Metodi di classificazione

Uno dei problemi piú frequenti nel data mining é quello di dover allocare un'unitá a una categoria o una classe tra K possibili alternative utilizzando le osservazioni relative a delle variabili rilevate su quella unitá.
Situazioni tipiche possono essere le seguenti:

- Una banca che deve attribuire i suoi clienti alla categoria dei 'debitori insolventi' o dei 'debitori solventi'. Per tale valutazione la banca ha a disposizione varie informazioni sui clienti, sia di natura anagrafica che della storia pregressa come clienti. Problema associato alle tematiche del credit scoring o del rating.
- Una societá di assicurazioni vuole valutare se un cliente che stipula una polizza RCA avrá $0, 1, 2$ o piú incidenti nel prossimo anno. Qui le informazioni disponibili sono quelle anagrafiche del cliente, quelle sulle caratteristiche dell'auto, i dati sulla storia assicurativa. In ambito aziendale tale problema é associato alla tematica dell'identificazione del prezzo (pricing).

Uno dei problemi piú frequenti nel data mining é quello di dover allocare un'unitá a una categoria o una classe tra K possibili alternative utilizzando le osservazioni relative a delle variabili rilevate su quella unitá.
Situazioni tipiche possono essere le seguenti:

- Una banca che deve attribuire i suoi clienti alla categoria dei 'debitori insolventi' o dei 'debitori solventi'. Per tale valutazione la banca ha a disposizione varie informazioni sui clienti, sia di natura anagrafica che della storia pregressa come clienti. Problema associato alle tematiche del credit scoring o del rating.
- Una societá di assicurazioni vuole valutare se un cliente che stipula una polizza RCA avrá $0, 1, 2$ o piú incidenti nel prossimo anno. Qui le informazioni disponibili sono quelle anagrafiche del cliente, quelle sulle caratteristiche dell'auto, i dati sulla storia assicurativa. In ambito aziendale tale problema é associato alla tematica dell'identificazione del prezzo (pricing).

- Una compagnia aerea vuole prevedere quali dei suoi clienti, possessori di carta fedeltá, effettuerá un volo intercontinentale verso una destinazione di villeggiatura entro i prossimi 12 mesi. Ai clienti con un'alta propensione in tal senso proporrá un'azione promozionale, cercando invece di evitare l'invio a persone non interessate. In questo caso le informazioni disponibili saranno tutte quelle presenti nel data-base aziendale. In ambito aziendale si parla di tematiche up sell e cross sell.
- Un'azienda di automobili vuole individuare dei potenziali clienti tra coloro che nei prossimi sei mesi acquisteranno una nuova auto del segmento 'alta gamma', in modo da spedire loro una brochure di presentazione del suo nuovo modello. Ha bisogno quindi di rivolgersi a una ditta specializzata per la fornitura di liste di potenziali clienti. Queste sono formate partendo da amplissime raccolte di dati ottenute da fonti diverse, ma che concorrono alla formazione di profili di comportamento economico individuale. In ambito aziendale questo tipo di problema é associato alla gestione dei prospect.

- Una compagnia aerea vuole prevedere quali dei suoi clienti, possessori di carta fedeltá, effettuerá un volo intercontinentale verso una destinazione di villeggiatura entro i prossimi 12 mesi. Ai clienti con un'alta propensione in tal senso proporrá un'azione promozionale, cercando invece di evitare l'invio a persone non interessate. In questo caso le informazioni disponibili saranno tutte quelle presenti nel data-base aziendale. In ambito aziendale si parla di tematiche up sell e cross sell.
- Un'azienda di automobili vuole individuare dei potenziali clienti tra coloro che nei prossimi sei mesi acquisteranno una nuova auto del segmento 'alta gamma', in modo da spedire loro una brochure di presentazione del suo nuovo modello. Ha bisogno quindi di rivolgersi a una ditta specializzata per la fornitura di liste di potenziali clienti. Queste sono formate partendo da amplissime raccolte di dati ottenute da fonti diverse, ma che concorrono alla formazione di profili di comportamento economico individuale. In ambito aziendale questo tipo di problema é associato alla gestione dei prospect.

L'obiettivo quindi é quello di costruire una regola per allocare un individuo in una delle classi, sulla base delle variabili a disposizione.

I diversi modelli si fondano sull'ipotesi che si disponga di un certo numero n di casi per i quali é nota la classe di appartenenza.

Si indicano con $y_1, \ldots, y_n$ le classi di appartenenza del campione e con n_k il numero di unitá appartenenti alla k-esima classe.

L'obiettivo quindi é quello di costruire una regola per allocare un individuo in una delle classi, sulla base delle variabili a disposizione.

I diversi modelli si fondano sull'ipotesi che si disponga di un certo numero n di casi per i quali é nota la classe di appartenenza.

Si indicano con $y_1, \dots, y_n$ le classi di appartenenza del campione e con n_k il numero di unitá appartenenti alla k-esima classe.

L'obiettivo quindi é quello di costruire una regola per allocare un individuo in una delle classi, sulla base delle variabili a disposizione.

I diversi modelli si fondano sull'ipotesi che si disponga di un certo numero n di casi per i quali é nota la classe di appartenenza.

Si indicano con $y_1, \ldots, y_n$ le classi di appartenenza del campione e con n_k il numero di unitá appartenenti alla k-esima classe.

