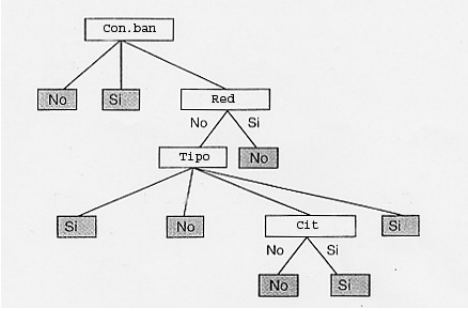

Figura 5.2: Albero di decisione indotto dall'insieme di addestramento di dodici esempi

Dopo che il test sul primo attributo ha suddiviso gli esempi, ciascuno dei risultati è in se stesso un nuovo problema di apprendimento di un albero di decisione, con meno esempi e un attributo in meno. Ci sono quattro casi da considerare per questi sottoproblemi ricorsivi:

1. Se ci sono sia esempi positivi che esempi negativi, si scelga l'attributo migliore per suddividerli. la figura (5.1c) mostra il caso in cui *Reddito* viene usato per suddividere i rimanenti esempi.

2. Se i rimanenti esempi sono tutti positivi (o tutti negativi), allora abbi-

amo terminato: possiamo rispondere sì o no. La figura (5.1c) mostra degli esempi nei casi in cui l'attributo *Conto bancario* assume i valori 2 e 3.

3. Se ci sono esempi rimasti significa che nessun esempio con quei valori per gli attributi è stato osservato e quindi restituiamo un valore di default calcolato in base alla maggioranza delle classificazioni relative al nodo progenitore.

4. Se non ci sono attributi rimasti ma ci sono ancora esempi sia positivi che negativi allora abbiamo un problema: significa che questi esempi hanno esattamente la stessa descrizione ma classificazioni differenti. Questo accade quando alcuni dei dati sono scorretti: diciamo in questo caso che c'è rumore nei dati. La stessa circostanza si può presentare anche quando gli attributi non danno abbastanza informazioni per descrivere completamente la situazione oppure quando il dominio è genuinamente non deterministico. Un modo semplice per liberarsi del problema consiste nel servirsi di un voto a maggioranza; cioè se ad un certo nodo vi sono più esempi negativi che positivi, verranno tutti classificati come negativi e viceversa.

Continuiamo ad applicare l'algoritmo finchè non otteniamo l'albero mostrato in figura (5.2). Si potrebbe credere che l'algoritmo non stia facendo un buon lavoro nell'apprendere la funzione corretta ma non è così; esso infatti tiene conto degli esempi non della funzione corretta e, in effetti, la sua ipotesi non solo concorda con tutti gli esempi, ma è anche considerevolmente più semplice di un albero costruito usando la prima metodologia spiegata. L'algoritmo di apprendimento non ha alcun motivo di includere test per *Crediti* o *Casa* dato

funzione *Apprendimento-Alberi-Decisione* (*esempi, attributi, default*) **returns** un albero di decisione

inputs *esempi*, insieme di esempi

attributi, insieme dei fattori

default, valore di default del predicato

if *esempi* è vuoto **return** *default*

else if tutti gli *esempi* hanno la stessa classificazione **then return** la classificazione

else if *attributi* è vuoto **then return** VALORE-MAGGIORANZA(*esempi*)

else

migliore←SCEGLI-ATTRIBUTO(*attributi, esempi*)

albero←un nuovo albero di decisione con radice con test *migliore*

for each valore v_i di *migliore* **do**

$esempi_i \leftarrow \{$ elementi di *esempi* con *migliore* $= v_i\}$

sottoalbero ← APPRENDIMENTO-ALBERI-DECISIONE ($esempi_i$,

, *attributi, migliore,*VALORE-MAGGIORANZA(*esempi*)

aggiungi un ramo ad *albero* con etichetta v_i e sottoalbero *sottoalbero*

end

return *albero*

Tabella 5.2: algoritmo di apprendimento degli alberi di decisione

che può classificare tutti gli esempi senza di essi.
Naturalmente più esempi si hanno e più l'albero è preciso; infatti l'albero in figura (5.2) non classificherà correttamente il caso in cui la durata del prestito è di 0-12 mesi e l'attributo *Conto bancario* ha il valore 1, dato che non ha mai visto un esempio del genere.

5.2 Procedura matematica

Ora viene spiegata la teoria matematica usata per costruire un albero decisionale. Mostreremo un modello matematico per scegliere l'attributo migliore. L'idea è quella di cercare di minimizzare la profondità dell'albero cercando di selezionare l'attributo che più di ogni altro riesce a fornire una classificazione esatta degli esempi. Un attributo perfetto è quello che suddivide gli esempi in sottoinsiemi che sono completamente positivi o negativi. Facendo riferimento all'esempio della concessione del mutuo, l'attributo *Conto bancario* non è perfetto ma è piuttosto buono. Un attributo veramente inutile come tipo lascia gli insiemi di esempi con, approssimativamente, la stessa proporzione tra esempi positivi e negativi che era presente nell'insieme originario.
Abbiamo bisogno di una misura formale di buono e di inutile: con essa possiamo implementare la funzione SCEGLI-ATTRIBUTO dell'algoritmo in tabella (5.2). La misura dovrebbe avere valore massimo quando l'attributo è perfetto e minimo quando è totalmente inutile. Una misura possibile è il valore atteso della quantità di informazione fornita dall'attributo. Per comprendere la nozione di informazione immaginiamo che essa fornisca la risposta, ad esempio, alla domanda : "se lancio una moneta uscirà testa o croce?"se già si ha qualche informazione sulla risposta allora la risposta vera e propria sarà

