Analiza discriminantă liniară (Linear Discriminant Analysis). Funcțiile Fisher

Mai este cunoscută și sub denumirea de analiză factorială discriminantă sau analiza canonică discriminantă. Perspectiva analizei este una asemănătoare celei de la analiza în componente principale. Analiza factorială discriminantă determină doi variabile predictor, numite variabile discriminante, astfel încât instanțele să fie cât mai bine separate după aceste variabile. Variabilele discriminante sunt, ca și la analiza în componente principale, combinații liniare de variabilele inițiale (din X) și necorelate două câte două.

Un criteriu natural de determinare a variabilelor discriminante este acela de a maximiza coeziunea claselor cu ajutorul varianței intra-clase și a varianței inter-clase. Raportul dintre varianța inter-clase a variabilei și varianța totală (sau varianța intra-clase) să fie cât mai mare.

Presupunem că matricea de observații, X, este centrată. Prima variabilă discriminantă se determină astfel:

$z_1 = \begin{bmatrix} z_{11} \\ z_{21} \\ \dots \\ z_{n1} \end{bmatrix} = X \cdot u_1$, unde u_1 este primul factor discriminant (coeficienții combinației liniare), $u_1 = \begin{bmatrix} u_{11} \\ u_{21} \\ \dots \\ u_{m1} \end{bmatrix}$. Centrii de grupă pentru variabila z_1 sunt: $\overline{z_1} = \begin{bmatrix} \overline{z}_{11} \\ \overline{z}_{21} \\ \dots \\ \overline{z}_{q1} \end{bmatrix} = G \cdot u_1$.

Dacă X nu este centrat, $z_1 = u_{01} + X \cdot u_1$.

Variabilele discriminante pot fi văzute și ca funcții discriminante (Figura 1), denumite și funcții Fisher, prin care se poate face o separare cât mai bună între instanțe. z^1 este un hiperplan de ecuație $u_{01} + X \cdot u_1$ care desparte instanțele (dreapta roșie continuă din figura 1). u_1 este normala la hiperplan sau axa discriminantă (dreapta roșie întreruptă din figura 1) care trece prin origine, în cazul în care datele sunt centrate, sau prin centrul de greutate (punctul negru din figura 1) dacă datele nu sunt centrate. Dacă datele sunt centrate, termenul liber u_{01} este 0. Axa u_1 este astfel aleasă încât să separe cât mai bine grupele, adică distanțele dintre proiecțiile centrilor de grupe pe axă să fie cât mai mari. Centrii de grupă sunt marcați cu roșu în figura 1.

Varianța totală a variabilei z_1 scrisă matriceal este: $VT_1 = \frac{1}{n}(z_1)^t z_1 = \frac{1}{n}(X \cdot u_1)^t X \cdot u_1 = \frac{1}{n}(u_1)^t X^t X \cdot u_1$ (simbolul t la exponent înseamnă transpus - $(u_1)^t$ este un vector linie, transpusa vectorului u_1)

Varianța inter-clasă a variabilei z_1 este: $VB_1 = \sum_{k=1}^{q} \frac{n_k}{n} (\overline{z}_{k1})^2$, unde $\overline{z}_{k1}$ este media variabilei pentru grupa k.

Matriceal varianța inter-clasă se poate scrie: $VB_1 = \frac{1}{n}\left(\overline{z}_1\right)^t D_G \cdot \overline{z}_1 = \frac{1}{n}(G \cdot u_1)^t D_G G \cdot u_1 = \frac{1}{n}(u_1)^t G^t D_G G \cdot u_1$.

Varianța intra-clasă este: $VW_1 = \sum_{k=1}^{q} \frac{n_k}{n} \cdot \frac{1}{n_k} \sum_{i \in k} \left(z_{i1} - \overline{z}_{k1}\right)^2$.

Relația dintre varianțe: $VT_1 = VB_1 + VW_1$.