Previsioni attraverso la regressione logistica

Dati relativi alle preferenze di 1070 consumatori tra due marchi di succhi di frutta (CH e MM):

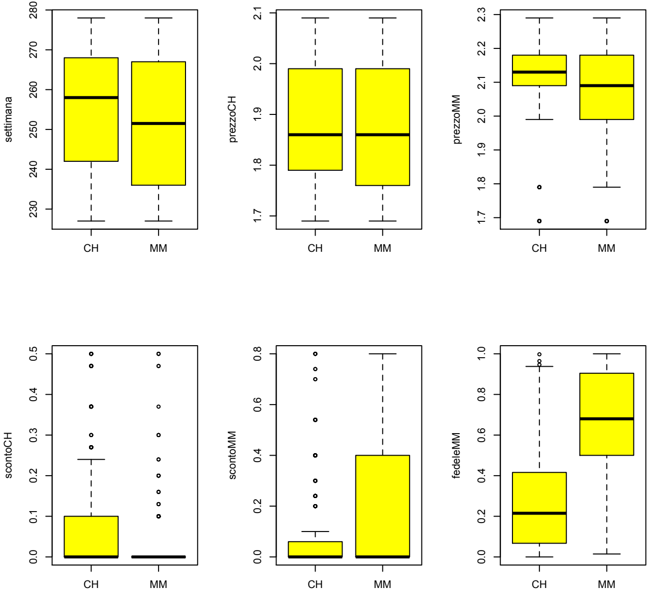
Variabile	Descrizione
scelta	marca prescelta
id. cliente	identificativo cliente
settimana	identificativo settimana di acquisto
prezzoCH	prezzo di riferimento per la marca CH
prezzoMM	prezzo di riferimento per la marca MM
scontoCH	sconto applicato per il prodotto CH
scontoMM	sconto applicato per il prodotto MM
fedeleCH	indicatore fedeltá per il prodotto CH
fedeleMM	indicatore fedeltá per il prodotto MM
negozio	identificativo del negozio

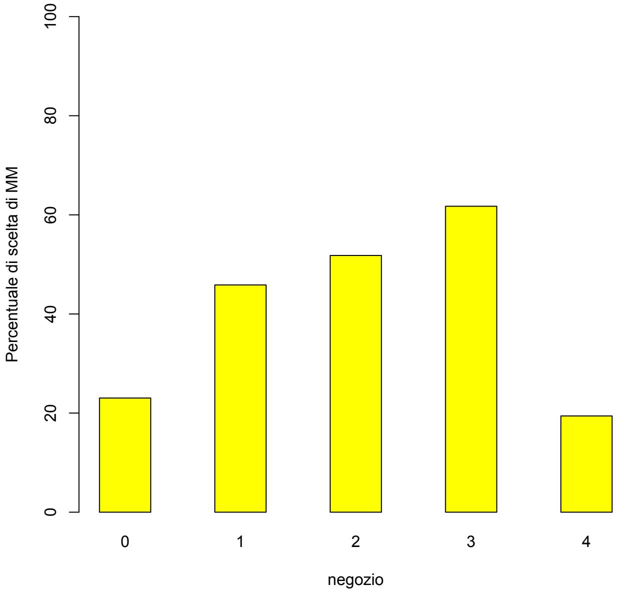
La variabile fedeleMM viene costruita partendo da un calore di 0.5 e aggiornandola ad ogni acquisto del medesimo cliente, con valore che aumenta del 20% della differenza fra il valore corrente e 1, se la scelta é stata MM, altrimenti diminuisce allo stesso modo. La variabile fedeleCH é calcolata come 1-fedeleMM.

Anche in questo caso, dividiamo il dataset nei due insiemi e vediamo il comportamento delle variabili sull'insieme di stima.

La variabile fedeleMM viene costruita partendo da un calore di 0.5 e aggiornandola ad ogni acquisto del medesimo cliente, con valore che aumenta del 20% della differenza fra il valore corrente e 1, se la scelta é stata MM, altrimenti diminuisce allo stesso modo. La variabile fedeleCH é calcolata come 1-fedeleMM.

Anche in questo caso, dividiamo il dataset nei due insiemi e vediamo il comportamento delle variabili sull'insieme di stima.


settimana
230
240
250
260
270
280
CH
MM
prezzoCH
1.7
1.8
1.9
2.0
2.1
CH
MM
prezzoMM
1.7
1.8
1.9
2.0
2.1
2.2
2.3
CH
MM
scontoCH
0.0
0.1
0.2
0.3
0.4
0.5
CH
MM
scontoMM
0.0
0.2
0.4
0.6
0.8
CH
MM
fedeleMM
0.0
0.2
0.4
0.6
0.8
1.0
CH
MM


Percentuale di scelta di MM
0
20
40
60
80
100
0
1
2
3
4
negozio

Il modello

$$logit(\pi) = \beta_0 + \beta_1 settimana + \beta_2 prezzoCH + \beta_3 prezzoMM + \beta_4 scontoCH +$$
$$\beta_5 scontoMM + \beta_6 fedeleMM + \beta_7^\top I_{negozio}$$

	Stima	Std. Error	valore-t	p-value
(Intercetta)	-3.81611	2.06284	-1.85	0.06432
settimana	-0.00173	0.01309	-0.13	0.89486
prezzoCH	4.43557	2.11851	2.09	0.03628
prezzoMM	-3.70577	1.00809	-3.68	0.00024
scontoCH	-3.64867	1.14489	-3.19	0.00144
scontoMM	2.09456	0.50162	4.18	3e-05
fedeleMM	5.86455	0.44978	13.04	$<$2e-16
negozio1	0.55120	0.31550	1.75	0.08062
negozio2	0.65630	0.28592	2.30	0.02171
negozio3	0.57354	0.36819	1.56	0.11930
negozio4	0.03861	0.42014	0.09	0.92677

D=631.63 con 791 g.d.l.