meno informativa. Supponiamo di voler scommettere un dollaro sul lancio di una moneta e di ritenere che la moneta sia truccata in modo che esca testa con probabilità 0.99. Ovviamente scommeteremo testa ed avremo un valore atteso di \$0,99 per la scommessa; vuol dire che saremmo disposti a pagare non più di \$0.01 per avere in anticipo informazioni sul risultato effettivo del lancio. Se la moneta non fosse stata truccata il valore atteso sarebbe nullo e saremmo stati disposti a pagare fino a \$1 per avere informazioni in anticipo. In breve: meno si conosce più l'informazione è preziosa.

La teoria dell'informazione si basa sullo stesso concetto appena espresso solo che misura il contenuto informativo in bit piuttosto che in dollari. Un bit di informazioni è sufficiente per rispondere sì o no ad una domanda sulla quale non si sa nulla, come l'esito di una moneta truccata. In generale, se le possibili risposte v_i hanno probabilità $P(v_i)$ allora il contenuto informativo I della risposta è dato da

$$I(P(v_1), P(v_2), \ldots, P(v_n)) = \sum_{i=1}^{n} -P(v_i) \log_2 P(v_i)$$

che altro non è se non la media dei contenuti informativi dei vari elementi(i termini $-\log_2 P(v_i)$ pesati con la probabilità degli eventi stessi. Nel caso di una moneta non truccata abbiamo

$$I(\frac{1}{2}\frac{1}{2}) = -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2} = 1 \;\; bit$$

mentre nel caso di quella truccata abbiamo

$$I(0.99, 0.1) = -0.99\log_2(o.99) - 0.01\log_2(0.01) = 0.08 \;\; bit$$

L'informazione contenuta nella risposta tende a zero al tendere a uno della probabilità di ottenere essa.

Per l'apprendimento di alberi di decisione la domanda cui dobbiamo dare risposta è: per un dato esempio, qual'è la classificazione corretta? Una stima delle probabilità di ciascuna delle risposte possibili è data dalla proporzione tra gli esempi positivi e quelli negativi nell'insieme di addestramento. Supponiamo che l'insieme di addestramento contenga h esempi positivi e k negativi. Allora una stima dell'informazione contenuta in una risposta corretta è data da:

$$I(\frac{h}{h+k}, \frac{k}{h+k}) = -\frac{h}{h+k}\log_2(\frac{h}{h+k}) - \frac{k}{h+k}\log_2(\frac{k}{h+k})$$

Per l'insieme di addestramento mostrato in tabella (5.1) abbiamo $k = h = 6$, quindi abbiamo bisogno di un bit di informazione.

Un test su un singolo attributo A non ci fornirà, in genere, tutta questa informazione ma ce ne darà una parte. Possiamo misurare esattamente quanta andando a vedere di quanta informazione abbiamo ancora bisogno dopo il test sull'attributo. Ogni attributo A divide l'insieme di addestramento E in sottoinsiemi $E_1, E_2, \ldots, E_v$ in base al valore degli esempi su A,in cui A può avere v valori distinti. Ciascun sottoinsieme E_i ha p_i esempi positivi e n_i negativi, in modo tale che se seguiamo quel ramo abbiamo bisogno di altri

$$I(\frac{h_i}{h_i + k_i}, \frac{k_i}{h_i + k_i})$$

bit di informazione per rispondere alla domanda. Un esempio a caso ha il valore i-esimo per l'attributo con probabilità $(h_i + k_i)/(h + k)$, così, in media, dopo aver effettuato il test sull'attributo A si ha ancora bisogno di:

$$Resto(A) = \sum_{i=1}^{v} \frac{h_i + k_i}{h + k} I(\frac{h_i}{h_i + k_i}, \frac{k_i}{h_i + k_i})$$

bit di informazione per classificare l'esempio. Il guadagno di informazione derivante dal test sull'attributo è definito come: la differenza tra la necessità

di informazione originaria e la nuova necessità

$$Guadagno(A) = I(\frac{h}{h+k}, \frac{k}{h+k}) - Resto(A)$$

La logica usata nella funzione SCEGLI-ATTRIBUTO consiste semplicemente nello scegliere l'attributo con il massimo guadagno.

Ritornando al nostro esempio e facendo riferimento alla figura (5.1) per quanto riguarda gli attributi *Conto bancario* e *Tipo* abbiamo:

$$Guadagno(Conto\ bancario) = 1-[\frac{2}{12}I(0,1)+\frac{4}{12}I(1,0)+\frac{6}{12}I(\frac{2}{6},\frac{4}{6})] \approx 0.541 \text{ bit}$$

$$Guadagno(Tipo) = 1 - [\frac{2}{12}I(\frac{1}{2},\frac{1}{2}) + \frac{2}{12}I(\frac{1}{2},\frac{1}{2}) + \frac{4}{12}I(\frac{2}{4},\frac{2}{4})+ \\ +\frac{4}{12}I(\frac{2}{4},\frac{2}{4})] = 0 \text{ bit}$$

In effetti *Conto bancario* presenta il guadagno massimo tra tutti gli attributi e sarebbe scelto come radice dall'algoritmo di apprendimento degli alberi di decisione.

