1.1. Classificatore di Bayes

1.1.1 Introduzione

Il metodo di Bayes rappresenta un fondamentale approccio al problema della classificazione; basato su una descrizione probabilistica delle classi e dei valori dei parametri delle varie classi.

Prima di passare alla trattazione generale si inizierà con un semplice esempio caratterizzato da due classi e da un solo parametro: si supponga di voler eseguire automaticamente la diagnosi di influenza virale in una popolazione basandosi unicamente sulla misurazione della temperatura corporea del paziente. A tale scopo si definiscono le seguenti due classi:

classe ω_1 = Pazienti sani

classe ω_2 = Pazienti affetti da influenza virale.

Indichiamo con $P(\omega_1)$ la probabilità che un generico paziente appartenga 'a priori' alla classe ω_1 e con $P(\omega_2)$ la probabilità che appartenga alla classe ω_2, le due classi sono mutuamente esclusive, e sicuramente ogni persona appartiene o all'una o all'altra classe, per cui sarà $P(\omega_1)+P(\omega_2) = 1$.

Il parametro che descrive l'evento da classificare e che costituisce l' ingresso al classificatore è la temperatura corporea del paziente (β) misurata in gradi centigradi.

Considerando il parametro β come una variabile aleatoria continua, si definiscono le funzioni $p(\beta \mid \omega_1)$ e $p(\beta \mid \omega_2)$ pari alle densità di probabilità di β nelle classi ω_1 e ω_2.

Un possibile andamento di tali funzioni è riportato in fig. 1.1.

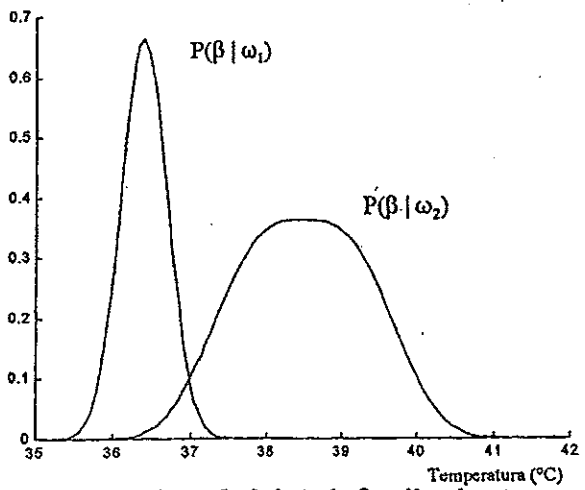


Fig. 1.1 Densità di probabilità di β nelle classi ω_1 e ω_2

A questo punto ci si pone la seguente domanda: dopo aver misurato la temperatura corporea del paziente da classificare e aver trovato che questo valore è pari ad un certo β , come questa misura influenzerà la probabilità che il soggetto sia affetto o meno da influenza virale? La risposta a questa domanda è fornita dalla regola di Bayes:

$$P(\omega_i|\beta) = \frac{p(\beta|\omega_i)P(\omega_i)}{p(\beta)} \text{ con i=1,2}$$

$$\text{dove } p(\beta) = \sum_{i=1}^{2} p(\beta|\omega_i)P(\omega_i)$$

Note le due probabilità a posteriori, il procedimento di classificazione segue la seguente regola:

$\beta \rightarrow \omega_1$	se	$P(\omega_1 \mid \beta) > P(\omega_2 \mid \beta)$	(*)
$\beta \rightarrow \omega_2$	se viceversa	$P(\omega_1 \mid \beta) < P(\omega_2 \mid \beta)$	

dove il simbolo $\rightarrow$ indica l'assegnazione ad una classe.

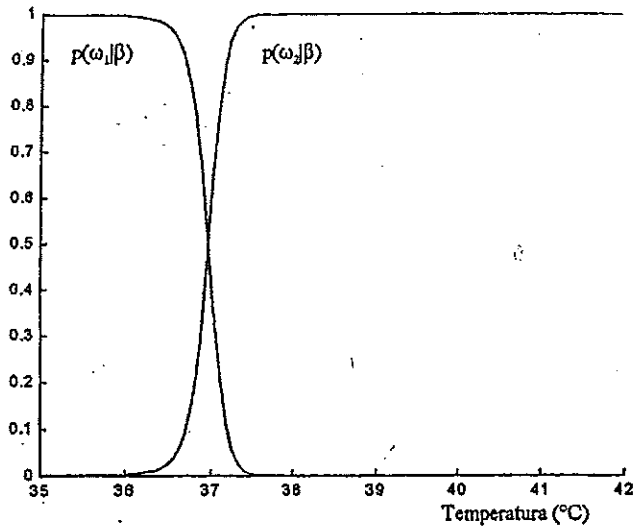


Fig. 1.2. Densità di probabilità a posteriori.

Indicato come in fig. 1.2 con H l'ascissa dell'intersezione delle due funzioni $p(\omega_1|\beta)$ e $p(\omega_2|\beta)$, la regola di decisione può essere riformulata come segue:

$\beta \rightarrow \omega_1$	se	$\beta < H$
$\beta \rightarrow \omega_2$	se	$\beta > H$

Il valore di H quindi rappresenta un valore di soglia (TH) per il parametro β da classificare.
Si dimostra ora che la scelta di tale soglia è quella che consente di minimizzare l'errore di classificazione. Infatti per una scelta qualsiasi (TH) della soglia si hanno due possibili situazioni di errore:

Caso 1:	$\beta \in \omega_1$	ma	$\beta >$TH, quindi $\beta \rightarrow \omega_2$
Caso 2:	$\beta \in \omega_2$	ma	$\beta <$ TH, quindi $\beta \rightarrow \omega_1$

Quindi:

$$P(errore) = \int_{-\infty}^{+\infty} P(errore|\beta)p(\beta)d\beta =$$

$$= \int_{-\infty}^{TH} P(\omega_2|\beta)p(\beta)d\beta + \int_{TH}^{+\infty} P(\omega_1|\beta)p(\beta)d\beta$$

Essendo p(β) una funzione che assume valori comunque positivi P(errore) è minima se, per ogni valore di β si integra la funzione probabilità a posteriori minima tra le due:

$$\min \text{P(errore)} = \int_{-\infty}^{+\infty} \min[P(\omega_1|\beta), P(\omega_2|\beta)]p(\beta)d\beta$$

Con riferimento alla fig. 1.3 è evidente, anche da un punto di vista grafico come la scelta di una soglia diversa da H comporta una probabilità di errore maggiore.
Quindi si è dimostrato che la regola di classificazione (*) è quella che minimizza la P_{errore}. Pertanto il classificatore basato su tale regola si chiama classificatore a errore minimo.

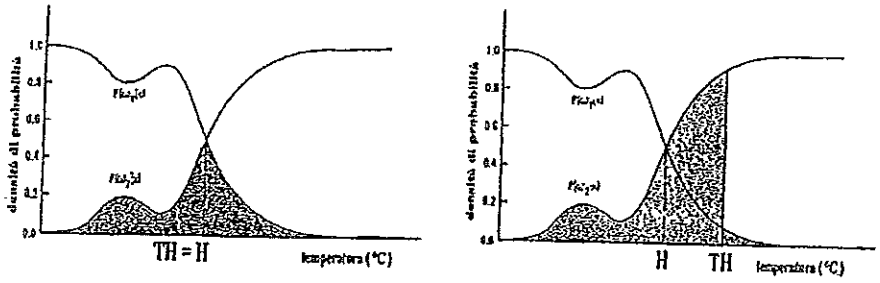


Fig. 1.3. - Rappresentazione grafica dell'errore di classificazione

1.1.2 Estensione del metodo di Bayes al caso generale

Considerando ora la situazione più generale in cui si hanno M classi $\omega_1, \omega_2,...,\omega_M$ e l'evento da classificare è rappresentato da un vettore $\beta=[\beta_1,\beta_2,...\beta_N]$ di dimensione N.
In analogia al caso discusso in precedenza, supponiamo note la probabilità a priori delle singole classi

$$P(\omega_i) \quad i = 1,...,M \qquad \text{con} \quad \sum_{i=1}^{M} P(\omega_i) = 1$$

e le densità di probabilità del vettore aleatorio β condizionato dalle classi stesse

$$p(\beta|\omega_i) \qquad \text{con } i = 1,...,M$$

Si introduce un ulteriore concetto che consente di dare un costo diverso alle varie situazioni di errore. Si assume cioè che, ad ogni decisione $\beta \rightarrow \omega_i$ segua un' azione α_i , e si definisce come $\lambda(\alpha_i|\ \omega_j)$ il costo dell' azione α_i quando la classe vera di appartenenza è ω_j. Essendo $P(\omega_j|\beta)$ la probabilità che la classe a cui appartiene il parametro β sia ω_j, il costo complessivo associato alla decisione α_i è rappresentato dalla seguente espressione:

$$R(\alpha_i|\beta) = \sum_{j=1}^{M} \lambda(\alpha_i|\omega_j) P(\omega_j|\beta)$$

Nella terminologia della teoria della decisione, tale costo viene chiamato *rischio condizionato*. Il rischio totale legato all' intero procedimento di decisione è dato dal seguente integrale:

$$R = \int_{R^N} R(\alpha_i|\beta)p(\beta)d\beta$$

La regola di decisione allora assegna l'elemento β alla classe che comporta il rischio minore:

$$\beta \rightarrow \omega_i \qquad \text{se} \quad R(\alpha_i \mid \beta) < R(\alpha_j \mid \beta) \quad \text{con } j = 1,...,M, \quad j \neq i$$

Per chiarire meglio il concetto di costo si prenda in considerazione l'esempio di classificazione tra due classi visto precedentemente.
Dopo l'applicazione della regola di Bayes si conosce la probabilità a posteriori $P(\omega_1 \mid \beta)$ e $P(\omega_2 \mid \beta)$ e quindi si può passare al calcolo dei rischi. Prima di tale calcolo è necessario definire le azioni α_1 e α_2, ad esempio

α_1 = terapia/astensione dal lavoro
α_2 = nessuna cura

Quindi utilizzando la notazione semplificata $\lambda(\alpha_i \mid \omega_j) = \lambda_{ij}$ si specificano i costi:

λ_{12} = costo della terapia/astensione dal lavoro per un soggetto sano
λ_{21} = costo di escludere dalla terapia un soggetto affetto da influenza

In generale λ_{11} e λ_{22} sono costi che si possono ritenere pari a zero in quanto si riferiscono ad una diagnosi corretta.
Seguono le espressioni dei costi condizionati:

$$R(\alpha_1 \mid \beta) = \lambda_{11} P(\omega_1 \mid \beta) + \lambda_{12} P(\omega_2 \mid \beta)$$
$$R(\alpha_2 \mid \beta) = \lambda_{21} P(\omega_1 \mid \beta) + \lambda_{22} P(\omega_2 \mid \beta)$$