	Stima	Std. Error	valore-t	p-value
(Intercetta)	-3.8127	2.0630	-1.85	0.06458
prezzoCH	4.2411	1.5233	2.78	0.00537
prezzoMM	-3.7444	0.9652	-3.88	0.00010
scontoCH	-3.6952	1.0892	-3.39	0.00069
scontoMM	2.0818	0.4918	4.23	2.3e-05
fedeleMM	5.8683	0.4489	13.07	$<$ 2e-16
negozio1	0.5433	0.3095	1.76	0.07923
negozio2	0.6513	0.2834	2.30	0.02155
negozio3	0.5927	0.3386	1.75	0.08001
negozio4	0.0548	0.4021	0.14	0.89168

D=631.64 con 792 g.d.l.

Appropriatezza della riduzione confermata dal test del rapporto di verosimiglianza.

Misure adeguatezza classificazione

Previsione	Risposta effettiva		
	CH	MM	Totale
CH	150	23	173
MM	19	76	95
Totale	169	99	268

Il tasso di corretta classificazione é pari a $(150 + 76)/268 = 0.843$.

Ci interessa solo questo?

Falsi positivi e falsi negativi.

Misure adeguatezza classificazione

Previsione	Risposta effettiva		
	CH	MM	Totale
CH	150	23	173
MM	19	76	95
Totale	169	99	268

Il tasso di corretta classificazione é pari a $(150 + 76)/268 = 0.843$.

Ci interessa solo questo?

Falsi positivi e falsi negativi.

Misure adeguatezza classificazione

Previsione	Risposta effettiva		
	CH	MM	Totale
CH	150	23	173
MM	19	76	95
Totale	169	99	268

Il tasso di corretta classificazione é pari a $(150 + 76)/268 = 0.843$.

Ci interessa solo questo?

Falsi positivi e falsi negativi.

Previsione	Risposta effettiva		
	$-$	$+$	
$-$	n_{11}	n_{12}	$n_{1.}$
$+$	n_{21}	n_{22}	$n_{2.}$
Totale	$n_{.1}$	$n_{.2}$	n

Previsione	Risposta effettiva	
	$-$	$+$
$-$	$1-\alpha$	β
$+$	α	$1-\beta$
Totale	1	1

Le due probabilitá α e β non sono note ma possono essere solo stimate attraverso $\hat{\alpha} = n_{21}/n_{.1}$ e $\hat{\beta} = n_{12}/n_{.2}$

Il costo di un'errata classificazione non é uguale nei due casi e a seconda dell'obiettivo si puó dare un peso maggiore a uno o all'altro tipo di errore.

Se per esempio si é interessati a individuare le caratteristiche di chi sceglie MM, si cercherá di minimizzare l'errore nell'identificazione di tali soggetti.

Le due probabilitá α e β non sono note ma possono essere solo stimate attraverso $\hat{\alpha} = n_{21}/n_{.1}$ e $\hat{\beta} = n_{12}/n_{.2}$
Il costo di un'errata classificazione non é uguale nei due casi e a seconda dell'obiettivo si puó dare un peso maggiore a uno o all'altro tipo di errore.
Se per esempio si é interessati a individuare le caratteristiche di chi sceglie MM, si cercherá di minimizzare l'errore nell'identificazione di tali soggetti.

Una delle possibilitá é quella di cambiare il valore soglia per la classificazione in un gruppo piuttosto che nell'altro. scegliere il valore 0.5 é una semplificazione, ma si puó graduare anche diversamente.

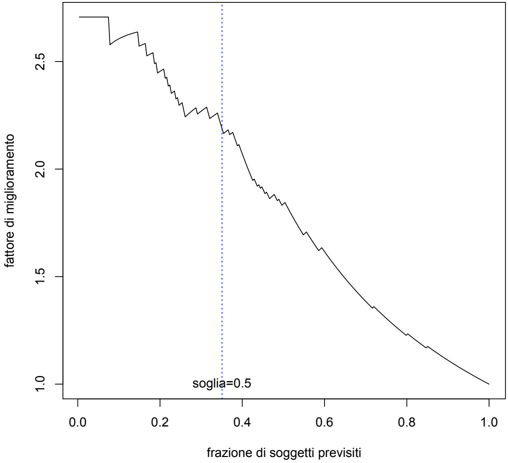
Spostando il valore soglia si possono assegnare pesi diversi alle due categorie.

Una delle possibilitá é quella di cambiare il valore soglia per la classificazione in un gruppo piuttosto che nell'altro. scegliere il valore 0.5 é una semplificazione, ma si puó graduare anche diversamente.