Capitolo 6

Implementazione numerica, studio di dataset reali

In questo capitolo mostriamo l'applicazione, tramite il paccheto statistico **R** e nel caso del perceptrone del programma **matlab**, dell'analisi discriminante di Fischer , della regressione logistica, del metodo degli alberi decisionali e appunto del perceptrone nel classificare in due classi D_1 e D_2 un campione di 1000 persone rivoltasi ad una determinata banca tedesca per la richiesta di un prestito [VR02],[IM03]. Nel nostro caso D_1 è la classe delle persone che hanno ottenuto il prestito mentre D_2 è la classe di coloro che non lo hanno ottenuto. I dati sono stati scaricati dal sito ftp.ics.edu/pub/machine-learning-databases/statlog/german/databases

Come ho detto ci troviamo in presenza di un campione di 1000 individui ognuno dei quali è individuato da $p = 24$ fattori diversi indicanti il tipo di prestito che si richiede , la nazionalità , il sesso etc etc...

Si riporta in seguito una breve descrizione dei fattori $(x_1, x_2, \ldots, x_p)$

Descrizione del dataset germania

1) *Informazioni sulla fonte* :

   Professor Dr. Hans Hofmann Institut fur Statistik und Okonometrie Universitat Hamburg FB Wirtschaftswissenschaften Von-Melle-Park 5 2000 Hamburg 13

2) *Numbero del campione*: 1000

3) *Numero attributi del dataset german*: 20 (7 numerici, 13 categorici)

   Numero attributi del dataset german.numer: 24 (24 numerici)

4) *Descrizione attributi per il dataset german* : Riportiamo ora una descrizione parziale del dataset per rendere un idea di quali sono gli attributi usati da una banca per decidere l'erogazione o meno del prestito ad un privato cittadino.

   a. *Attribute* 1: (qualitative)

      Status of existing checking account

      A11 : ... < 0 DM

      A12 : 0 <= ... < 200 DM

      A13 : ... >= 200 DM

      salary assignments for at least 1 year

      A14 : no checking account

b. *Attribute* 2: (numerical)

   Duration in month

c. *Attribute* 3: (qualitative)

   Credit history

   A30 : no credits taken , all credits paid back duly

   A31 : all credits at this bank paid back duly

   A32 : existing credits paid back duly till now

   A33 : delay in paying off in the past

   A34 : critical account , other credits existing (not at this bank)

d. Attribute 4: (qualitative)

   Purpose

   A40 : car (new)

   A41 : car (used)

   A42 : furniture/equipment

   A43 : radio/television

   A44 : domestic appliances

   A45 : repairs

   A46 : education

   A47 : (vacation - does not exist?)

   A48 : retraining

   A49 : business

   A410 : others

5) *Matrice di verifica*

Questo dataset richiede l'uso di una matrice di verifica (tabella 6.1)

	0	1
0	0	1
1	5	0

Tabella 6.1: dati mutuo

le righe rappresentano la vera classificazione mentre le colonne quella predetta;

é peggio classificare un cliente come buono quando non lo sono (5) , che classificarlo il contrario(1).

6.1 Implementazione al calcolatore

Mostriamo ora l'implementazione al calcolatore, tramite il pacchetto statistico **R** e del software **Matlab**, dei quattro metodi di classificazione analizzati nei capitoli precedenti: l'analisi discriminante di Fischer, la regressione logistica, il perceptrone e il metodo degli alberi decisionali. Nell'eseguire l'analisi al calcolatore ho diviso l'intero dataset numerico german.num, costituito da 1000 esempi, in due parti: la prima costituita da 800 individui usata come *training set* e la seconda costituita da 200 individui usata come *validation set* per testare la bontà del modello. Inoltre, come detto in precedenza nel dataset german.num la decisione finale presa dalla banca è stata descritta con 1 in caso di erogazione del prestito e con 2 in caso negativo; nell'analisi ho cambiato 2 in 0. Le due classi in cui è diviso il campione sono $D_1 = individui$

a cui è concesso il prestito e $D_2 =$ *individui a cui non è concesso il prestito*

6.1.1 Analisi discriminante

Iniziamo con l'analisi discriminante di Fischer; la funzione predefinita in **R** che implementa questo metodo di classificazione è `lda` ; l'analisi viene, quindi , eseguita tramite la seguente linea di comando:

```
germania.l<-lda(g  .,prior=c(1,1)/2,data=ger[200:1000,])
```

dove `germania.l` è il nome assegnato all'analisi, "`g` "è la variabile decisionale che contiene l'informazione sull'erogazione o meno del prestito da parte della banca , le altre variabili sono state chiamate `V1,V2,...,V24` ; `g`$\sim$ `.` è una forma abbreviata che si usa per indicare che si vuole analizzare la variabile `g` utilizzando tutte le altre variabili presenti nel nostro dataset; `prior=c(1,1)/2` indica le probabilità a priori assegnate rispettivamente alle due classi D_1 e D_2 , in questo caso è stata assegnata ad entrambe le classi una probabilità a priori di 0.5; `data=ger[200:1000]` indica che si stanno utilizzando gli ultimi 800 esempi contenuti nel campione , il dataset german.num è stato importato in **R** con il nome di `ger` .

