4.6 Klasifikace objektů

Hledáním struktury a vzájemných vazeb v objektech se zabývají klasifikační metody vícerozměrné statistické analýzy. *Klasifikační metody* jsou postupy, pomocí kterých se jeden objekt zařadí do jedné existující třídy (*diskriminační analýza DA*), nebo pomocí nichž lze neuspořádanou skupinu objektů uspořádat do několika vnitřně sourodých tříd či shluků (*analýza shluků CLU*). Postup klasifikace je založen na určitých předpokladech o vlastnostech klasifikovaných objektů, např. normální rozdělení náhodného vektoru, charakterizujícího objekty; pak hovoříme o *parametrických klasifikačních metodách.* Není-li klasifikace založena na znalostech rozdělení náhodného vektoru, mluvíme o *neparametrických klasifikačních metodách.* Významnou roli při hledání struktury a vazeb mezi objekty na základě jejich podobnosti tvoří také *vícerozměrné škálování MDS.*

4.6.1 Diskriminační analýza DA

Diskriminační analýza patří mezi metody zkoumání závislosti mezi skupinou *p* nezávisle proměnných, nazvaných *diskriminátory,* tj. sloupců zdrojové matice na jedné straně a jednou kvalitativní závisle proměnnou na druhé straně. Umožňuje zařazení objektu do jedné z již existujících tříd. Ve vstupních datech jsou svými hodnotami diskriminátorů u všech objektů dány *zařazené objekty do primárních tříd*. Dále jsou dány *nezařazené objekty*, pro které budeme hledat zařazení do třídy. Objekt zařadíme do třídy na základě jeho největší míry podobnosti, např. nejmenší Mahalanobisovy vzdálenosti.

Diskriminační (zařazovací) pravidla: při diskriminační analýze se snažíme vyčíslit hodnotu *diskriminační funkce*, která nám usnadní zařazení do primární třídy. Takto vyčíslené hodnoty funkce používáme také ke třídění *nezařazených objektů* do předem známých primárních tříd, a to na základě *p* diskriminátorů $x_1 , x_2 , ..., x_p$. Každá primární třída je charakterizována svou funkcí hustoty pravděpodobnosti $f_j(\boldsymbol{x})$, kde $\boldsymbol{x}^T = [x_1 , x_2 , ..., x_p]$. Existuje citlivé pravidlo pro zařazení, diskriminaci objektu vektoru $\boldsymbol{x}$, do třídy G_j

$$f_j(\boldsymbol{x}) = \max_{i = 1,\ldots,g} f_i(\boldsymbol{x}) \quad .$$

Uveďme příklady diskriminace:

1. Existuje jednoduchá binární proměnná x a dvě třídy G_1 a G_2. Nejprve předpokládejme, že pravděpodobnost $P(x = 0) = P(x = 1) = 1/2$ a dále pravděpo-dobnost $P(x = 0) = 1/4$ a pravděpodobnost $P(x = 1) = 3/4$. Pravidlo zařadí objekt $x = 0$ do G_1 a objekt $x = 1$ do G_2.

2. Předpokládejme spojitou jednoduchou proměnnou x a opět dvě třídy G_1 a G_2. Ve třídě G_1 má proměnná normální rozdělení se střední hodnotou μ_1 a rozptylem σ^2_1 a ve třídě G_2 má proměnná rovněž normální rozdělení se střední hodnotou μ_2 a rozptylem σ^2_2, přičemž budeme předpokládat $\mu_1 < \mu_2$ a $\sigma^2_1 > \sigma^2_2$. Pomocí diskriminačního pravidla $f_j(\boldsymbol{x})$ bude objekt o skóre $\boldsymbol{x}$ zařazen do třídy G_1, když bude platit $f_1(\boldsymbol{x}) > f_2(\boldsymbol{x})$. Nahrazením skutečnou hustotou pravděpodobnosti normálního rozdělení dostaneme pravidlo k zařazení objektu x do třídy G_1:

$$\frac{F_1}{F_2} \exp\left\{-\frac{1}{2}\left[\frac{(x - \mu_1)^2}{F_1^2} - \frac{(x - \mu_2)^2}{F_2^2}\right]\right\} > 1$$

a po zlogaritmování a úpravě bude toto pravidlo ve tvaru

$$x^2\left[\frac{1}{F_1^2} - \frac{1}{F_1^2}\right] - 2x\left[\frac{\mu_1}{F_1^2} - \frac{\mu_2}{F_2^2}\right] + \left[\frac{\mu_1^2}{F_1^2} - \frac{\mu_2^2}{F_2^2}\right] < 2 \ln \frac{F_1}{F_2} \quad .$$

