Analiza discriminantă

Sub numele de analiză discriminantă, sunt reunite o serie de metode explicative, descriptive şi predictive, destinate studierii unei populaţii împărţite în clase. Fiecare individ este caracterizat printr-un ansamblu de variabile independente predictor şi o variabilă calitativă identificând clasa din care face parte. Mulțimea de indivizi se împarte în două submulțimi:

- *eşantionul de bază*, pentru care se cunoaşte valoare variabilei calitative, deci încadrarea pe clase a indivizilor;
- *eşantionul neinvestigat*, pentru care nu se cunoaşte încadrarea indivizilor în clase, deci nici valoarea variabilei calitative.

Analiza discriminantă urmăreşte pe de o parte să identifice regulile pe baza cărora toţi indivizii din mulţimea investigată să poată fi încadraţi în clase, iar pe de altă parte să reducă numărul de variabile necesare realizării împărţirii în clase (discriminării). Primul aspect urmărit scoate în evidenţă caracterul predictiv, decizional al analizei discriminante pe când al doilea aspect reliefează caracterul ei descriptiv. Analiza discriminantă se aplică în mod frecvent în variate domenii în care recunoașterea formelor este utilă (previzionarea comportamentelor solicitanților de credit, detecția fraudelor , diagnostic medical, prognoze meteo, recunoaștere vocală etc.

Notații

Matricea de observații: $X = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1m} \\ x_{21} & x_{22} & \dots & x_{2m} \\ \dots & & & \\ x_{n1} & x_{n2} & \dots & x_{nm} \end{bmatrix}$,

unde n este numărul de observații, m - numărul de variabile predictor (variabile independente).

Variabila discriminantă: $Y = \begin{bmatrix} y_1 \\ \dots \\ y_n \end{bmatrix}$. Este o variabilă calitativă. O valoare y_i, i =1,n reprezintă clasa din care face parte observația i.

Vectorii observațiilor sunt notați cu w_i, i=1,n. Un vector X_i este linia i din matricea X.

Vectorii de variabile sunt notați cu X_j, j =1, m și sunt constituiți din coloanele matricei X.

Matricea centrilor de grupă: $G = \begin{bmatrix} g_{11} & g_{12} & \dots & g_{1m} \\ g_{21} & g_{22} & \dots & g_{2m} \\ \dots & & & \\ g_{q1} & g_{q2} & \dots & g_{qm} \end{bmatrix}$, unde q este numărul de grupe. O valoare g_{kj} reprezintă media variabilei predictor j pentru grupa k. Vectorii centrilor de grupă sunt notați cu G_k, k=1,q, $G_k = \begin{bmatrix} g_{k1} \\ \dots \\ g_{km} \end{bmatrix}$. Media generală: $\overline{X} = \begin{bmatrix} \bar{x}_1 \\ \dots \\ \bar{x}_m \end{bmatrix}$.

Maricea diagonală a frecvențelor grupelor: $D_G = \begin{bmatrix} n_1 & 0 & \dots & 0 \\ 0 & n_2 & \dots & 0 \\ \dots & & & \\ 0 & 0 & \dots & n_q \end{bmatrix}$.

Indicatori de variabilitate și împrăștiere

Discriminarea între grupe se realizează cu ajutorul indicatorilor de variabilitate și împrăștiere.

1. ***Matricele de împrăștiere*** (*sum of square and cross product*) reflectă împrăștierea la nivelul întregii colectivități (*SST*), în interiorul grupelor (*SSW*) și împrăștierea grupelor intre ele (*SSB*).

SST este matricea de împrăștiere la nivelul întregii colectivități și arată gradul de împrăștiere în jurul mediei generale. Termenul general al matricei este:

$$SST_{jl} = \sum_{i=1}^{n}\left(x_{ij} - \bar{x}_j\right)\left(x_{il} - \bar{x}_l\right) = \sum_{k=1}^{q}\sum_{i\in k}\left(x_{ij} - \bar{x}_j\right)\left(x_{il} - \bar{x}_l\right) = \sum_{k=1}^{q}\sum_{i\in k}\left(x_{ij} - g_{kj} + g_{kj} - \bar{x}_j\right)\left(x_{il} - g_{kl} + g_{kl} - \bar{x}_l\right) =$$

$$= \sum_{k=1}^{q}\sum_{i\in k}\left(x_{ij} - g_{kj}\right)\left(x_{il} - g_{kl}\right) + \sum_{k=1}^{q}\sum_{i\in k}\left(g_{kj} - \bar{x}_j\right)\left(g_{kl} - \bar{x}_l\right) + \sum_{k=1}^{q}\sum_{i\in k}\left(x_{ij} - g_{kj}\right)\left(g_{kl} - \bar{x}_l\right) + \sum_{k=1}^{q}\sum_{i\in k}\left(g_{kj} - \bar{x}_j\right)\left(x_{il} - g_{kl}\right) =$$

$$= \sum_{k=1}^{q}\sum_{i\in k}\left(x_{ij} - g_{kj}\right)\left(x_{il} - g_{kl}\right) + \sum_{k=1}^{q} n_k\left(g_{kj} - \bar{x}_j\right)\left(g_{kl} - \bar{x}_l\right) + \sum_{k=1}^{q}\left(g_{kl} - \bar{x}_l\right)\sum_{i\in k}\left(x_{ij} - g_{kj}\right) + \sum_{k=1}^{q}\left(x_{il} - g_{kl}\right)\sum_{i\in k}\left(g_{kj} - \bar{x}_j\right)$$

Prima sumă, $\sum_{k=1}^{q}\sum_{i\in k}\left(x_{ij}-g_{kj}\right)\left(x_{il}-g_{kl}\right)$, reprezintă termenul general al matricei de împrăștiere în interiorul grupelor, *SSW*. A doua sumă,

$\sum_{k=1}^{q} n_k\left(g_{kj}-\bar{x}_j\right)\left(g_{kl}-\bar{x}_l\right)$, este termenul general al matricei împrăștierii intre grupe, *SSB*. Sumele trei și patru au valoarea 0 deoarece

sumele simple ale abaterilor față de mediile grupelor sunt 0, $\sum_{i\in k}\left(x_{ij}-g_{kj}\right)=0$, $\sum_{i\in k}\left(x_{il}-g_{kl}\right)=0$. Deci:

$SST_{jl} = SSW_{jl} + SSB_{jl}$, j=1,m, l=1,m.
La nivel matriceal: $SST = SSW + SSB$.