L'output principale della funzione `lda` sono i pesi α_i relativi a ciascun fattore poi riutilizzati dal nostro pacchetto statistico per costruire il decision boundary. I pesi stimati dal calcolatore sono i seguenti:

```
V1 0.456770438
V2 -0.024774493
V3 0.319326884
V4 -0.004599671
```

```
V5 0.182750015
V6 0.085640822
V7 0.083560886
V8 -0.011729762
V9 -0.150384976
V10 0.008540286
V11 0.312551896
V12 -0.118111822
V13 0.063182508
V14 0.280403017
V15 0.882824363
V16 -0.669380676
V17 0.576370317
V18 -0.940749810
V19 -1.001588786
V20 -0.449175914
V21 0.050903322
V22 0.200362277
V23 -0.036214161
V24 -0.148550685
```

La classificazione dei restanti 199 individui tramite il modello `germania.l` è fatta tramite la funzione `predict` nel seguente modo:

```
predict(germania.l,ger[1:199,])
```

Questa funzione classifica i primi 199 individui del campione `ger` tramite il modello `germania.l`; il decision boundary è costruito tramite l'uso della

sigmoide logistica $g(u)$ considerando come appartenenti a D_i gli individui con una probabilità di assegnazione alla classe in questione maggiore di 0.5. Per creare una tabella di verifica si usa il comando `table` nel seguente modo

```
table.l<-table(predict(germania.l,ger[1:199,])$class,ger[1:199,1])
```

dove `table.l` è il nome della variabile in cui viene memorizzata la tabella; `table` crea una tabella sulle cui righe c'è la classificazione predetta dalla funzione `predict` mentre sulle colonne la classificazione vera contenuta nella prima colonna del dataset `ger`; l'output è il seguente: dalla tabella si evince

	0	1
0	41	40
1	15	103

Tabella 6.2: tablella di verifica lda

che 15 su 55 individui a cui non è stato concesso il mutuo sono stati classificati, in maniera errata, come persone in grado di ripagare la banca; mentre 40 persone su 143 in grado di ripagare la banca sono state erroneamente classificate come individui potenzialmente insolventi.
Inoltre è possibile tramite la linea di comando

```
prop.table(table.l)
```

riscrivire la tabella (6.2) in termini probabilistici dividendo ogni valore al suo interno per il numero totale del campione in questione; applicando quanto detto si ottiene la tabella (6.3) dalla quale si evince che circa il 72% è classificato correttamente inoltre nonostante più del 20% del validation set

	0	1
0	0.20603015	0.20100503
1	0.07537688	0.51758794

Tabella 6.3: tablella percentuale di verifica lda

sia classificato in maniera errata, l'errore più grave (cioè classificare come individuo a cui erogare il prestito un soggetto al quale andrebbe negato) si verifica solo nello 0.075% dei casi.

6.1.2 Regressione logistica

Trattiamo ora lo stesso set di dati utilizzando la regressione logistica; questo modello si implementa tramite la funzione predefinita `glm` nel seguente modo:

```
germania.g<-glm(g ~ .,family=binomial,data=ger[200:1000,])
```

dove `germania.g` è la variable in cui viene memorizzato il modello `glm` è la funzione in grado di implementare il metodo in questione; `g ~.` e `data=ger[200:1000,]` hanno la stessa funzione assunta in precedenza; mentre `family = binomial` specifica che fra i possibili modelli implementati dalla funzione `glm` si vuole usare quello relativo alla regressione logistica. Anche in questo caso l'autput principale è costituito dai pesi α_i, relativi ai vari fattori :

```
Coefficients:
(Intercept) -3.33525
V1 0.55568
V2 -0.02718
```

V3 0.38167

V4 -0.00620

V5 0.27586

V6 0.10262

V7 0.08877

V8 -0.02397

V9 -0.20028

V10 0.01435

V11 0.39808

V12 -0.11202

V13 0.09744

V14 0.35429

V15 1.39325

V16 -0.80088

V17 0.96788

V18 -1.05377

V19 -1.01444

V20 -0.47709

V21 0.10256

V22 0.16430

V23 -0.07481

V24 -0.17614

anche in questo caso per costruire il decision boundary si ricorre all'uso della funzione `predict` nel modo seguente

```
predict(germania.g,ger[1:199,],type=response)->pr
```

il vettore pr contiene ora le probabilità stimate per ciascun individuo; utilizzando l'opzione `type=response` la funzione calcola direttamente la trasformata inversa del logit . Per creare la tabella di verifica si usa la seguente procedura:
si crea un dataframe (tabella) nel quale si sceglie un determinato valore di `pr` come soglia (ad esempio 0.5) per poi classificare ogni individuo in base alla propria probabilità stimata di ripianare il debito; la scelta del valore soglia dipende da scelte interne alla banca, ad esempio un valore soglia di 0.5 significa che una determinata banca decide di concedere il prestito ad un privato cittadino se la probabilità stimata , in base alle informazioni ad esso relative, di ripianare completamente il debito è superiore a 0.5. Quanto detto finora si attua in **R** tramite la seguente linea di comando:

```
germania.g2<-data.frame(ger[1:199,],pr=pr,ok=(pr>0.5))
```

dove `germania.g2` rappresenta la variabile in cui vengono memorizzati i dati della tabella creata dalla funzione `data.frame`, la variabile ok è ritenuta vera quando $\texttt{pr} > \texttt{0.5}$ e quindi gli individui etichettati con true sono assegnati dal modello alla classe D_1 . La tabella di verifica (6.4) è creata con la seguente linea di comando:

```
table.g<-table(germania.g2$ok,ger[1:199,1])
```

l'output è riportato nella tabella (6.4) questo risultato non è molto buono infatti vengono classificati, come 32 individui su 56 come soggetti in grado di ripagere il debito quando in realtà non lo sono! Per migliorare questa tabella si cambia il valore soglia; ad esempio con un valore soglia di 0.7 si ottiene la tabella (6.5)

