Tecniche di classificazione della statistica multivariata e dell'analisi dei *cluster*: un *tutorial*

Giulio Agostini

30 marzo 2000

Sommario

In questa relazione, dopo aver esposto formalmente la teoria della classificazione e riepilogato alcuni concetti fondamentali della statistica multivariata, vengono esaminati e discussi i principali metodi di riconoscimento automatico basati su tecniche statistiche. Si illustra nel seguito la teoria dell'analisi dei *cluster* e i più noti algoritmi di raggruppamento, piatti e gerarchici.

Il problema della classificazione automatica, o *pattern recognition*, trova i suoi prodromi nei lavori pionieristici di *statistica multivariata* di Pearson [20] di inizio secolo, sviluppati da Fisher [7], Mahalanobis [15] e Hotelling [13] negli anni 30. In seguito, parallelamente all'approccio statistico classico (*analisi discriminante*, o *discriminant analysis*—una branca della statistica multivariata), nascono e si sviluppano le tecniche di classificazione non-parametrica. L'obiettivo è quello di identificare la natura dell'oggetto in esame all'interno di una rosa di classi possibili, siano esse note a priori o meno. Perché il processo sia automatizzabile, è necessario effettuare una serie di misurazioni di caratteristiche significative sui campioni (*feature extraction*), in modo da avere a disposizione dei vettori numerici, visti come realizzazione di un vettore casuale, facilmente manipolabili dal calcolatore, che fornisce in uscita la classe di appartenenza stimata (figura 1). La classificazione è detta assistita (*supervised*) se vengono forniti al calcolatore esempi di osservazioni di cui sia nota la classe. L'analisi dei *cluster* (*cluster analysis*, *analisi dei grappoli*) e

le reti neurali a mappa auto-organittante (*Self-Organizing Map*, SOM) sono tipici esempi di tecniche di classificazione non-assistita.

La classificazione automatica è strettamente correlata, e per certi versi sovrapposta, all'apprendimento automatico [16]. Esso presenta infatti una simile tassonomia delle tecniche, che possono essere parametriche o non parametriche, mentre l'apprendimento può essere assistito e non-assistito.

Si espongono in questo capitolo i fondamenti della teoria della classificazione, e vengono illustrati in dettaglio l'approccio statistico al problema (sezione 2) e le principali tecniche di analisi dei *cluster* (sezione 3).

Si assume che il lettore conosca i fondamenti della statistica e dell'algebra lineare (si vedano, ad esempio, [19] e [3]). [9] fornisce una gradevole introduzione alla statistica multivariata e alle sue applicazioni, con un encomiabile ricchezza di esempi, e una serie di istruttivi algoritmi disponibili via `ftp`. Un'ottima monografia sull'analisi discriminante, ricca di risultati e spunti stimolanti (incluse le regole di *k-nearest neighbor* e le *kernel rules*), contenente una imponente bibliografia di oltre 1200 titoli, è [18]. Per una panoramica molto tecnica e articolata sulle principali tecniche parametriche e non-parametriche esistenti e sulla classificazione automatica in generale (seppure limitata a due classi), si rimanda a [4].

L'analisi dei *cluster* è coperta da una vecchia, ma autorevole monografia [11], e da due testi di impostazione applicativa, completi di algoritmi [21, 22].

1 Formalizzazione del problema

In questo paragrafo si fornisce una definizione rigorosa del problema della classificazione. La notazione è mutuata principalmente da [9].

Una *osservazione* (o *oggetto*, o *caso*, o *entità*) è determinata da un vettore p-dimensionale di variabili casuali, che rappresenta le misurazioni effettuate sull'oggetto da classificare. La natura (nota o incognita) di un'osservazione è detta *classe*. Nella presente trattazione si assume che le variabili siano con-

Figura 1. Schema a blocchi del processo di classificazione automatica.

tinue, o comunque misurate su una metrica razionale. Per una efficace esposizione dei quattro tipi di variabili esistenti (nominali, ordinali, intervallari e razionali), si veda [22].

Le osservazioni già classificate, se esistono, sono raccolte in k matrici di dati $\mathbf{D}_j$, che hanno tante righe (N_j) quante osservazioni appartenenti a quella classe, e tante colonne (p) quante variabili. Si definisce $N \stackrel{\text{def}}{=} \sum_{j=1}^{k} N_j$ il numero totale di osservazioni.

Data una generica osservazione $\mathbf{y}$ di natura incognita ed un numero k di classi si definisce *classificatore* (o *regola di classificazione*) una funzione $g(\mathbf{y}) : \mathbb{R}^p \to \{1, \ldots, k\}$. Se $\mathbf{y}$ appartiene alla classe j e $g(\mathbf{y}) \neq j$ si dice che il classificatore g commette un *errore* nella classificazione di $\mathbf{y}$.

Sia X una variabile casuale a valori discreti $\{1, \ldots, k\}$ che indichi l'appartenenza ad una classe. Si definiscono le *probabilità a priori* di appartenenza alla j-esima classe le quantità

$$\pi_j \stackrel{\text{def}}{=} \Pr[X = j] \qquad 1 \leq j \leq k. \tag{1}$$

Naturalmente $\sum_{j=1}^{k} \pi_j = 1$. Sia $\mathbf{Y}$ un vettore casuale continuo di dimensione p, e sia

$$X_g \stackrel{\text{def}}{=} g(\mathbf{Y}) \tag{2}$$

la variabile casuale che rappresenta la classificazione di g per $\mathbf{Y}$. Si definisce inoltre la *probabilità* o *tasso di errore* di un classificatore

$$\gamma_g \stackrel{\text{def}}{=} \Pr[X_g \neq X]. \tag{3}$$

Si definisce *classificatore ottimo*, *classificatore a minimo tasso di errore*, o *classificatore di Bayes* una funzione

$$g^*(\cdot) \stackrel{\text{def}}{=} \operatorname*{arg\,min}_{g:\mathbb{R}^p \to \{1,\ldots,k\}} \gamma_g, \tag{4}$$

ovvero un classificatore che renda minima la probabilità di errore per la generica osservazione $\mathbf{Y}$.

Ogni classificatore $g(\cdot)$ ripartisce lo spazio campionario $\mathbb{R}^p$ in *regioni di classificazione* $C_1, \ldots, C_k$, tali per cui

$$g(\mathbf{y}) = j \Leftrightarrow \mathbf{y} \in C_j \qquad 1 \leq j \leq k.$$

Un classificatore può essere espresso da un insieme di k funzioni

$$g_i : \mathbb{R}^p \to \{1, \ldots, p\} \qquad 1 \leq i \leq k$$

scelte in modo tale che

$$g(\mathbf{y}) = j \quad \Rightarrow \quad g_j(\mathbf{y}) > g_i(\mathbf{y}) \qquad \mathbf{y} \in \mathbb{R}^p, \quad j \neq i \in \{1, \ldots, k\}. \tag{5}$$

Questo consente di effettuare la decisione basandosi sul calcolo delle k funzioni ed attribuendo all'osservazione in esame la classe per cui la $g_i(\cdot)$ relativa assume in quel punto il valore massimo rispetto alle altre. Le $g_i(\cdot)$ vengono dette *funzioni discriminanti*. Si osservi che non sono univocamente definite per un determinato classificatore, dal momento che, dato un insieme $\{g_1(\cdot), \ldots, g_k(\cdot)\}$ ed una funzione monotona crescente $f(\cdot)$, la (5) continua a valere anche per l'insieme $\{f(g_1(\cdot)), \ldots, f(g_k(\cdot))\}$. Pur dando luogo alle stesse regioni di classificazione, le funzioni trasformate possono essere più facili da calcolare. Questa proprietà viene sfruttata ad esempio nella dimostrazione del risultato della sezione 2.4, in cui si fa uso dell'addizione di costanti positive, la moltiplicazione per costanti strettamente positive e la funzione logaritmo.

Sia $f_j(\cdot)$ la funzione densità di probabilità (*probability density function*, PDF) del vettore $\mathbf{Y}$ data la sua appartenenza alla classe j, o, per brevità, PDF della classe j. Facendo uso della formula di Bayes si ottiene che la probabilità che un'osservazione $\mathbf{y}$ faccia parte della classe j (*probabilità a posteriori*) è data da

$$\pi_{j\mathbf{y}} \stackrel{\text{def}}{=} \Pr[X = j | \mathbf{Y} = \mathbf{y}] = \frac{\pi_j f_j(\mathbf{y})}{\sum_{h=1}^{k} \pi_h f_h(\mathbf{y})} \qquad 1 \leq j \leq k. \tag{6}$$

La probabilità di classificare come appartenente alla classe i un'osservazione di natura j è data da

$$q_{ij} = \Pr[\mathbf{Y} \in C_i | X = j] = \int_{C_i} f_j(\mathbf{y})\, d\mathbf{y}, \tag{7}$$

e si verifica facilmente che la probabilità di errore di un classificatore è pari a

$$\gamma_g = \sum_{j=1}^{k} \pi_j (1 - q_{jj}) = 1 - \sum_{j=1}^{k} \pi_j q_{jj}. \tag{8}$$

Si dimostra che ogni classificatore $\check{g}(\cdot)$ avente come funzioni discriminanti

$$\check{g}_j(\mathbf{y}) = \pi_j f_j(\mathbf{y}) \qquad 1 \leq j \leq k, \tag{9}$$

è un classificatore ottimo[1].

[1]Dal punto di vista strettamente formale, in realtà, questo è scorretto, perché nei punti

La figura 2 esemplifica quanto detto finora in un semplice caso ($p = 1$, $k = 2$).

1.1 Fattori di costo e teoria delle decisioni

È possibile generalizzare [6, 18] la teoria suesposta tenendo conto di condizioni particolari in cui alcuni errori di classificazione danno luogo a costi diversi da altri. Si pensi al classico esempio delle diagnosi mediche, in cui classificare un paziente sano come malato è molto meno grave di classificarne uno malato come sano.

Si introduce la matrice dei costi $\mathbf{C}$, il cui generico elemento c_{ij} rappresenta il *fattore di costo* della classificazione di un'osservazione come elemento della classe i ($X_g = i$) quando è di natura j ($X = j$). Si noti che non è necessario imporre la condizione $c_{ii} = 0$. Nel paragrafo precedente si era implicitamente assunto $\mathbf{C} = 1 - \mathbf{I}$. In altre parole, si attribuiva costo 1 ad un qualsiasi errore di classificazione, e costo zero in caso di classificazione corretta.

Si definisce *rischio condizionato* $R_i(\mathbf{y})$ il valore atteso del costo di una classificazione di $\mathbf{y}$ nella classe i. Si deriva

$$R_i(\mathbf{y}) = \sum_{j=1}^{k} c_{ij} \Pr[X = j | \mathbf{Y} = \mathbf{y}] = \sum_{j=1}^{k} c_{ij} \pi_{j\mathbf{y}}. \tag{10}$$

Il *rischio totale* del classificatore è definito come

$$R_g \stackrel{\text{def}}{=} \int_{\mathbb{R}^p} R_{X_g}(\mathbf{y}) f_{\mathbf{Y}}(\mathbf{y})\, d\mathbf{y}, \tag{11}$$

in cui $f_{\mathbf{Y}}(\cdot)$ è la PDF marginale di $\mathbf{Y}$, data da

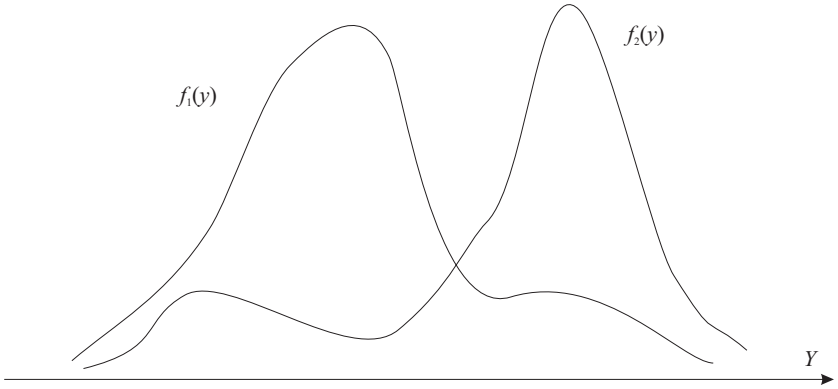
$$f_{\mathbf{Y}}(\mathbf{y}) = \sum_{j=1}^{k} \pi_j f_j(\mathbf{y}) \qquad \mathbf{y} \in \mathbb{R}^p \tag{12}$$

Si dimostra che ogni classificatore $\check{g}(\cdot)$ avente come funzioni discriminanti

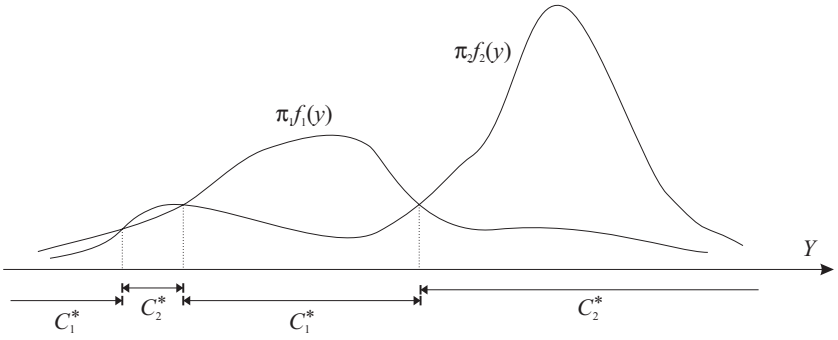
$$\check{g}_j(\mathbf{y}) = \sum_{j=1}^{k} c_{ij} \pi_j f_j(\mathbf{y}) \qquad 1 \leq j \leq k, \tag{13}$$

---

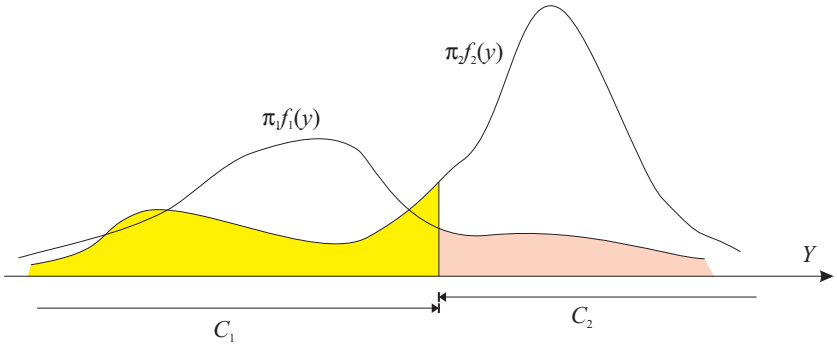
in cui due o più $\check{g}_j(\mathbf{y})$ assumono lo stesso valore, alcune disuguaglianze della (5) vengono violate. Tuttavia si preferisce non appesantire ulteriormente la notazione, e assumere che, in questi casi, la decisione è arbitraria.



(a) Funzioni di distribuzione di due variabili casuali.



(b) Funzioni discriminanti e regioni di classificazione del classificatore ottimo assumendo $\pi_1 = \frac{1}{3}$ e $\pi_2 = \frac{2}{3}$.