Spostando il valore soglia si possono assegnare pesi diversi alle due categorie.

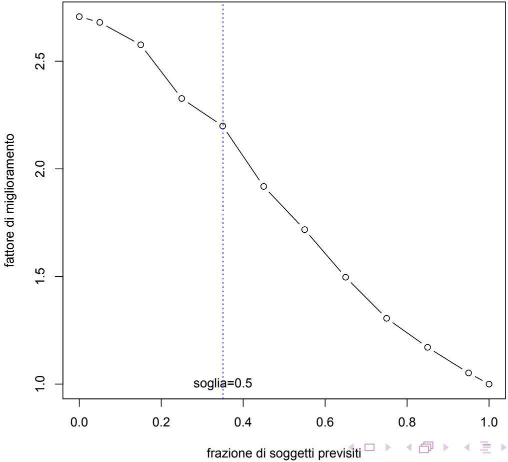
La funzione lift

Tra i vari strumenti in uso, uno particolarmente diffuso é la cosiddetta funzione lift.



La funzione lift

Questa fornisce una misura di miglioramento ottenuto dal modello in considerazione rispetto alla classificazione casuale con probabilitá uniforme pari alla frazione osservata nell'insieme di verifica.



La parte a sinistra di entrambe le curve corrisponde a porzioni di soggetti per i quali la probabilitá stimata era piú alta. L'ordinata rappresenta la proporzione di acquirenti di MM osservata per quei soggetti divisa per la proporzione media calcolata su tutti i dati.
Nel primo grafico sono considerati tutti i casi, mentre nel secondo c'é stato un lisciamento della curva.

Per estrarre l'ordinata della curva si calcola la seguente quantitá:

$$\frac{n_{22}/n_{2.}}{n_{.2}/n}$$

Tale valore indicherá il miglioramento atteso a seguito dell'utilizzo della regressione logistica rispetto alla scelta casuale.

Un'ulteriore osservazione riguarda la asimmetria del comportamento del lift rispetto alla scelta di evento 'favorevole' o 'sfavorevole'. Scambiando la scelta dell'evento di interesse si otterranno grafici diversi.

La parte a sinistra di entrambe le curve corrisponde a porzioni di soggetti per i quali la probabilitá stimata era piú alta. L'ordinata rappresenta la proporzione di acquirenti di MM osservata per quei soggetti divisa per la proporzione media calcolata su tutti i dati.
Nel primo grafico sono considerati tutti i casi, mentre nel secondo c'é stato un lisciamento della curva.

Per estrarre l'ordinata della curva si calcola la seguente quantitá:

$$\frac{n_{22}/n_{2.}}{n_{.2}/n}$$

Tale valore indicherá il miglioramento atteso a seguito dell'utilizzo della regressione logistica rispetto alla scelta casuale.

Un'ulteriore osservazione riguarda la asimmetria del comportamento del lift rispetto alla scelta di evento 'favorevole' o 'sfavorevole'. Scambiando la scelta dell'evento di interesse si otterranno grafici diversi.

La parte a sinistra di entrambe le curve corrisponde a porzioni di soggetti per i quali la probabilitá stimata era piú alta. L'ordinata rappresenta la proporzione di acquirenti di MM osservata per quei soggetti divisa per la proporzione media calcolata su tutti i dati.
Nel primo grafico sono considerati tutti i casi, mentre nel secondo c'é stato un lisciamento della curva.

Per estrarre l'ordinata della curva si calcola la seguente quantitá:

$$\frac{n_{22}/n_{2.}}{n_{.2}/n}$$

Tale valore indicherá il miglioramento atteso a seguito dell'utilizzo della regressione logistica rispetto alla scelta casuale.

Un'ulteriore osservazione riguarda la asimmetria del comportamento del lift rispetto alla scelta di evento 'favorevole' o 'sfavorevole'. Scambiando la scelta dell'evento di interesse si otterranno grafici diversi.

La curva ROC

La curva ROC (Receiving Operating Characteristic) é un altro strumento per valutare l'adeguatezza di un criterio di classificazione. Particolarmente usata nella statistica medica e nel controllo della qualitá.

Considerando la tabella ottenuta con la scelta della soglia pari a 0.5, si riesca a quantificare la proporzione di falsi positivi e di falsi negativi, pari rispettivamente a

$$19/169 \qquad e \qquad 23/99$$

La curva ROC

La curva ROC (Receiving Operating Characteristic) é un altro strumento per valutare l'adeguatezza di un criterio di classificazione. Particolarmente usata nella statistica medica e nel controllo della qualitá.

Considerando la tabella ottenuta con la scelta della soglia pari a 0.5, si riesca a quantificare la proporzione di falsi positivi e di falsi negativi, pari rispettivamente a

$$19/169 \qquad e \qquad 23/99$$

Facendo variare il valore della soglia e calcolando le proporzioni associate di falsi positivi e falsi negativi si avranno le due quantitá in genere indicate come segue:

- sensibilitá: proporzione di previsti positivi rispetto al numero di positivi effettivi;
- specificitá: proporzione di previsti negativi rispetto al numero di negativi effettivi.