	0	1
FALSE	24	17
TRUE	32	126

Tabella 6.4: tablella di verifica della regressione logistica; valore soglia= 0.5

	0	1
FALSE	41	43
TRUE	15	100

Tabella 6.5: tablella di verifica della regressione logistica; valore soglia = 0.7

Si nota un netto miglioramento della classificazione infatti pur aumentando il numero di individui a cui si potrebbe concedere il prestito ma classificati incorrettamente diminuisce il numero dei soggetti a cui viene scorrettamente erogato il prestito ; chiaramente è meglio non erogare il prestito ad un individuo a cui lo si potrebbe concedere piuttosto che prestare denaro ad un soggetto che poi non lo restituirebbe!
Come nel capitolo precedente usando il comando `prop.table` applicato a `table.g` è possibile ottenere una tabella di verifica di tipo probabilistico. La tabella in questione è la (6.6). Come si poteva notare anche dalla tabella

	0	1
FALSE	0.20603015	0.21608040
TRUE	0.07537688	0.50251256

Tabella 6.6: tablella percentuale di verifica della regressione logistica; valore soglia = 0.7

(6.5) i risultati della regressione logistica prendendo come valore soglia 0.7 non si discostano di molto da quelli dell'analisi discriminante lineare; anche qui infatti, facendo riferimento alla tabella (6.6) circa il 70% del validation set è classificato correttamente e della percentuale di individui classificati in maniera errata solo con circa lo 7.5% si commette l'errore di classificazione più grave e cioè classificare come individuo a cui erogare il prestito un soggetto al quale andrebbe negato.

6.1.3 Alberi di decisione

Il terzo metodo implementato tramite il pacchetto statistico **R** è quello degli alberi decisionali; questo metodo è implementato con la funzione `rpart` tramite la linea di comando

```
germania.r<-rpart(formula = g   ., data = ger[200:1000, ], method = class)
```

dove il parametro `method= class` permette di costruire un albero di classificazione.La funzione `rpart` agisce in maniera leggermente diversa rispetto al modo di procedere mostrato dall'esempio trattato nel capitolo degli alberi decisionali; infatti se un possibile fattore può assumere tre valori come l'attributo *Conto Bancario*, relativo all'esempio del prestito bancario nel capitolo 5, la funzione `rpart` divide il nodo in due anziche in tre ed effettuando l'ulteriore divisione solo in caso di necessità. La visualizzazione dell'albero è resa possibile possibile dalle seguenti due righe di comando:

```
plot(germania.r)
text(germania.r)
```

Il risultato è riportato in figura (6.1)

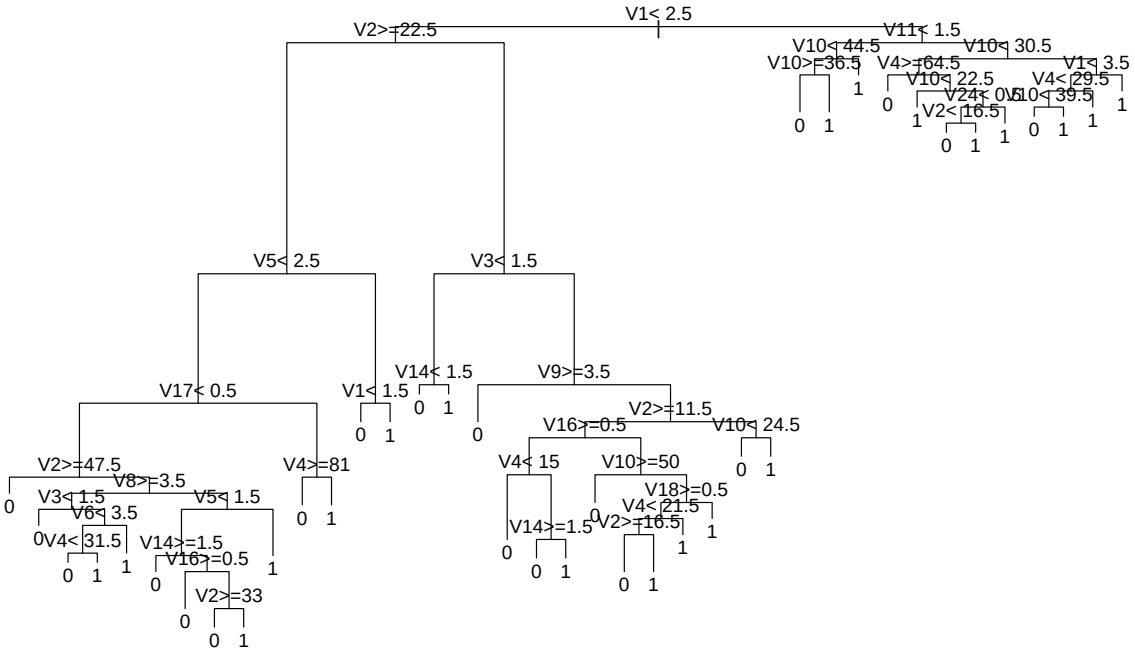

Figura 6.1: albero decisionale

dove `Vi=` *i-esimo attributo*;