(c) Regioni di classificazione di un classificatore *non* ottimo e relativo tasso di errore (somma delle aree colorate). Si noti che anche per il classificatore ottimo il tasso di errore non è nullo.

Figura 2. Esempio di classificazione ottima e di errore di classificazione per $k = 2$ variabili casuali ($p = 1$).

rende minimo il rischio totale $R_{\tilde{g}}$.

Nel seguito, se non verrà esplicitato il contrario, si assumerà $\mathbf{C} = 1 - \mathbf{I}$, ricercando il classificatore ottimo introdotto nel paragrafo 1.

2 L'approccio statistico

La (9) del paragrafo precedente sembrerebbe fornire, insieme, il problema e la soluzione. Purtroppo, però, le quantità di quella equazione sono ignote, e ne è richiesta una stima a partire dai dati disponibili, forniti al calcolatore nella *fase di addestramento* (o *training*).

Per quanto riguarda le probabilità a priori π_j si profilano tre alternative:

1. $\hat{\pi}_j = \frac{1}{k}$;
2. una semplice stima basata sulla sequenza di *training*, data da $\hat{\pi}_j = \frac{N_j}{N}$;
3. una stima fornita da esperti (per esempio basandosi su popolazioni più significative della sequenza di *training*, o sulla mera esperienza).

Il problema della stima delle PDF di un vettore casuale è un problema vastissimo ed ancora oggetto di ricerca nella sua formulazione generale. In questa trattazione ci si occuperà del caso più studiato in letteratura in cui le popolazioni seguono una distribuzione multinormale. Nel seguito, perciò, a volte si assumerà tacitamente che le distribuzioni trattate obbediscono a questa legge. Per una trattazione comprendente risultati ed esempi riguardanti anche altre distribuzioni, si rimanda a [9].

L'ipotesi di multinormalità dei dati può essere suffragata da opportuni test statistici, illustrati nella sezione 2.10.

2.1 Distanza standard

La maggior parte delle tecniche di classificazione e raggruppamento analizzate in questo capitolo coinvolge un processo di misurazione. La distanza standard riveste un ruolo centrale in questo ambito, e viene quindi introdotta insieme ad altri concetti di base, come le distribuzioni di probabilità multinormali e la stima dei parametri, in questi primi paragrafi.

Un generico vettore casuale $\mathbf{Y}$ di dimensione p avente matrice di covarianza $\mathbf{\Sigma}$ induce sullo spazio campionario $\mathbb{R}^p$ la metrica

$$\|\mathbf{y}\|_{\mathbf{\Sigma}^{-1}} = \sqrt{\mathbf{y}'\mathbf{\Sigma}^{-1}\mathbf{y}}, \tag{14}$$

da cui si deriva la *distanza standard*, o *distanza di Mahalanobis*[2], definita come

$$\Delta_{\mathbf{Y}}(\mathbf{y}_1, \mathbf{y}_2) \stackrel{\text{def}}{=} \sqrt{(\mathbf{y}_1 - \mathbf{y}_2)' \, \boldsymbol{\Sigma}^{-1} \, (\mathbf{y}_1 - \mathbf{y}_2)}. \tag{15}$$

La distanza standard può essere interpretata come una generalizzazione della familiare distanza euclidea, in cui vale $\boldsymbol{\Sigma} = \mathbf{I}$. Per altri tipi di generalizzazione della metrica euclidea, si veda [22].

Di particolare interesse per l'analisi discriminante si rivela la distanza standard di un punto dalla media $\boldsymbol{\mu}$ della distribuzione in questione. La quantità $\Delta_{\mathbf{Y}}(\mathbf{y}, \boldsymbol{\mu})$ fornisce infatti una comoda misura dello scostamento di un'osservazione dagli "standard" di una determinata classe, permettendo di individuare immediatamente le osservazioni atipiche (*outlier*). La visualizzazione della distanza standard è un passo fondamentale per la comprensione dei concetti che seguiranno. Si dimostra che, per $p = 2$, i luoghi dei punti aventi distanza standard costante dalla media $\boldsymbol{\mu}$ sono delle ellissi[3] centrate attorno a $\boldsymbol{\mu}$. La figura 3 mostra i luoghi dei punti a distanza standard costante dalla media per una distribuzione bivariata. Le ellissi più esterne corrispondono a costanti più grandi. Si noti che i punti $\mathbf{d}_1$ e $\mathbf{d}_2$ hanno la medesima distanza euclidea dalla media, ma, in termini di distanza standard, $\mathbf{d}_1$ è più vicino a $\boldsymbol{\mu}$ di $\mathbf{d}_2$.

Siano λ_i gli autovalori della matrice di covarianza ordinati in modo che

$$|\lambda_i| \geq |\lambda_j| \qquad 1 \leq i < j \leq p.$$

Essi sono reali, in quanto $\boldsymbol{\Sigma}$ è reale e simmetrica. Siano $\mathbf{u}_i$ gli autovettori normalizzati associati ai λ_i. In seguito si assumerà che gli autovalori sono distinti, ipotesi ragionevole nel caso in cui si debba stimare la matrice di covarianza da dati reali. Si dimostra [1, 3] che in queste condizioni gli autovettori formano un sistema ortonormale. Il valore assoluto degli autovalori quantifica l'estensione della distribuzione nella direzione dell'autovettore corrispondente. Ad esempio, per la distribuzione trattata nella figura 3, si ottengono i seguenti risultati, facilmente interpretabili alla luce di quanto detto:

$$\lambda_1 = 10 \qquad \mathbf{u}_1 = \begin{bmatrix} 1/2 \\ \sqrt{3}/2 \end{bmatrix}, \qquad \lambda_2 = 4 \qquad \mathbf{u}_2 = \begin{bmatrix} \sqrt{3}/2 \\ -1/2 \end{bmatrix}.$$

[2] Alcuni autori definiscono distanza di Mahalanobis il quadrato della quantità indicata nella (15). In questo caso, tuttavia, essa non soddisfa più gli assiomi che caratterizzano una metrica. Si spera che il numero sfortunato capitato a questa nota a pié di pagina non generi ulteriore confusione.

[3] Per $p = 3$ sono ellissoidi, per $p > 3$ iperellissoidi.

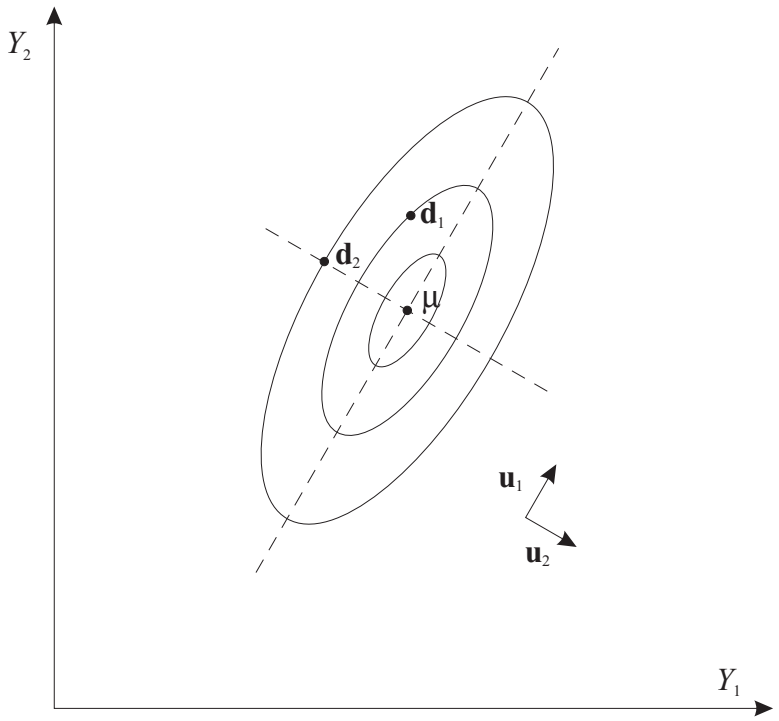


Figura 3. Luoghi di distanza standard costante dalla media $\boldsymbol{\mu}$ per una distribuzione bivariata avente matrice di covarianza $\boldsymbol{\Sigma} = \begin{bmatrix} 5{,}5 & 2{,}5981 \\ 2{,}5981 & 8{,}5 \end{bmatrix}$. $\mathbf{d}_1$ e $\mathbf{d}_2$ rappresentano due osservazioni aventi la stessa distanza euclidea dalla media. $\mathbf{u}_1$ e $\mathbf{u}_2$ sono gli autovettori della matrice di covarianza.

Un altro modo di leggere questo risultato è il seguente: l'autovettore associato all'autovalore dominante fornisce i coefficienti della combinazione lineare delle variabili che presenta la massima variabilità (a meno di una costante moltiplicativa, che è un grado di libertà nella scelta dei coefficienti). Le combinazioni lineari associate agli altri autovettori di conseguenza ordinati possiedono una variabilità via via decrescente, ma massima, sotto il vincolo di indipendenza dalle combinazioni lineari precedenti. Si potrebbe dire che l'informazione relativa alla distribuzione del vettore casuale è maggiormente concentrata nelle combinazioni lineari delle variabili studiate associate (attraverso i relativi autovettori) agli autovalori dominanti. Considerazioni di questo tipo sono alla base dell'Analisi delle Componenti Principali (*Principal Component Analysis*, PCA), che si preoccupa di sfruttare la correlazione delle variabili al fine di condensare una buona parte dell'informazione in un numero inferiore di variabili ortogonali.

Un'altra interessante proprietà della distanza standard è la sua invarianza alle trasformazioni lineari. Data cioè una matrice non singolare $\mathbf{T}$, si ha

$$\Delta_{\mathbf{Y}}(\mathbf{y}_1, \mathbf{y}_2) = \Delta_{\mathbf{Y}}(\mathbf{T}\mathbf{y}_1, \mathbf{T}\mathbf{y}_2). \tag{16}$$

Questo è un ulteriore vantaggio sulla distanza euclidea, invariante solo rispetto alle trasformazioni rigide (rototraslazioni e riflessioni).

2.2 Distribuzioni multinormali

In questo paragrafo si richiamano le definizioni di distribuzione normale e multinormale, e le relative proprietà.

Si dice che una variabile casuale continua Y ha una una distribuzione *normale*, o *gaussiana* (univariata) se possiede una PDF del tipo

$$f_Y(y) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-\frac{1}{2}\frac{(y-\mu)^2}{\sigma^2}\right), \tag{17}$$

e si verifica che essa ha media μ e varianza σ^2. Per brevità, si scriverà $Y \sim \mathcal{N}(\mu, \sigma^2)$.

Generalizzando, un vettore casuale p-variato $\mathbf{Y}$ ha una una distribuzione *multinormale* (o normale multivariata) se possiede una PDF del tipo

$$f_{\mathbf{Y}}(\mathbf{y}) = (2\pi)^{-p/2}(\det \boldsymbol{\Sigma})^{-1/2} \exp\left(-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{y}-\boldsymbol{\mu})\right), \tag{18}$$

con $\boldsymbol{\Sigma}$ matrice definita positiva, e si verifica che esso ha vetore media $\boldsymbol{\mu}$ e matrice di covarianza $\boldsymbol{\Sigma}$. Per brevità, si scriverà $\mathbf{Y} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Si ricorda che la *matrice di covarianza* di un vettore casuale $\mathbf{Y}$ avente media $\boldsymbol{\mu}$ è definita come

$$\mathrm{Cov}[\mathbf{Y}] \stackrel{\text{def}}{=} \mathrm{E}\left[(\mathbf{y}-\boldsymbol{\mu})(\mathbf{y}-\boldsymbol{\mu})'\right].$$

Sulla diagonale di questa matrice si leggono le varianze delle rispettive variabili (con un pasticcio notazionale, $\sigma_{ii} = \sigma_i^2 \stackrel{\text{def}}{=} \mathrm{Var}[Y_i]$), e sul generico elemento σ_{ij} la covarianza tra le variabili Y_i e Y_j.

È possibile dimostrare che, se $\mathbf{Y}$ è un vettore multinormale di dimensione p, per ogni matrice $\mathbf{A}$ di dimensioni $k \times p$ ed ogni vettore $\mathbf{b}$ di dimensione k, il vettore casuale $\mathbf{Y} = \mathbf{A}\mathbf{Y} + \mathbf{b}$ è un vettore multinormale di dimensione k.

Sia $\mathbf{Y} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Si dimostra che la variabile casuale

$$\Delta^2_{\mathbf{Y}}(\mathbf{y}, \boldsymbol{\mu}), \tag{19}$$

ovvero il quadrato della distanza standard di $\mathbf{Y}$ dalla media, segue una distribuzione chi-quadrato con p gradi di libertà (χ^2_p).

Il vettore media e la matrice di covarianza costituiscono statistiche sufficienti per una distribuzione multinormale. Da un semplice confronto tra la (15) e la (18) segue che nel caso di distribuzioni multinormali i luoghi dei punti per cui la PDF ha valore costante presentano la stessa forma ellittica di cui si è parlato nel paragrafo precedente. In altri termini, qualora di una distribuzione si conoscano solo media e matrice di covarianza, la distanza standard fornisce informazioni tanto più aderenti alla realtà quanto più la distribuzione si avvicina ad una multinormale.

Potrebbe trasparire fin d'ora un semplice procedimento per classificare osservazioni all'interno di k distribuzioni multinormali: una volta stimati i parametri relativi a queste classi, e tenendo in debito conto le probabilità a priori, si attribuisce l'osservazione alla classe la cui distanza standard tra media e osservazione è minima. Sebbene questo procedimento porti al classificatore che abbiamo definito ottimo, sarà chiaro alla fine del paragrafo 2.4 che, paradossalmente, i risultati ottenuti in questo modo sono talvolta drogati, in quanto dipendono fortemente dalla sequenza di *training*.

2.3 Stima dei parametri

Si introducono in questa sezione i principali metodi di stima dei due parametri che caratterizzano una distribuzione multinormale: il vettore media e la matrice di covarianza.

Si definiscono formalmente le matrici delle osservazioni della j-esima classe introdotte a pagina 3, ciascuna contenente N_j vettori osservazione. Sia

$$\mathbf{D}_j \stackrel{\text{def}}{=} \begin{bmatrix} \mathbf{d}'_{j1} \\ \vdots \\ \mathbf{d}'_{jN_j} \end{bmatrix} \qquad 1 \leq j \leq k. \tag{20}$$

Si introduce inoltre la matrice

$$\mathbf{D} \stackrel{\text{def}}{=} \begin{bmatrix} \mathbf{d}'_1 \\ \vdots \\ \mathbf{d}'_N \end{bmatrix} \stackrel{\text{def}}{=} \begin{bmatrix} \mathbf{d}'_{11} \\ \vdots \\ \mathbf{d}'_{1N_1} \\ \mathbf{d}'_{21} \\ \vdots \\ \mathbf{d}'_{kN_k} \end{bmatrix}, \tag{21}$$

dove $N \stackrel{\text{def}}{=} \sum_{j=1}^{k} N_j$, ottenuta incolonnando le matrici $\mathbf{D}_j$. Si noti che i $\mathbf{d}_i$ sono vettori colonna.