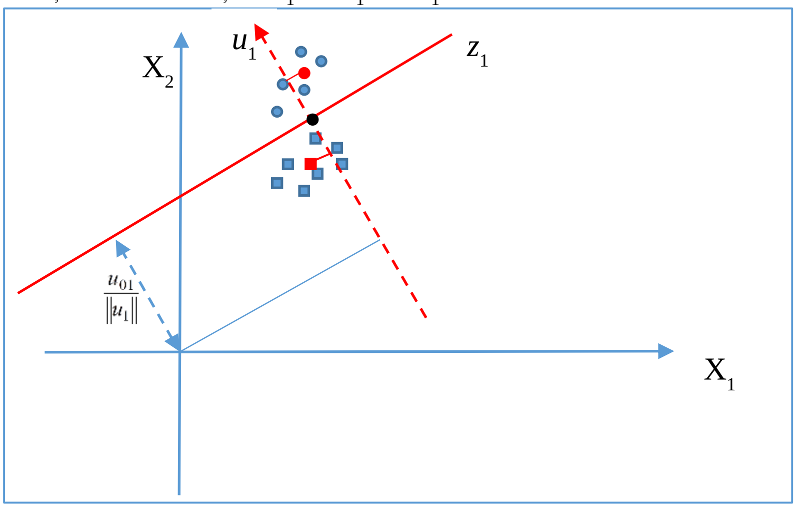

Figura 1. Funcțiile discriminante Fisher. Cazul bidimensional cu două clase

Pentru o valoare a varianței totale dată, o variabilă va discrimina cu atât mai bine grupele cu cât sunt îndeplinite mai bine condițiile:

- indivizii aparținând unei clase au valori cât mai apropiate cu putință, adică varianța intra-clase este cât mai mică;
- mediile claselor sunt cât mai distanțate, adică varianța inter-clase este mai mare.

Maximizarea distanței dintre proiecțiile centrilor pe axa discriminantă u_1 este echivalentă cu maximizarea raportului $\frac{VB_1}{VT_1} = \frac{\frac{1}{n}(u_1)^t G^t D_G G \cdot u_1}{\frac{1}{n}(u_1)^t X^t X \cdot u_1}$ pentru variabila discriminantă z_1.

Problema de optim care se pune este:

$$\begin{cases} \underset{u_1}{Maxim} \dfrac{\frac{1}{n}(u_1)^t G^t D_G G \cdot u_1}{\frac{1}{n}(u_1)^t X^t X \cdot u_1}, \\ u_1^t u_1 = 1 \end{cases}$$

unde u_1 este factorul discriminant și vector axă (versor de lungime 1), $B = \frac{1}{n} G^t D_G G$ este matricea de covarianță inter-clase iar $T = \frac{1}{n} X^t X$ este matricea de covarianță totală.

Soluția problemei, u_1, obținută prin rezolvarea problemei de optim, este vectorul propriu al matricei $T^{-1}B$ corespunzător celei mai mari valori proprii. Această valoare proprie este chiar α_1, raportul dintre varianța inter-clase și varianța totală:

$$T^{-1}B \cdot u_1 = \alpha_1 u_1 \Rightarrow B \cdot u_1 = \alpha_1 T \cdot u_1 \Rightarrow u_1^t B \cdot u_1 = \alpha_1 u_1^t T \cdot u_1 \Rightarrow \quad \alpha_1 = \frac{u_1^t B \cdot u_1}{u_1^t T \cdot u_1} = \frac{\frac{1}{n} u_1^t G^t D_G G \cdot u_1}{\frac{1}{n} u_1^t X^t X \cdot u_1} = \frac{VB_1}{VT_1}$$

Următoarele variabile discriminante se obțin la fel ca prima variabilă, având condiții suplimentare: să fie absolut necorelate cu celelalte variabile discriminante, $\frac{1}{n}(z_k)^t z_j = 0$, $j = 1,k-1$, deci axele lor să fie ortogonale. Problema de optim devine:

$$\begin{cases} \underset{u^k}{Maxim} \dfrac{\frac{1}{n}(u_k)^t G^t D_G G \cdot u_k}{\frac{1}{n}(u_k)^t X^t X \cdot u_k} \\ u_k^t u_k = 1, \quad u_k^t u_j = 0, \quad j = 1, k-1 \end{cases}$$

Soluția problemei, factorul u_k, este vectorul propriu al matricei $T^{-1}B$ corespunzător valorii proprii de ordin k (în sensul descrescător al valorilor) α_k.

