

Capitolo 8

Materiale didattico integrativo

Analisi discriminante lineare

1. La distribuzione normale multivariata

Si consideri una variabile casuale normale multivariata \mathbf{X} a p dimensioni e due popolazioni G_0 e G_1 . Si ponga che le distribuzioni della variabile casuale \mathbf{X} (vettore colonna) differiscano per le due popolazioni solo per il vettore delle medie:

$$\mathbf{X}/G_0 \sim MN(\boldsymbol{\mu}_0; \boldsymbol{\Sigma}), \quad \mathbf{X}/G_1 \sim MN(\boldsymbol{\mu}_1; \boldsymbol{\Sigma})$$

dove:

$\boldsymbol{\mu}_k$ è il vettore (colonna) della media della popolazione G_k , ($k=0,1$) e $\boldsymbol{\Sigma}$ è la matrice di varianza e covarianza comune ai due gruppi (le due popolazioni hanno la stessa matrice di varianza e covarianza). Inoltre:

$$f(\mathbf{x} | G_k) = \frac{1}{(2\pi)^{p/2} |\boldsymbol{\Sigma}|^{1/2}} \exp \left\{ -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}_k)' \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}_k) \right\} \quad (1)$$

è la funzione di densità congiunta delle p variabili nella popolazione G_k , $k=0,1$.

2. La distribuzione normale multivariata e le regole decisionali

Si consideri una determinazione \mathbf{x}_i della v.c. \mathbf{X} , corrispondente all'unità di osservazione U_i . La regola della verosimiglianza massima ci dice che (v. formula (20) Paragrafo 8.6.3 del testo):

$$\text{se } \ln f(\mathbf{x}_i | G_1) - \ln f(\mathbf{x}_i | G_0) > 0 \quad U_i \text{ va in } G_1 \quad (2)$$

dove \ln indica l'operatore logaritmo naturale.

Sostituendo la (1) nella (2), si ricava:

$$\text{se } -\frac{1}{2} (\mathbf{x}_i - \boldsymbol{\mu}_1)' \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_1) > -\frac{1}{2} (\mathbf{x}_i - \boldsymbol{\mu}_0)' \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_0) \quad U_i \text{ va in } G_1$$

moltiplicando i due membri per -2 si ha:

$$\text{se } (\mathbf{x}_i - \boldsymbol{\mu}_1)' \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_1) < (\mathbf{x}_i - \boldsymbol{\mu}_0)' \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_0) \quad U_i \text{ va in } G_1 \quad (3)$$

Svolgendo le varie operazioni si ricava:

se

$$\cancel{\mathbf{x}_i' \boldsymbol{\Sigma}^{-1} \mathbf{x}_i} - \cancel{\mathbf{x}_i' \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_1} - \cancel{\boldsymbol{\mu}_1' \boldsymbol{\Sigma}^{-1} \mathbf{x}_i} + \boldsymbol{\mu}_1' \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_1 < \cancel{\mathbf{x}_i' \boldsymbol{\Sigma}^{-1} \mathbf{x}_i} - \cancel{\mathbf{x}_i' \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_0} - \cancel{\boldsymbol{\mu}_0' \boldsymbol{\Sigma}^{-1} \mathbf{x}_i} + \boldsymbol{\mu}_0' \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_0 \quad U_i \text{ va in } G_1$$

$$\text{se } -2\mathbf{x}_i' \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_1 + \boldsymbol{\mu}_1' \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_1 < -2\mathbf{x}_i' \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_0 + \boldsymbol{\mu}_0' \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_0 \quad U_i \text{ va in } G_1$$

poiché $\mathbf{x}_i' \Sigma^{-1} \boldsymbol{\mu}_k = \boldsymbol{\mu}_k' \Sigma^{-1} \mathbf{x}_i$ essendo uno scalare ($k=0,1$).

Dividendo i due membri per -2 si ha:

$$\text{se } \mathbf{x}_i' \Sigma^{-1} \boldsymbol{\mu}_1 - \frac{1}{2} \boldsymbol{\mu}_1' \Sigma^{-1} \boldsymbol{\mu}_1 > \mathbf{x}_i' \Sigma^{-1} \boldsymbol{\mu}_0 - \frac{1}{2} \boldsymbol{\mu}_0' \Sigma^{-1} \boldsymbol{\mu}_0 \quad U_i \text{ va in } G_1$$

$$\text{se } \mathbf{x}_i' (\Sigma^{-1} \boldsymbol{\mu}_1 - \Sigma^{-1} \boldsymbol{\mu}_0) - \frac{1}{2} \boldsymbol{\mu}_1' \Sigma^{-1} \boldsymbol{\mu}_1 + \frac{1}{2} \boldsymbol{\mu}_0' \Sigma^{-1} \boldsymbol{\mu}_0 > 0 \quad U_i \text{ va in } G_1 \quad (4)$$