Si supponga inizialmente $k = 1$, ovvero $\mathbf{D}_1 = \mathbf{D}$. Si assume che le $\mathbf{d}_i$ siano realizzazioni indipendenti del medesimo vettore casuale $\mathbf{Y}$.

La media è stimata dalla *media campionaria*, o *centroide*

$$\mathbf{m} \stackrel{\text{def}}{=} \frac{1}{N} \sum_{i=1}^{N} \mathbf{d}_i. \tag{22}$$

È raro che questa statistica necessiti di un'alternativa, godendo delle principali proprietà auspicabili per uno stimatore e coincidendo peraltro con lo stimatore di massima verosimiglianza[4].

Per la matrice di covarianza, esistono due alternative principali: lo stimatore *plug-in*

$$\mathbf{S}_{\text{P}} \stackrel{\text{def}}{=} \frac{1}{N} \sum_{i=1}^{N} (\mathbf{d}_i - \mathbf{m})(\mathbf{d}_i - \mathbf{m})' = \frac{1}{N} \sum_{i=1}^{N} \mathbf{d}_i \mathbf{d}_i' - \mathbf{m}\mathbf{m}', \tag{23}$$

e la *matrice di covarianza campionaria*

$$\mathbf{S} \stackrel{\text{def}}{=} \frac{1}{N-1} \sum_{i=1}^{N} (\mathbf{d}_i - \mathbf{m})(\mathbf{d}_i - \mathbf{m})' = \frac{N}{N-1} \mathbf{S}_{\text{P}}. \tag{24}$$

Il vantaggio della $\mathbf{S}_{\text{P}}$ è che coincide con lo stimatore di massima verosimiglianza, ma è uno stimatore distorto. Al contrario, si dimostra che $\mathbf{S}$ è non-distorto. Anche se, evidentemente, la differenza tra i due è minima per N sufficientemente grande, verrà utilizzata l'una o l'altra formula a seconda delle necessità. Alcuni risultati dipendono infatti dalle proprietà della particolare statistica adottata.

Condizione necessaria per la definita positività delle matrici $\mathbf{S}$ ed $\mathbf{S}_{\text{P}}$ è $N \geq p+1$. Se le osservazioni $\mathbf{d}_i$ sono effettivamente indipendenti, comunque, la condizione è anche sufficiente con probabilità 1.

Data una coppia di vettori casuali p-variati $\mathbf{X}$ e $\mathbf{Y}$ aventi medie $\boldsymbol{\mu}_{\mathbf{X}} \neq \boldsymbol{\mu}_{\mathbf{Y}}$, e comune matrice di covarianza $\boldsymbol{\Sigma}$, si dimostra che lo stimatore

$$\mathbf{S}_{\text{pooled}} \stackrel{\text{def}}{=} \frac{(N_{\mathbf{X}} - 1)\mathbf{S}_{\mathbf{X}} + (N_{\mathbf{Y}} - 1)\mathbf{S}_{\mathbf{Y}}}{(N_{\mathbf{X}} + N_{\mathbf{Y}} - 2)}, \tag{25}$$

denominato *matrice di covarianza campionaria comune* (*pooled sample covariance matrix*), è non-distorto per $\boldsymbol{\Sigma}$. Si tratta di una specie di media pesata

[4] Per una trattazione esauriente sulla teoria degli stimatori di massima verosimiglianza si rimanda a [9, sezione 4.3]

delle due matrici, e viene talvolta utilizzata anche nel caso in cui l'ipotesi di uguale matrice di covarianza non è verificata. Essa è infatti la versione non distorta dello stimatore di massima verosimiglianza della matrice di covarianza comune, data dalla media pesata delle stime *plug-in* delle singole distribuzioni. La (25) è infine facilmente generalizzabile per $k > 2$ vettori casuali $\mathbf{X}_i$, definendo

$$\mathbf{S}_{\text{pooled}} \stackrel{\text{def}}{=} \frac{\sum_{j=1}^{k}(N_i - 1)\mathbf{S}_{\mathbf{X}_i}}{(N-k)}. \tag{26}$$

Sia ora $k \geq 1$. Il seguente notevole risultato, noto in letteratura sotto il nome di equazioni MANOVA (*Multivariate ANalysis Of VAriance*), esprime media campionaria e varianza *plug-in* ($\mathbf{m}_{\text{Tot}}$ e $\mathbf{S}_{\text{Tot P}}$) di un insieme di k popolazioni, in funzione delle loro medie campionarie e varianze *plug-in* ($\mathbf{m}_j$ e $\mathbf{S}_{\text{P}j}$). Posto

$$\hat{\pi}_j = \frac{N_j}{N} \qquad 1 \leq j \leq k, \tag{27}$$

si ha

$$\mathbf{m}_{\text{Tot}} = \sum_{j=1}^{k} \hat{\pi}_j \mathbf{m}_j \tag{28}$$

$$\begin{aligned}
\mathbf{S}_{\text{P Tot}} &= \sum_{j=1}^{k} \hat{\pi}_j \mathbf{S}_{\text{P}j} + \sum_{j=1}^{k} \hat{\pi}_j (\mathbf{m}_j - \mathbf{m}_{\text{Tot}})(\mathbf{m}_j - \mathbf{m}_{\text{Tot}})' \qquad (29)\\
&= \frac{1}{N} \sum_{j=1}^{k} \sum_{i=1}^{N_j} (\mathbf{d}_{ji} - \mathbf{m}_j)(\mathbf{d}_{ji} - \mathbf{m}_j)' \\
&\quad + \frac{1}{N} \sum_{j=1}^{k} N_j (\mathbf{m}_j - \mathbf{m}_{\text{Tot}})(\mathbf{m}_j - \mathbf{m}_{\text{Tot}})' \\
&\stackrel{\text{def}}{=} \mathbf{S}_{\text{W}} + \mathbf{S}_{\text{B}}.
\end{aligned}$$

La matrice di covarianza totale consta quindi di due termini: il primo, $\mathbf{S}_{\text{W}}$, rende conto della variabilità all'interno (*within*) di ciascuna classe; il secondo, $\mathbf{S}_{\text{B}}$, è una misura della dispersione tra (*between*) le classi[5]. Questo rappresenta un notevole vantaggio computazionale rispetto alle definizioni (22) e (23),

[5]Le stesse matrici moltiplicate per N prendono il nome di *matrici di dispersione*.

in quanto non è necessario riconsiderare le k matrici dei dati, le cui dimensioni non sono limitate superiormente. Un'immediata applicazione di queste equazioni è la seguente. Si supponga di possedere una matrice di dati per una classe, di cui sono disponibili le stime $\mathbf{m}$ e $\mathbf{S}_{\mathrm{P}}$. Se in un secondo momento si rendono disponibili atri dati per quella classe, è possibile aggiornare le stime senza doverle ricalcolare.

In [9] sono analizzati in dettaglio metodi di stima più generali, come gli stimatori di massima verosimiglianza, e avanzati, come gli stimatori *bootstrap*. Alcuni studiosi (si veda [18, sezione 5.7] per approfondimenti) propongono metodi di stima robusta rispetto alle aberrazioni delle distribuzioni multinormali che saranno esposte nella sezione 2.8. In sostanza esse danno automaticamente un peso inferiore alle osservazioni atipiche.

2.4 Classificatore ottimo per classi multinormali

Si torna ora a ricercare il classificatore ottimo del paragrafo 1 per popolazioni aventi densità multinormali, partendo dal risultato (9) di pagina 4, che permette di classificare un'osservazione in modo da rendere massima la probabilità a posteriori $\Pr[X = j|\mathbf{Y} = \mathbf{y}]$.

L'ipotesi di multinormalità consente di calcolare in forma chiusa, a partire dalle sole medie $\boldsymbol{\mu}_j$ e matrici di covarianza $\boldsymbol{\Sigma}_j$, l'espressione $\breve{g}_j(\mathbf{y}) = \pi_j f_j(\mathbf{y})$ per ciascuna classe. Ottenuti questi valori, è sufficiente trovarne il massimo e decidere per la classe relativa, con la certezza di avere effettuato la scelta con la minima probabilità di errore. Sfruttando inoltre la proprietà delle funzioni discriminanti illustrata a pagina 4, è possibile effettuare calcoli più efficienti. Si considerano infatti le

$$g_j^{\mathrm{QDA}}(\mathbf{y}) \stackrel{\text{def}}{=} \log(\breve{g}_j(\mathbf{y})) + C \stackrel{\text{def}}{=} \log(\pi_j f_j(\mathbf{y})) + \frac{p}{2}\log(2\pi), \tag{30}$$

dove la costante serve ad eliminare il fattore $(2\pi)^{-p/2}$ comune a tutte le distribuzioni (si veda la (18)). Si verifica che

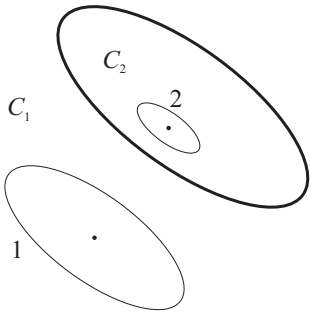
$$g_j^{\mathrm{QDA}}(\mathbf{y}) = \mathbf{y}'\mathbf{A}_j\mathbf{y} + \mathbf{b}_j'\mathbf{y} + c_j \qquad 1 \leq j \leq k, \tag{31}$$

dove

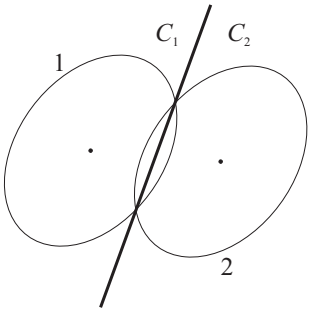
$$\mathbf{A}_j = -\frac{1}{2}\boldsymbol{\Sigma}_j^{-1} \tag{32}$$

$$\mathbf{b}_j = \boldsymbol{\Sigma}_j^{-1}\boldsymbol{\mu}_j, \tag{33}$$

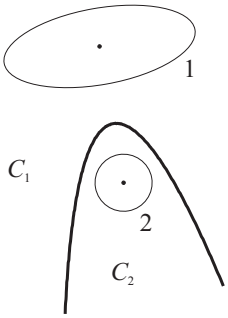
$$c_j = \log \pi_j - \frac{1}{2}\log(\det \boldsymbol{\Sigma}_j) - \frac{1}{2}\boldsymbol{\mu}_j'\boldsymbol{\Sigma}_j^{-1}\boldsymbol{\mu}_j. \tag{34}$$



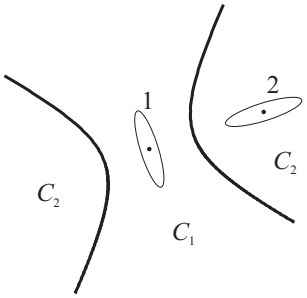
(a) Un'ellisse.



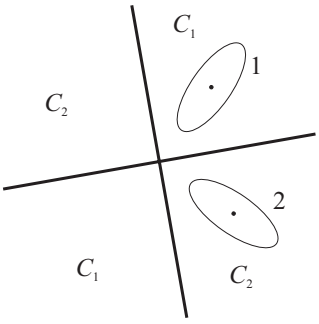
(b) Una linea.



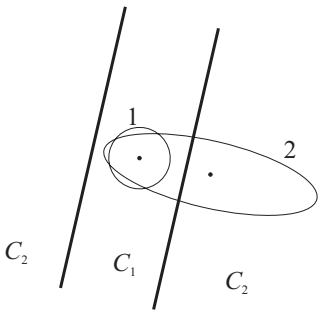
(c) Una parabola.



(d) Un'iperbole.



(e) Due linee.



(f) Due linee parallele.

Figura 4. Alcune regioni di classificazione quadratiche per $k = 2$ e $p = 2$.

La (31) è una forma quadratica, e per questo motivo si parla di *analisi discriminante quadratica* (*quadratic discriminant analysis*, QDA). Per $k = 2$ e $p = 2$ i bordi delle regioni di classificazione assumono la forma, relativamente semplice, di sezioni coniche. Nella figura 4, tratta da [6], ne vengono illustrati alcuni casi, assumendo $\pi_1 = \pi_2 = \frac{1}{2}$. Si noti che, in alcuni casi, la sezione conica degenera in linee rette. In particolare, se $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$, la curva è una retta diretta come l'autovettore dominante e posta esattamente tra $\boldsymbol{\mu}_1$ e $\boldsymbol{\mu}_2$.

È possibile ricavare le probabilità a posteriori introdotte nella (6),

$$\pi_{j\mathbf{y}} = \frac{\exp\left(g_j^{\text{QDA}}(\mathbf{y})\right)}{\sum_{h=1}^{k} \exp\left(g_h^{\text{QDA}}(\mathbf{y})\right)} \qquad 1 \leq j \leq k. \tag{35}$$

Sostituendo nelle (32) i parametri $\boldsymbol{\mu}_j$ e $\boldsymbol{\Sigma}_j$ con le relative stime campionarie, si ottiene, a partire dalle matrici di *training* $\mathbf{D}_j$, un classificatore automatico ottimo. Nonostante il nome promettente, però, un classificatore di questo tipo rischia di aderire troppo ai dati, che sono generalmente affetti da rumore e che possono non provenire da distribuzioni esattamente multinormali. Si rischia così di tracciare confini artificiosi tra le classi, dando origine ad errori di *overfitting*, o eccessiva aderenza ai dati. Per ovviare a questi fenomeni, si può ricorrere ad una tecnica altrettanto semplice, ma meno sensibile alle sequenze di addestramento, illustrata nel seguente paragrafo.

2.5 Classificazione lineare

Innanzitutto, qualche definizione e risultato riguardante la combinazione lineare di variabili casuali. Sia $\mathbf{a}$ un vettore reale di dimensione p. Il prodotto scalare $Z_{\mathbf{a}} = \mathbf{a}'\mathbf{Y}$ proietta il vettore casuale $\mathbf{Y}$ in una variabile (monodimensionale). In altri termini, $Z_{\mathbf{a}}$ è una combinazione lineare delle componenti di $\mathbf{Y}$. Come illustrato dalla figura 5, se $\mathbf{Y}$ è un vettore multinormale, $Z_{\mathbf{a}}$ è, per quanto esposto nel paragrafo 2.2, una variabile casuale normale. Si deduce inoltre dal paragrafo 2.1 che, se si considerano solamente vettori $\mathbf{a}$ normalizzati ($\mathbf{a}'\mathbf{a} = 1$), $Z_{\mathbf{a}}$ ha varianza massima per $\mathbf{a} = \mathbf{u_1}$.