Per costruire la tabella di verifica si procede come nel caso della regressione logistica con le seguenti linee di comando:

```
pr<-predict(germania.r,ger[1:199,])
germania.r2<-data.frame(ger[1:199,],pr=pr,ok.1=(pr>0.5))
table.r<-table(germania.r2$ok.1,germania.r2$g)
```

nella seconda e nella terza riga appena riportate appare `ok.1` dato che nel caso degli alberi di decisione la funzione predict da come output,in maniera separata per ogni classe in questione, la probabilità stimata di appartenenza ad una di queste. Quindi bisogna specificare di quale probabilità deve essere maggiore `pr`: se di `ok.1` o di `ok.0`; in questo caso è stato scelto `ok.1` dato che la classe etichettata con 1 è quella degli individui a cui è stato concesso il prestito (cioè D_1).

La verifica è riportata nella tabella (6.7): come nei due casi precedenti anche

	0	1
FALSE	32	26
TRUE	24	117

Tabella 6.7: tablella di verifica albero decisionale

qui volendo fare la una tabella percentuale di verifica si usa la seguente linea di comando

```
prop.table(table.r)
```

e si ottiene la tabella (6.8) dalla quale si evince che nonostante il 74% del

	0	1
FALSE	0.1608040	0.1306533
TRUE	0.1206030	0.5879397

Tabella 6.8: tablella percentuale di verifica albero decisionale

campione sia classificato correttamente (più di quanto accadeva con l'analisi

discriminante lineare e con la regressione logistica) questo modello può essere considerato peggiore, in questo caso, rispetto ai due precedenti perchè ben con 12% del validation set è classificato in maniera errata commettendo, per quanto riguarda il nostro problema, un "grave "errore di classificazione: classificare come individuo a cui erogare il prestito un soggetto al quale andrebbe negato.

6.1.4 Perceptrone

Il metodo del perceptrone è stato implementato con l'aiuto del programma **Matlab**; questo programma possiede un tool grafico attivato con il comando

```
nntool
```

grazie al quale è possibile implementare il metodo di classificazione del perceptrone tramite la finestra grafica in figura (6.2). Come prima cosa bisogna suddividere, come è stato fatto precedentemente, il data set da analizzare (nel nostro caso è il dataset germania.num) in due sottoinsiemi: un training set con il quale settare il modello e un validtion set con il quale testarne la validità; come nell'implementazione degli altri modelli di classificazione il validation set è costituito dai primi 199 individui del dataset mentre il training set è costituito dai rimanenti 801.Oltre ai due insiemi appena descritti è necessario immettere anche i *targets*, ovvero due vettori con l'esatta classificazione degli esempi contenuti nel *training set* e nel *validation set* Una volta immessi questi dati bisogna specificare, come mostrato in figura (6.4), che tipo di rete neurale si vuole implementare e su che set di dati settare i pesi della reti, cliccando sul tasto `crate new network` e poi scegliendo *perceptron* e nella sezione `input` selezionare il training set precedentemente importato;

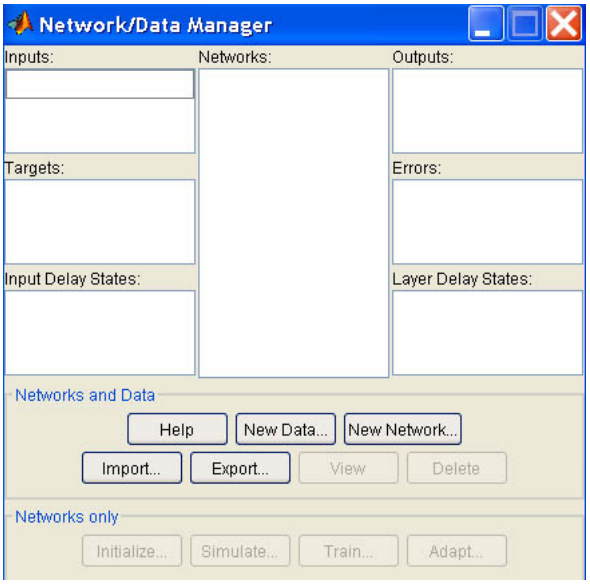

Figura 6.2: finestra di comando per implementare il perceptrone

Una volta creato il perceptrone bisogna decidere il numero delle volte che si vuole far analizzare l'intero training set al perceptrone per settare i pesi del modello cliccando su *initialize-Train-Training Parameters*. Ad esempio se si vogliono stimare i pesi in base ad una sola analisi del training set porremo il parametro `epochs` uguale a 1, se invece vogliamo una stima dei pesi basata su più analisi del training set in questione porremo `epochs` uguale a 20,30 e cosi via. Nel nostro caso non sappiamo se i dati che possediamo sono classificati in due classi linearmente separabili ed inoltre non conosciamo il vettore dei pesi $\underline{\widehat{\alpha}}$, nel caso in cui i dati siano divisi in due classi linearmente separabili, con il quale costruire l'iperpiano separatore fra le due classi; quindi a priori non ci possiamo aspettare che il perceptrone classifichi correttamente ogni individuo descritto dal generico vettore di informazioni p dimensionale $\underline{x}_i$. Nell'implementare il metodo del perceptrone è stato scelto di far analizzare il