Dacă ținem cont de gradele de libertate, matricele de împrăștiere sunt calculate ca valori medii astfel: $MST = \frac{SST}{n-1}$, MSW = $\frac{SSW}{n-q}$,

MSB = $\frac{SSB}{q-1}$.

2. *Matricele de covarianță*. Termenii generali ai matricelor de covarianță diferă de cei ai matricelor de împrăștiere prin faptul că se calculează ca valori medii.

$T_{jl} = \frac{1}{n}\sum_{i=1}^{n}\left(x_{ij}-\bar{x}_j\right)\left(x_{il}-\bar{x}_l\right)$, j=1,m, l=1,m - covarianța totală

$W_{jl} = \sum_{k=1}^{q}\frac{n_k}{n}\cdot\frac{1}{n_k}\sum_{i\in k}\left(x_{ij}-g_{kj}\right)\left(x_{il}-g_{kl}\right)$, j=1,m, l=1,m - covarianța intra-gupă (în interiorul grupelor)

$B_{jl} = \sum_{k=1}^{q}\frac{n_k}{n}\left(g_{kj}-\bar{x}_j\right)\left(g_{kl}-\bar{x}_l\right)$, j=1,m, l=1,m - covarianța inter-gupă (intre grupe).

Relația intre termeni este aceeași ca și la matricele de împrăștiere:
$T_{jl} = W_{jl} + B_{jl}$, j=1,m, l=1,m.
La nivel matriceal:
$T = W + B$.

3. *Varianța totală*. Varianța totală este reflectată de valorile de pe diagonalele principale ale matricei de covarianță și se calculează ca suma variațelor celor m variabile:
VT = Trace(T) - varianța totală generală,
VW = Trace(W) - varianța totală intra-grupă,
VB = Trace(B) - varianța totală inter-grupă,
Varianța totală se poate calcula și pe baza matricelor de împrăștiere:
VT = Trace(SST), VW = Trace(SSW), VB = Trace(SSB).

4. *Varianța generalizată*. Varianța generalizată se calculează ca determinant al matricelor de covarianță sau de împrăștiere:
$VGT = |T|$, $VGW = |W|$, VGB = $|B|$, sau $VGT = |SST|$, $VGW = |SSW|$, VGB = $|SSB|$.

Semnificația modelului. Teste statistice

Testarea modelului se face în două etape:
- un test F (Fisher) bazat pe statistica Wilks care arată dacă ansamblul de variabile predictor poate face discriminarea pe grupe a instanțelor
- teste statistice individuale pentru fiecare variabilă predictor prin care se decide dacă o variabilă poate fi sau nu un bun predictor

Testul F global

H0: $g_1 = g_2 = \ldots = g_q$

H1: există două grupe i, k, astfel încât $g_i \neq g_k$

Se calculează un indicator $\Lambda = \frac{|SSB|}{|SSB+SSW|}$. Cu cât valoarea lui Λ este mai mare cu atât șansele ca ipoteza H0 să fie respinsă

sunt mai mari. Statistica testului este o valoare F calculată astfel: $F = \frac{1-\Lambda^{1/b}}{\Lambda^{1/b}}\frac{ab-c}{m(q-1)}$, unde a, b și c se calculează astfel:

$$a = n-q-\frac{m-q+2}{2}, b = \begin{cases}\sqrt{\frac{m^2(q-1)-4}{m^2+(q-1)^2-5}} & \text{dacă } m^2+(q-1)^2-5>0\\ 1 & \text{daca } m^2+(q-1)^2-5\le 0\end{cases}, c = \frac{m(q-1)-2}{2}.$$

Dacă $F < F^{Critic}_{m(q-1),ab-c;\alpha}$ ipoteza nulă este respinsă cu un nivel de încredere 1-α.

Teste statistice la nivel de variabile predictor

O variabilă predictor poate fi considerată bun predictor dacă poate separa cât mai clar grupele. Deci raportul dintre varianța medie intergrupe și varianța medie intragrupe este cât mai mare. Variațele sunt preluate din matricele de împrăștiere/covarianță de pe diagonala principală. Fiind vorba de un raport între două variațe, testul aplicat este testul *F*. Astfel pentru variabila predictor X_j, ipotezele nulă și alternativă sunt:

H0: $g_{1j} = g_{2j} = ... = g_{qj}$

H1: $\exists\ k,i$ două grupe astfel încât $g_{kj} \neq g_{ij}$

Statistica testului: $F_j = \dfrac{MSB_{jj}}{MSW_{jj}}$.

Valoarea critică pentru q-1 și n-q grade de libertate și un prag de semnificație α este $F^{Critic}_{q-1;n-q;\alpha}$.

Dacă $F_j > F^{Critic}_{q-1;n-q;\alpha}$ ipoteza nulă este respinsă cu un nivel de încredere 1-α.