La regola di decisione, nel caso specifico, segue lo schema:

se $R(\alpha_1 \mid \beta) < R(\alpha_2 \mid \beta)$ allora $\beta \rightarrow \omega_1$

altrimenti $\beta \rightarrow \omega_2$

Esplicitando la funzione rischio, scegliere la classe ω_1 significa che la seguente disequazione è soddisfatta

$$R(\alpha_1 \mid \beta) = (\lambda_{11} P(\omega_1 \mid \beta) + \lambda_{12} P(\omega_2 \mid \beta) < R(\alpha_2 \mid \beta) = \lambda_{21} P(\omega_1 \mid \beta) + \lambda_{22} P(\omega_2 \mid \beta)$$

che equivale, dopo alcuni semplici passaggi, alla disequazione

$$\frac{P(\omega_1|\beta)}{P(\omega_2|\beta)} > \frac{\lambda_{12} - \lambda_{22}}{\lambda_{21} - \lambda_{11}}$$

dove il secondo membro dipende solo dai costi ed è detto *rapporto di verosimiglianza.*
Tornando ad M classi e N parametri, è interessante esaminare il caso particolare in cui si scelgono i costi nel modo seguente:

$$\lambda(\alpha_i|\omega_i) = \lambda_{ij} = \begin{cases} 0 & se \quad i = j \\ 1 & se \quad i \neq j \end{cases}$$

che significa assegnare costo nullo alle decisioni corrette e costi tutti uguali e pari a 1 alle decisioni sbagliate. L'espressione del rischio condizionato relativo all'azione α_i allora diventa:

$$R(\alpha_i|\beta) = \sum_{\substack{j=1 \\ j \neq i}}^{M} P(\omega_j|\beta)$$

Essendo: $\sum_{j=1}^{M} P(\omega_j|\beta) = 1$, il rischio diventa:

$$R(\alpha_i \mid \beta) = 1 - P(\omega_i \mid \beta)$$

e quindi, essendo il rischio pari al complemento della probabilità a posteriori della classe, scegliere la classe a rischio minimo equivale a scegliere la classe con la massimo probabilità a posteriori (classificatore ad errore minimo).

Quindi, il classificatore a errore minimo può essere visto come caso particolare del classificatore a rischio minimo per la scelta particolare dei pesi.

1.1.3 Funzioni discriminanti per il classificatore di Bayes

Un modo generale per rappresentare il classificatore di Bayes è tramite un insieme di funzioni del vettore dei parametri β dette *funzioni discriminanti*. Nel caso di M classi, tale insieme sarà formato da M funzioni

$$g_i(\beta) \qquad \text{con } i = 1,\ldots,M$$

ciascuna associata alle varie classi.

Il classificatare assegna β alla classe ω_i se e solo se:

$$g_i(\beta) > g_k(\beta) \qquad \text{per ogni } i \neq k$$

quindi può essere visto come un dispositivo che riceve come dato di input il vettore dei parametri β, calcola le M funzioni discriminanti, ed infine tramite un comparatore, sceglie la funzione che ha fornito il valore massimo ed assegna β alla classe corrispondente.

Un possibile schema a blocchi è mostrato in fig. 1.4.

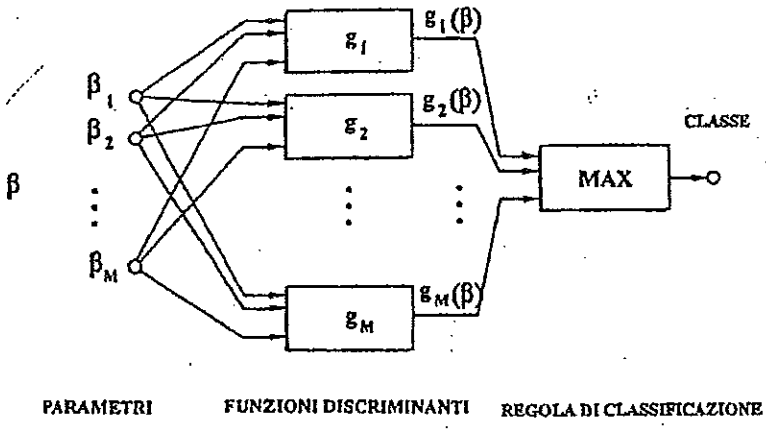


Fig. 1.4. - Schema a blocchi di un classificatore che utilizza le funzioni discriminanti.

Per il classificatore di Bayes a errore minimo, è immediato identificare la funzione discriminante con la probabilità a posteriori ponendo:

$$g_i(\beta) = P(\omega_i \mid \beta) \qquad i = 1,...,M$$

Tenendo conto che:

$$P(\omega_i|\beta) = \frac{p(\beta|\omega_i)P(\omega_i)}{p(\beta)} \qquad \text{con i} = 1,...,M$$

ed osservando che il termine p(β) è positivo e non dipende dall'indice i (quindi non discrimina) si può anche scrivere:

$$g_i(\beta) = p(\beta \mid \omega_i)P(\omega_i)$$

Per classificatori a costo minimo, le funzioni discriminanti possono essere definite come l'inverso delle funzioni di rischio:

$$g_i(\beta) = -R(\alpha_1 \mid \beta)$$

Si possono anche considerare trasformazioni monotone per ottenere altre funzioni discriminanti, ad esempio utilizzando la funzione logaritmo si ottiene:

$$g_i(\beta) = \ln [p(\beta \mid \omega_i)] + \ln [P(\omega_i)]$$

La scelta di una di queste funzioni non inciderà sul risultato della classificazione, ma sulla realizzazione del classificatore stesso.

Un classificatore, in generale, ha come effetto quello di suddividere lo spazio dei parametri R^N in regioni contigue, separate da iperpiani di dimensione (N-1) che sono il luogo dei punti per io quali vale:

$$g_i(\beta)=g_k(\beta) \quad \text{per qualche } i \neq k$$

Due esempi sono riportati in fig 1.5a e 1.5b: il primo considera il caso di tre classi ed un solo parametro, mentre il secondo il caso di due classi e di due parametri.

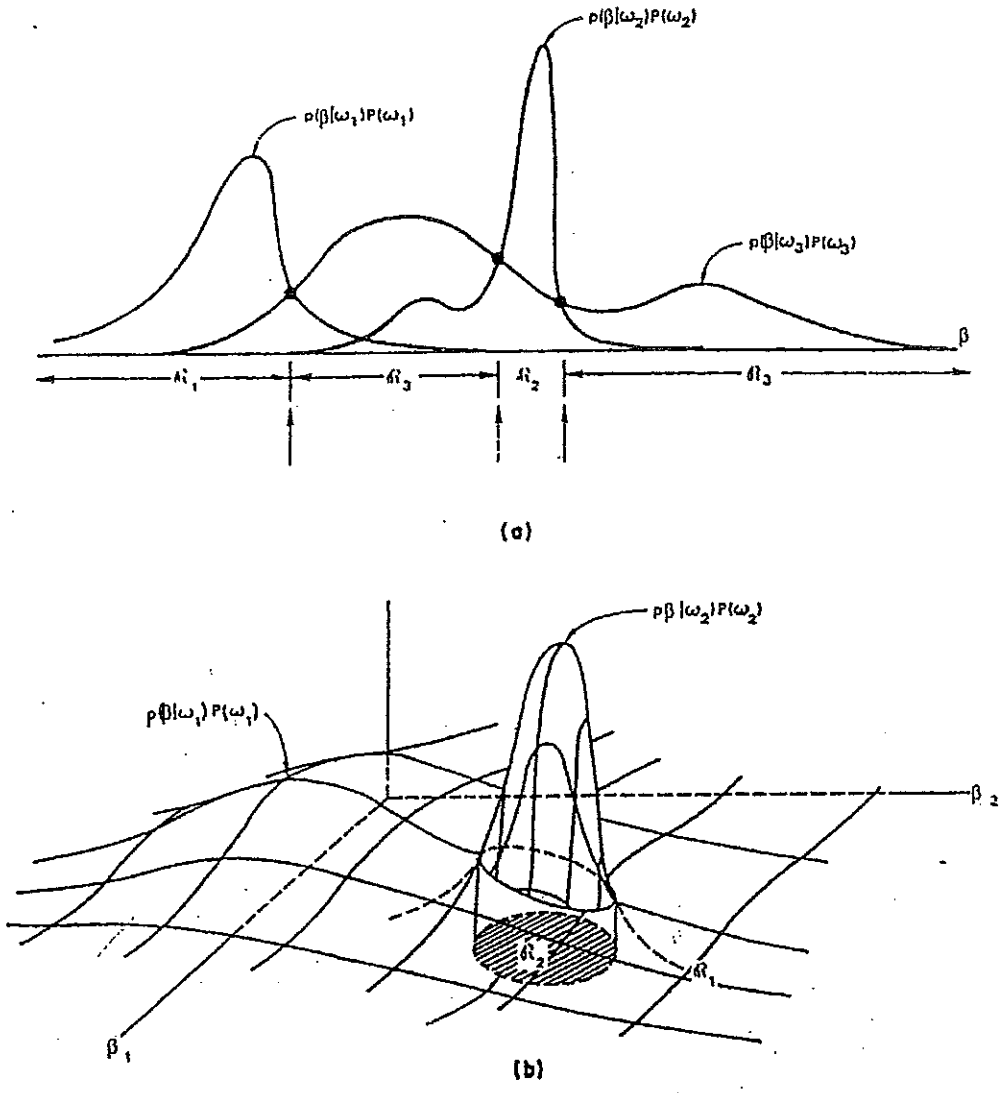


Fig. 1.5. - Esempi di superfici di decisione e di regioni di decisione

1.1.4 Classificatore ad errore minimo nel caso di densità di probabilità gaussiana

Per applicare il metodo di Bayes è necessario conoscere le probabilità a priori e le densità di probabilità condizionate $p(\beta \mid \omega_i)$. La struttura del classificatore dipende dalla descrizione statistica di tale funzione. Supponiamo allora che β sia un vettore aleatorio gaussiano, e quindi:

$$p(\beta|\omega_i) \in N(\mu_i, \Sigma_i)$$

per cui si ha:

$$p(\beta|\omega_i) = \frac{1}{(2\pi)^{\frac{N}{2}}|\Sigma_i|^{\frac{1}{2}}} e^{-\frac{1}{2}(\beta-\mu_i)^T \Sigma_i^{-1}(\beta-\mu_i)}$$

dove:

$\mu_i = E(\beta)$ è l' aspettazione del vettore dei parametri β nella classe ω_i

$\Sigma_i = E[(\beta - \mu_i)(\beta - \mu_i)^T]$ è la matrice di covarianza N×N di β nella classe ω_i

Nel caso bidimensionale l' andamento di tale funzione è mostrata in fig. 1.6a. Le curve di livello cioè il luogo dei punti a probabilità costante sono della ellissi le cui direzioni degli assi sono individuate dagli autovettori della matrice di covarianza Σ.