Numărul de variabile discriminante este dat de numărul valorilor proprii nenule ale matricei $T^{-1}B$ (vezi mai jos analiza canonică discriminantă). Acest număr este r = minim (m, q-1).

Soluțiile modelului:

$T^{-1}B \cdot u_k = \alpha_k \cdot u_k$, $k = 1, r$

$z_k = X \cdot u_k$, $k = 1, r$

Puterea de discriminare a unei variabile discriminante se poate calcula ca raport între varianța medie inter-clase și varianța medie intra-clase. Dacă notăm cu λ_k puterea de discriminare a variabilei discriminante z_k, obținem:

$$\lambda_k = \frac{\frac{VB_k}{q-1}}{\frac{VW_k}{n-q}} = \frac{n-q}{q-1} \cdot \frac{VB_k}{VT_k - VB_k} = \frac{n-q}{q-1} \cdot \frac{VB_k}{VT_k \cdot \left(1 - \frac{VB_k}{VT_k}\right)} = \frac{n-q}{q-1} \cdot \frac{\alpha_k}{1-\alpha_k}.$$

Funțiile de clasificare

Axele discriminante sunt determinate prin calculul vectorilor proprii ai matricei $T^{-1}B$, unde $B = \frac{1}{n} G^t D_G G$, deci $T^{-1} \frac{1}{n} G^t D_G G$. După cum s-a văzut, în analiza în componente principale pe un tabel de observații X, axele principale (axele componentelor principale) se calculează ca vectori proprii normați ai matricei $\frac{1}{n} X^t X$. Dacă distanța dintre instanțele descrise în tabelul X nu ar fi una euclidiană, ci o distanță oarecare, de metrică M, atunci axele principale s-ar calcula ca vectori proprii normați ai matricei $\frac{1}{n} M X^t X$. Mai mult, dacă ponderea instanțelor nu ar fi egală, ci am folosii o matrice diagonală a ponderilor, D (matrice care ar avea pe diagonală ponderile indiviziilor - vezi analiza în componente principale ponderată), axele principale s-ar calcula ca vectori proprii ai matricei $\frac{1}{n}MX^tDX$.

Prin urmare, analiza factorială discriminantă este o analiză în componente principale în care metrica M este inversa matricei de covarianță totală, T^{-1}, ponderile sunt date de matricea DG (matricea diagonală a frecvențelor grupelor), iar tabelul de observații este G, tabelul centrilor de grupă. Deci este vorba de o analiză în componente principale ponderată a centrilor de grupe, cu o metrică specială, metrica $M = T^{-1}$ (metrica Mahalanobis). Axele discriminante sunt axele principale ale acestui tip particular de analiză în componente principale.

Analiza factorială discriminantă se utilizează atunci când numărul de variabile predictor este foarte mare (sau suficient de mare), iar aplicarea modelului liniar de discriminare devine greoaie. În această situație se utilizează pentru clasificare variabilele discriminante, numite și funcțiile discriminante Fisher, care sunt semnificative ca putere de discriminare. Testul de semnificație este un test F, identic celui făcut variabilelor predictor (vezi cursul 9), valorile calculate fiind valorile λ_k (vezi cursul 10).

În situația în care modelul nu este suficient de complex, numărul de predictori este mic, clasificarea se face pe baza distanțelor Mahalanobis (distanța de metrica T^{-1}). Adică o instanță oarecare este clasificată în acea grupă, de care este mai apropiată ca distanță. Acest principiu este standardizat sub forma unor ***funcții de clasificare***, prin care se furnizează niște ***scoruri***, scoruri folosite mai apoi pentru clasificare.