Dle tohoto pravidla dojde k rozdělení hodnot x do dvou tříd: první třída G_1 bude obsahovat malé hodnoty x a druhá třída G_2 velké hodnoty x. Ve zvláštním případě $\sigma_1 = \sigma_2$ dostaneme pravidlo pro zařazení do třídy G_1 ve znění $|x - \mu_2| > |x - \mu_1|$. Bude-li navíc $\mu_1 < \mu_2$, objekt se skóre x padne do třídy G_1, když bude platit, že
$x < (\mu_1 + \mu_2)/2$.

Zobecnění diskriminačního pravidla: G_1 je třída objektů s vícerozměrným normálním rozdělením a střední hodnotou μ_1 a G_2 obdobně třída objektů se střední hodnotou μ_2. Předpokládejme, že kovarianční matice obou tříd jsou stejné a užijeme proto pro ně společné označení $\boldsymbol{S}$. Obecné pravidlo zařazení objektu o vektoru $\boldsymbol{x}$ do třídy G_1 bude

$$(\boldsymbol{\mu_1} - \boldsymbol{\mu_2})\, \boldsymbol{S}^{-1} \left(\boldsymbol{x} - \frac{\boldsymbol{\mu_1} + \boldsymbol{\mu_2}}{2} \right) > 0 \quad .$$

Když třídy mají známé hustoty pravděpodobnosti rozličných rozdělení π_1, π_2, ..., π_p, bude pravidlo o zařazení do třídy upraveno takto: jde-li o 2 třídy, bude pravidlo ve tvaru

$$(\boldsymbol{\mu_1} - \boldsymbol{\mu_2})\, \boldsymbol{S}^{-1} \left(\boldsymbol{x} - \frac{\boldsymbol{\mu_1} + \boldsymbol{\mu_2}}{2} \right) > \ln \frac{B_1}{B_2} \quad .$$

Lineární diskriminační funkce (LDA): z diskriminačních funkcí je neznámější *Fisherova lineární diskriminační funkce* tvaru

$$z_i = a_{i1}\, x_1 + a_{i2}\, x_2 + a_{i3}\, x_3 + \ldots + a_{ip}\, x_p \ ,$$

kde p je počet proměnných primárních tříd čili počet diskriminátorů a x_1 , x_2 , ..., x_p jsou standardizované hodnoty těchto proměnných. Parametry z_i nazýváme *standardi-zované klasifikační koeficienty* Fisherovy diskriminační funkce $\boldsymbol{a}^T = [a_1 , a_2 , ..., a_p]$, které byly nalezeny tak, že poměr rozptylu mezi třídami $\boldsymbol{B}$ a rozptylu uvnitř tříd $\boldsymbol{S}$

$$V = \boldsymbol{a}^T \boldsymbol{B}\, \boldsymbol{a} / (\boldsymbol{a}^T \boldsymbol{S}\, \boldsymbol{a})$$

je maximální. Zde $\boldsymbol{B}$ je kovarianční matice třídních průměrů a $\boldsymbol{S}$ je celková kovarianční matice uvnitř tříd. Vektor $\boldsymbol{a}$, který maximalizuje poměr V, se vypočte ze vztahu

$$(\boldsymbol{B} - 8\, \boldsymbol{S})\, \boldsymbol{a} = 0.$$

V případě pouze dvou tříd budou klasifikační koeficienty diskriminační funkce $\boldsymbol{a}^T = [a_1 , a_2 , ..., a_p]$ vypočteny jednoduchým vztahem $\boldsymbol{a} = \boldsymbol{S}^{-1}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)$.