Le due quantitá vengono in genere stimate come:

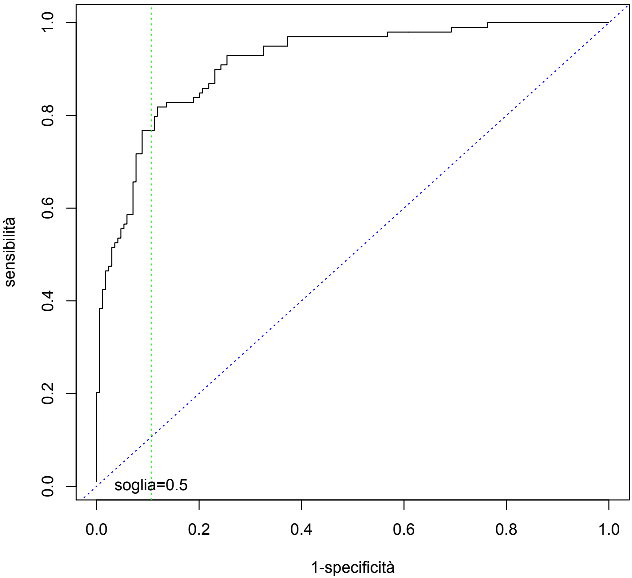
$$sensibilit\acute{a} = \frac{n_{22}}{n_{22} + n_{12}}; \qquad specificit\acute{a} = \frac{n_{11}}{n_{11} + n_{21}}$$

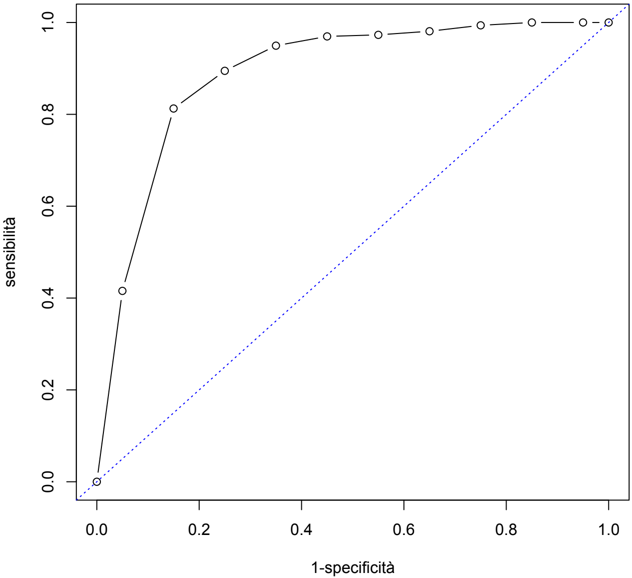
Facendo variare il valore della soglia e calcolando le proporzioni associate di falsi positivi e falsi negativi si avranno le due quantitá in genere indicate come segue:

- sensibilitá: proporzione di previsti positivi rispetto al numero di positivi effettivi;
- specificitá: proporzione di previsti negativi rispetto al numero di negativi effettivi.

Le due quantitá vengono in genere stimate come:

$$sensibilit\acute{a} = \frac{n_{22}}{n_{22} + n_{12}}; \qquad specificit\acute{a} = \frac{n_{11}}{n_{11} + n_{21}}$$


1.0
0.8
0.6
0.4
0.2
0.0
sensibilità
soglia=0.5
0.0
0.2
0.4
0.6
0.8
1.0
1-specificità


sensibilità
1-specificità
0.0
0.2
0.4
0.6
0.8
1.0

Gli indici e i grafici ottenuti possono essere estesi al caso di piú categorie. In tal caso lo strumento da utilizzare sará la regressione logistica multinomiale.

Anche la regressione lineare potrebbe essere usata per gli stessi fini, ma sarebbe una forzatura, anche se spesso porta a risultati simili.

Gli indici e i grafici ottenuti possono essere estesi al caso di piú categorie. In tal caso lo strumento da utilizzare sará la regressione logistica multinomiale.

Anche la regressione lineare potrebbe essere usata per gli stessi fini, ma sarebbe una forzatura, anche se spesso porta a risultati simili.

Analisi discriminante

Si abbia una popolazione complessiva composta da K subpopolazioni (classi), aventi la funzione di densitá di probabilitá rispettiva pari a $p_1(x), \ldots, p_K(x)$, per la distribuzione di X (variabile casuale p-dimensionale assunta per ora continua). Il peso di ogni subpopolazione é pari a $\pi_1, \ldots, \pi_K$, con somma dei pesi pari a 1. La densitá per la popolazione complessiva sará quindi:

$$p(x) = \sum_{k=1}^{K} \pi_k p_k(x)$$

A priori la probabiltiá che un soggetto non ancora classificato appartenga alla k-ma popolazione é data da π_k. Se per quel soggetto é noto il valore assunto dalla variabile X (x_0) allora per il teorema di Bayes la probabilitá a posteriori che quel soggetto appartenga al gruppo k é data da:

$$P\{y = k | X = x_0\} = \frac{\pi_k p_k(x_0)}{p(x_0)}$$

Equivalentemente il confronto di probabilitá fra la classe k e la classe m sará dato da:

$$log \frac{P\{y = k | X = x_0\}}{P\{y = m | X = x_0\}} = log \frac{\pi_k}{\pi_m} + log \frac{p_k(x_0)}{p_m(x_0)}$$

Le varie classi possono quindi essere confrontate secondo la rispettiva funzione discriminante:

$$d_k(x_0) = log \pi_k + log p_k(x_0)$$

che é legata alla probabilitá a posteriori delle classi. Il valore di k che massimizza la funzione discriminante individua il gruppo al quale verrá attribuito il nuovo soggetto.