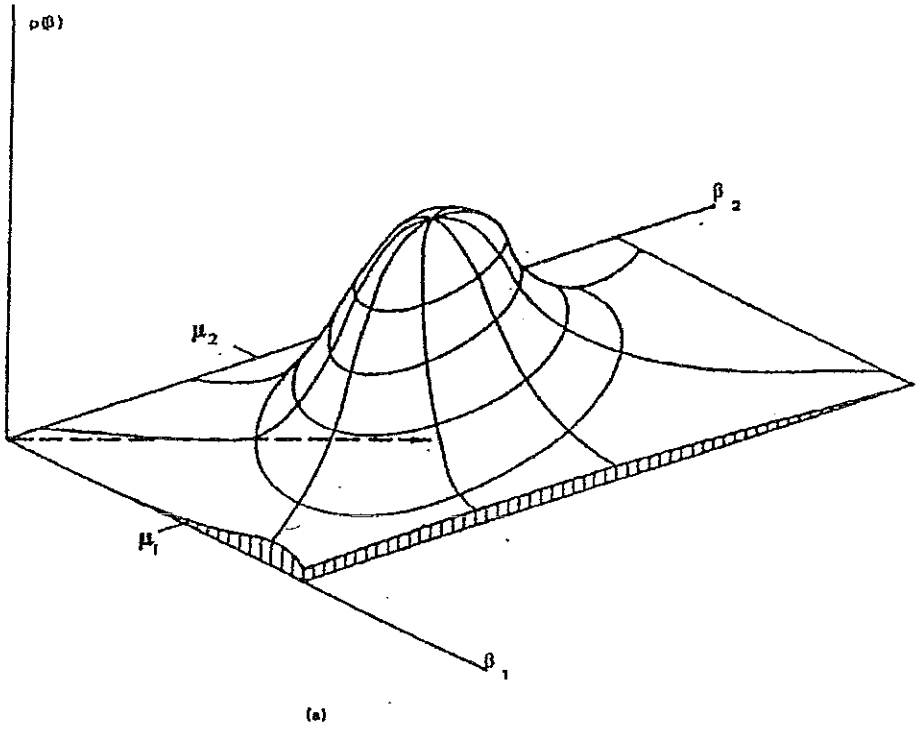


Fig. 1.6.-Densita di probabilità gaussiana

In questo caso è conveniente scegliere come funzione discriminante la seguente funzione:

$$g_i(\beta) = \ln [p(\beta \mid \omega_i)] + \ln [P(\omega_i)]$$

quindi

$$g_i(\beta) = -\frac{N}{2}\ln(2\pi) - \frac{1}{2}\ln|\Sigma_i| - \frac{1}{2}[(\beta - \mu_i)^T \Sigma_i^{-1}(\beta - \mu_i)] + \ln[P(\omega_i)]$$

Pertanto, le funzioni discriminanti sono funzioni quadratiche del vettore β.

Si prenderanno ora in considerazione alcuni casi particolari:

1. $\Sigma_i = \sigma^2 I$ per $\forall$ i

La condizione è soddisfatta nel caso di parametri statisticamente indipendenti e tutti con la stessa varianza, allora geometricamente questo significa che i valori delle varie classi si distribuiscono all' interno di ipersfere uguali per tutte le classi i cui centri sono i punti di R^N individuati dai vettori delle medie μ_i

Nella espressione della funzione discriminante i termini che non dipendono da una particolare classe non hanno un peso nel discriminare e quindi possono essere trascurati.

Si ottiene allora:

$$g_i(\beta) = -\frac{1}{2\sigma^2}[(\beta - \mu_i)^T(\beta - \mu_i)] + \ln[P(\omega_i)]$$

$$= -\frac{1}{2\sigma^2}\|\beta - \mu_i\|^2 + \ln[P(\omega_i)]$$

dove la norma si intende euclidea.

Se si introduce l'ulteriore ipotesi di classi equiprobabili, le funzioni discriminanti assumono la forma:

$$g_i(\beta) = -\frac{1}{2\sigma^2}\|\beta - \mu_i\|^2$$

Si ottiene così il classificatore di Bayes a minima distanza euclidea. Infatti, il classificatore associa il punto nello spazio dei parametri β alla classe la cui media si trova alla minima distanza da β.

In fig 1.7 è rappresentato un caso bidimensionale: due sono le classi con valori medi μ_1, μ_2, e matrice di covarianza $\Sigma = \sigma^2 I$ (le ellissi diventano cerchi).

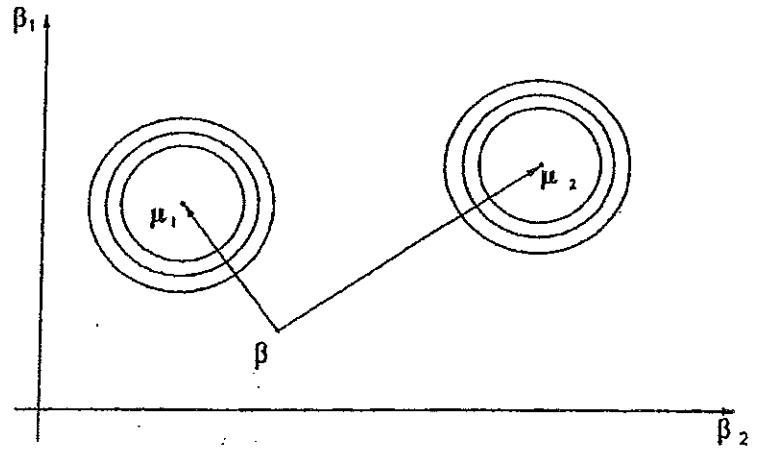


Fig. 1.7. - Classificatore a minima distanza euclidea nel caso di due classi.

Il classificatore che che soddisfa alle suddette ipotesi viene detto *classificatore di Bayes lineare* in quanto le funzioni discriminanti possono essere espresse come funzioni lineari del vettore dei parametri β come evidenziato da quanto segue:

$$g_i(\beta) = -\frac{1}{2\sigma^2}\|\beta - \mu_i\|^2 \Rightarrow$$

$$g_i(\beta) = -\frac{1}{2\sigma^2}(\beta^T\beta - 2\mu_i^T\beta + \mu_i^T\mu_i)$$

Trascurando il termine $\beta^T\beta$ che essendo indipendente dalla classe non discrimina, si ottiene:

$$g_i(\beta) = a_i^T\beta + b_i$$

con: $a_i = \frac{1}{\sigma^2}\mu_i$, $\qquad b_i = -\frac{1}{2\sigma^2}\mu_i^T\mu_i$

Le superfici di separazione, che identificano le regioni nelle quali è diviso lo spazio dei parametri, sono degli iperpiani di dimensione N-1. L' iperpiano I_{ik} sarà ortogonale al vettore differenza $\mu_i - \mu_k$ e, nel caso di classi equiprobabili, passante per il punto di mezzo della congiungente delle medie. Infatti le superfici di separazione sono il luogo dei punti per i quali $g_i(\beta) = g_k(\beta)$ e quindi, nel caso specifico, l' equazione diventa:

$$a_i^T \beta + b_i = a_k^T \beta + b_k$$

che equivale a

$$(a_i^T - a_k^T)\beta + (b_i - b_k) = 0$$

Sostituendo si ottiene:

$$\frac{1}{\sigma^2}(\mu_i^T - \mu_k^T)\beta + (b_i - b_k) = 0$$

quindi β appartiene all'iperpiano ortogonale al vettore $(\mu_i - \mu_k)$.

2. $\Sigma_i = \Sigma \quad \forall i$

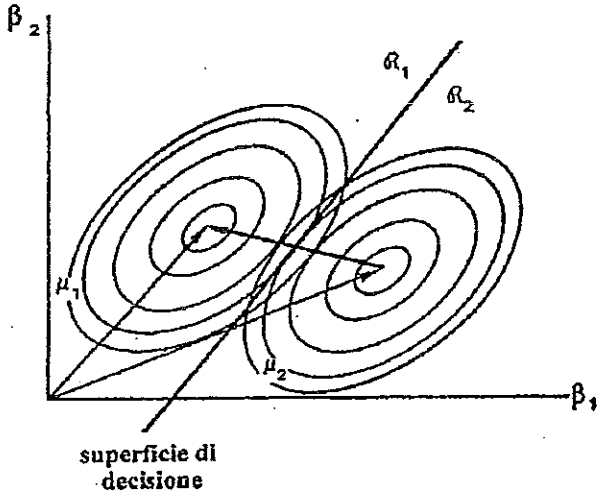
In questo caso tutte le classi hanno la stessa matrice di covarianza. Dal punto di vista geometrico questo significa che i campioni cadono all'interno di iperellissoidi tutti uguali i cui centri sono individuati dai vettori media μ_i.
Trascurando ancora i termini che non dipendono dalla classe, tutti uguali tra loro, le funzioni discriminanti hanno la forma:

$$g_i(\beta) = -\frac{1}{2}[(\beta - \mu_i)^T \Sigma^{-1}(\beta - \mu_i)] + \ln[P(\omega_i)]$$

Il primo termine di $g_i(\beta)$ è la distanza secondo una norma non euclidea (*distanza di Mahalanobis*) tra il vettore β da classificare ed i vettori delle medie μ_i:

$$[(\beta - \mu_i)^T \Sigma^{-1}(\beta - \mu_i)]$$

Se le classi sono equiprobabili si ottiene il *classificatore di Bayes a minima distanza di Mahalanobis*. Un esempio con due classi e due parametri è riportato in Fig. 1.8.



Anche per questo classificatore si può dimostrare che le funzioni discriminanti sono funzioni lineari del vettore dei parametri β:

$$g_i(\beta) = a_i^T \beta + b_i$$

con:

$$a_i = \Sigma_i^{-1}\mu_i,\ b_i = -\frac{1}{2}\mu_i^T \Sigma_i^{-1}\mu_i + \ln[P(\omega_i)]$$

Le superfici di separazione che partizionano lo spazio dei parametri sono quindi ancora degli iperpiani

di dimensione N-1. Siano $\mathcal{R}_i$ ed $\mathcal{R}_k$ due superfici contigue, l' equazione dell' iperpiano che le separa si ricava nello stesso modo visto sopra:

$$a_i^T\beta + b_i = a_k^T\beta + b_k$$

che equivale a

$$(a_i^T - a_k^T)\beta + (b_i - b_k) = 0$$

Dopo semplici passaggi si ottiene:

$$a^T(\beta - \beta_0) = 0$$

con:

$$a = \Sigma^{-1}(\mu_i - \mu_i)$$

$$\beta_0 = \tfrac{1}{2}(\mu_i + \mu_k) - \frac{\log\frac{P(\omega_i)}{P(\omega_k)}}{(\mu_i - \mu_k)^T\Sigma^{-1}(\mu_i - \mu_k)}(\mu_i - \mu_k)$$

Si possono fare alcune considerazioni: generalmente il vettore $\mathbf{K^{-1}}(\mu_i - \mu_k)$ non ha la direzione del vettore $(\mu_i - \mu_k)$ e quindi l' iperpiano di separazione tra $\mathcal{R}_i$ ed $\mathcal{R}_k$ non è ortogonale alla linea congiungente le due medie. Tuttavia, l' iperpiano interseca tale linea nel punto di coordinate β_0 che si trova alla metà tra le medie solo nel caso di classi equiprobabili.