Dalla linearità dell'operatore valore atteso $\mathrm{E}[\cdot]$ segue che

$$\mu_{Z_{\mathbf{a}}} \stackrel{\text{def}}{=} \mathrm{E}[Z_{\mathbf{a}}] = \mathbf{a}'\mathrm{E}[\mathbf{Y}] \stackrel{\text{def}}{=} \mathbf{a}'\boldsymbol{\mu}_{\mathbf{Y}}, \tag{36}$$

mentre per la varianza vale

$$\sigma^2_{Z_{\mathbf{a}}} = \mathbf{a}'\boldsymbol{\Sigma}_{\mathbf{Y}}\mathbf{a}. \tag{37}$$

Più in generale, considerando un insieme di l combinazioni lineari i cui coefficienti stanno in una matrice $\mathbf{A}$ di dimensioni $l \times p$, per il vettore $\mathbf{Z} = \mathbf{A}'\mathbf{Y}$

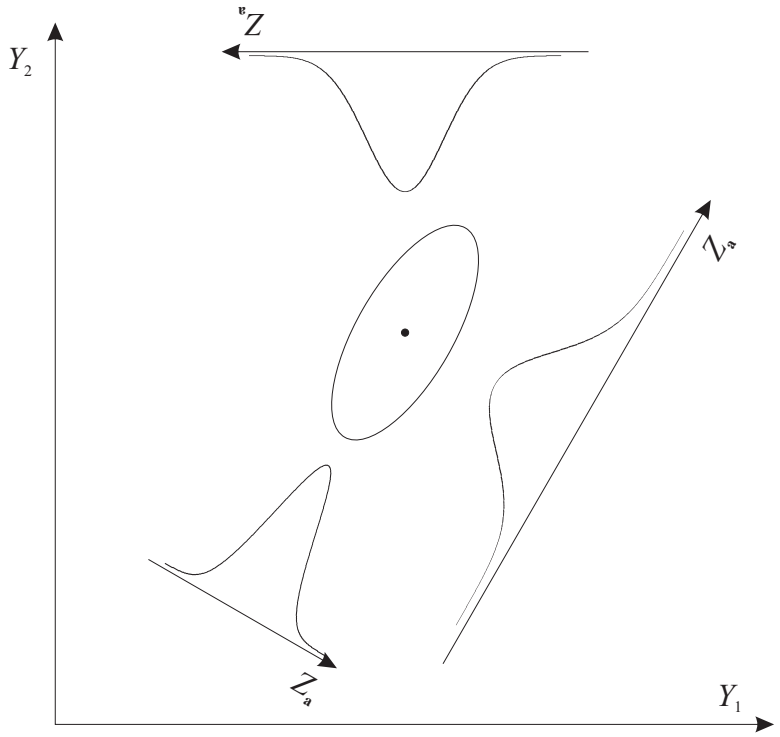


Figura 5. Visualizzazione di combinazioni lineari di una distribuzione normale bivariata.

vale

$$\boldsymbol{\mu}_{\mathbf{Z}} = \mathbf{A}'\boldsymbol{\mu}_{\mathbf{Y}} \tag{38}$$

$$\boldsymbol{\Sigma}_{\mathbf{Z}} = \mathbf{A}'\boldsymbol{\Sigma}_{\mathbf{Y}}\mathbf{A}. \tag{39}$$

Ad esempio, se $\mathbf{A}$ è la matrice $\mathbf{U}$ degli autovettori di $\boldsymbol{\Sigma}_{\mathbf{Y}}$, $\boldsymbol{\Sigma}_{\mathbf{Z}} = \mathbf{U}'\boldsymbol{\Sigma}_{\mathbf{Y}}\mathbf{U}$, e si dimostra che è diagonale (ovvero le variabili sono incorrelate), ed i suoi elementi sono gli autovalori.

Naturalmente i risultati fin qui esposti valgono anche per le relative statistiche campionarie $\mathbf{m}$ e $\mathbf{S}$.

Si consideri ora un problema di discriminazione tra $k = 2$ classi, di cui si dispongono i campioni $\mathbf{D}_1$ e $\mathbf{D}_2$. Per semplicità si assume inizialmente che le due classi abbiano uguale probabilità a priori, $\pi_1 = \pi_2 = \frac{1}{2}$. L'*analisi discriminante lineare* (*linear discriminant analysis*, LDA) si propone di trovare il vettore $\mathbf{a}$ per cui la cifra di merito

$$D(\mathbf{a}) = \frac{|\mathbf{a}'\mathbf{m}_1 - \mathbf{a}'\mathbf{m}_2|}{(\mathbf{a}'\mathbf{S}_{\text{pooled}}\mathbf{a})^{1/2}} \tag{40}$$

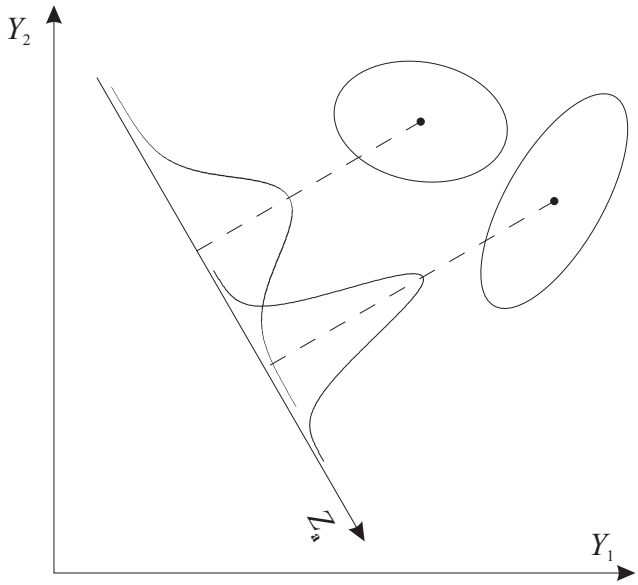


Figura 6. Una proiezione di due distribuzioni normali bivariate[7].

è massima, dove $\mathbf{S}_{\text{pooled}}$ è la matrice di covarianza comune delle due popolazioni, definita nella (25). Aiutandosi con la figura 6 non è difficile convincersi che la (40) cresce al crescere della distanza euclidea tra le medie delle proiezioni monodimensionali delle due popolazioni, e decresce al crescere della loro variabilità. Si dimostra che il vettore cercato è un qualsiasi vettore proporzionale a

$$\mathbf{a}_0 = \mathbf{S}_{\text{pooled}}^{-1}(\mathbf{m}_1 - \mathbf{m}_2), \tag{41}$$

in corrispondenza del quale

$$D(\mathbf{a}_0) = ((\mathbf{m}_1 - \mathbf{m}_2)\mathbf{S}_{\text{pooled}}^{-1}(\mathbf{m}_1 - \mathbf{m}_2))^{1/2}, \tag{42}$$

che equivale alla distanza standard[8] tra $\mathbf{m}_1$ ed $\mathbf{m}_2$ per una distribuzione avente covarianza $\mathbf{S}_{\text{pooled}}$.

Si introduce, con parziale sovrapposizione rispetto alla definizione di pagina 4, la *funzione discriminante lineare*

$$z(\mathbf{y}; a, b) = a \cdot \mathbf{a}_0'\mathbf{y} + b, \tag{43}$$

[7]Sebbene nel testo ci si riferisca a due popolazioni (campioni di vettori casuali), in questa figura si visualizzano per semplicità le relative distribuzioni.

[8]Alcuni autori la definiscono proprio come valore massimo di $D(\mathbf{a})$.

definita a meno dei due gradi di libertà a e b, dove $\mathbf{y}$ rappresenta l'osservazione da classificare.

L'intenzione è quella di trasformare il vettore di osservazione secondo la (43), e classificarla nell'uno o nell'altro gruppo a seconda che lo scalare risultante sia maggiore o minore di un certo valore di soglia f, da determinare. Dal punto di vista geometrico, questo corrisponde a fissare la posizione dell'iperpiano[9] normale alla direzione definita da $\mathbf{a}_0$ in modo che separi il meglio possibile le due classi. Esso è infatti chiamato *iperpiano separatore*. In altri termini, il classificatore cercato ha la forma

$$g^{\text{LDA}}(\mathbf{y}) = \begin{cases} 1 & \text{se } a \cdot \mathbf{a}_0'\mathbf{y} + b > f, \\ 2 & \text{altrimenti.} \end{cases} \tag{44}$$

Contrariamente alla scelta di a, quella di b (e congiuntamente di f) è determinante per la regola di decisione, e dipende dalle distribuzioni delle proiezioni delle due classi

$$Z_i = z(\mathbf{Y_i}; a, b) \qquad i = 1, 2. \tag{45}$$

Esse sono una combinazione lineare di variabili casuali, e quindi, per il teorema limite centrale, ha senso assumere che siano normali, anche se la distribuzione multivariata delle due popolazioni si discosta da quella multinormale. Un metodo per determinare la soglia potrebbe essere quello di determinare analiticamente il punto di intersezione "più significativo" delle due gaussiane[10] a partire dalle stime dei loro parametri ricavate grazie alla (36) e alla (37). Nell'analisi discriminante lineare, tuttavia, si assume che le due normali abbiano uguale varianza e si pone così la soglia a metà delle medie

$$f = \frac{m_{Z_1} + m_{Z_2}}{2}. \tag{46}$$

Una possibile scelta di a e b è

$$a = 1, \tag{47}$$

$$b = -\frac{1}{2}\mathbf{a}_0'(\mathbf{m}_1 + \mathbf{m}_2), \tag{48}$$

per cui il punto medio tra m_{Z_1} e m_{Z_2} è pari a zero, e viene scelto come punto di soglia f. Per inciso, dividendo i coefficienti così trovati per $D(\mathbf{a}_0)$, la varianza di Z è unitaria.

---

[9] Si tratta di un piano se $p = 3$, di una retta se $p = 2$, di una costante se $p = 1$.

[10] Si ricorda infatti che, a meno che abbiano uguale varianza, esse si intersecano in due punti.

Se le probabilità a priori π_j sono disuguali, ovvero se è più probabile osservare esemplari di una delle due classi, la regola della LDA viene generalizzata ponendo la soglia $f = \log(\pi_2/\pi_1)$. Riassumendo

$$g^{\text{LDA}}(\mathbf{y}) = \begin{cases} 1 & \text{se } \mathbf{a}_0'\mathbf{y} - \frac{1}{2}\mathbf{a}_0'(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) > \log\left(\frac{\pi_2}{\pi_1}\right), \\ 2 & \text{altrimenti.} \end{cases} \tag{49}$$

Equivalentemente, è possibile porre

$$a = 1, \tag{50}$$

$$b = -\frac{1}{2}\mathbf{a}_0'(\mathbf{m}_1 + \mathbf{m}_2) + \log\left(\frac{\pi_1}{\pi_2}\right), \tag{51}$$

lasciando la soglia $f = 0$. Con quest'ultima scelta dei coefficienti, si ricavano le probabilità a posteriori (6) stimate da questo metodo

$$\hat{\pi}_{1\mathbf{y}}^{\text{LDA}} = \frac{\mathrm{e}^{z(\mathbf{y})}}{1 + \mathrm{e}^{z(\mathbf{y})}} \tag{52}$$

$$\hat{\pi}_{2\mathbf{y}}^{\text{LDA}} = 1 - \hat{\pi}_{1\mathbf{y}}. \tag{53}$$

Si dimostra, come ci si poteva aspettare dalla figura 4, che l'analisi discriminante lineare corrisponde al classificatore ottimo solo nel caso in cui le due distribuzioni multinormali abbiano uguale matrice di covarianza (*omoschedasticità*).

L'analisi discriminante lineare, come evidenziato da Fisher, è strettamente imparentata con la regressione lineare. È infatti possibile mostrare che i due problemi sono riconducibili uno all'altro, e quindi equivalenti. Non stupisce perciò che si possano adattare risultati di un campo già disponibili in letteratura per applicarli nell'altro.

2.6 Analisi discriminante canonica

È possibile estendere la teoria della LDA ad un numero $k \geq 2$ di classi. Si parla in questo caso di *analisi discriminante canonica* (*canonical discriminant analysis*, CDA), o *analisi discriminante lineare multipla* (*multiple linear discriminant analysis*), e la quantità che si cerca di rendere massima, con riferimento alle quantità introdotte nella (29), è

$$D(\mathbf{a}) = \frac{\mathbf{a}'\mathbf{S}_{\text{P Tot}}\mathbf{a}}{\mathbf{a}'\mathbf{S}_{\text{W}}\mathbf{a}}. \tag{54}$$

Altrimenti detto, si cerca la combinazione lineare (avente come coefficienti gli elementi di $\mathbf{a}$) che massimizzi il rapporto tra la varianza della proiezione della popolazione totale e la varianza delle proiezioni delle singole classi.

La soluzione è piuttosto complessa, e fa uso di concetti e risultati avanzati di algebra lineare, e ci si limita ad esporre per completezza l'algoritmo per il calcolo di $\mathbf{a}$, riportandolo impudentemente da [9].

La *diagonalizzazione simultanea* (o *scomposizione simultanea*) di una matrice simmetrica e definita positiva $\mathbf{W}$ ed una matrice simmetrica $\mathbf{A}$ (aventi uguali dimensioni) è definita come una coppia di matrici $(\mathbf{H}, \mathbf{\Lambda})$, aventi uguali dimensioni e con $\mathbf{\Lambda}$ diagonale, tale che

$$\mathbf{W} \stackrel{\text{def}}{=} \mathbf{H}\mathbf{H}', \tag{55}$$

$$\mathbf{A} \stackrel{\text{def}}{=} \mathbf{H}\mathbf{\Lambda}\mathbf{H}'. \tag{56}$$

Siano $\widetilde{\mathbf{\Lambda}}$ e $\widetilde{\mathbf{U}}$ le matrici di scomposizione spettrale di

$$\mathbf{W} = \widetilde{\mathbf{U}}\widetilde{\mathbf{\Lambda}}\widetilde{\mathbf{U}} = \sum_{j=1}^{p} \tilde{\lambda}_i \tilde{\mathbf{u}}_i \tilde{\mathbf{u}}_i', \tag{57}$$

cioè, rispettivamente, la matrice degli autovalori (diagonale) e degli autovettori di $\mathbf{W}$. Un metodo efficiente per il calcolo della matrice radice quadrata di una matrice simmetrica definita positiva è

$$\mathbf{W}^{1/2} = \widetilde{\mathbf{U}}\widetilde{\mathbf{\Lambda}}^{1/2}\widetilde{\mathbf{U}} = \sum_{j=1}^{p} \tilde{\lambda}_i^{1/2} \tilde{\mathbf{u}}_i \tilde{\mathbf{u}}_i'. \tag{58}$$

Siano $\widehat{\mathbf{\Lambda}}$ e $\widehat{\mathbf{U}}$ le matrici di scomposizione spettrale della matrice simmetrica $\mathbf{W}^{-1/2}\mathbf{A}\mathbf{W}^{-1/2}$. Si dimostra che le matrici cercate sono date da

$$\mathbf{H} = \mathbf{W}^{1/2}\widehat{\mathbf{U}}, \tag{59}$$

$$\mathbf{\Lambda} = \widehat{\mathbf{\Lambda}}. \tag{60}$$

Siano ora $\mathbf{\Lambda}$ e $\mathbf{H}$ le matrici di scomposizione simultanea delle matrici $\mathbf{S}_{\mathrm{W}}$ (simmetrica e definita positiva) e $\mathbf{S}_{\mathrm{B}}$, e si riordinino le loro colonne secondo l'ordine decrescente degli autovalori presenti sulla diagonale di $\mathbf{\Lambda}$. Sia

$$\mathbf{\Gamma} \stackrel{\text{def}}{=} \left[\boldsymbol{\gamma}_1 \dots \boldsymbol{\gamma}_p\right] \stackrel{\text{def}}{=} \left(\mathbf{H}'\right)^{-1}. \tag{61}$$

Definendo

$$m \stackrel{\text{def}}{=} \min(p, k-1), \tag{62}$$

si dice *j-esima variata canonica* ognuna delle m combinazioni lineari delle osservazioni date dalle

$$z_j(\mathbf{y}) \stackrel{\text{def}}{=} \boldsymbol{\gamma}_j \mathbf{y} \qquad 1 \leq j \leq m, \tag{63}$$

ed è unica a meno di una costante proporzionale. Secondo la notazione matriciale

$$\check{\boldsymbol{\Gamma}} \stackrel{\text{def}}{=} [\boldsymbol{\gamma}_1 \dots \boldsymbol{\gamma}_m], \tag{64}$$

$$\mathbf{z}(\mathbf{y}) \stackrel{\text{def}}{=} \check{\boldsymbol{\Gamma}}' \mathbf{y}. \tag{65}$$

Si è perciò proiettato lo spazio campionario su uno spazio m-dimensionale, in modo che la separazione tra le classi sia massima. Si dimostra infatti che la quantità (54) è massima per $\mathbf{a} = \boldsymbol{\gamma}_1$. Inoltre, sotto il vincolo di ortogonalità di $\mathbf{a}$ rispetto alle $\boldsymbol{\gamma}_1, \dots, \boldsymbol{\gamma}_{j-1}$, la (54) è massima per $\mathbf{a} = \boldsymbol{\gamma}_j$.