Kvadratická diskriminační funkce (QDA). Jsou-li střední hodnoty dvou souborů μ_1 a μ_2 shodné, ale soubory se liší v kovariančních maticích $\boldsymbol{S}_1$ a $\boldsymbol{S}_2$, nelze použít lineární diskriminační funkci, což dokumentuje příklad

Soubor G_1: $\boldsymbol{\mu_1}^T = [0, 0]$, $\boldsymbol{S}_1 = \begin{pmatrix} 1.0 & 0.0 \\ 0.0 & 1.0 \end{pmatrix}$,

Soubor G_2: $\boldsymbol{\mu_2}^T = [0, 0]$, $\boldsymbol{S}_2 = \begin{pmatrix} 4.0 & 0.0 \\ 0.0 & 4.0 \end{pmatrix}$.

Užije se kvadratická diskriminační funkce. Objekt o vektoru $\boldsymbol{x}$ bude patřit do třídy G_1, když bude splněna nerovnost

$$\boldsymbol{\mu_1}^T (\boldsymbol{S_2}^{-1} - \boldsymbol{S_1}^{-1}) \boldsymbol{x} - 2 \boldsymbol{x}^T (\boldsymbol{S_2}^{-1}\, \boldsymbol{\mu_2} - \boldsymbol{S_1}^{-1}\, \boldsymbol{\mu_1}) +$$
$$+ (\boldsymbol{\mu_2}^T\, \boldsymbol{S_2}^{-1}\, \boldsymbol{\mu_2} - \boldsymbol{\mu_1}^T\, \boldsymbol{S_1}^{-1}\, \boldsymbol{\mu_1}) \leq \ln \frac{|\boldsymbol{S_1}|}{|\boldsymbol{S_2}|} + 2 \ln \frac{B_1}{B_2},$$

kde $\boldsymbol{S}_1$ a $\boldsymbol{S}_2$ jsou kovarianční matice pro 1. a 2. třídu, G_1 a G_2.

Diskriminace mezi více než dvěma třídami. Pro tři třídy budou tři lineární diskriminační funkce nabývat následujících tvarů:

$$h_{12} = (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)^T\, \boldsymbol{S}^{-1} \left[\boldsymbol{x} - \frac{\bar{\boldsymbol{x}}_1 + \bar{\boldsymbol{x}}_2}{2} \right],$$
$$h_{13} = (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_3)^T\, \boldsymbol{S}^{-1} \left[\boldsymbol{x} - \frac{\bar{\boldsymbol{x}}_1 + \bar{\boldsymbol{x}}_3}{2} \right],$$
$$h_{23} = (\bar{\boldsymbol{x}}_2 - \bar{\boldsymbol{x}}_3)^T\, \boldsymbol{S}^{-1} \left[\boldsymbol{x} - \frac{\bar{\boldsymbol{x}}_2 + \bar{\boldsymbol{x}}_3}{2} \right].$$

kde $\boldsymbol{S}$ je vážená kovarianční matice všech tříd. Klasifikační pravidla zařazení objektu do dotyčné třídy jsou

umístění objektu do první třídy G_1 nastane, když $h_{12}(\boldsymbol{x}) > 0$ a $h_{13}(\boldsymbol{x}) > 0$,
umístění objektu do druhé třídy G_2 nastane, když $h_{12}(\boldsymbol{x}) < 0$ a $h_{23}(\boldsymbol{x}) > 0$,
umístění objektu do třetí třídy G_3 nastane, když $h_{13}(\boldsymbol{x}) > 0$ a $h_{23}(\boldsymbol{x}) < 0$.

Kvalita zařazení objektů do tříd (diskriminace). Předpokládejme, že máme data o K třídách s N_k , $k = 1, ..., K,$ objekty v každé třídě, N představuje celkový počet objektů (např. $N = N_1 + N_2 + N_3 = 150$). Každý objekt je popsán p diskriminátory. Každý i-tý objekt je prezentován prvkem x_{ki} . Nechť $\bar{\boldsymbol{x}}$ představuje vektor průměrů těchto diskriminátorů ve všech třídách a $\bar{\boldsymbol{x}}_{\boldsymbol{k}}$ pak vektor průměrů objektů v k-té třídě. Definujme sumy čtverců $\boldsymbol{S_T}$, $\boldsymbol{S_W}$, $\boldsymbol{S_B}$ odchylek od středních hodnot vztahy