Per poter sviluppare il procedimento é ovviamente necessario conoscere (o stimare) i dati. Per quanto riguarda π_k, se non si hanno altre informazioni, sono stimati semplicemente come $\hat{\pi}_k = n_k/n$.

Per ció che concerne invece $p_k(x)$ ci sono diverse strade divise nei due gruppi dell'approccio parametrico e dell'approccio non parametrico.

Nel primo caso ci sono diverse opzioni sulla famiglia di funzioni di densitá da considerare, mentre in quello non parametrico diverse alternative fra i metodi di stima.

Per poter sviluppare il procedimento é ovviamente necessario conoscere (o stimare) i dati. Per quanto riguarda π_k, se non si hanno altre informazioni, sono stimati semplicemente come $\hat{\pi}_k = n_k/n$.

Per ció che concerne invece $p_k(x)$ ci sono diverse strade divise nei due gruppi dell'approccio parametrico e dell'approccio non parametrico.

Nel primo caso ci sono diverse opzioni sulla famiglia di funzioni di densitá da considerare, mentre in quello non parametrico diverse alternative fra i metodi di stima.

Analisi discriminante lineare

L'ipotesi parametrica piú semplice é quella in cui ciascuna densitá é normale multipla con parametri dipendenti da k ($N_p(\mu_k, \Sigma_k)$), tale per cui:

$$\frac{1}{(2\pi)^{p/2} det(\Sigma_k)^{1/2}} exp\left[-\frac{1}{2}(x-\mu_k)^T \Sigma_k^{-1}(x-\mu_k)\right]$$

con $k = 1, \ldots, K$

Nel caso in cui tutte la matrici di varianza siano uguali (Σ), la funzione discriminante prende la seguente forma

$$d_k(x) = log \pi_k - \frac{1}{2}\mu_k^T \Sigma^{-1}\mu_k + x^T \Sigma^{-1}\mu_k$$

che é una funzione lineare in x, da cui il nome analisi discriminante lineare.

Analisi discriminante lineare

L'ipotesi parametrica piú semplice é quella in cui ciascuna densitá é normale multipla con parametri dipendenti da k ($N_p(\mu_k, \Sigma_k)$), tale per cui:

$$\frac{1}{(2\pi)^{p/2} det(\Sigma_k)^{1/2}} exp\left[-\frac{1}{2}(x-\mu_k)^T \Sigma_k^{-1}(x-\mu_k)\right]$$

con $k = 1, \ldots, K$

Nel caso in cui tutte la matrici di varianza siano uguali (Σ), la funzione discriminante prende la seguente forma

$$d_k(x) = log\pi_k - \frac{1}{2}\mu_k^T \Sigma^{-1}\mu_k + x^T \Sigma^{-1}\mu_k$$

che é una funzione lineare in x, da cui il nome analisi discriminante lineare.

La stima dei parametri sará data da:

$$\hat{\mu}_k = \frac{1}{n_k} \sum_{i:y_i=k} x_i$$

$$\hat{\Sigma} = \frac{1}{n-K} \sum_{k=1}^{K} \sum_{i:y_i=k} (x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^T$$

x_i é la determinazione della variabile X relativa alla i-esima unitá.
Il numero di parametri stimati sará pari a $pK + p(p+1)/2$
Consideriamo due predittori lineari del seguente tipo:

$$x_i = (z_{i1}, z_{i2})$$

$$x_i = (z_{i1}, z_{i2}, z_{i1}^2, z_{i1}z_{i2}, z_{i2}^2)$$

con quindi $p = 2$ e $p = 5$ componenti.

La stima dei parametri sará data da:

$$\hat{\mu}_k = \frac{1}{n_k} \sum_{i:y_i=k} x_i$$

$$\hat{\Sigma} = \frac{1}{n-K} \sum_{k=1}^{K} \sum_{i:y_i=k} (x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^T$$

x_i é la determinazione della variabile X relativa alla i-esima unitá.
Il numero di parametri stimati sará pari a $pK + p(p+1)/2$
Consideriamo due predittori lineari del seguente tipo:

$$x_i = (z_{i1}, z_{i2})$$

$$x_i = (z_{i1}, z_{i2}, z_{i1}^2, z_{i1}z_{i2}, z_{i2}^2)$$

con quindi $p = 2$ e $p = 5$ componenti.