Siano

$$\mathbf{n}_j \stackrel{\text{def}}{=} \check{\boldsymbol{\Gamma}}' \mathbf{m}_j \qquad 1 \leq j \leq k \tag{66}$$

le proiezioni delle medie campionarie delle classi in questo spazio. La CDA, tenendo conto anche delle probabilità a priori π_j, adotta le seguenti funzioni discriminanti

$$g_j^{\text{CDA}}(\mathbf{y}) \stackrel{\text{def}}{=} \mathbf{n}_j' \left(\mathbf{z}(\mathbf{y}) - \frac{1}{2} \mathbf{n}_j \right) + \log \pi_j \qquad 1 \leq j \leq k, \tag{67}$$

cioè classifica l'osservazione $\mathbf{y}$ nel gruppo j-esimo per cui $g_j^{\text{CDA}}(\mathbf{y})$ è maggiore rispetto alle altre valutazioni $g_i^{\text{CDA}}(\mathbf{y})$, per $i \neq j$. Al solito, le stime delle probabilità a posteriori sono date da

$$\hat{\pi}_{j\mathbf{y}}^{\text{CDA}} = \frac{\exp\left(g_j^{\text{CDA}}(\mathbf{y})\right)}{\sum_{h=1}^{k} \exp\left(g_h^{\text{CDA}}(\mathbf{y})\right)} \qquad 1 \leq j \leq k. \tag{68}$$

Anche in questo caso più generale si dimostra che l'approccio lineare è ottimo solo se le matrici di covarianza delle classi sono identiche. Tuttavia, come si analizzerà più in dettaglio nel paragrafo 2.8, il metodo della CDA si comporta dignitosamente anche in condizioni di eteroschedasticità.

2.7 Stime del tasso di errore

Una volta ottenuto un classificatore, è prassi comune verificarne la validità stimando la probabilità di errore introdotta nella

$$\gamma_g \stackrel{\text{def}}{=} \Pr[X_g \neq X]. \tag{3}$$

Si riportano due semplici stime basate sui dati di *training*. Ricordando le definizioni (20) e (21), si introduce il vettore

$$\mathbf{x} \stackrel{\text{def}}{=} \begin{bmatrix} x_1 \\ \vdots \\ x_N \end{bmatrix}, \tag{69}$$

che memorizza nella sua i-esima componente il numero associato alla classe di appartenenza dell'i-esima osservazione.

Si definisce infine la quantità e_i, che vale 0 se il classificatore identifica la i-esima osservazione correttamente, 1 altrimenti

$$e_i \stackrel{\text{def}}{=} \begin{cases} 0 & \text{se } x_i = g(\mathbf{d}_i) \\ 1 & \text{se } x_i \neq g(\mathbf{d}_i). \end{cases} . \tag{70}$$

La stima più semplice del tasso di errore γ_g, chiamata stima *plug-in*, o *tasso di errore apparente*, è data da

$$\hat{\gamma}_{\text{plug-in}} \stackrel{\text{def}}{=} \frac{1}{N} \sum_{i=1}^{N} e_i. \tag{71}$$

Essa si rivela spesso ottimistica, poiché, di fatto, confonde i dati di *training* con quelli di convalida (o *assessment*), aderendo troppo ai dati osservati. In questi casi, invece, si preferisce generalmente utilizzare una parte dei dati disponibili per la fase di addestramento, e la rimanente parte per la convalida (una ripartizione molto popolare è 70% e 30%, rispettivamente). Si parla in questo caso di convalida incrociata, o *cross-validation*. Estremizzando questo argomento, si perviene alla definizione della stima che in letteratura prende il nome di *leave-one-out*, letteralmente "lasciane fuori uno." Senza definirlo formalmente, si introduce il classificatore $g_{-i}(\cdot)$, ottenuto escludendo dalla sequenza di addestramento $\mathbf{D}$ la sola osservazione i-esima. Il passo successivo è la definizione di

$$e_{i,-i} \stackrel{\text{def}}{=} \begin{cases} 0 & \text{se } x_i = g_{-i}(\mathbf{d}_i) \\ 1 & \text{se } x_i \neq g_{-i}(\mathbf{d}_i). \end{cases} . \tag{72}$$

È ora possibile battezzare

$$\hat{\gamma}_{\text{leave-one-out}} \stackrel{\text{def}}{=} \frac{1}{N} \sum_{i=1}^{N} e_{i,-i}, \tag{73}$$

ovvero la somma degli errori di classificazione di ciascun elemento sul classificatore addestrato senza di esso. Evidentemente, il computo di questa stima

è enormemente più oneroso di quello della stima *plug-in*, in quanto prevede N addestramenti diversi del classificatore in esame. Generalmente, specie se serve una convalida in tempi brevi, si preferisce adottare soluzioni intermedie come quella accennata poc'anzi.

2.8 Classificazione lineare e quadratica a confronto

Come era stato anticipato alla fine del paragrafo 2.4, anche per classi aventi distribuzioni multinormali, le prestazioni della LDA (o della CDA) possono essere superiori rispetto al classificatore "ottimo" della QDA. Il paradosso è generato dal fatto che le stime dei parametri delle distribuzioni vengono effettuate su un numero *finito* di campioni, e sono quindi affette da errore. Quanto detto può essere suffragato da evidenze sperimentali, calcolando $\hat{\gamma}_{\text{leave-one-out}}$ per entrambi i classificatori (si veda ad esempio [9, esempio 7.2.4]).

La scelta tra i due metodi si basa su diversi fattori [18]. Se ne elencano alcuni:

1. Dal punto di vista della complessità computazionale della sola fase di identificazione (la più critica per applicazioni in tempo reale) per la QDA è $\Theta\left(k(p^2+2p)\right)$, cioè $\Theta(kp^2)$, mentre per la CDA è $\Theta\left(k(p+mp+m)\right)$, cioè $\Theta(kpm)$, con m definita nella (62). Come si vede, sotto questo aspetto i metodi sono pressoché equivalenti.

2. La CDA si comporta meglio della QDA se i dati disponibili per la fase di addestramento sono pochi, in quanto aderisce meno al rumore inevitabilmente presente.

3. La CDA è ottima in caso di omoschedasticità. Se le matrici di covarianza sono "molto diverse" (si veda [8] per una trattazione monografica sulla comparazione tra matrici di covarianza) conviene quindi spostarsi verso la QDA.

4. L'approssimazione introdotta dalla CDA è più marcata se le classi sono poco separate. Viceversa, in queste situazioni la QDA è preferibile, in quanto disegna regioni di classificazione più accurate.

5. La robustezza rispetto alla non-normalità delle distribuzioni è un elemento di decisione fondamentale in presenza di aberrazioni ("la mappa non è il territorio").

   Code corte — Se le classi sono più compatte rispetto ad una multinormale, le prestazioni non peggiorano in ogni caso.

Code lunghe Questo tipo di deformazione compromette i risultati di entrambe le tecniche. Tuttavia, se sono disponibili parecchi dati di addestramento e non ci sono evidenti asimmetrie, conviene scegliere la QDA.

Curtosi Se le popolazioni hanno forme concave ("a fagiolo") molto pronunciate, la QDA fornisce risultati migliori.

Asimmetrie (*skewness*) Entrambe le tecniche si comportano bene, ma la CDA è da preferirsi, specie se il difetto non è accentuato.

Esistono in letteratura numerose proposte di compromesso tra i due metodi esposti. Ad esempio Friedman [10] propone l'*analisi discriminante regolarizzata*, operando una sorta di combinazione lineare convessa tra QDA e CDA. Infine, una alternativa piuttosto popolare nello stesso ambito è rappresentata dalla *regressione logistica* [5].

2.9 Cenni ad altre tecniche di classificazione

Si accennano in questo paragrafo ad alcune altre tecniche di classificazione, rimandando alla letteratura per eventuali approfondimenti. Per semplicità, si assume di voler discriminare tra $k = 2$ classi.

***k*-nearest neighbor (*k*-NN)** Sia k un numero dispari[11]. Data l'osservazione di natura incognita $\mathbf{y}$, siano

$$\mathbf{d}_j^{\mathrm{NN}} \qquad 1 \leq j \leq k \tag{74}$$

le osservazioni appartenenti alla sequenza di *training* che sono più vicine ad $\mathbf{y}$. La regola di decisione attribuisce $\mathbf{y}$ alla classe che contiene il maggior numero di queste osservazioni. L'implementazione banale di questa tecnica, ovvero quella che calcola la distanza tra $\mathbf{y}$ e tutti dati della sequenza di addestramento, non brilla per efficienza, e per questo sono stati messi a punto algoritmi migliori. La scelta della particolare metrica adottata dipende dal problema in esame e dalla morfologia delle classi.

Kernel rules È la tecnica duale della *k-nearest neighbor*. Si consideri una regione dello spazio p-dimensionale centrata in $\mathbf{y}$ e di forma

[11] La collisione con il nome della variabile utilizzata nel resto del capitolo per indicare il numero di classi non è stata risolta, a causa della diffusione del nome di questa tecnica.

arbitraria[12], e siano

$$\mathbf{d}_j^{\mathrm{KR}} \qquad 1 \leq j \leq t \tag{75}$$

le t osservazioni appartenenti alla sequenza di *training* che cadono in questa regione. La regola di decisione attribuisce $\mathbf{y}$ alla classe che contiene il maggior numero di queste osservazioni. Valgono le stesse considerazioni di efficienza e flessibilità esposte per la tecnica di *k-nearest neighbor*.

Support vector machines, SVM

Similmente alla LDA, esposta in dettaglio nel paragrafo 2.5, questa tecnica separa le classi attraverso degli iperpiani (uno solo, nel semplice caso di due classi). Essa, però, non richiede alcuna assunzione riguardo alle loro distribuzioni. L'iperpiano viene fissato in modo da rendere massima la sua distanza con l'osservazione più vicina. Se le due classi non sono separabili da un iperpiano, questo semplice criterio non porta ad alcuna soluzione, e quindi lo si modifica introducendo delle penalizzazioni in caso di classificazione erronea. L'estensione del metodo a $k > 2$ classi non è univoca. Per un *tutorial* su questa tecnica si veda [2].

2.10 Test per le proprietà dei campioni

Un elemento importante nella realizzazione di un classificatore è rappresentato dai test statistici e da quantificazioni di determinate proprietà dei campioni. Ad esempio, ci si potrebbe chiedere quanto "normali" sono i dati, o da che grado di curtosi sono affetti, per adottare un modello adeguato di classificazione. Si riportano di seguito una serie di procedure di test e statistiche per campioni multinormali tratte da [9, 18].

Test 1 $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$ [9, sezione 6.2]

Si vuole testare l'ipotesi che la media del vettore casuale di cui il campione in esame è una realizzazione abbia media $\boldsymbol{\mu}_0$. Si supponga che il campione sia composto di N osservazioni, abbia media $\mathbf{m}$ e matrice di covarianza $\mathbf{S}$. Hotelling [14] propone la seguente statistica, generalizzazione della t di Student[13] del caso monovariato

$$T^2 \stackrel{\text{def}}{=} N \cdot (\boldsymbol{\mu}_0 - \mathbf{m})' \, \mathbf{S}^{-1} \, (\boldsymbol{\mu}_0 - \mathbf{m}) = N \cdot D^2(\boldsymbol{\mu}_0, \mathbf{m}), \tag{76}$$

[12] Generalmente di utilizzano ipersfere, iperellissoidi, o parallelepipedi in p dimensioni.

[13] Pseudonimo che W. S. Gossett fu costretto ad usare nel 1908 per poter pubblicare i propri risultati.

dove

$$D(\mathbf{y}_1, \mathbf{y}_2) \stackrel{\text{def}}{=} \sqrt{(\mathbf{y}_1 - \mathbf{y}_2)' \, \mathbf{S}^{-1} \, (\mathbf{y}_1 - \mathbf{y}_2)} \tag{77}$$

rappresenta la versione campionaria della distanza standard.

Sia f il $(1-\alpha)$-quantile della distribuzione F con p ed $(N-p)$ gradi di libertà. Si accetti l'ipotesi H_0 se

$$T^2 \leq \frac{p(N-1)}{N-p} f \tag{78}$$

o, equivalentemente, se

$$D(\boldsymbol{\mu}_0, \mathbf{m}) \leq \sqrt{\frac{p(N-1)}{N(N-p)} f} \, . \tag{79}$$

Si dimostra che, se le classi sono multinormali e possiedono la stessa matrice di covarianza, la probabilità condizionale di rifiutare l'ipotesi nulla nel caso in cui sia vera è α (*test di livello* α).

Test 2 $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ [9, sezione 6.4]

Si desidera testare l'ipotesi che due campioni hanno media coincidente. Si accetti l'ipotesi se

$$T^2 \leq \frac{p(N_1 + N_2 - 2)}{N_1 + N_2 - p - 1} f \tag{80}$$

o, equivalentemente, se

$$D(\boldsymbol{\mu}_0, \mathbf{m}) \leq \sqrt{\frac{p(N_1 + N_2)(N_1 + N_2 - 2)}{N_1 N_2 (N_1 + N_2 - p - 1)} f} \, , \tag{81}$$

con N_1 ed N_2 pari al numero di osservazioni delle due popolazioni, ed f pari all'$(1-\alpha)$-quantile della distribuzione F con p ed $(N_1 + N_2 - p - 1)$ gradi di libertà. Anche in questo caso si dimostra che, se le classi sono multinormali e possiedono la stessa matrice di covarianza, la probabilità condizionale di rifiutare l'ipotesi nulla nel caso in cui sia vera è α.

Questo test ha diverse applicazioni. Ad esempio, può servire per verificare che due classi distinte non differiscano solo per "rumore statistico" (test di significatività globale), oppure per assicurarsi che due sessioni di campionamento distinte della stessa classe presentino medie campionarie "sufficientemente vicine."