Funcțiile de clasificare sunt determinate deci, cu ajutorul distanțelor Mahalanobis. Distanța Mahalanobis dintre două instanțe a și b este:

$$d^2(a,b) = (a-b)^t T^{-1}(a-b).$$

Să luăm în considerare cazul simplificat în care avem două grupe, una cu centrul g_1 și una cu centrul g_2. O instanță oaecare $x = \begin{bmatrix} x_1 \\ \dots \\ x_m \end{bmatrix}$ este clasificată în prima grupă dacă $d^2(x, g_1) < d^2(x, g_2)$ sau în a doua grupă dacă $d^2(x, g_1) \geq d^2(x, g_2)$.

Înlocuind în prima inegalitate cu relația de calcul a distanței obținem:

$$(x-g_1)^t T^{-1}(x-g_1) < (x-g_2)^t T^{-1}(x-g_2)$$
$$(x^t - g_1^t)T^{-1}(x-g_1) < (x^t - g_2^t)T^{-1}(x-g_2)$$
$$x^tT^{-1}x - g_1^tT^{-1}x - x^tT^{-1}g_1 + g_1^tT^{-1}g_1 < x^tT^{-1}x - g_2^tT^{-1}x - x^tT^{-1}g_2 + g_2^tT^{-1}g_2$$

Simplificând și ținând cont că $x^tT^{-1}g_1 = g_1^tT^{-1}x$ și $x^tT^{-1}g_2 = g_2^tT^{-1}x$, matricea T^{-1} fiind simetrică, se obține:

$$-2 \cdot g_1^tT^{-1}x + g_1^tT^{-1}g_1 < -2 \cdot g_2^tT^{-1}x + g_2^tT^{-1}g_2$$
$$\Rightarrow 2 \cdot g_1^tT^{-1}x - g_1^tT^{-1}g_1 > 2 \cdot g_2^tT^{-1}x - g_2^tT^{-1}g_2$$

Dacă se notează cu $F_1 = g_1^tT^{-1}$ și $F_2 = g_2^tT^{-1}$ și se simplifică cu 2 inegalitatea anterioară, rezultă că instanța x va fi clasificată în prima clasă dacă:

$$(F_1)^t \cdot x - \frac{1}{2}(F_1)^t \cdot g_1 \geq (F_2)^t \cdot x - \frac{1}{2}(F_2)^t \cdot g_2$$

și în cea de a doua clasă dacă:

$$(F_1)^t \cdot x - \frac{1}{2}(F_1)^t \cdot g_1 < (F_2)^t \cdot x - \frac{1}{2}(F_2)^t \cdot g_2.$$

Cu alte cuvinte instanța x va fi clasificată în acea clasă pentru care cantitatea $(F_k)^t \cdot x - \frac{1}{2}(F_k)^t \cdot g_k$, $k=\overline{1,2}$ este mai mare.

Pentru cazul general cu q clase, instanța x va fi clasificată în acea clasă pentru care cantitatea:

$$(F_k)^t \cdot x - \frac{1}{2}(F_k)^t \cdot g_k,\ k=\overline{1,q}$$

este maximă, $F_k = g_k^tT^{-1}$.

Vectorii F_k se numesc funcții de clasificare, iar valorile calculate cu ajutorul lor, scoruri.

Deci o instanță se clasifică în acea clasă pentru care obține scorul maxim.

Relațiile de calcul pentru scoruri, prin intermediul funcțiilor de clasificare sunt:

$S_k(x) = F_{k0} + F_{k1} \cdot x_1 + F_{k2} \cdot x_2 + \dots + F_{km} \cdot x_m$, $k = 1,q$,

unde:

$F_k = \begin{bmatrix} F_{k1} \\ \dots \\ F_{km} \end{bmatrix} = g_k^tT^{-1}$, $x = \begin{bmatrix} x_1 \\ \dots \\ x_m \end{bmatrix}$ este instanța clasificată, iar $F_{k0} = -\frac{1}{2}(F_k)^t \cdot g_k$.

Instanța x va fi clasificată în acea grupă, k, pentru care $S_k(x) > S_i(x)$, $i = 1,q$, $i \neq k$.