La stima dei parametri sará data da:

$$\hat{\mu}_k = \frac{1}{n_k} \sum_{i:y_i=k} x_i$$

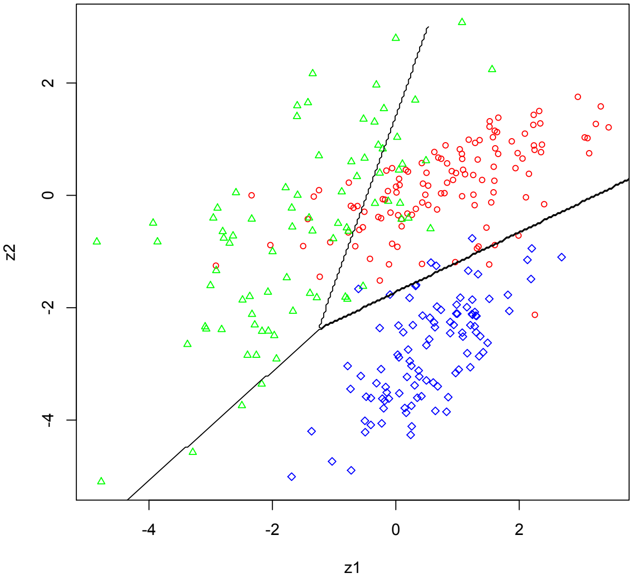
$$\hat{\Sigma} = \frac{1}{n-K} \sum_{k=1}^{K} \sum_{i:y_i=k} (x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^T$$

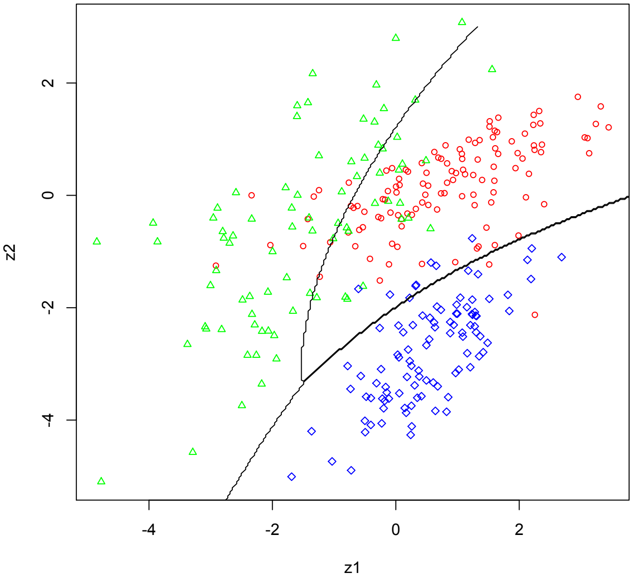
x_i é la determinazione della variabile X relativa alla i-esima unitá.
Il numero di parametri stimati sará pari a $pK + p(p+1)/2$
Consideriamo due predittori lineari del seguente tipo:

$$x_i = (z_{i1}, z_{i2})$$

$$x_i = (z_{i1}, z_{i2}, z_{i1}^2, z_{i1}z_{i2}, z_{i2}^2)$$

con quindi $p = 2$ e $p = 5$ componenti.


z2
z1
2
0
-2
-4
-4
-2
0
2


z2
z1
2
0
-2
-4
-4
-2
0
2

Analisi discriminante quadratica

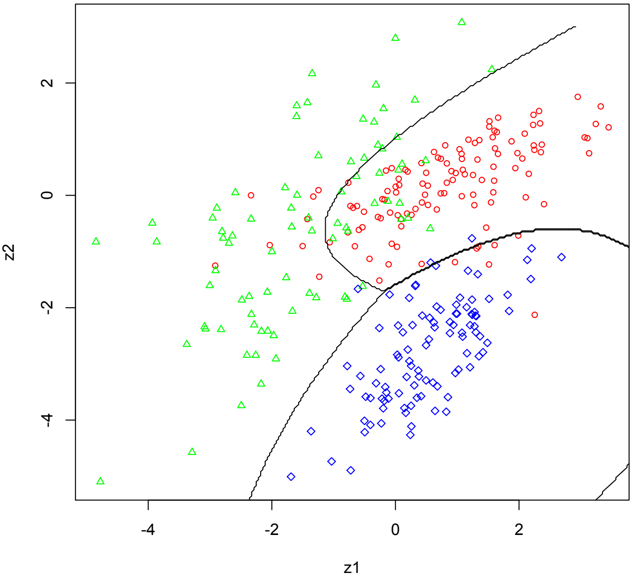
Se si elimina la condizione di omoschedasticitá si ha la seguente funzione discriminante:

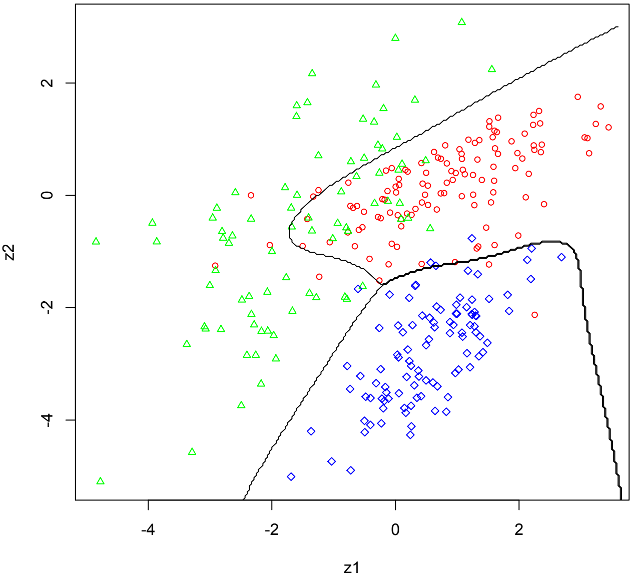
$$\delta_k(x) = log\pi_k - \frac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k) - \frac{1}{2}logdet(\Sigma_k)$$

che é quadratica in x.
La stima dei vettori medi non cambia mentre la varianza sará:

$$\hat{\Sigma}_k = \frac{1}{n_k-1}\sum(x_i-\hat{\mu}_k)(x_i-\hat{\mu}_k)^T$$

I parametri da stimare in questo caso saranno $Kp + Kp(p+1)/2$.