Test 3 $H_0 : \boldsymbol{\mu}_1 = \ldots = \boldsymbol{\mu}_k$ [9, sezione 7.4]

Si desidera testare l'ipotesi che k classi aventi distribuzioni multinormali ed omoschedastiche hanno media coincidente, generalizzando il test 2. Con riferimento alle quantità introdotte nella (29), si dimostra che la statistica dei rapporti di massima verosimiglianza logaritmica (*log-likelihood ratio statistic*) è data da

$$\mathrm{LLRS_m} = N \log\det(\mathbf{S}_\mathrm{W}^{-1}\mathbf{S}_{\mathrm{P\,Tot}}). \tag{82}$$

Conseguentemente, dalla teoria degli stimatori di massima verosimiglianza (si veda, ad esempio, [9, sezione 4.3]) si ha che, asintoticamente, $\mathrm{LLRS_m} \sim \chi^2_d$, dove $d = p(k-1)$. Operativamente, si accetti l'ipotesi di ridondanza se e solo se $\mathrm{LLRS_m} \leq c$, dove c è il $(1-\alpha)$-quantile della distribuzione χ^2_d: la decisione è corretta con probabilità approssimativamente pari ad α.

Si dimostra che la (82) può essere anche espressa come

$$\mathrm{LLRS_m} = N \sum_{j=1}^{m} \log(1+\hat{\lambda}_j), \tag{83}$$

dove le $\hat{\lambda}_j$ sono gli autovalori della matrice $\mathbf{S}_\mathrm{W}^{-1}\mathbf{S}_\mathrm{B}$, e m è definito nella (62).

Test 4 $H_0 : \mathbf{Y} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

Si desidera testare la normalità della variabile casuale p-variata $\mathbf{Y}$ dato un suo campione. Si calcolino le quantità

$$c_j \overset{\mathrm{def}}{=} \frac{(N-p-1)ND(\mathbf{d}_j, \mathbf{m})}{p\left[(N-1)^2 - ND(\mathbf{d}_j, \mathbf{m})\right]}. \tag{84}$$

Si dimostra che le c_j seguono una distribuzione F con p e $(N-p-1)$ gradi di libertà. Se a_j denota l'area alla destra di c_j sottesa dalla $F_{p,N-p-1}$, sotto l'ipotesi nulla si ha che

$$a_1, \ldots, a_N \overset{iid}{\sim} \mathcal{U}(0,1), \tag{85}$$

ovvero le a_j sono indipendenti ed uniformemente distribuite in $(0,1)$.

Test 5 $H_0 : \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \ldots = \boldsymbol{\Sigma}_k$ [9, sezione 6.6 esercizio 13]

Una misura di omoschedasticità tra k distribuzioni normali p-variate, tra le tante disponibili, è data da

$$\mathrm{LLRS_h} \overset{\mathrm{def}}{=} N \log(\det \mathbf{S}_\mathrm{W}) - \sum_{j=1}^{k} N_j \log(\det \mathbf{S}_{\mathrm{P}j}), \tag{86}$$

con $\mathbf{S}_{\mathrm{P}}$ e $\mathbf{S}_{\mathrm{W}}$ definiti nella (23) e nella (29), rispettivamente. Si dimostra che la (86) corrisponde alla statistica dei rapporti di massima verosimiglianza logaritmica. Dalla teoria della stima di massima verosimiglianza si deriva quindi che, asintoticamente, LLRS $\sim \chi^2_d$, con $d = (k-1)p(p+1)/2$.

Test 6 $H_0 : (\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) = (\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$ [9, sezione 6.6 esercizio 14]

Si desidera testare l'uguaglianza di media e matrice di covarianza di due distribuzioni normali p-variate. Si dimostra che

$$\mathrm{LLRS}_{\mathrm{eq}} \stackrel{\mathrm{def}}{=} N \log(\det \mathbf{S}_{\mathrm{P\,Tot}}) - \sum_{j=1}^{2} N_j \log(\det \mathbf{S}_{\mathrm{P}j}), \tag{87}$$

dove $\mathbf{S}_{\mathrm{P\,Tot}}$ è la matrice di covarianza complessiva introdotta nella (29), corrisponde alla statistica dei rapporti di massima verosimiglianza logaritmica. Si deriva quindi che, asintoticamente, LLRS $\sim \chi^2_d$, con $d = p(p+3)/2$.

Test 7 $H_0 : \mathbf{Y}_i \sim \mathcal{N}_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$ [18, sezione 6.2.2]

Si desidera testare la normalità e, congiuntamente, l'omoschedasticità delle variabili casuali p-variate $\mathbf{Y}_i$ dati i loro campioni. Si tratta di una generalizzazione del test 4 [12].

Sia

$$\nu \stackrel{\mathrm{def}}{=} N - p - k. \tag{88}$$

Si calcolino le quantità

$$c_{ij} \stackrel{\mathrm{def}}{=} \frac{\nu N_i D(\mathbf{d}_{ij}, \mathbf{m}_i)}{p\left[(\nu + p)(N_i - 1) - N_i D(\mathbf{d}_{ij}, \mathbf{m}_i)\right]} \quad 1 \le i \le k, \tag{89}$$

con $D(\cdot, \cdot)$ distanza standard campionaria basata sulla matrice di covarianza comune $\mathbf{S}_{\mathrm{pooled}}$, definita nella (25). Si dimostra che le c_{ij} seguono una distribuzione F con p e ν gradi di libertà. Se a_{ij} denota l'area alla destra di c_{ij} sottesa dalla $F_{p,\nu}$, sotto l'ipotesi nulla si ha che

$$a_{i1}, \ldots, a_{iN_i} \stackrel{iid}{\sim} \mathcal{U}(0,1) \quad 1 \le i \le k. \tag{90}$$

Test 8 Misura di asimmetria (*skewness*) [18, sezione 6.2.3]

Si calcoli la quantità [17]

$$S \stackrel{\mathrm{def}}{=} \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \left[(\mathbf{d}_i - \mathbf{m})' \mathbf{S}^{-1} (\mathbf{d}_j - \mathbf{m})\right]^3. \tag{91}$$

Sotto le ipotesi di multinormalità si ha che, asintoticamente,

$$\frac{N}{6}S \sim \chi^2_d, \tag{92}$$

con $d = (p/6)(p+1)(p+2)$. Si osservi che questo indice di asimmetria, al contrario del corrispondente indice scalare ($p = 1$), non fornicse alcuna informazione riguardo alla direzione della eventuale asimmetria.

Test 9 Misura di curtosi [18, sezione 6.2.3]

Si calcoli la quantità [17]

$$K \stackrel{\text{def}}{=} \frac{1}{N}\sum_{j=1}^{N}\left[(\mathbf{d}_j - \mathbf{m})'\mathbf{S}^{-1}(\mathbf{d}_j - \mathbf{m})\right]^2. \tag{93}$$

Sotto le ipotesi di multinormalità si ha che, asintoticamente,

$$\frac{K - p(p+2)}{\sqrt{p(p+2)(8/N)}} \sim \mathcal{N}_p(0,1). \tag{94}$$

I test 8 e 9 possono vedersi come test alternativi per la normalità.

2.11 Test per la ridondanza di variabili

Nell'ambito di una analisi discriminante, o di una procedura di classificazione automatica in generale, può aver senso chiedersi se alcune variabili, si supponga in un numero q, non siano ridondanti, cioè se non aggiungano alcuna informazione utile alla classificazione rispetto all'insieme delle rimanenti $p - q$. La risposta a questa domanda si rivela di notevole interesse se alcune variabili sono particolarmente costose o difficili da ottenere. La restrizione del numero di variabili si rivela infine necessaria se è disponibile una casistica limitata per la fase di addestramento. Si dimostra infatti [4] che, a parità di tasso di errore, un classificatore necessita di una sequenza di *training* che cresce esponenzialmente con il numero delle variabili p.

Un possibile modo di procedere è quello di valutare, in presenza e in assenza delle variabili in esame, una cifra di merito, per esempio una stima del tasso di errore, o una misura della separazione tra le classi, come la (42) per la LDA, o la prima variata canonica per la CDA.

Paradossalmente, l'eliminazione di variabili che singolarmente presentano un basso indice di separazione può rivelarsi una pessima idea. Viceversa, è possibile che una variabile con un alto contenuto informativo ai fini della classificazione sia superflua se utilizzata con altre variabili. Per un'illuminante illustrazione di questo fenomeno si veda [9, esempio 5.3.3].

Nel semplice caso in cui si abbiano $k = 2$ classi multinormali e con uguale matrice di covarianza, sono disponibili due test di ridondanza di variabili.

Test 10 H_0 : la variabile Y_j è ridondante [9, teorema 6.5.1]

Siano, al solito, N_1 ed N_2 il numero di osservazioni per ogni classe; sia D la distanza standard tra le due medie definita dalla (42), e sia D_{-j} la distanza standard calcolata senza fare uso della variabile Y_j. Si calcoli

$$|t_j| = \sqrt{(N_1 + N_2 - p - 1) \cdot \frac{D^2 - D_{-j}^2}{m + D_{-j}^2}}, \tag{95}$$

dove

$$m \stackrel{\text{def}}{=} \frac{(N_1 + N_2)(N_1 + N_2 - 2)}{N_1 N_2}. \tag{96}$$

Si dimostra che, per classi multinormali e omoschedastiche, la quantità t_j segue una distribuzione t con $(N_1 + N_2 - p - 1)$ gradi di libertà. Operativamente, si accetti l'ipotesi di ridondanza se e solo se $|t_j| \leq c$, dove c è il $(1-\alpha/2)$-quantile della distribuzione t con (N_1+N_2-p-1) gradi di libertà: la decisione è corretta con probabilità α.

È possibile calcolare D_{-j} senza dover affrontare una analisi discriminante da zero. Sia infatti a_{0j} la j-esima componente del vettore $\mathbf{a}_0$ calcolato nella (41), e sia $\check{s}_{jj}$ il j-esimo elemento diagonale della matrice $\mathbf{S}_{\text{pooled}}^{-1}$. Si verifica che

$$D_{-j}^2 = D^2 - \frac{a_{0j}^2}{\check{s}_{jj}}. \tag{97}$$

Test 11 H_0 : le ultime $(p-q)$ variabili Y_j $(q+1 \leq j \leq p)$ sono ridondanti [9, teorema 6.5.2]

Sia D_p la distanza standard tra le due medie definita dalla (42), e sia D_q la distanza standard ottenuta utilizzando solo le prime q variabili. Si calcoli

$$R_q = \frac{N_1 + N_2 - p - 1}{p - q} \cdot \frac{D_p^2 - D_q^2}{m + D_q^2}, \tag{98}$$

con m definita come nella (96). Si dimostra che, per classi multinormali e omoschedastiche, la quantità R_q segue una distribuzione F con $(p - q)$ e $(N_1 + N_2 - p - 1)$ gradi di libertà.

Quest'ultimo test riguarda sottoinsiemi di variabili, ed è una generalizzazione di entrambi i test 2 e 10, in cui si consideravano rispettivamente singoletti e l'insieme vuoto. Si osserva che una ricerca esaustiva su tutti i

sottoinsiemi ha complessità $\Theta(2^p)$, e può diventare facilmente intrattabile al crescere di p. Volendo trovare il sottoinsieme di cardinalità $q < p$ dell'insieme delle variabili disponibili, tale che il suo tasso di errore sia minimo, si dimostra [4, teorema 32.1] che è necessario, nel caso generale, cercarlo esaustivamente tra tutti i $\binom{q}{p}$ sottoinsiemi di cardinalità q. L'unica cosa certa è che, aumentando il numero di variabili, il tasso di errore del classificatore *ottimo* non decresce (ma, come già detto, il tasso di errore può aumentare per una determinata regola). Si rendono perciò necessarie delle procedure euristiche di selezione delle variabili, riportate alla fine del seguente test.

Test 12 H_0 : la variabile Y_j è ridondante (per la classificazione tra $k > 2$ classi) [18]

Si intende generalizzare il test 10 per un numero arbitrario di classi, sempre in condizioni di omoschedasticità. Sia $j = p$, ovvero, senza perdita di generalità, si riordinino le variabili in modo che quella in esame sia l'ultima. Siano

$$\mathbf{W} \stackrel{\text{def}}{=} (N-k)\mathbf{S}_{\text{pooled}}, \tag{99}$$

$$\mathbf{B} \stackrel{\text{def}}{=} (k-1)\mathbf{S}_{\text{B}}, \tag{100}$$

dove $\mathbf{S}_{\text{pooled}}$ è la matrive di covarianza comune generalizzata definita nella (26). Si suddividano inoltre queste matrici

$$\mathbf{W} \stackrel{\text{def}}{=} \begin{bmatrix} \mathbf{W}_{11} & \mathbf{W}_{12} \\ \mathbf{W}_{21} & \mathbf{W}_{22} \end{bmatrix}, \tag{101}$$

$$\mathbf{B} \stackrel{\text{def}}{=} \begin{bmatrix} \mathbf{B}_{11} & \mathbf{B}_{12} \\ \mathbf{B}_{21} & \mathbf{B}_{22} \end{bmatrix}, \tag{102}$$

isolando le componenti relative alla p-esima variabile (gli scalari $\mathbf{W}_{22}$ e $\mathbf{B}_{22}$). Si definiscono le quantità scalari[14]

$$\mathbf{W}_{2.1} \stackrel{\text{def}}{=} \mathbf{W}_{22} - \mathbf{W}_{21}\mathbf{W}_{11}^{-1}\mathbf{W}_{12}, \tag{103}$$

$$\mathbf{B}_{2.1} \stackrel{\text{def}}{=} (\mathbf{B}_{22} + \mathbf{W}_{22}) - (\mathbf{B}_{21} + \mathbf{W}_{21})(\mathbf{B}_{11} + \mathbf{W}_{11})^{-1}(\mathbf{B}_{12} + \mathbf{W}_{12}) - \mathbf{W}_{2.1}, \tag{104}$$

$$\Lambda_{-p} \stackrel{\text{def}}{=} \frac{|\mathbf{W}_{2.1}|}{|\mathbf{W}_{2.1} + \mathbf{B}_{2.1}|}. \tag{105}$$

[14] Si preferisce mantenere la notazione matriciale per uniformarsi a quella di [18], che tratta il caso più generale di ridondanza di $q \geq 1$ variabili.

Si dimostra che, sotto l'ipotesi di ridondanza, la statistica

$$\Lambda \stackrel{\text{def}}{=} \frac{(N-k-p+1)(1-\Lambda_{-p})}{(k-1)\Lambda_{-p}} \tag{106}$$

segue una distribuzione F con $(k-1)$ e $(N-k-p-1)$ gradi di libertà.

A partire da questo test, è possibile costruire delle procedure euristiche passo-passo per la determinazione di un sottoinsieme subottimo di variabili [18, sezione 12.3.3]. Le due tecniche fondamentali sono la selezione in avanti (*forward selection*) e l'eliminazione all'indietro (*backward elimination*). La prima, partendo dall'insieme vuoto, aggiunge all'insieme corrente la variabile che presenta il valore massimo per la (106), a patto che superi un valore di soglia, ad esempio il $(1-\alpha)$-quantile di $F_{k-1,N-k-p+1}$. Viceversa, nell'eliminazione all'indietro, partendo dall'insieme di tutte le variabili vengono eliminate una ad una quelle che presentano il valore minimo per la (106), a patto che sia sotto il valore di soglia.