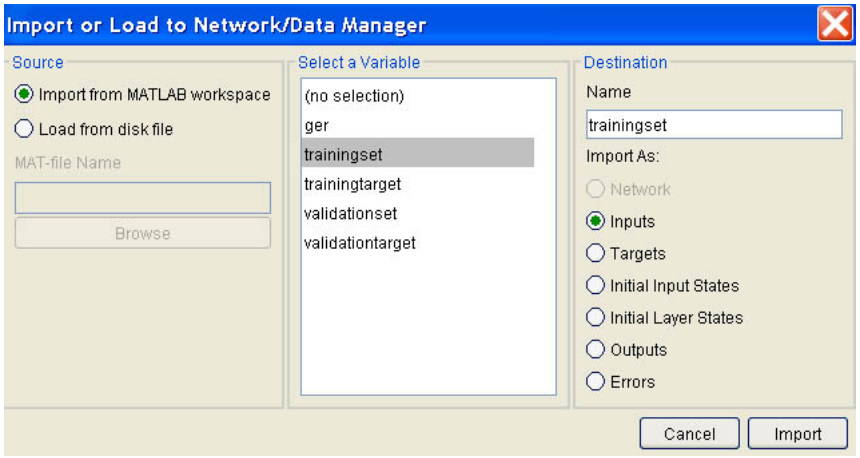

Figura 6.3: finestra di comando per immettere i dati

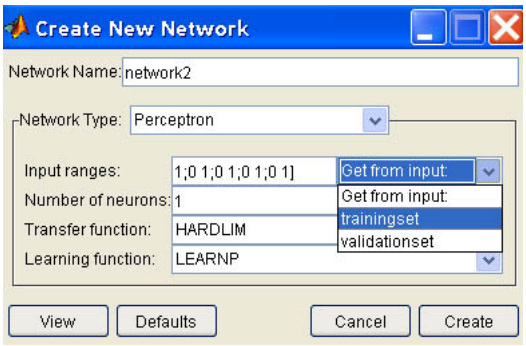

Figura 6.4: finestra di comando per implementare il perceptrone

campione in questione per 200 volte; quindi `epochs`=200. Una volta settati i pesi, cliccando su *train* dopo aver immesso i dati di imput e di target relativi al training set, è possibile, per testare la validità del modello, far classificare in base al modello appena creato gli elementi del validation set, cliccando su *simulate* ed immettendo i dati di input e di target relativi al validation set. L'output della classificazione dei vettori appartenenti al validation set è scritto su di un file il cui nome è possibile specificare nella finestra *simulate*

al passo precedente (è possibile eseguire le operazioni di settaggio e verifica del modello tramite la finestra grafica mostrata in figura (6.5)); l'output,

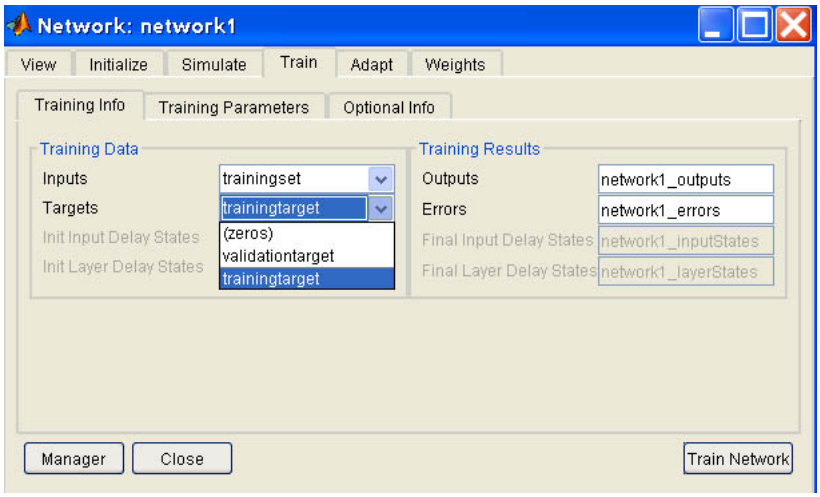

Figura 6.5: finestra di comando per settaggio e verifica del modello

stampato nel file *network1_errors* come mostrato in figura (6.6), si interpreta nel seguente modo: le due classi D_1 e D_2 sono etichettate rispettivamente con 1 e 0; il perceptrone prende un vettore di osservazioni $\underline{x}_i$ appartenente al validation set, lo classifica in base ai pesi stimati precedentemente e poi verifica il risultato ottenuto facendo la sottrazione fra classificazione da lui ottenuta e quella vera; quindi nel nostro caso se un vettore è correttamente classificato avremo come risposta 0 infatti la sottrazione sarà effettuata fra due quantità identiche, in questo caso si avrebbe 1-1=0 oppure 0-0=0; se, invece, un vettore è classificato in maniera errata si ha 1-0=1 se si classifica un individuo appartenente alla classe D_1, come appartenente a quella D_0, al contrario si ottiene 0-1=-1 se si classifica un individuo appartenente alla classe D_0, come appartenente a quella D_1,. Il tipo di errore più **grave** è, come abbiamo ripetuto nell'implementazione degli altri modelli, classificare