1.2. Stima delle probabilità a priori

Il metodo di Bayes impone la conoscenza a priori di $p(\beta \mid \omega_i)$ e di $P(\omega_i)$ con i = 1,...,M. Nella pratica raramente ci si trova in questa condizione ideale e diventa necessario mettere a punto un procedimento di stima di tali grandezze da utilizzare in una fase precedente a quella della progettazione del classificatore vero e proprio.

Un approccio al problema è quello di utilizzare un insieme di campioni che costituiscono il *training set* (T.S.). Tale approccio è denominato stima con supervisore.

Il *training set* è in sostanza un insieme di elementi simili agli elementi da classificare di cui però si conosce a priori ed in modo indipendente la classe di appartenenza, cioè:

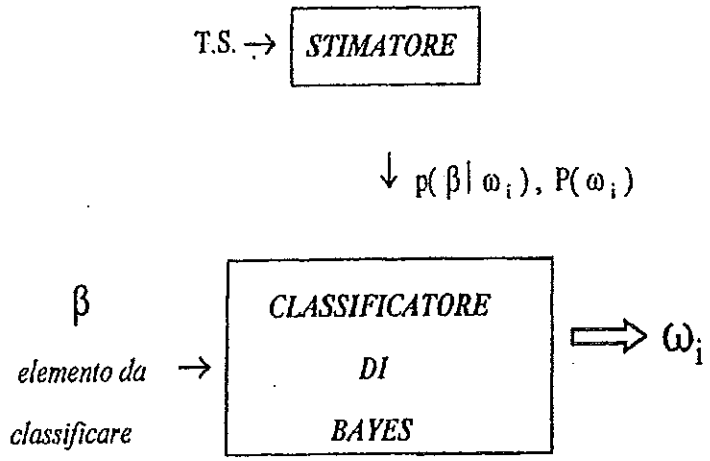
$$\text{T.S.}: \quad \beta_i^{(1)} \quad i = 1,...,Q_1 \in \omega_1$$
$$\beta_i^{(2)} \quad i = 1,...,Q_2 \in \omega_2$$
$$\vdots$$
$$\beta_i^{(M)} \quad i = 1,...,Q_M \in \omega_M$$

Quindi:

1. A partire dal training set si stimano $p(\beta \mid \omega_i)$ e $P(\omega_i)$.

2. Assumendo queste stime come i valori veri di $p(\beta \mid \omega_i)$ e di $P(\omega_i)$, si mette a punto il classificatore.

Schematizzando le due fasi si ottiene:



La stima di $P(\omega_i)$ è un problema più semplice nel senso che spesso succede di possedere effettivamente tale informazione, ad esempio l' incidenza patologia ω_i può essere nota a priori da dati sulla popolazione. In alternativa, utilizzando il training set, ammesso che questo descriva un campione rappresentativo della popolazione, si può stimare la probabilità delle singole classi dalle frequenze campionarie, cioè se Q è il numero di campioni che costituiscono il T.S. e di questi $Q_1 \in \omega_1$, $Q_2 \in \omega_2$, ..., $Q_M \in \omega_M$, le probabilità a priori possono essere stimate come:

$$\hat{P}(\omega_1) = \frac{Q_1}{Q} \qquad \hat{P}(\omega_2) = \frac{Q_2}{Q} \quad \ldots \qquad \hat{P}(\omega_M) = \frac{Q_M}{Q}$$

Più difficile è in generale la stima di $p(\beta \mid \omega_i)$.

Si possono distinguere due situazioni di partenza:

1. La struttura probabilistica è nota a priori, ma i parametri sono incogniti e quindi oggetto di stima.

2. Non si possiedono informazioni a priori, quindi struttura e parametri sono incogniti

1.2.1 Struttura probabilistica nota a priori

Se la struttura probabilistica è nota a priori, si riporta il problema della stima della densità di probabilità ad un problema di stima parametrica. Ad esempio, nel caso gaussiano ($p(\beta \mid \omega_i) \in N(\mu_i, \Sigma_i)$), i parametri da stimare sono il vettore media μ_i e la matrice covarianza Σ_i. Allora, $p(\beta \mid \omega_i)$ può essere considerata funzione del vettore dei parametri φ, (considerando che Σ_i è matrice simmetrica di dimensione N×N) di dimensione $\frac{N(N-1)}{2} + 2N$.

Quindi:

$$\varphi = [\, \mu_1, \mu_2, \ldots, \mu_N, \Sigma_{11}, \Sigma_{21}, \Sigma_{22}, \ldots, \Sigma_{NN} \,]$$

Per semplificare la trattazione del problema, assumeremo che i campioni del training set siano vettori aleatori statisticamente indipendenti. Prendendo in considerazione una classe alla volta, si divide il problema in M sottoproblemi simili tra di loro.

Consideriamo come esempio la stima di $p(\omega_1 \mid \beta)$: dati Q_1 vettori indipendenti $\beta_1, \beta_2, \ldots, \beta_{Q1}$ appartenenti alla classe ω_1, nota la struttura della funzione densità di probabilità $p(\beta \mid \omega_1)$, si può scrivere che

$$P(\beta_1 \beta_2 \ldots \beta_{Q_1} \mid \omega_1, \varphi) = \prod_{j=1}^{Q_1} p(\beta_j \mid \omega_1, \varphi)$$

Se ora si calcola tale funzione in corrispondenza dei Q_1 campioni appartenenti alla classe 1 del T.S. si ottiene una funzione dei soli parametri φ. La stima parametrica di massima verosimiglianza consiste nello scegliere i vari parametri φ in modo da massimizzare tale funzione, cioè di scegliere i parametri φ che rendono le misure le più verosimili possibili: e quindi

$$\hat{\varphi} = \arg\left[\max_{\varphi} P(\beta_1 \beta_2 \ldots \beta_{Q_1} \mid \omega_j, \varphi) \right]$$

Risolvendo tale problema, nel caso particolare di distribuzione gaussiana, si ottiene il seguente

G43

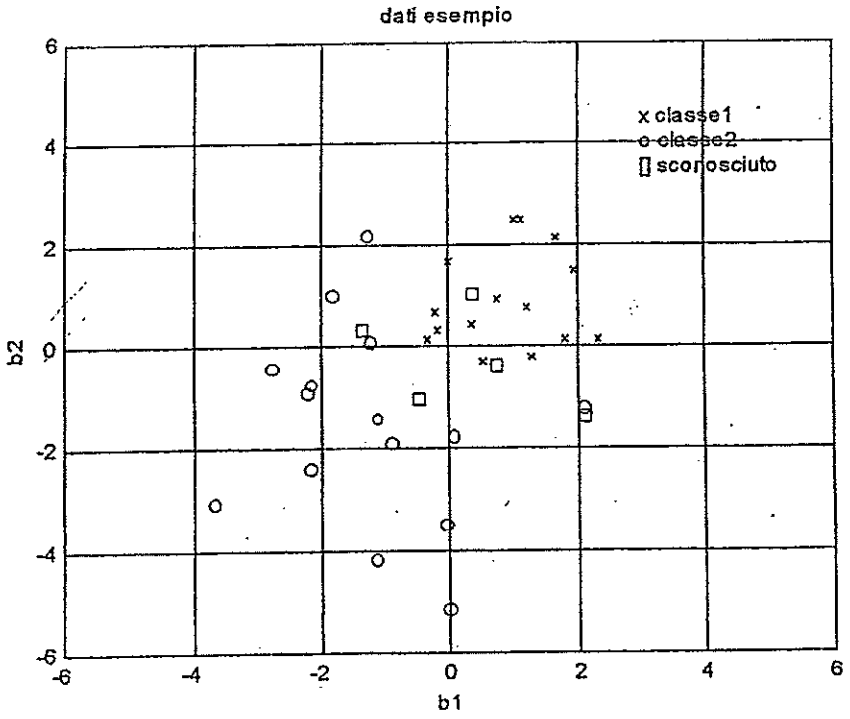
risultato:

$$\hat{\mu}_i = \frac{1}{d}\sum_{j=1}^{d}\beta_j ; \qquad \hat{\Sigma}_i = \frac{1}{d}\sum_{j=1}^{d}[(\beta_j - \hat{\mu}_i)(\beta_j - \hat{\mu}_i)^T]$$

Nel caso di distribuzioni diverse da quella gaussiana non sempre si giunge ad una soluzione analitica e quindi il calcolo delle stime viene spesso eseguito per via numerica.

Esempio 1

Siano date due classi ω1 e ω2 e un vettore dei parametri bidimensionale $\beta=(\beta_1, \beta_2)$. Il training set è costituito da quindici esempi per ogni classe. Si assume che i parametri abbiano distribuzione gaussiana. La probabilità a priori delle due classi sia P[ω1]= P[ω2]=0.5



Fig, e1.1 Dati dell'esempio 1

Siano $\beta_{1,x}$ e $\beta_{2,x}$ i vettori appartenenti alle classi ω_1 e ω_2 rispettivamente:

$\beta_{1,1} = 1.1286 \quad 2.4789$

$\beta_{1,2} = 1.6565 \quad 2.1380$

$\beta_{1,3} = -0.1678 \quad 0.3159$

$\beta_{1,4} = 0.5394 \quad -0.2919$

$\beta_{1,5} = 0.7376 \quad 0.9271$

$\beta_{1,6} = -0.2132 \quad 0.6694$

$\beta_{1,7} = -0.3194 \quad 0.1564$

$\beta_{1,8} = 1.9312 \quad 1.4978$

$\beta_{1,9} = 1.0112 \quad 2.4885$

$\beta_{1,10} = 0.3549 \quad 0.4535$

$\beta_{1,11} = 1.8057 \quad 0.1532$

$\beta_{1,12} = 1.2316 \quad 0.7537$

$\beta_{1,13} = 0.0102 \quad 1.6630$

$\beta_{1,14} = 2.3396 \quad 0.1458$

$\beta_{1,15} = 1.2895 \quad -0.2013$

$\beta_{2,1} = -1.2398 \quad 0.1058$

$\beta_{2,2} = -1.1306 \quad -1.4074$

$\beta_{2,3} = -0.0294 \quad -5.1086$

$\beta_{2,4} = -2.1910 \quad -0.7348$

$\beta_{2,5} = -1.2994 \quad 2.1858$

$\beta_{2,6} = -1.8696 \quad 1.0368$

$\beta_{2,7} = -1.1586 \quad -4.1608$

$\beta_{2,8} = 2.0704 \quad -1.1574$

$\beta_{2,9} = -2.2130 \quad -2.3634$

$\beta_{2,10} = -3.6948 \quad -3.0492$

$\beta_{2,11} = -0.0612 \quad -3.4688$

$\beta_{2,12} = -2.8072 \quad -0.4224$

$\beta_{2,13} = -0.9282 \quad -1.8586$

$\beta_{2,14} = -2.2550 \quad -0.8884$

$\beta_{2,15} = 0.0708 \quad -1.7358$

$\beta_{x,x}$ il vettore dei parametri da classificare:

$\beta_{x,1} = -0.4650 \quad -1.0226$

$\beta_{x,2} = 0.3710 \quad 1.0378$

$\beta_{x,3} = 0.7283 \quad -0.3898$

$\beta_{x,4} = 2.1122 \quad -1.3813$

$\beta_{x,5} = -1.3573 \quad 0.3155$

Si ottengono le seguenti stime

$$\hat{\mu}_1 = \frac{1}{15}\sum_{i=1}^{15} \beta_{i1} = [\ 0.8890 \quad 0.8899]$$

$$\hat{\mu}_2 = \frac{1}{15}\sum_{i=1}^{15} \beta_{i2} = [-1.2491 \quad -1.5351]$$

$$\hat{\Sigma}_1 = \left[\frac{1}{15}\sum_{i=1}^{15} (\beta_{i1} - \hat{\mu}_1)(\beta_{i1} - \hat{\mu}_1)^T\right] = \begin{bmatrix} 0.7118 & 0.1162 \\ 0.1162 & 0.8868 \end{bmatrix}$$

$$\hat{\Sigma}_1^{-1} = \begin{bmatrix} 1.4357 & -0.1881 \\ -0.1881 & 1.1522 \end{bmatrix}$$

$$\hat{\Sigma}_2 = \left[\frac{1}{15}\sum_{i=1}^{15} (\beta_{i2} - \hat{\mu}_2)(\beta_{i2} - \hat{\mu}_2)^T\right] = \begin{bmatrix} 1.9579 & -0.4570 \\ -0.4570 & 3.7239 \end{bmatrix}$$

$$\hat{\Sigma}_2^{-1} = \begin{bmatrix} 0.5258 & 0.0645 \\ 0.0645 & 0.2765 \end{bmatrix}$$

La funzione discriminante ha la forma:

$$g_i(\beta) = \ln\left[P(\beta|\omega_i)\right] + \ln\left[(P(\omega_i)\right]$$

che nel caso gaussiano diventa:

$$g_i = -\frac{N}{2}\ln(2\pi) - \frac{1}{2}\ln|\Sigma_i| - \frac{1}{2}\left[(\beta - \mu_i)^T \Sigma_i^{-1}(\beta - \mu_i)\right] + \ln\left[P(\omega_i)\right]$$

trascurando i termini comuni alle due classi (si ricordi P[ω1] = P[ω2]):

$$g_i = -\frac{1}{2}\ln|\Sigma_i| - \frac{1}{2}\left[(\beta - \mu_i)^T \Sigma_i^{-1}(\beta - \mu_i)\right]$$

utilizzando come validation set gli stessi dati utilizzati per il training set si ottiene

	$g_1(\beta_{1,x})$	$g_2(\beta_{1,x})$	classe		$g_1(\beta_{2,x})$	$g_1(\beta_{2,x})$	classe
$\beta_{1,1}$	-1.1835	-5.3081	ω_1	$\beta_{2,1}$	-3.0526	-1.3520	ω_2
$\beta_{1,2}$	-0.8993	-5.7520	ω_1	$\beta_{2,2}$	-4.8551	-0.9857	ω_2
$\beta_{1,3}$	-0.6366	-1.8889	ω_1	$\beta_{2,3}$	-20.0583	-2.8538	ω_2
$\beta_{1,4}$	-0.5738	-2.1769	ω_1	$\beta_{2,4}$	-7.1486	-1.2519	ω_2
$\beta_{1,5}$	0.2225	-3.1701	ω_1	$\beta_{2,5}$	-4.6980	-2.8812	ω_2
$\beta_{1,6}$	-0.6136	-2.0800	ω_1	$\beta_{2,6}$	-5.3106	-1.8914	ω_2
$\beta_{1,7}$	-0.9507	-1.7030	ω_1	$\beta_{2,7}$	-15.5204	-1.9185	ω_2
$\beta_{1,8}$	-0.6326	-5.5318	ω_1	$\beta_{2,8}$	-3.6305	-3.9765	ω_1*
$\beta_{1,9}$	-1.2055	-5.1466	ω_1	$\beta_{2,9}$	-10.8663	-1.3694	ω_2
$\beta_{1,10}$	-0.0298	-2.4077	ω_1	$\beta_{2,10}$	-20.3857	-3.1071	ω_2
$\beta_{1,11}$	-0.8019	-4.1590	ω_1	$\beta_{2,11}$	-10.5735	-1.7184	ω_2
$\beta_{1,12}$	0.1372	-3.6872	ω_1	$\beta_{2,12}$	-9.6463	-1.6763	ω_2
$\beta_{1,13}$	0.7857	-3.0694	ω_1	$\beta_{2,13}$	-5.5425	-1.0136	ω_2
$\beta_{1,14}$	-1.7915	-5.1446	ω_1	$\beta_{2,14}$	-7.6254	-1.2606	ω_2
$\beta_{1,15}$	-0.6424	-3.1375	ω_1	$\beta_{2,15}$	-3.8075	-1.4253	ω_2

* classificato non correttamente

Se $g_2 > g_1$ il parametro viene classificato appartenente alla classe ω_2 e viceversa se $g_1 > g_2$ appartenente alla classe ω_1.

Classificando i termini incogniti si ottiene:

	$g_1(\beta_{x,x})$	$g_2(\beta_{x,x})$	classe
$\beta_{x,1}$	-2.6954	-1.2027	ω_2
$\beta_{x,2}$	0.0212	-2.8529	ω_1
$\beta_{x,3}$	-0.6824	-2.3343	ω_1
$\beta_{x,4}$	-4.3273	-3.9859	ω_2
$\beta_{x,5}$	-3.3288	-1.4424	ω_2

Se si classificano gli stessi parametri introducendo dei costi diversi per una classificazione sbagliata per esempio:

$\lambda_{11}=0$, $\lambda_{22}=0$, $\lambda_{12}=2$, $\lambda_{21}=1$, [si penalizza maggiormente l'assegnazione alla classe ω_1 OK otteniamo:

$$R(\alpha_1|\beta) = \lambda_{12}P(\omega_2|\beta)$$

$$R(\alpha_2|\beta) = \lambda_{21}P(\omega_1|\beta)$$

minimizzare il rischio nel caso di sole classi ~~vuol dire~~ equivale a minimizzare

$$g'_i(\beta) = \ln\left[P(\beta|\omega_j)\right] + \ln\left[(P(\omega_j)\right] + \ln\lambda_{ij} \quad \text{con i=1,2; j=1,2; i}\neq\text{j.}$$

	$g_1(\beta_{x,x})$	$g_2(\beta_{x,x})$	classe
$\beta_{x,1}$	-2.0023	-2.6954	ω_2
$\beta_{x,2}$	0.7144	0.0212	ω_2*
$\beta_{x,3}$	0.0107	-0.0824	ω_2*
$\beta_{x,4}$	-3.6341	-4.3273	ω_2
$\beta_{x,5}$	-2.6357	-3.3288	ω_2

* indica i parametri classificati diversamente rispetto al primo metodo.

1.2.2 Struttura probabilistica incognita

In questo caso non si hanno informazioni e quindi dal training set si cerca di stimare l'intera descrizione descrizione della funzione $p(\beta \mid \omega_1)$. Si consideri ancora la classe 1: il vettore dei parametri β è un vettore aleatorio di dimensione N la cui descrizione statistica $p(\beta)$ rappresenta l'oggetto della stima. Si omette per semplicità la classe sottointendendo ω_1.
Sia R una regione dello spazio R^N; si può allora calcolare la probabilità P_R che β appartenga a questa regione :

$$P_R = P[\beta \in R] = \int_R p(\beta) d\beta$$

Se $p(\beta)$ è una funzione continua, P_R può essere intesa come il valore medio di $p(\beta)$ nell' intervallo R moltiplicato per il volume della regione e quindi:

$$P_R = \int_R p(\beta) d\beta \cong \overline{p}_{R'} \quad V \Rightarrow \overline{p}_{R'} \cong \frac{P_R}{V}$$

Quindi per ottenere la stima di $p(\beta)$ nella regione R è necessario calcolare una stima di P_R. Per far ciò si utilizza il T.S. e si fanno le seguenti considerazioni. Il T.S. può essere visto come risultato di un esperimento in cui siano misurati m campioni indipendenti $\beta_1, \beta_2, \ldots, \beta_{Q1}$ di un vettore di variabili aleatorie che ha $p(\beta)$ come densità di probalità. La probabilità che k di questi appartengano alla regione dello spazio dei parametri R ha una distribuzione binomiale:

$$\Pr ob\left[\frac{k\_elementi \in R}{(Q_1 - k)\_elementi \notin R}\right] = \binom{Q_1}{k} (P_R)^k (1 - P_R)^{Q_1 - k}$$

La variabile aleatoria k ha media $E[k]=Q_1P_R$ e quindi:

$$P_R = \frac{E[k]}{Q_1}$$

In teoria , allora per valutare P_R, si dovrebbe valutare E[k], quindi ripetere l' esperimento molte volte. In pratica si ha a disposizione un solo esperimento rappresentato dal T.S. Contando allora quanti elementi del T.S. appartengono a ω_1e cadono nella regione R, indicando con k tale numero, si può solo ottenere una stima (istogramma):

$$\hat{P}_R = \frac{k}{Q_1}$$

Passando alla stima della funzione densità nella regione R si ottiene:

$$\hat{p}_R(\beta) = \frac{k}{Q_1 V}$$

Ci sono alcuni problemi teorici e pratici sui quali è opportuno soffermarsi. Se si fissa il volume della regione R e si hanno a disposizione moltissimi campioni, il rapporto k/Q converge in probabilità come desiderato, ma solo verso il valore medio (nella regione R) di p(β). Se si vuole ottenere una stima del valore puntuale di p(β) si deve scegliere il valore di V il più possibile tendente a zero.
Tuttavia se il numero di campioni disponibili è limitato, come avviene nella pratica, la regione potrebbe non comprendere alcun campione ($\hat{p}(\beta) = 0$) rendendo la stima inutile. Quindi il volume V non può essere scelto arbitrariamente piccolo. E' necessario allora prevedere comunque un certo errore di stima, legato al fatto che il numero di campioni è finito; ed un certo errore di approssimazione, legato al fatto che la regione ha dimensioni finite.
Dal punto di vista teorico è interessante chiedersi se queste limitazioni possono essere superate qualora il numero di campioni sia illimitato. In tal caso per la stima della densità nel punto β, si individuano una serie di regioni R_1, R_2,...,
Detto V_Q il volume della regione R_Q , k_Q il numero dei campioni che cadono nella regione R_Q e $p_Q(\beta)$ la Q-esima stima di $p_Q(\beta)$.

Si ha:

$$\hat{p}_Q(\beta) = \frac{k_Q}{QV_Q}$$

Tre condizioni sono necessarie perché $\hat{p}_Q(\beta)$ converga a $p(\beta)$:

1. $\lim_{Q\to\infty} V_Q = 0$

2. $\lim_{Q\to\infty} k_Q = \infty$

3. $\lim_{Q\to\infty} \frac{k_Q}{Q} = 0$

Ci sono essenzialmente due modi per ottenere sequenze di regioni che soddisfano queste condizioni. Uno di questi è ridurre il volume della regione in funzione di Q, ad esempio nel seguente modo:

$$V_Q \approx \frac{1}{\sqrt{Q}}$$

Su questo principio si basa il metodo delle finestre di Parzen. Il secondo metodo è quello di fissare la regione in modo che contenga k_Q elementi, ad esempio nel seguente modo:

$$k_Q = \sqrt{Q}$$

Questo principio dà origine al metodo K-NN.