3 Analisi dei cluster

L'*analisi dei cluster* (*cluster analysis*, a volte tradotta come *analisi dei grappoli*) è una tencica nata e diffusasi negli anni 60 e 70, mirata all'individuazione di agglomerati di dati all'interno di una popolazione nota. Gli obiettivi finali possono essere i più disparati, ad esempio l'individuazione o la convalida di un'ipotesi di ricerca a partire dai dati, l'isolamento di *pattern* caratteristici in determinate sotto-popolazioni, o la classificazione dei dati. In questa sezione verranno esposti i principali strumenti dell'analisi dei *cluster* con quest'ultimo scopo in mente.

L'analisi dei *cluster* si basa su procedure semplici e facilmente automatizzabili, fa largo uso di euristiche e poggia su una matematica piuttosto elementare. Per questi motivi, essa è spesso snobbata dagli statistici, che la vedono come il "fratello povero" [9, pagina 123] dell'*analisi delle misture finite*, strettamente connessa all'analisi discriminante, dalle basi teoriche certamente più solide e rigorose. D'altra parte, proprio la sua semplicità ne ha favorito la diffusione tra i ricercatori delle scienze naturali, e la leggibilità dei suoi risultati, l'alto potenziale euristico (appunto) e la disponibilità di numerosi strumenti di analisi automatica ne fanno uno strumento valido e meritevole di considerazione.

Nella tradizione di questa disciplina, ma la definizione potrebbe essere estesa alla classificazione in generale, si distingue tra tecniche di tipo Q e tecniche di tipo R. Nel primo caso, come si assume in questo intero capitolo se

non specificato diversamente, vengono analizzate e classificate le *osservazioni*, mentre nelle tecniche di tipo R vengono esaminate le variabili, ad esempio, come si è fatto nella sezione 2.11, per eliminare variabili superflue o dallo scarso contenuto informativo.

Le tecniche di analisi dei *cluster* possono suddividersi in due ampie categorie: i metodi di ripartizione e i metodi gerarchici. Prima della loro trattazione si introducono le principali procedure di trasformazione delle variabili.

3.1 Trasformazione delle variabili e normalizzazione

Sebbene la normalizzazione, e in generale la trasformazione delle variabili, possa essere utilizzata anche nei metodi statistici, viene introdotta in questo contesto, in quanto nei metodi presentati nel paragrafo 2 non era indispensabile. Viceversa, nell'analisi dei *cluster* il suo utilizzo è fortemente consigliato, poiché rende il risultato indipendente dalle unità di misura adottate per le variabili. Inoltre, la normalizzazione fa sì che tutte le variabili contribuiscano in ugual misura alla classificazione.

Per *trasformazione* di una variabile, o attributo, si intende la derivazione di nuove variabili attraverso l'applicazione di funzioni a quelle originarie. In formula

$$Y_i' \stackrel{\text{def}}{=} f_i(Y_i) \qquad 1 \leq i \leq p. \tag{107}$$

In alcuni casi può essere utile applicare ad alcune variabili delle trasformazioni non lineari, al fine di correggerne la distorsione. Le più usate sono $\log(\cdot)$, $\log(\cdot + 1)$, $\sqrt{\cdot}$, $\arctan(\cdot)$ e $\cosh(\cdot)$.

Tra le trasformazioni lineari, la più usata è senz'altro la

$$Y_i' = \frac{Y_i - \mathrm{E}[Y_i]}{\sqrt{\mathrm{Var}[Y_i]}} \qquad 1 \leq i \leq p, \tag{108}$$

spesso denominata *normalizzazione*. Naturalmente, nella pratica si utilizzano le stime campionarie di queste quantità. Si verifica facilmente che, le variabili così trasformate, hanno media (campionaria) nulla e varianza (campionaria) unitaria. Si noti infine che le nuove variabili sono adimensionali.

Esistono forme alternative di normalizzazione. Ad esempio, nel caso in cui i valori delle variabili siano non negative, si può far uso della

$$Y_i' = \frac{Y_i}{\max\limits_{1 \leq j \leq N} \mathbf{e}_i' \mathbf{d}_j} \qquad 1 \leq i \leq p, \tag{109}$$

dove si è indicato con $\mathbf{e}_i$ l'i-esimo versore, e quindi il prodotto scalare $\mathbf{e}_i' \mathbf{d}_j$ rappresenta la componente i-esima dell'osservazione j-esima (si ricordi che

$\mathbf{d}_j$ è stato definito come un vettore colonna). In sostanza, si dividono i dati rilevati di ciascuna variabile per il valore massimo, in modo che i tutti i valori delle nuove variabili siano comprese nell'intervallo unitario. Affinché quest'ultimo sia il *più piccolo* intervallo contenente tutti i nuovi valori, si può ricorrere alla

$$Y_i' = \frac{\left(Y_i - \min\limits_{1\leq j\leq N} \mathbf{e}_i'\mathbf{d}_j\right)}{\left(\max\limits_{1\leq j\leq N} \mathbf{e}_i'\mathbf{d}_j - \min\limits_{1\leq j\leq N} \mathbf{e}_i'\mathbf{d}_j\right)} \qquad 1 \leq i \leq p. \tag{110}$$

Se si vuole applicare una tecnica di tipo R ai dati, le tecniche di normalizzazione vanno applicate alle osservazioni, anziché alle variabili. In pratica, vengono utilizzate le stesse formule, previa trasposizione della matrice dei dati $\mathbf{D}$.

Se il metodo adottato prevede sia una analisi Q che una analisi R, è opportuno decidere se normalizzare rispetto alle variabili o alle osservazioni, in quanto effettuare entrambe le operazioni in cascata fornisce risultati poco interpretabili, dipendenti peraltro dall'ordine in cui le due normalizzazioni vengono applicate.

Nei paragrafi a seguire si assumerà di lavorare con una matrice delle osservazioni con variabili normalizzate.

3.2 Metodi di ripartizione

L'obiettivo di questa classe di algoritmi è la ripartizione dei dati disponibili in n sottoinsiemi (*cluster*) $C_1, \ldots, C_n$, quindi tali per cui

$$C_1 \cup \ldots \cup C_n = \{\mathbf{d}_i | 1 \leq i \leq N\} \tag{111}$$

$$C_j \cap C_k = \emptyset \qquad j \neq k, \tag{112}$$

in modo che gli elementi di ogni sottoinsieme siano "il più compatti possibile." È l'interpretazione e la formalizzazione di questa proprietà alquanto sfumata che caratterizza i singoli algoritmi. Alcuni di essi procedono euristicamente, mentre altri cercano di ottimizzare una determinata funzione obiettivo.

In questa sezione verrà analizzato l'algoritmo kmeans, di gran lunga il più noto ed utilizzato. Esso utilizza come funzione obiettivo da minimizzare la somma dei quadrati delle distanze tra i punti e la media campionaria del *cluster* a cui appartengono. In formula

$$s_{\mathrm{W}} \stackrel{\mathrm{def}}{=} \sum_{j=1}^{n} \sum_{i=1}^{N_j} (\mathbf{d}_{ji} - \mathbf{m}_j)'(\mathbf{d}_{ji} - \mathbf{m}_j). \tag{113}$$

Come suggerisce il simbolo adottato per questa cifra di demerito, esiste un parallelo scalare, a meno del fattore $\frac{1}{N}$, dell'equazione MANOVA introdotta nella (29). È infatti

$$\begin{aligned} s_{\text{Tot}} &\stackrel{\text{def}}{=} \text{tr}(\mathbf{S}_{\text{Tot P}}) \\ &= \sum_{i=1}^{N} (\mathbf{d}_i - \mathbf{m}_{\text{Tot}})'(\mathbf{d}_i - \mathbf{m}_{\text{Tot}}) \\ &= \sum_{j=1}^{n} \sum_{i=1}^{N_j} (\mathbf{d}_{ji} - \mathbf{m}_j)'(\mathbf{d}_{ji} - \mathbf{m}_j) \\ &\quad + \sum_{j=1}^{n} N_j (\mathbf{m}_j - \mathbf{m}_{\text{Tot}})'(\mathbf{m}_j - \mathbf{m}_{\text{Tot}}) \\ &= \text{tr}(\mathbf{S}_{\text{W}}) + \text{tr}(\mathbf{S}_{\text{B}}) \stackrel{\text{def}}{=} s_{\text{W}} + s_{\text{B}}. \end{aligned} \tag{114}$$

In questo contesto, però, la s_{Tot} è una costante, mentre la suddivisione in classi è, per così dire, l'incognita. Questo giustifica la scelta della funzione obiettivo, che può essere vista come misura di variabilità intraspecifica, e mostra che è possibile scegliere equivalentemente di rendere massima la misura di separazione tra i gruppi s_{B}.

Prima di presentare l'algoritmo si fa notare che il numero di configurazioni possibili degli n *cluster* sugli N dati si dimostra [22] essere pari a

$$\frac{1}{n!} \sum_{j=1}^{n} (-1)^{n-j} \binom{n}{j} j^N, \tag{115}$$

che esplode facilmente per valori non banali dei due parametri[15]. Una ricerca esaustiva della configurazione ottima è quindi improponibile, e ci si accontenta di algoritmi subottimi.

Sia $\mathbf{x}$ il vettore di lunghezza N che conserva i codici associati ai *cluster* di appartenenza di ciascun dato. Il metodo kmeans, partendo da un assegnamento iniziale $\mathbf{x}_0$ e scandendo i dati uno ad uno, ad ogni passo calcola le medie e la funzione obiettivo, e assegna l'osservazione in esame al *cluster* per cui la nuova valutazione della funzione obiettivo è minima. Il procedimento si arresta allorquando $\mathbf{x}$ rimane invariato per N cicli consecutivi.

Questo algoritmo è ottimo ad ogni passo, ma non trova necessariamente la soluzione ottima cercata. È consigliabile pertanto ripetere la procedura con

[15] Già per $N = 25$ ed $n = 3$ si hanno 141.197.991.025 configurazioni.

diverse configurazioni iniziali. Si tenga in considerazione, comunque, che la funzione obiettivo s_{W} soffre di alcune limitazioni, e fornisce risultati scadenti per *cluster* non sufficientemente compatti e separati, o aventi cardinalità molto diverse tra loro.

In [22] sono riportate numerose varianti di kmeans, con i listati in linguaggio FORTRAN a corredo. Ad esempio, hmeans ricalcola la funzione obiettivo solo dopo aver completato il ciclo su tutti i dati, e dà la possibilità di ridurre automaticamente il numero di *cluster* durante l'esecuzione. Per funzioni obiettivo alternative, sempre basate sulle matrici di dispersione, si veda [6, sezione 6.8.3].

3.3 Metodi gerarchici

In questa sezione verranno esaminati gli algoritmi gerarchici, che più di altri hanno riscosso successo all'interno delle comunità scientifiche di fisici, naturalisti e sociologi, tanto che alcune pubblicazioni (ad esempio [21]) si riferiscono con il termine *cluster analysis* alla sola analisi gerarchica dei *cluster*.

L'obiettivo di questi algoritmi è l'organizzazione dei dati in una struttura gerarchica, che raggruppa osservazioni molto simili in piccoli *cluster* ai livelli più bassi, e osservazioni più lascamente collegate in *cluster* più grandi e generici ai livelli più alti, fino ad arrivare all'insieme di tutti i dati. Formalmente, si ottiene una sequenza di h partizioni di cardinalità strettamente crescente degli N dati. Sia n_i la cardinalità della i-esima partizione. Sarà allora

$$1 = n_1 < \ldots < n_h \leq N. \tag{116}$$

In altre parole, la prima partizione della sequenza è rappresentata da un solo insieme $C_1 = \{\mathbf{d}_i | 1 \leq i \leq N\}$, comprendente tutte le osservazioni; la seconda partizione prevede $n_2 \geq 2$ sottoinsiemi disgiunti e complementari di C_1, e così via, fino all'ultima partizione, che si noti non prevede necessariamente la frammentazione dei dati in N singoletti (*culster* degeneri).

I metodi di analisi gerarchica si distinguono in due categorie: le procedure *divisive*, che, al j-esimo passo, ripartiscono uno o più *cluster* del $(j-1)$-esimo livello in due o più *cluster* di dimensioni inferiori; le procedure *agglomerative*, decisamente le più utilizzate, accorpano al contrario i *cluster* più piccoli (solitamente partendo da quelli degeneri), fino ad arrivare all'insieme di tutti i dati.

3.3.1 Metodi agglomerativi binari

In questa sezione viene analizzata l'ampia classe delle procedure agglomerative, che raggruppano ad ogni passo i due cluster più vicini, partendo da quelli

degeneri (le singole osservazioni). Con terminologia strettamente locale, si battezzano queste procedure agglomerative *binarie*. In [22] si trovano due esempi di procedure divisive, ed un algoritmo agglomerativo alternativo, qui non illustrato, basato sul concetto di *minimo albero di copertura*.

Vengono ora descritti i passi del generico algoritmo binario, e si discuteranno in seguito le scelte possibili per i suoi parametri.

Dopo aver provveduto alla normalizzazione della matrice dei dati, si costruisce la *matrice di somiglianza* $\mathbf{R}$ (*resemblance matrix*), il cui generico elemento r_{ij} rappresenta il valore di un *coefficiente di somiglianza* calcolato sulle osservazioni $\mathbf{d}_i$ e $\mathbf{d}_j$. La matrice è evidentemente simmetrica. Ci sono due tipi di coefficienti di somiglianza: i coefficienti di *similarità* (*similarity*), tanto più grandi quanto più gli oggetti esaminati sono simili, e i coefficienti di *diversità* (*dissimilarity*), caratterizzati dalla proprietà opposta. Si esaminano gli elementi della matrice, e la coppia di elementi maggiormente affini tra loro viene promossa a *cluster* (proprio). A questo punto la matrice di somiglianza deve essere aggiornata. In particolare, le due righe e le due colonne relative alle osservazioni scelte vanno eliminate, e deve essere calcolata una misura di somiglianza tra il *cluster* trovato e tutti gli altri oggetti. Questo implica l'esistenza di una generalizzazione del coefficiente di somiglianza, detto *coefficiente cofenetico*, che misura l'affinità tra coppie di *cluster* qualsiasi, propri o degeneri. Il tipo di generalizzazione adottato è un altro grado di libertà del ricercatore, prende spesso l'ambiguo nome di *metodo di raggruppamento* (*clustering method*) e verrà discusso più avanti. La procedura riprende considerando la nuova matrice fino a che essa si riduce ad uno scalare. Ci si riferirà nel seguito a questa matrice con il nome di *matrice di lavoro* $\mathbf{W}$.