2
0
-2
-4
z2
-4
-2
0
2
z1


2
0
-2
-4
z2
-4
-2
0
2
z1

Come scegliere fra le due?

Previsione con LDA	Risposta effettiva		
	CH	MM	Totale
CH	147	20	167
MM	22	79	101
Totale	169	99	268

Previsione con LDA	Risposta effettiva		
	CH	MM	Totale
CH	145	22	167
MM	24	77	101
Totale	169	99	268

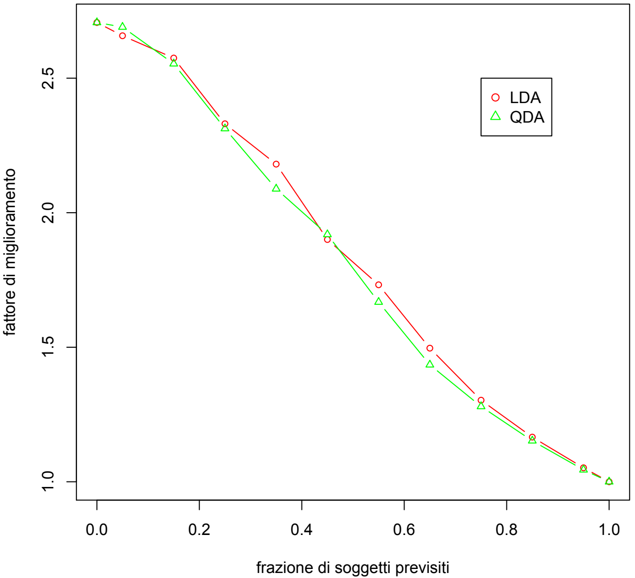
Errata classificazione pari quindi a $42/268 = 0.157$ nel primo caso e $46/268 = 0.172$ nel secondo caso.

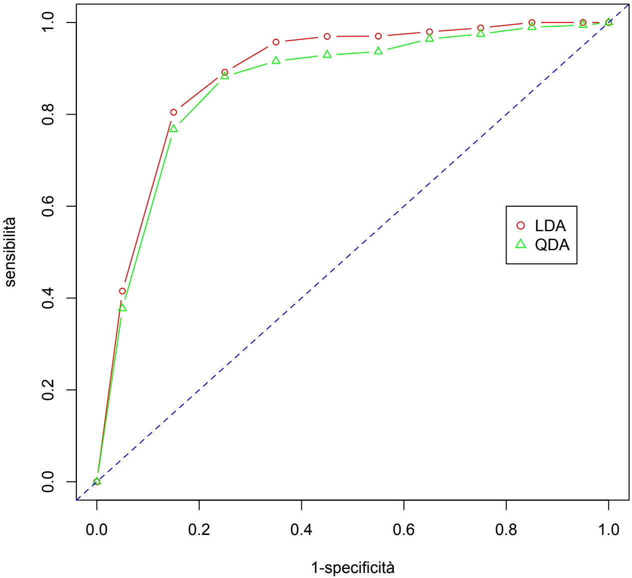
Come scegliere fra le due?

Previsione con LDA	Risposta effettiva		
	CH	MM	Totale
CH	147	20	167
MM	22	79	101
Totale	169	99	268

Previsione con LDA	Risposta effettiva		
	CH	MM	Totale
CH	145	22	167
MM	24	77	101
Totale	169	99	268

Errata classificazione pari quindi a $42/268 = 0.157$ nel primo caso e $46/268 = 0.172$ nel secondo caso.


LDA
QDA
fattore di miglioramento
frazione di soggetti previsiti
1.0
1.5
2.0
2.5
0.0
0.2
0.4
0.6
0.8
1.0



LDA
QDA
sensibilità
1-specificità
0.0
0.2
0.4
0.6
0.8
1.0

Vantaggi

- Appropriatezza metodo: costruito specificamente per il problema di classificazione.
- Informazioni a priori: se sono disponibili possono essere inglobate facilmente nella stima della probabilitá a priori attraverso ad esempio metodi bayesiani.
- Semplicitá di calcolo: sia la stima dei parametri che il calcolo delle funzioni discriminanti sono estremamente semplici dal punto di vista computazionale.
- Qualitá e stabilitá dei risultati: risultati stabili in un gran numero di casi e stabili rispetto all'ingresso di nuovi dati.
- Robustezza rispetto alle ipotesi: anche se le ipotesi del modello sono violate, tende a produrre risultati validi.

Svantaggi

- Ipotesi restrittive: Il metodo é costruito sotto ipotesi abbastanza dettagliate.
- Selezione e graduatoria delle variabili: non ci sono tecniche semplici per valutare quali siano le variabili piú influenti per la classificazione.
- Numero di parametri: Nel caso che p e/o K non siano piccoli, nell'analisi quadratica ci sono tanti parametri da stimare e le stima di alcune matrici di covarianza potrebbero essere singolari.
- Non robustezza stima: stime non robuste rispetto a valori anomali, anche se si puó pensare all'inserimento di procedure robuste.