Esempio 2

Utilizzando gli stessi dati dell' esempio 1 si ripete ora la classificazione stimando le densità di probabilità con il metodo dell'istogramma. Si considerino nello spazio dei parametri 36 regioni di forma rettangolare (in figura delimitate dalla griglia), ciascuna delle quali verrà identificata dalle coordinate cartesiane del vertice in basso a sinistra. Indicando con Q_1 e Q_2 rispettivamente il numero totale di campioni appartenenti alle classi ω_1 e ω_2 e con k in numero di campioni appartenenti alla regione considerata si ottengono le seguenti stime:

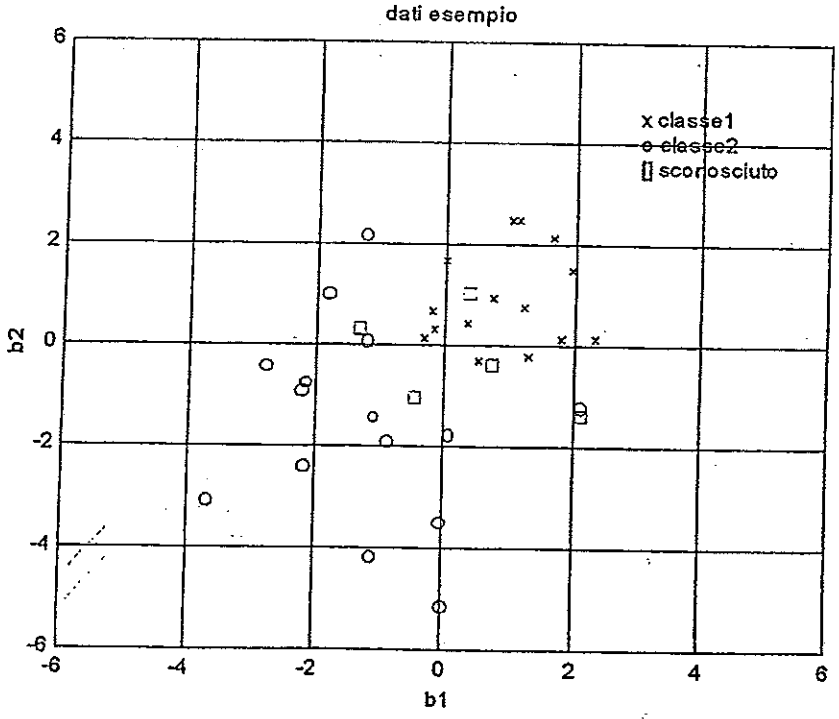


Fig. e2.1 Rappresentazione dei dati sul piano β_1, β_2.

$$\hat{p}(\beta \in R_{0,0}|\omega_1) = \frac{k}{Q_1} = \frac{5}{15} \qquad \hat{p}(\beta \in R_{0,0}|\omega_2) = \frac{k}{Q_2} = \frac{0}{15}$$

$$\hat{p}(\beta \in R_{-1,0}|\omega_1) = \frac{k}{Q_1} = \frac{3}{15} \qquad \hat{p}(\beta \in R_{-1,0}|\omega_2) = \frac{k}{Q_2} = \frac{2}{15}$$

ecc. ...

Supponiamo ora di voler classificare l'elemento $\beta_{x,1}$ =[-0.4650 -1.0226] che appartiene alla regione $R_{-1,-1}$. In tale regione si ottiene la seguente stima della densità di probabilità di $p(\beta \mid \omega_i)$:

$$\hat{p}(\beta \quad |\omega_1) = \frac{k}{Q_1 \cdot V} = 0 \qquad \hat{p}(\beta \quad |\omega_2) = \frac{k}{Q_2 \cdot V} = \frac{2}{60}$$

Essendo per ipotesi $P(\omega_1) = P(\omega_2)$ possiamo considerare $g_i(\beta) = \hat{p}(\beta \mid \omega_i)$ da cui si ottiene

$\beta_{x,1} \rightarrow \omega_2$

dove il simbolo $\rightarrow$ indica assegnazione alla classe.

La tabella completa è la seguente:

	$g_1(\beta)$	$g_2(\beta)$	classe
$\beta_{x,1}$	0	2/60	ω_2
$\beta_{x,2}$	5/60	0	ω_1
$\beta_{x,3}$	2/60	1/60	ω_1
$\beta_{x,4}$	0	1/60	ω_2
$\beta_{x,5}$	4/60	2/60	ω_1*

* indica parametro classificato diversamente dall'esempio 1

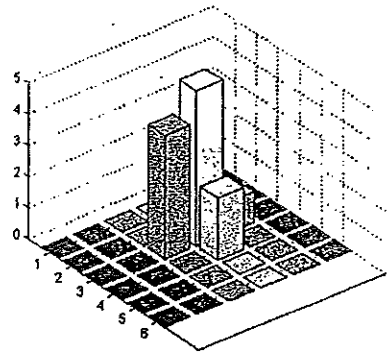


Fig. e2.2.a istogramma della classe ω_1

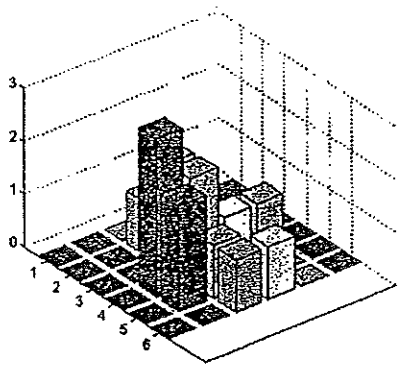


Fig. e2.2.b istogramma della classe ω_2

1.2.3 Metodo di Parzen

E' una generalizzazione del metodo visto in precedenza. Data l'osservazione $\beta \in \omega_1$ è possibile asserire che $p(\beta \mid \omega_1)$ è non nulla e assumendo $p(\beta \mid \omega_1)$ continua è possibile dedurre che sia non nulla pure nelle immediate vicinanze. E' naturale quindi esprimere questa informazione con una funzione che assuma valore massimo in corrispondenza dell'osservazione che prende il nome di kernel. Fra i kernel più comunemente usati si ricordano il kernel gaussiano, triangolare, ipersferico, ovunque centrati in $\beta_i^{(k)}$, in generale esprimibile con la funzione:

$$\varphi(\beta, \beta_i^{(k)})$$

che soddisfa alla condizione

$$\int_{R^N} \varphi(\beta, \beta_i^{(k)}) d\beta = 1$$

Ad esempio, nel caso dim $\beta = 1$, il kernel rettangolare ha la seguente espressione

$$\varphi(\beta, \beta_i^{(k)}) = \begin{cases} \dfrac{1}{h} & \left|\beta - \beta_i^{(k)}\right| \le \dfrac{h}{2} \\ 0 & \text{altrove} \end{cases}$$

dove h è un parametro che deve essere specificato. Ogni kernel infatti ha un grado di libertà h che sempre nel caso di dim $\beta = 1$ è, per kernel triangolari la base del triangolo, per kernel gaussiani la varianza della gaussiana, ecc..

Ogni kernel soddisfa alla condizione:

$$\int_{R^N} \varphi(\beta, \beta_i^{(k)}) d\beta = 1$$

La stima della funzione densità di probabilità è allora data da:

$$\hat{p}(\beta|\omega_1) = \frac{1}{Q_i} \sum_{i=1}^{Q_i} \varphi[(\beta, \beta_i^{(1)})]$$

In ogni caso la stima dipende dalla scelta della forma del kernel e data una certa scelta, dal parametro h. Scegliere un valore elevato di h significa avere una stima a bassa risoluzione, mentre un valore troppo basso implica un'alta variabilità statistica. Si deve cercare un giusto compromesso in relazione anche al numero di campioni e quindi in sostanza alla dimensione del training set.

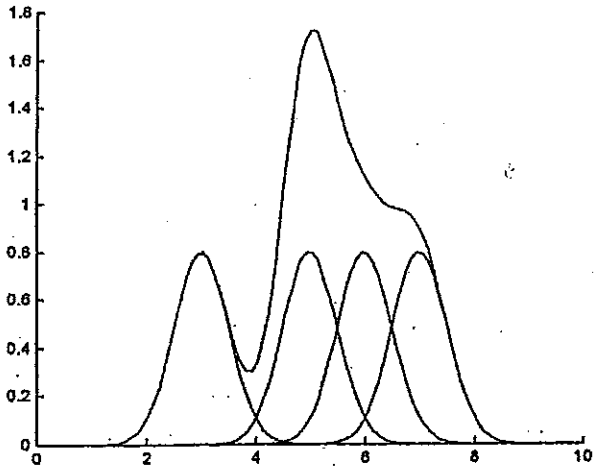


Fig. 2.3.1 Esempio di costruzione della funzione di densità col metodo del kernel gaussiano

Esempio 3

Si consideri l'esempio 2: si ripete qui la classificazione utizzando il metodo di Parzen.

Si divide lo spazio dei parametri in 36 regioni delimitate dalla griglia quindi si definisce il seguente kernel bidimensionale:

$$\varphi(\beta,\beta_{i,j}^{(k)})=\begin{cases}1 & \text{se } \beta \in R(\beta_{i,j}^{(k)}) \\ \frac{1}{2} & \text{se } \beta \in R(\beta_{l,m}^{(k)}) \text{ limitrofa} \\ 0 & \text{altrimenti}\end{cases}$$

Definendo :

$$k_{R_{i,j}}^{(1)}=\sum_{n=1}^{Q_1}\varphi(\beta,\beta_{i,j}^{(1)}) \qquad k_{R_{i,j}}^{(2)}=\sum_{n=1}^{Q_2}\varphi(\beta,\beta_{i,j}^{(2)})$$

otteniamo per esempio

$$k_{R_{0,0}}^{(1)}=10.5 \qquad k_{R_{0,0}}^{(2)}=3.5$$

$$\hat{P}^{(1)}{}_{R_{0,0}}=\frac{k_{R_{0,0}}}{Q_1}=\frac{10.5}{15} \qquad \hat{P}^{(2)}{}_{R_{0,0}}=\frac{k_{R_{0,0}}}{Q_2}=\frac{3.5}{15}$$

ecc.