$$\boldsymbol{S_T} = \sum_{k=1}^{K} \sum_{i=1}^{N_k} (x_{ki} - \bar{x})(x_{ki} - \bar{x})^T \ ,$$

$$\boldsymbol{S_W} = \sum_{k=1}^{K} \sum_{i=1}^{N_k} (x_{ki} - \bar{x}_k)(x_{ki} - \bar{x}_k)^T \ ,$$

$$\boldsymbol{S_B} = \boldsymbol{S_T} - \boldsymbol{S_W}$$

a definujme stupně volnosti, *df1* a *df2,* vztahy $df1 = K - 1$ a $df2 = N - K$. Diskriminační funkce je váženým průměrem hodnot nezávisle proměnných. Váhy jsou přitom voleny tak, že výsledný vážený průměr rozděluje objekty do tříd. Vysoké hodnoty průměru pocházejí z jedné třídy, nízké hodnoty průměru pocházejí z jiné třídy. Problém spočívá v nalezení vah tak, aby dobře diskriminovaly objekty do tříd. Řešení spočívá v nalezení vlastních vektorů $\boldsymbol{V}$ matice $\boldsymbol{S}_W^{-1} \boldsymbol{S}_B$. Kanonické koeficienty jsou totiž prvky těchto vlastních vektorů. *Mírou těsnosti proložení* je potom Wilkovo kritérium λ, definované vztahem

$$\lambda = \frac{|\boldsymbol{S_W}|}{|\boldsymbol{S_T}|} = \prod_{j=1}^{m} \frac{1}{1+\lambda_j} \ ,$$

kde λ_j je j-té vlastní číslo, odpovídající vlastnímu vektoru, a m je minimum ze dvou čísel, K-*1* a p.

Kanonická korelace mezi j-tou diskriminační funkcí a nezávisle proměnnými čili diskriminátory je vztažena k těmto vlastním číslům

$$r_{cj} = \sqrt{\frac{\lambda_j}{1 + \lambda_j}} \ .$$

Řada rozličných matic potřebných v diskriminační analýze je definována vztahy:

celková kovariančí matice $\boldsymbol{T} = \frac{1}{N-1} \boldsymbol{S_T}$,

kovariančí matice uvnitř tříd $\boldsymbol{W} = \frac{1}{N-K} \boldsymbol{S_W}$,

kovariančí matice mezi třídami $\boldsymbol{B} = \frac{1}{K-1} \boldsymbol{S_B}$,

lineární diskriminační funkce $z_k = \boldsymbol{W^{-1}} \bar{\boldsymbol{x}}_{\boldsymbol{k}}$,

standardizované kanonické koeficienty $v_{ij} \sqrt{w_{ij}}$,

kde v_{ji} jsou prvky $\boldsymbol{V}$ a w_{ij} prvky matice $\boldsymbol{W}$. Korelace mezi nezávisle proměnnými a kanonickými proměnnými jsou dány vztahem

$$\mathrm{Cor}_{jk} = \frac{1}{\sqrt{w_{jj}}} \sum_{i=1}^{p} v_{ik} \, w_{ji} \ .$$

Logistická diskriminace. Fisherova lineární diskriminace je optimální, když dva soubory mají vícerozměrné normální rozdělení se stejnými kovariančními maticemi. Tato diskriminační funkce se jeví také dostatečně robustní na odchylky od normality. Existuje však řada případů silné nenormality, např. přítomnost binárních proměnných. Pak je možné užít logistický model k výpočtu pravděpodobnosti, že objekt je členem dotyčné třídy:

$$P(G_1^* \boldsymbol{x}) = \frac{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p)}{1 + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p)},$$

$$P(G_2^* \boldsymbol{x}) = \frac{1}{1 + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p)}.$$

Neznámé parametry $\beta_0, \beta_1, \beta_2, ..., \beta_p$ jsou odhadovány na základě maximální věrohodnosti. Důležité je, že odhad je zcela nezávislý na funkci hustoty třídní pravděpodobnosti. Po vyčíslení odhadů b_0 , b_1 , b_2 , ..., b_p neznámých parametrů $\beta_0, \beta_1, \beta_2, ..., \beta_p$ se uplatní klasifikační pravidlo zařazení objektu do třídy G_1, platí-li

$$b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_p x_p > 0,$$

což odpovídá pravděpodobnosti $P(G_1^* \boldsymbol{x}) > P(G_2^* \boldsymbol{x})$.