Si ponga $\mathbf{a} = (\Sigma^{-1} \boldsymbol{\mu}_1 - \Sigma^{-1} \boldsymbol{\mu}_0)$; \mathbf{a} è un vettore colonna a p dimensioni e quindi

$$\mathbf{x}_i' \mathbf{a} = \mathbf{x}_i' (\Sigma^{-1} \boldsymbol{\mu}_1 - \Sigma^{-1} \boldsymbol{\mu}_0)$$

è uno scalare.

Si ponga inoltre $b = -\frac{1}{2} \boldsymbol{\mu}_1' \Sigma^{-1} \boldsymbol{\mu}_1 + \frac{1}{2} \boldsymbol{\mu}_0' \Sigma^{-1} \boldsymbol{\mu}_0$ che è anch'esso uno scalare.

La regola decisionale in (4) diventa pertanto la seguente:

$$\text{se } \mathbf{x}_i' \mathbf{a} + b > 0 \quad U_i \text{ va in } G_1 \quad (5)$$

E' chiaro che $\mathbf{x}_i' \mathbf{a} + b$ è la funzione discriminante che produce lo score dell'unità i -esima e 0 il cut-off corrispondente. Si noti che la funzione discriminante è una **combinazione lineare** delle variabili originarie (**funzione discriminante lineare**).

La stima delle medie mediante le medie di gruppo $\bar{\mathbf{x}}_1$ e $\bar{\mathbf{x}}_0$ e la stima di Σ mediante la matrice di varianza e covarianza interna \mathbf{W} (detta anche *pooled*) ci conduce a:

$\hat{\mathbf{a}} = (\mathbf{W}^{-1} \bar{\mathbf{x}}_1 - \mathbf{W}^{-1} \bar{\mathbf{x}}_0)$ è un vettore riga a p dimensioni che contiene i coefficienti stimati della funzione discriminante (i coefficienti da moltiplicare per le variabili, per calcolare lo score discriminante);

$\hat{b} = -\frac{1}{2} \underline{\boldsymbol{\mu}}_1' \mathbf{W}^{-1} \underline{\boldsymbol{\mu}}_1 + \frac{1}{2} \underline{\boldsymbol{\mu}}_0' \mathbf{W}^{-1} \underline{\boldsymbol{\mu}}_0$ è la stima della costante della funzione discriminante.

Se i coefficienti sono applicati alle variabili espresse come scarti dalla media, la costante della funzione discriminante è zero.

Consideriamo ora la regola della massima probabilità a posteriori, indicando con ϕ_0 e ϕ_1 le probabilità a priori associate a gruppi G_0 e G_1 .

$$\text{se } \ln f(\mathbf{x}_i | G_1) + \ln \phi_1 - \ln f(\mathbf{x}_i | G_0) - \ln \phi_0 > 0 \quad U_i \text{ va in } G_1$$

$$\text{se } \ln f(\mathbf{x}_i | G_1) - \ln f(\mathbf{x}_i | G_0) > \ln \phi_0 - \ln \phi_1 \quad U_i \text{ va in } G_1 \quad (6)$$

Confrontando la (6) con la (2) si vede che cambia solo il cut-off (ovvero la costante della funzione discriminante, se si usa il cut-off zero).

Nella stima dei parametri della distribuzione normale si procede nel modo visto prima (stimando le medie di gruppo e la matrice di varianza e covarianza *pooled* sui dati campionari), pertanto non cambiano i coefficienti della funzione discriminante da applicare alle variabili ma cambia il cut-off della regola (ovvero la costante della funzione discriminante, se si usa il cut-off zero).

Un'ultima considerazione riguarda l'ipotesi di normalità. Questa non è verificata in generale dalle distribuzioni empiriche degli indici di bilancio. Tuttavia l'impostazione della discriminante lineare consente di proporre una **interpretazione geometrica** della regola che si ricava dall'ipotesi di multinormalità (v. anche l'Esempio 8.6 del testo). Infatti l'espressione (3) può essere letta come:

se l'unità U_i è meno distante dal centroide del gruppo G_1 questa viene assegnata al gruppo G_1 , in base da una distanza generalizzata che tiene conto delle varianze e covarianze fra le variabili discriminanti.