Si riportano i risultati in tabella:

	$\hat{P}_R^{(1)}$	$\hat{P}_R^{(2)}$	classe
$\beta_{x,1}$	2.5/15	6.5/15	ω_2
$\beta_{x,2}$	10.5/15	3.5/15	ω_1
$\beta_{x,3}$	7/15	4/15	ω_1
$\beta_{x,4}$	4	1.5/15	ω_1*
$\beta_{x,5}$	7.5/15	6/15	ω_1

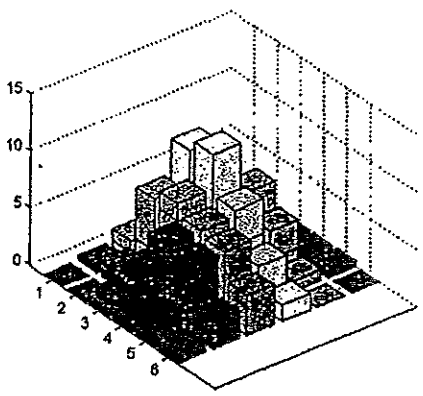


Fig e3.1.a istogramma della classe ω_1 ottenuto col metodo di Parzean

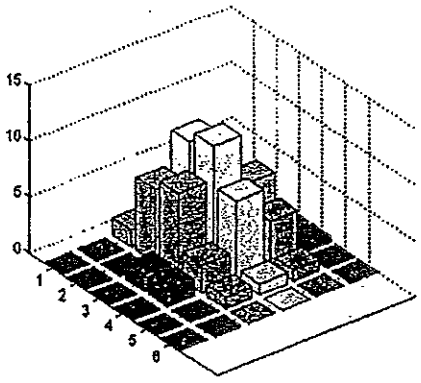


Fig e3.1b istogramma della classe ω_2 ottenuto col metodo di Parzen

Si osservi come viene classificato diversamente rispetto all'esempio 2 li termine $\beta_{x,4}$. Questa nuova classificazione appare più verosimile.

1.2.4 Metodo K-NN

Si è precedentemente visto che, nel metodo di Parzen, la scelta del volume delle celle è un punto delicato che incide sensibilmente sull' andamento della stima di p(β). Un approccio alternativo consiste nell adattare il volume della cella in modo che al suo interno cada comunque un certo numero k di campioni del training set ad esempio $k = \sqrt{Q}$.

Più precisamente dato un evento β da classificare, si dimensiona la cella attorno al campione in modo che k elementi del training set cadano all'interno di essa. La stima di p(β) è ancora data dalla solita formula $\hat{p}(\beta) = \frac{k/Q}{V}$ dove k e Q sono costanti mentre V dipende da β

Si può ora pensare di applicare delle considerazioni simili a queste per la stima di p(β) e mettere poi a punto le funzioni discriminanti del classificatore errore minimo. Per far ciò, per un particolare valore β da classificare si consideri una regione che contenga un numero k complessivo di elementi del T.S., indipendentemente dalla classe di appartenenza, e si indichi al solito con V il volume della regione.
Si dividano ora i k elementi del training set per classe di appartenenza:

k_1 numero che appartengono alla classe ω_1
k_2 numero che appartengono alla classe ω_2

.

.

.

k_M numero che appartengono alla classe ω_M

Vale ovviamente $\sum_{i=1}^{M} k_i = k$

Sia il training set composto nel seguente modo:

Q_1 elementi della classe ω_1

Q_2 elementi della classe ω_2

.

.

.

Q_M elementi della classe ω_M

Si ottengono le seguenti stime locali delle probabilità a priori:

$$\hat{p}(\beta|\omega_i) = \frac{k_i}{Q_i V} \quad \text{con i} = 1,...,M$$

Per quanto riguarda la stima della probabilità a priori delle classi, supponendo che il training set rifletta la distribuzione di elementi nelle varie classi vale:

$$p(\omega_i) = \frac{Q_i}{Q_{tot}} \quad \text{con i} = 1,...,M$$

Applicando il classificatore ad errore minimo scegliendo come funzione discriminante

$$g_i(\beta) = p(\beta|\omega_i)P(\omega_i)$$

e sostituendo nell'equazione le stime si ottiene:

$$g_i(\beta) = \frac{k_i}{Q_i V} \cdot \frac{Q_i}{Q_{tot}} = \frac{k_i}{V Q_{tot}}$$

Eliminando i termini che non discriminano la funzione discriminante diventa:

$$g_i(\beta) = k_i$$

e la regola di decisione sarà la seguente:

$$\beta \rightarrow \omega_i \quad \text{se} \quad k_i > k_j \quad \text{per} \quad \forall j = 1,...,M \quad \text{con} \quad j \neq i.$$

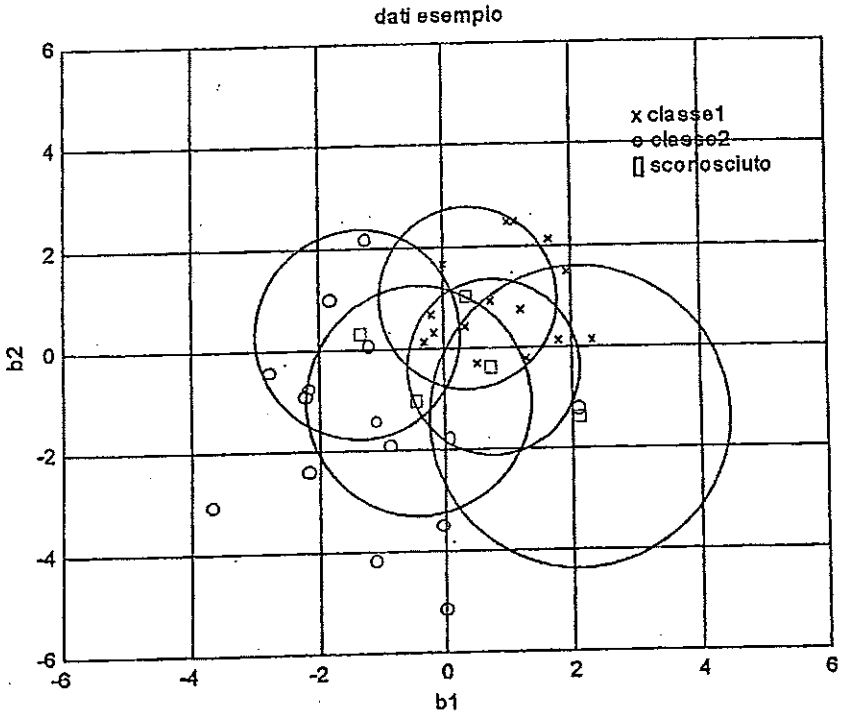
Quindi si assegna β alla classe che presenta più campioni all'interno della regione.

Mentre con i metodi visti in precedenza il training set serviva solo nella fase di progetto del classificatore, con questo metodo lo si usa in continuazione per il fatto di dover calcolare per ogni punto da classificare i valori di $k_1, k_2, ..., k_M$.

Questo costringe a tenere in memoria l'intero training set. Quindi il metodo K-NN è vantaggioso soprattutto se si prevede in futuro la possibilità di aggiornare il training set in quanto basterà aggiungere i nuovi dati ai vecchi senza modificare l' algoritmo di decisione.

Esempio 4

Utilizzando gli stessi dati dell'esempio 1 si ripete la classificazione dei termini $\beta_{x,x}$ utilizzando il metodo K-NN. Si dimensiona la cella, che si è il questo caso assunta circolare, in modo tale che 10 campioni del training set, indipendentemente dalla classe di appartenenza, cadano all'interno di essa. Assumendo in questo caso la funzione discriminante la semplice espressione $g_i(\beta) = k_i$ si ricava facilmente il risultato riportato in tabella.



	$g_1(\beta)$	$g_2(\beta)$	classe
$\beta_{x,1}$	5	5	?
$\beta_{x,2}$	10	0	ω_1
$\beta_{x,3}$	8	2	ω_1
$\beta_{x,4}$	8	2	ω_1
$\beta_{x,5}$	3	7	ω_2

Selezione di parametri

Il problema della messa a punto di un classificatore bayesiano diventa più complesso, e richiede training set di dimensione elevata, quanto più elevato è il numero di parametri.

Conviene lavorare con pochi parametri, che siano però in grado di discriminare tra le classi

E' difficile stabilire a priori quali parametri siano discriminanti, per cui generalmente si estraggono molti parametri, anche tra loro dipendenti, e poi si lascia alla fase della selezione il compito di stabilire quali di essi (o quali loro combinazioni) siano in grado di discriminare tra le classi.

Selezione:
dal vettore di parametri $\beta = [\beta_1 \ldots \beta_N]^T$

ad un nuovo vettore di parametri $\mathbf{y} = [y_1 \ldots y_D]^T$

con $N=\dim(\beta)<D=\dim(\mathbf{y})$

Selezione diretta : $y_i = \beta_j$

Trasformazioni : $y_i = F(\beta_1 \ldots \beta_N)$

Trasformazione lineare di Fisher

E' una trasformazione lineare dal vettore di parametri
$\beta = [\beta_1 \ldots \beta_N]^T$
al vettore di parametri $\mathbf{y} = [y_1 \ldots y_D]^T$
con D=M-1 M=numero di classi
La trasformazione è scelta in modo che y sia la scelta "ottima" rispetto ad un criterio di separabilità tra le classi

Caso 1 M=2
$y = \rho_1 \beta_1 + \rho_2 \beta_2 + \ldots \rho_N \beta_N = \rho^T \beta$

y è uno scalare che rappresenta la proiezione di β su una retta che ha $[\rho_1 \rho_2 \ldots \rho_N]^T$ come versore.

Definiamo:

Classe	n° elementi	v.medio	dispersione intra-
ω_1	Q_1	μ_1	W_1
ω_2	Q_2	μ_2	W_2

dispersione inter = $\mathbf{B} = (\mu_1 - \mu_2)(\mu_1 - \mu_2)^T$

Per le proiezioni

Classe	n° elementi	v.medio	dispersione intra-
ω_1	Q_1	$\rho^T \mu_1$	$\rho^T W_1 \rho$
ω_2	Q_2	$\rho^T \mu_1$	$\rho^T W_1 \rho$

dispersione inter = $b = \rho^T \mathbf{B} \rho$

Indice di separabilità

$$J = \rho^T B \rho / \rho^T (W_1 + W_2) \rho$$

J assume valore elevato se le medie delle due classi sono "lontane" tra loro (numeratore è elevato) e i valori nelle due classi sono concentrate attorno ai valori medi (denominatore è basso).

Problema: trovare ρ che rende massimo J

Soluzione : ρ è autovettore della matrice $W^{-1}B$
quindi
$W^{-1}B\,\rho = \lambda\rho$
ovvero $W^{-1}(\mu_1 - \mu_2)(\mu_1 - \mu_2)^T\rho = \lambda\rho$
quindi ρ nella direzione di $W^{-1}(\mu_1 - \mu_2)$
In quanto $(\mu_1 - \mu_2)^T\rho$ è uno scalare
di ρ interessa la direzione e non il modulo!

