Lineární diskriminační analýza

(Předpoklad – varianční matice jsou ve všech skupinách shodné.)

Odvození bayesovského rozhodovacího pravidla pro dvě skupiny objektů

Nechť v 1. skupině je n_1 objektů, ve 2. skupině n_2 objektů. Každý objekt je charakterizován p-rozměrným vektorem pozorování $\mathbf{X} = (X_1, \ldots, X_p)'$.

Předpokládáme, že v h-té skupině má náhodný vektor $\mathbf{X}$ hustotu $\varphi_h(\mathbf{x})$, h = 1, 2.

Nechť H_h je jev „objekt patří do h-té skupiny“.

Apriorní pravděpodobnost $P(H_h)$ příslušnosti objektu k h-té skupině označíme π_h, h = 1, 2.

Známe-li u nějakého objektu vektor pozorování $\mathbf{x}$, můžeme podle Bayesova vzorce vypočítat aposteriorní pravděpodobnost příslušnosti objektu ke skupině:

$$P(H_h / \mathbf{X} = \mathbf{x}) = \frac{\pi_h \varphi_h(\mathbf{x})}{\pi_1 \varphi_1(\mathbf{x}) + \pi_2 \varphi_2(\mathbf{x})}, h = 1, 2$$

Rozhodovací pravidlo: nový objekt zařadíme do té skupiny, u níž je aposteriorní pravděpodobnost větší.

Objekt s vektorem pozorování **x** zařadíme do 1. skupiny, když $\pi_1\varphi_1(\mathbf{x}) > \pi_2\varphi_2(\mathbf{x})$, jinak ho zařadíme do 2. skupiny.

Součin $\pi_h\varphi_h(\mathbf{x})$ se nazývá diskriminační skór pro h-tou skupinu.

Lze ukázat, že bayesovské rozhodovací pravidlo je optimální v tom smyslu, že minimalizuje celkovou pravděpodobnost mylné klasifikace.

Konstrukce Fisherovy lineární diskriminační funkce pro dvě skupiny objektů

V diskriminační analýze se předpokládá, že hustota v h-té skupině je normální a má parametry $\boldsymbol{\mu}_h$, $\boldsymbol{\Sigma}_h$, tj.

$$\varphi_h(\mathbf{x}) = \frac{1}{\sqrt{\det(2\pi\boldsymbol{\Sigma}_h)}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_h)'\boldsymbol{\Sigma}_h^{-1}(\mathbf{x}-\boldsymbol{\mu}_h)\right), \quad h = 1, 2.$$

Jestliže zlogaritmujeme diskriminační skór $\pi_h\varphi_h(\mathbf{x})$ a vynecháme člen $-\frac{p}{2}\ln(2\pi)$, který je společný pro obě skupiny, dostaneme tzv. kvadratický diskriminační skór pro h-tou skupinu ve tvaru $-\frac{1}{2}\ln(\det\boldsymbol{\Sigma}_h) - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_h)'\boldsymbol{\Sigma}_h^{-1}(\mathbf{x}-\boldsymbol{\mu}_h) + \ln\pi_h$, h = 1, 2.

Jsou-li varianční matice v obou skupinách stejné (společnou varianční matici označíme $\boldsymbol{\Sigma}$), obsahují oba kvadratické diskriminační skóry týž člen $-\frac{1}{2}\ln(\det\boldsymbol{\Sigma}) - \frac{1}{2}\mathbf{x}'\boldsymbol{\Sigma}^{-1}\mathbf{x}$. Po jeho vynechání obdržíme lineární diskriminační skór pro h-tou skupinu – tzv. Andersonovu diskriminační statistiku - ve tvaru $\lambda_h(\mathbf{x}) = \boldsymbol{\mu}_h'\boldsymbol{\Sigma}^{-1}\mathbf{x} - \frac{1}{2}\boldsymbol{\mu}_h'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_h + \ln\pi_h$, h = 1, 2.

Objekt s vektorem pozorování $\mathbf{x}$ tedy zařadíme do 1. skupiny, když $\lambda_1(\mathbf{x}) > \lambda_2(\mathbf{x})$, jinak ho zařadíme do 2. skupiny.