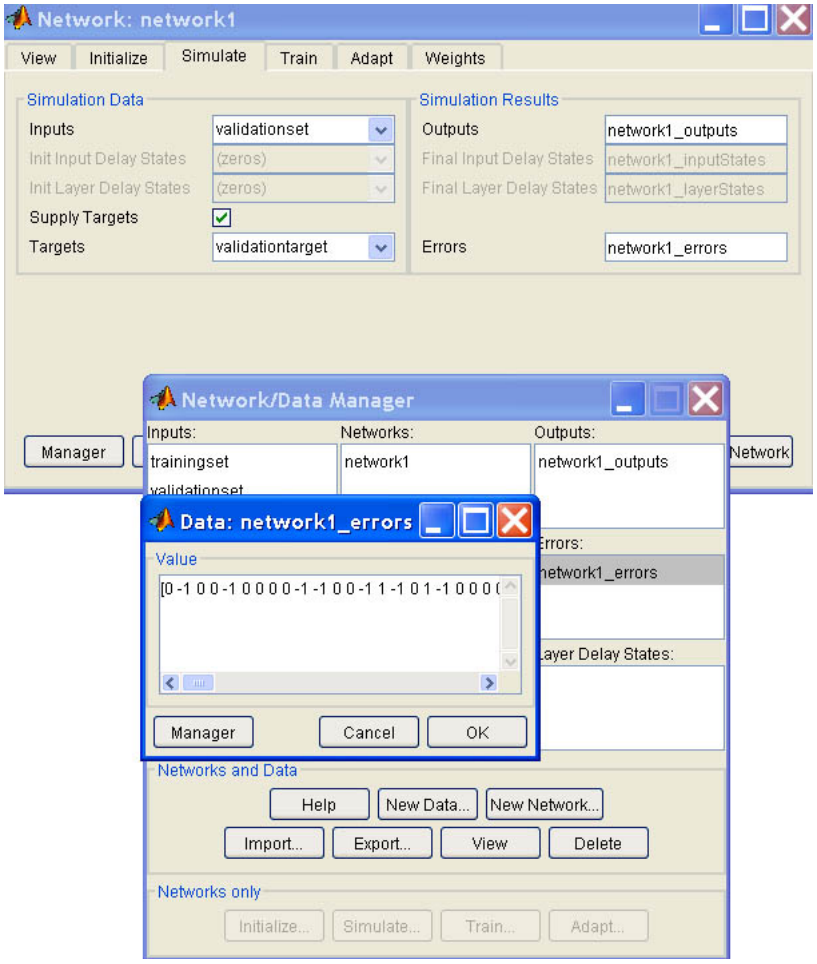

Figura 6.6: finestra di comando per settaggio e verifica del modello

come un individuo a cui erogare il prestito un soggetto al quale andrebbe negato; quindi nel nostro caso otteremo come risposta 1. La verifica dell'esattezza del modello del perceptrone ha portato al seguente risultato: 147 individui sono stati classificati correttamente; dei restanti 52 individui non correttamente classificati solo con 7 è stato commesso il tipo di errore ritenuto nel nostro caso **grave**. In percentuale si ha che circa il 74% degli individui appartenenti al validation set sono classificati correttamente mentre solo cir-

ca il 3.5% degli individui sono stati classificati commettendo l'errore da noi ritenuto **grave**. Per quanto riguarda il dataset da noi analizzato il miglior medoto di classificazione risulta essere il perceptrone che, rispetto al tipo di errore di classificazione da evitare, come si è visto nel corso di questo capitolo risulta avere un grado di precisione doppio rispetto dell'analisi discriminante lineare e della regressione logistica che pur non comportandosi male non raggiungono in questo caso la precisione del perceptrone. Non si comporta bene il metodo degli alberi di decisione infatti pur classificando correttamente il 74% del validation set, commette l'errore di classificazione **grave** con il 12% del campione e risulta essere, riguardo a questo aspetto della classificazione, quasi 4 volte più impreciso del perceptrone.

Bibliografia

[CB95] Christopher M. Bishop (1995) *Neural Networks for Pattern Recognition* Clarendon Press Oxford.

[CAN98] John B. Caouette, Edward I. Altman, Paul Narayanan (1998) *Managing Credit Risk* John wiley & Sons,Inc.

[VR02] W.N. Venables, B.D. Ripley (2002) *Modern Applied Statistics with S* Springer

[DA01] Richard O. Duda, Peter E. Hart (2001) *Pattern classification and scene analysis* Wiley-Interscience New York.

[IM03] Stefano M. Iacus, Guido Masarotto (2003) *Laboratorio di statistica con R* McGraw-Hill.

[PD02] Peter Dalgaard (2002) *Introductory Statistics with R* Springer.

[HS03] Wolfgang Hardle, Léopold Simar (2003) *Applied Multivariate Statistical Analysis* Springer.

[AS02] Anthony Saunders (2002) *Credit Risk Measurement* John wiley & Sons,Inc.

[SH98] Simon S. Haykins (1998) *Neural Networks: A Comprehensive Foundation* Hardcover.

[RN98] Stuart J. Russel, Peter Norvig (1998) *Intelligenza Artificiale: Un approccio moderno* Prentice Hall International.