Si può generalizzare al caso di M classi

$$y = T^T\beta$$

$J = \det(T^TBT) / \det(T^T (W_1 + .. W_M) T)$
oppure $J = tr(T^TBT) / tr(T^T (W_1 + .. W_M) T)$
Problema: trovare la matrice T che rende massimo J
Soluzione : nel primo caso, T ha come colonne gli autovettori della matrice $W^{-1}B$

Trasformazione K-L

E' una trasformazione lineare dal vettore di parametri
$\beta = [\beta_1 . . . . \beta_N]^T$

al vettore di parametri $y = [y_1 . . . . y_N]^T$

con $y = T\beta$

e T matrice che ha come righe vettori ortonormali Φ, per cui $T^T = T^{-1}$ ovvero

$$\beta = T^{-1} y = T^T y = \Phi_1 y_1 + \Phi_2 y_2 . . . + \Phi_N y_N$$

$y_1 = \Phi_1^T \beta$
. . . . .
$y_N = \Phi_N^T \beta$

Problema : trovare la trasformazione T in modo che, definito con β_d l'approssimazione ottenuta con i primi con i primi D termini :

$$\beta_d = \Phi_1 y_1 + \Phi_2 y_2 . . . + \Phi_D y_D + \Phi_{D+1} b_{D+1} . . . + \Phi_{N\,1} b_N$$

($b_{D+1} . . . b_N$ sono delle costanti che non dipendono da β)

sia minimo l'errore quadratico medio di approssimazione

$$\text{Errore} = E[(\beta - \beta_d)^2]$$

Soluzione:
Indicata con Σ_β la matrice di varianza-covarianza di β e con λ_1 λ_2 .. λ_N i suoi autovalori ordinati da quello di valore più elevato via via a quelli di valore più basso $\lambda_1 > \lambda_2 > .. \lambda_N$.

Φ_1 Φ_2 .. Φ_N sono gli autovettori corrispondenti

Si può dimostrare che:

Errore = λ_{D+1} + .. λ_N

$\Sigma_y = T\,\Sigma_\beta\,T^{-1} = \text{diag}\,(\lambda_1\ \lambda_2\ ..\ \lambda_N)$ Le componenti principali sono tra loro indipendenti!!

$Tr(\Sigma_y) = tr\,(\Sigma_\beta)$

Pertanto :
I° componente principale:

$y_1 = \Phi_1{}^T\beta$ è lo scalare che approssima al meglio β
ha varianza pari a λ_1
spiega $\lambda_1 /(\lambda_1+\lambda_2$ .. $+\lambda_N)$ di varianza totale

I° e 2°componente principale:

$y_1 = \Phi_1{}^T\beta$
$y_2 = \Phi_2{}^T\beta$ è il vettore di dim 2 che approssima l meglio β
ha varianza pari a $\lambda_1 +\lambda_2$
spiega $\lambda_1 +\lambda_2 /(\lambda_1+\lambda_2$ .. $+\lambda_N)$ di varianza totale
Ecc..

Selezione diretta: ricerca esaustiva

La tecnica della *ricerca esaustiva* prevede di :

1. considerare le possibili combinazioni di D parametri presi a partire dagli N originari
2. calcolare per ciascuno di essi un indice legato al loro potere discriminante, pari all'accuratezza (=1-probabilità di errore) del classificatore messo a punto sulla combinazione di parametri
3. selezionare la combinazione in corrispondeza della quale l'indice assume valore massimo

Risulta evidente che se si hanno a disposizione molti parametri, il numero di combinazioni da confrontare può risultare elevatissimo, ad es. con N=40, D=10, tale numero è 8.477×10^7). La messa a punto di un così alto numero di classificatori può risultare assai onerosa da un punto di vista computazionale.

Allora si può semplificare il punto 1 (considerare solo alcune combinazioni) e/o l punto 2 (considerare un indice più semplice da calcolare)

Per semplificare il punto 2:

si possono considerare degli indici di separabilità, ad es. quello definito per il classificatore di Fisher:

$J = \det(T^T B T) / \det(T^T (W_1 + \ldots W_M) T)$

dove T rappresenta ora la matrice che consente di passare da β a **y**:

$\mathbf{y} = T^T \beta$

NB. T^T è una matrice che contiene solo 1 o 0)

Per semplificare il punto 2: Ricerca sequenziale in avanti

1. Si prende ciascuno degli N parametri a disposizione e li si confronta secondo il criterio scelto.
2. La variabile migliore viene presa come componente fissa delle N-1 coppie realizzabili con i parametri restanti. Tali coppie vengono confrontate secondo lo stesso criterio.
3. La coppia migliore diventa componente fissa delle N-2 terne realizzate con i parametri restanti. Si confronta tra loro tali terne e si sceglie la migliore.
4. Si ripete il procedimento fino a quando si raggiunde il numero D di parametri, oppure se questo non è stato prefissato fino a quando le prestazioni ottenute con tutti gli insiemi di (N-k) elementi non migliorano rispetto quelle ottenute con un particolare insieme di (N-k-1) elementi.

Il numero di combinazioni risulta:

$$N + (N-1) + (N-2) + \ldots + (N-D+1)$$

Con N=40 e D=10 risultano 355 combinazioni. Confrontando questo numero con quello ottenuto con la tecnica esaustiva risulta chiaro il vantaggio in termini di tempo.
Va sottolineato però come questo sia un metodo sub-ottimo, cioè non garantisce di trovare la scelta migliore; esso rappresenta però un buon compromesso tra tempo computazionale e prestazioni ottenibili.

1.3 Validazione del classificatore

In sede di progetto, prima di passare ad un suo utilizzo, è molto importante validare il classificatore, cioè avere una misura della sue prestazioni attraverso una stima dell'errore di classificazione. Questo risulta molto utile anche per comparare le prestazioni dei vari classificatori.
Si può pensare di calcolare l'errore di classificazione in modo analitico, ad esempio esprimendo l'errore come visto a proposito del classificatore a errore minimo, ma questo può risultare difficoltoso. Inoltre si conoscono delle stime delle grandezze che caratterizzano la struttura probabilistica del problema, e si stima la P_{errore} senza mai sottoporre il classificatore ad una validazione indipendente, su campioni che non stati usati per mettere a punto il classificatore stesso. La situazione ottimale è quella di poter disporre di due training set: il primo viene utilizzati per mettere a punto il classificatore, mentre il secondo per stimare l'errore di classificazione. E' possibile successivamente calcolare la percentuale di classificazioni corrette e costruire delle tabelle da utilizzare per stimare la bontà del classificare. Per valutare la bontà di tale stima si esegue il test con n elementie se come risultato si ottiene un numero di classificazioni errate pari a k, la stima della probabilità di errore del classificatore è data dalla seguente equazione:

$$\hat{P}(\text{errore}) = \frac{k}{n}$$

tanto più vicina al valore vero quanto $n \to \infty$. Per dare una descrizione statistica a tale stima, supponiamo che la vera probabilità d'errore sia P(errore) = p. Allora la probabilità di allocazione corretta del campione nella classe di appartenenza è data dall'espressione:

$$P(\text{all. corretta}) = 1-p$$

La probabilità che in un insieme di Q elementi, k siano classificati in modo errato e Q-k siano classificati correttamente è descritta dalla seguente distribuzione binomiale:

$$P\begin{bmatrix} \text{k errati} \\ \text{(Q - k) corretti} \end{bmatrix} = \binom{Q}{k} p^k (1-p)^{n-k}$$

da cui

$$\frac{E[k]}{Q} = p$$

e quindi

$$\hat{p} = \frac{k}{Q} \quad \text{(la stima migliora al crescere di Q)}$$

Si può quindi usare questa descrizione statistica per costruire a partire dal nomogramma di Fig. 2.5 gli intervalli di confidenza delle stime. Ad esempio, con i seguenti valori: Q = 50, k = 10, si ottiene $\hat{P}(\text{errore}) = 20\ \%$. Questa è però una singola misura su di una distribuzione binomiale.

Ora, consultando il grafico di Fig. 2.5, con un livello di confidenza del 95% la vera probabilità di errore cade nell'intervallo 10% ÷ 35%

$$P[\ 10\% < P(\text{errore}) < 35\%\ ] > 95\%$$

Con n=250, k = 50, si ottiene sempre $\hat{P}(\text{errore}) = 20\ \%$, ma considerando lo stesso livello di confidenza si ha:

$$P[12\% < P(\text{errore}) < 25\%\ ] > 95\%$$

che evidenzia il fatto che più numeroso è il training set utilizzato per testare il classificatore, più l'intervallo di confidenza si stringe e quindi la stima dell'errore di classificazione è buona.
Avere a disposizione due training set significa avere un trainig set molto grande da dividere in due parti, ma come fare tale divisione nel caso in cui il T.S. non sia particolarmente numeroso? Se si decide di mettere più dati nel primo, si ottiene un buon classificatore, ma la stima dell'errore di

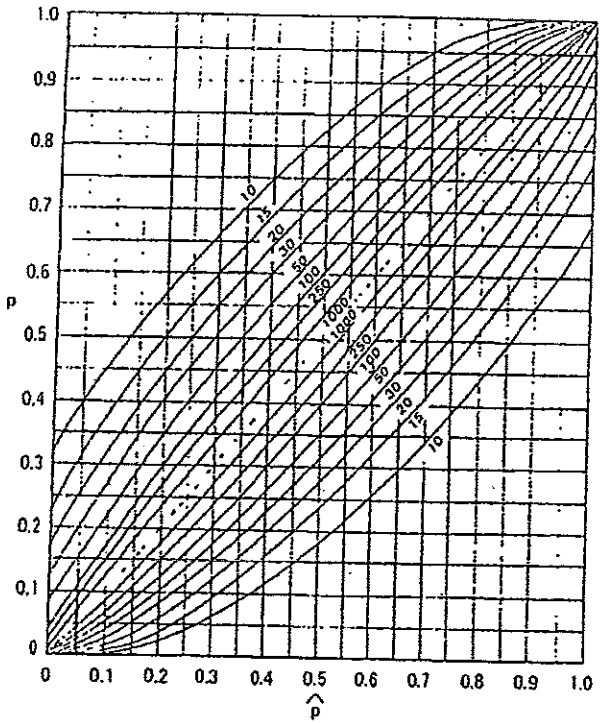


Fig. 2.5. - Livelli di confidenza per la stima della probabilità d' errore.

classificazione risulta poco realistica. Viceversa, se si predilige il secondo si riesce ad avere una buona stima dell'errore, ma su di un classificatore poco attendibile.

Una tecnica spesso usata nel caso di T.S. non particolarmente numeroso è basata sul metodo *leave one out* che opera la stima nel modo seguente: dato un training set composto dagli elementi

$$\beta_1, \beta_2, \ldots, \beta_Q,$$

Si consideri un primo training set composto dagli elementi $\beta_2, \ldots, \beta_Q$ con cui si mette a punto un classificatore e si usa l'elemento β_1 per testarlo. Fatto questo, si ripete la procedura assegnando al primo training set gli elementi $\beta_1, \beta_3, \ldots, \beta_Q$ usando l'elemento β_2 per il test. Si ripete quindi per Q volte e si contano quanti elementi sono stati classificati in modo errato e quindi al solito, se k è il numero:

$$\hat{p} = \frac{k}{Q}$$

Si riesce ad ottenere in questo modo una stima basata su Q test che hanno usato elementi diversi da quelli che sono stati utilizzati per progettare il classificatore.