I risultati degli algoritmi gerarchici sono facilmente rappresentabili graficamente attraverso alberi. Nel caso degli algoritmi binari si parla di *dendrogrammi*, o *fenogrammi*, di cui viene dato un esempio nella figura 7. Essi forniscono un colpo d'occhio attraente della soluzione trovata e ne favoriscono l'interpretazione. L'altezza a cui i due *cluster* si fondono rappresenta il valore del loro coefficiente cofenetico. Ad esempio, con riferimento alla figura, γ è il coefficiente cofenetico tra i *cluster* $\{\mathbf{d}_2, \mathbf{d}_5, \mathbf{d}_7\}$ e $\{\mathbf{d}_1, \mathbf{d}_3\}$. Si osservi che γ è in generale diverso dal coefficiente di somiglianza originario relativo a una qualsiasi coppia di osservazioni appartenenti rispettivamente al primo e al secondo *cluster*. Sempre riferendosi alla figura 7, i valori γ e r_{35}, ad esempio, possono essere differenti. Solo nel caso di *cluster* composti di due sole osservazioni vale, per costruzione, $c_{ij} = r_{ij}$. Il valore del coefficiente cofenetico viene esteso a ciascuna di queste coppie, simboleggiandolo con c_{ij}. Nel caso in esame è $c_{12} = c_{15} = c_{17} = c_{23} = c_{35} = c_{37} = \gamma$. Si ottiene così, ad algoritmo terminato, la *matrice cofenetica* $\mathbf{C}$, dal punto di vista informativo equivalente al dendrogramma, che, per quanto detto, è generalmente

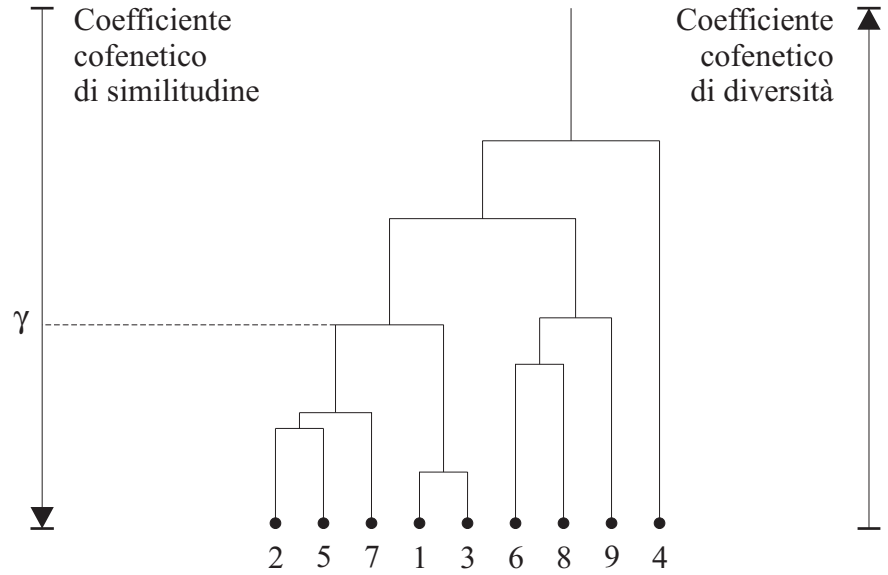


Figura 7. Dendrogramma relativo ad un'analisi di nove osservazioni.

diversa dalla matrice di somiglianza. Questo dà origine ad una distorsione, il cui grado dipende dal coefficiente di somiglianza e dal metodo di raggruppamento adottato. È possibile quantificare questa distorsione confrontando le due matrici, ad esempio attraverso il coefficiente di correlazione di Pearson, introdotto di seguito tra i coefficienti di somiglianza.

Scelta del coefficiente di somiglianza Il più semplice coefficiente di somiglianza è la *distanza euclidea*

$$r_{ij}^{\mathrm{e}} \stackrel{\mathrm{def}}{=} \|\mathbf{d}_i - \mathbf{d}_j\|_2 \stackrel{\mathrm{def}}{=} \sqrt{(\mathbf{d}_i - \mathbf{d}_j)'\,(\mathbf{d}_i - \mathbf{d}_j)}. \tag{117}$$

Una sua variante, la *distanza euclidea media*

$$r_{ij}^{\mathrm{d}} \stackrel{\mathrm{def}}{=} \frac{r_{ij}^{\mathrm{e}}}{\sqrt{p}}, \tag{118}$$

ha il vantaggio di poter essere usata anche in caso di valori mancanti. Il coefficiente di *differenza di forma* (*shape difference*) è definito da

$$r_{ij}^{\mathrm{z}} \stackrel{\mathrm{def}}{=} \sqrt{\frac{p}{p-1}} \left[d_{ij}^2 - \frac{1}{p^2} \left(\sum_{k=1}^{p} \mathbf{e}_k' \mathbf{d}_i - \sum_{k=1}^{p} \mathbf{e}_k' \mathbf{d}_j \right)^2 \right], \tag{119}$$

ed è nullo in caso di osservazioni coincidenti o le cui variabili differiscono tutte per la stessa quantità. Quelli introdotti fin qua sono tutti coefficienti di diversità, con valore minimo pari a zero e illimitati superiormente.

Tra i coefficienti di similitudine figura il *coefficiente del coseno*

$$r_{ij}^{\mathrm{c}} \stackrel{\mathrm{def}}{=} \frac{\mathbf{d}_i'\mathbf{d}_j}{\|\mathbf{d}_i\|_2 \, \|\mathbf{d}_j\|_2}, \tag{120}$$

che misura appunto il coseno dell'angolo tra le due linee che collegano le osservazioni con l'origine, e il *coefficiente di correlazione*, introdotto da Pearson

$$r_{ij}^{\mathrm{p}} \stackrel{\mathrm{def}}{=} \frac{\mathbf{d}_i'\mathbf{d}_j - \frac{1}{p}(\mathbf{1}_p'\mathbf{d}_i)(\mathbf{1}_p'\mathbf{d}_j)}{\sqrt{\left[\mathbf{d}_i'\mathbf{d}_i - \frac{1}{p}\left(\mathbf{1}_p'\mathbf{d}_i\right)^2\right]\left[\mathbf{d}_j'\mathbf{d}_j - \frac{1}{p}\left(\mathbf{1}_p'\mathbf{d}_j\right)^2\right]}}, \tag{121}$$

dove con $\mathbf{1}_p$ si è indicato il vettore di lunghezza p avente tutti gli elementi unitari. Per questi due coefficienti di similitudine vale $-1 \leq r_{ij} \leq 1$.

In [21], da cui sono stati tratti i precedenti, si trova qualche altro coefficiente, mentre in [22] vengono elencate numerose misure di distanza tra vettori di variabili nominali e ordinali.

Scelta del metodo di raggruppamento Dati due *cluster* C_i e C_j, eventualmente degeneri, si pone il problema della misura della distanza tra essi, detta coefficiente cofenetico. Le prime tre possibilità elencate si basano direttamente sui valori del coefficiente di somiglianza tra le coppie di elementi appartenenti ai due diversi *cluster*. Il metodo detto di *legame singolo* (*single linkage*, o *nearest neighbor*) pone

$$c_{ij}^{\mathrm{s}} \stackrel{\mathrm{def}}{=} \begin{cases} \min\limits_{\mathbf{d}_i \in C_i,\ \mathbf{d}_j \in C_j} r_{ij} & \text{se } r_{ij} \text{ è un coefficiente di diversità,} \\ \max\limits_{\mathbf{d}_i \in C_i,\ \mathbf{d}_j \in C_j} r_{ij} & \text{se } r_{ij} \text{ è un coefficiente di similitudine.} \end{cases} \tag{122}$$

Per il metodo di *legame completo* (*complete linkage*, o *farthest neighbor*) vale

$$c_{ij}^{\mathrm{c}} \stackrel{\mathrm{def}}{=} \begin{cases} \max\limits_{\mathbf{d}_i \in C_i,\ \mathbf{d}_j \in C_j} r_{ij} & \text{se } r_{ij} \text{ è un coefficiente di diversità,} \\ \min\limits_{\mathbf{d}_i \in C_i,\ \mathbf{d}_j \in C_j} r_{ij} & \text{se } r_{ij} \text{ è un coefficiente di similitudine.} \end{cases} \tag{123}$$

Un metodo tra questi due estremi calcola la media aritmetica

$$c_{ij}^{\mathrm{a}} \stackrel{\mathrm{def}}{=} \frac{1}{|C_i| + |C_j| - 1} \sum_{\mathbf{d}_i \in C_i,\ \mathbf{d}_j \in C_j} r_{ij}, \tag{124}$$

e fornisce alberi di compattezza intermedia rispetto a quelli forniti dai primi due.

È possibile esprimere queste distanze in un modo che rende l'algoritmo più efficiente, riferendosi direttamente alle quantità w_{ij} della matrice di lavoro. Si supponga di dover aggiornare questa matrice in seguito all'accorpamento di due *cluster* C_p e C_q in un unico *cluster* $C_i = C_p \cup C_q$. È necessario calcolare i coefficienti cofenetici c_{ij} tra C_i e tutti gli altri *cluster*. Si ha

$$c_{ij}^{\mathrm{s}} \stackrel{\mathrm{def}}{=} \begin{cases} \min\{w_{jp}, w_{jq}\} & \text{se } r_{ij} \text{ è un coefficiente di diversità,} \\ \max\{w_{jp}, w_{jq}\} & \text{se } r_{ij} \text{ è un coefficiente di similitudine;} \end{cases} \tag{125}$$

$$c_{ij}^{\mathrm{c}} \stackrel{\mathrm{def}}{=} \begin{cases} \max\{w_{jp}, w_{jq}\} & \text{se } r_{ij} \text{ è un coefficiente di diversità,} \\ \min\{w_{jp}, w_{jq}\} & \text{se } r_{ij} \text{ è un coefficiente di similitudine;} \end{cases} \tag{126}$$

$$c_{ij}^{\mathrm{a}} \stackrel{\mathrm{def}}{=} \frac{|C_p| w_{jp} + |C_q| w_{jq}}{|C_i|}. \tag{127}$$

Un metodo ancora più semplice, proposto da Sokal e Sneath, è

$$c_{ij}^{\mathrm{SS}} \stackrel{\mathrm{def}}{=} \frac{1}{2}(w_{jp} + w_{jq}). \tag{128}$$

Il metodo di Ward utilizza il seguente coefficiente cofenetico

$$c_{ij}^{\mathrm{W}} \stackrel{\mathrm{def}}{=} \frac{|C_i||C_j|}{|C_i| + |C_j|} \|\mathbf{m}_i - \mathbf{m}_j\|_2^2, \tag{129}$$

ove $\mathbf{m}_i$ rappresenta il centroide di C_i. A partire dal coefficiente di somiglianza

$$r_{ij}^{\mathrm{W}} \stackrel{\mathrm{def}}{=} \frac{1}{2}\|\mathbf{d}_i - \mathbf{d}_j\|_2^2, \tag{130}$$

è possibile anche in questo caso aggiornare la matrice di lavoro ricorsivamente, ponendo

$$c_{ij}^{\mathrm{W}} \stackrel{\mathrm{def}}{=} \frac{(|C_p| + |C_j|) w_{jp} + (|C_q| + |C_j|) w_{jp} - |C_j| w_{pq}}{|C_i| + |C_j|}. \tag{131}$$

Si dimostra che questo metodo rende minima la (113) ad ogni passo. Nel caso si utilizzi come coefficiente cofenetico il quadrato della distanza euclidea tra i centroidi, la formula ricorsiva diventa

$$c_{ij}^{\mathrm{GB}} \stackrel{\mathrm{def}}{=} \frac{|C_p|}{|C_i|} w_{jp} + \frac{|C_q|}{|C_i|} w_{jq} - \frac{|C_p||C_q|}{|C_i|^2} w_{pq}. \tag{132}$$

Si può pensare di adottare queste ultime formule ricorsive anche con coefficienti di somiglianza diversi da quelli per cui sono stati studiati, con il rischio di ottenere dendrogrammi con *inversioni* (*reversal*), che presentano cioè fusioni a livelli inferiori rispetto a quelli dei *cluster* componenti. Questo fenomeno dà origine a dendrogrammi di difficile interpretazione.

Un altro effetto collaterale dell'analisi dei *cluster* è detto *concatenazione* (*chaining*), e si verifica quando un grosso *cluster* si forma accorpando un'osservazione alla volta, spostando progressivamente il proprio centro di massa lontano dagli elementi che lo hanno originato. Questo fenomeno è più o meno accentuato a seconda del particolare metodo di raggruppamento adottato.

Riferimenti bibliografici

[1] K. E. Atkinson. *An Introduction to Numerical Analysis.* Seconda Edizione, John Wiley & Sons, New York, 1989.

[2] C. Burges. "A tutorial on Support Vector Machines for pattern recognition." *Data Mining and Knowledge Discovery* **2**(2), 1998.

[3] E. Dedò, A. Varisco. *Algebra Lineare—Elementi ed esercizi.* CittàStudi, Milano, 1988.

[4] L. Devroye, L. Györfi, G. Lugosi. *A Probabilistic Theory of Pattern Recognition.* Springer-Verlag, New York, 1996.

[5] A. J. Dobson. *An Introduction to Generalized Linear Models.* Chapman and Hall, New York, 1990.

[6] R. O. Duda, P. E. Hart. *Pattern Classification and Scene Analysis.* John Wiley & Sons, New York, 1973.

[7] R. A. Fisher. "The use of multiple measurements in taxonomic problems." *Ann. Eugen.* **7**, 179–188, 1936.

[8] B. Flury. *Common Principal Components and Related Multivariate Models.* John Wiley & Sons, New York, 1988.

[9] B. Flury. *A First Course in Multivariate Statistics.* Springer-Verlag, New York, 1997.
Tabelle di dati e *routine* di classificazione disponibili presso `ftp://129.79.94.6/pub/flury`.

[10] J. H. Friedman. "Regularized Discriminant Analysis." *Journal of American Statistical Association* **84**, 165–175, 1989.

[11] J. A. Hartigan. *Clustering Algorithms.* John Wiley & Sons, New York, 1975.

[12] D. M. Hawkins. "A new test for multivariate normality and homoscedasticity." *Technometrics* **23**, 105–110, 1981.

[13] H. Hotelling. "Relations between two sets of variates." *Biometrika* **28**, 321–377, 1936.

[14] H. Hotelling. "A generalized T test and measure of multivariate dispersion." Proceedings of the Second Berkeley Symposium, Berkeley. University of California Press, 23–41, 1951.

[15] P. C. Mahalanobis. "On the generalized distance in statistics." *Proceedings of the National Institute of Sciences India* **2**, 49–55, 1936.

[16] V. Maniezzo. *Algoritmi di apprendimento automatico.* Esculapio, Bologna, 1995.

[17] K. V. Mardia. "Applications of some measures of multivariate skewness and kurtosis in testing normality and robustness studies." *Sankhyā B* **36**, 115–128, 1974.

[18] G. J. McLachlan. *Discriminant Analysis and Statistical Pattern Recognition.* John Wiley & Sons, New York, 1992.

[19] A. M. Mood, F. A. Graybill, D. C. Boes. *Introduction to the Theory of Statistics.* McGraw-Hill, Inc., 1974.

[20] K. Pearson. "On lines and planes of closest fit to systems of points in space." *Philosophical Magazine* ser. 6, **2**, 559–572, 1901.

[21] C. H. Romesburg. *Cluster Analysis for Researchers.* Robert E. Krieger Publishing, Malabar, Florida, 1990.

[22] H. Späth. *Cluster Analysis Algorithms.* Ellis Horwood Ltd., Chichester, 1980.