Postup klasifikace diskriminační analýzou

1. *Bodové odhady parametrů polohy a rozptýlení všech diskriminátorů:* vyčíslí se (a) aritmetické průměry ve třídách, (b) směrodatné odchylky ve třídách, (c) celková korelační a kovarianční matice všech diskriminátorů, (d) mezitřídní korelace a kovariance za použití průměrů místo hodnot objektů, (e) vnitrotřídní korelace a kovariance za použití dat, ve kterých byly odečteny průměry tříd a provede se zhodnocení dosažených výsledků.
2. *Vyšetření vlivu jednotlivých diskriminátorů:* vliv jednotlivých diskriminátorů na výsledky diskriminační analýzy se sleduje pomocí testačních statistik při odstranění odpovídajícího diskriminátoru.
3. *Odhady neznámých parametrů b_0 , b_1 , ..., b_p lineární diskriminační funkce pro každou třídu:* odhady neznámých parametrů b_0 , b_1 , ..., b_p jsou mezivýpočtem k vyčíslení diskriminačního skóre.
4. *Odhady regresních parametrů b_0 , b_1 , ... , b_p lineárního regresní modelu pro každou třídu*: těmito regresními parametry predikované hodnoty budou ležet mezi nulou a jedničkou. Zařazení se provede na základě třídy s nejvyšším skóre, blízkým jedničce.
5. *Klasifikace objektů diskriminační funkcí (diskriminace do tříd):* provede se (a) vyčíslení klasifikačních počtů objektů v jednotlivých třídách po diskriminaci do tříd, (b) přehled chybně klasifikovaných objektů tak, že vedle skutečné třídy je predikovaná třída a procento pravděpodobnosti výskytu objektu v predikované třídě, (c) přehled klasifikovaných objektů - skutečná (primární) třída, predikovaná třída všech objektů a procento pravděpodobnosti výskytu objektu v predikované třídě.
6. *Kanonická korelační analýza:* (a) analýza kanonických proměnných: první soubor obsahuje diskriminátory a druhý soubor třídní proměnné, (b) odhady parametrů u kanonických proměnných, (c) kanonické proměnné u třídních průměrů, (d) standardizované kanonické koeficienty slouží k výpočtu kanonického skóre, což jsou vážené průměry objektů, (e) korelace původních a kanonických proměnných představuje zátěže (korelace) původních proměnných na kanonické proměnné. Tím se usnadní vysvětlení dotyčné kanonické proměnné.
7. *Lineární diskriminační skóre všech objektů:* jsou vyčísleny hodnoty predikovaných skóre lineárních diskriminačních proměnných pro všechny objekty.
8. *Regresní skóre všech objektů:* hodnoty predikovaných skóre regresních proměnných pro všechny objekty jsou založeny na regresních koeficientech.
9. *Kanonické skóre:* hodnoty predikovaných skóre kanonických proměnných pro všechny objekty jsou založeny na kanonických koeficientech.
10. *Volba proměnných:* z velké palety diskriminátorů se vybírají pouze ty, které jsou dostatečně účinné, maximálně 8 proměnných. Výběr se provádí krokově: k nejlepšímu diskriminátoru se nalezne druhý nejlepší tak, že se prověří zda diskriminace bude tak dokonalá, jako když byl jeden diskriminátor odebrán. U nové proměnné se ověřuje, zda její F má hodnotu pravděpodobnosti menší než $\alpha = 0.05$.
11. *Výklad grafů:* výsledkem diskriminační analýzy je grafické zařazení do tříd. Zobrazení se provede na třech grafech: (a) zobrazení lineárních diskriminačních skóre, (b) zobrazení regresního skóre, a (c) zobrazení kanonického skóre.