Vzhledem k tomu, že máme jen dvě skupiny objektů, lze rozhodnutí o zařazení objektu do skupiny učinit na základě rozdílu

$$\lambda(\mathbf{x}) = \lambda_1(\mathbf{x}) - \lambda_2(\mathbf{x}) = \left(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\right)' \boldsymbol{\Sigma}^{-1}\mathbf{x} - \frac{1}{2}\left(\boldsymbol{\mu}_1' \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2' \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2\right) + \ln \pi_1 - \ln \pi_2 .$$

Funkce $\lambda(\mathbf{x})$ se nazývá Fisherova lineární diskriminační funkce. Označíme-li

$$\boldsymbol{\beta}' = \left(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\right)' \boldsymbol{\Sigma}^{-1}, \gamma = -\frac{1}{2}\,\boldsymbol{\beta}'(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) + \ln \pi_1 - \ln \pi_2,$$

můžeme Fisherovu lineární diskriminační funkci psát ve tvaru

$\lambda(\mathbf{x}) = \boldsymbol{\beta}'\mathbf{x} + \gamma$.

Znamená to, že jsme našli takovou lineární kombinaci vektoru pozorování **x**, která nám umožní minimalizovat celkovou pravděpodobnost mylného zařazení objektu do skupiny. Objekt s vektorem pozorování **x** tedy zařadíme do 1. skupiny, když $\lambda(\mathbf{x}) > 0$, jinak ho zařadíme do 2. skupiny.

Modifikace pro případ neznámých parametrů

Při praktickém použití diskriminační analýzy většinou neznáme parametry $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$, $\boldsymbol{\Sigma}$ ani apriorní pravděpodobnosti π_1, π_2. V takovém případě používáme odhady:

$\boldsymbol{\mu}_h \rightarrow \mathbf{M}_h$, h = 1, 2

$$\boldsymbol{\Sigma} \rightarrow \mathbf{S} = \frac{(n_1 - 1)\mathbf{S}_1 + (n_2 - 1)\mathbf{S}_2}{n_1 + n_2 - 2}$$

$$\pi_h \rightarrow \frac{n_h}{n}, \; h = 1, 2.$$

Odhad Fisherovy lineární diskriminační funkce $\lambda(\mathbf{x}) = \boldsymbol{\beta}'\mathbf{x} + \gamma$:

$L(\mathbf{x}) = \mathbf{b}'\mathbf{x} + g$, kde

$$\mathbf{b}' = (\mathbf{M}_1 - \mathbf{M}_2)'\mathbf{S}^{-1}, \; g = -\frac{1}{2}\,\mathbf{b}'(\mathbf{M}_1 + \mathbf{M}_2) + \ln p_1 - \ln p_2.$$

Posouzení účinnosti diskriminace resubstituční metodou

Resubstituční metoda spočívá v uplatnění zkonstruovaného rozhodovacího pravidla na objekty se známou příslušností ke skupině. Uvažujeme postupně všechny tyto objekty a jejich zařazení podle rozhodovacího pravidla porovnáme se skutečnou příslušností ke skupině. Stanovíme podíl správně a mylně zařazených objektů.

skutečnost	zařazení		součet
	1. skupina	2. skupina	
1. skupina	n_{11}	n_{12}	$n_{1.} = n_1$
2. skupina	n_{21}	n_{22}	$n_{2.} = n_2$
součet	$n_{.1}$	$n_{.2}$	n

Podíl správně zařazených objektů:

$$\frac{n_{11} + n_{22}}{n}$$

Podíl mylně zařazených objektů:

$$\frac{n_{12} + n_{21}}{n}$$

Postup při lineární diskriminační analýze

1. Vzhledem k povaze úlohy určíme veličiny $X_1, ..., X_p$ a pořídíme $n_1 + n_2$ p-rozměrných pozorování tak, aby n_1 objektů pocházelo z 1. skupiny a n_2 objektů z 2. skupiny.
2. Na zvolené hladině významnosti α testujeme hypotézy o normalitě rozložení v obou skupinách a orientačně posoudíme linearitu vztahů mezi sledovanými proměnnými v obou skupinách.
3. Vypočteme odhady $\mathbf{M}_1$, $\mathbf{M}_2$, $\mathbf{S}_1$, $\mathbf{S}_2$, $\mathbf{S}$, p_1, p_2.
4. Na zvolené hladině významnosti α testujeme hypotézy o shodě variančních matic a vektorů středních hodnot v obou skupinách.
5. Vypočteme odhad $L(\mathbf{x})$ Fisherovy lineární diskriminační funkce. Objekt s vektorem pozorování $\mathbf{x}$ přiřadíme k 1. skupině, když $L(\mathbf{x}) > 0$, jinak ho přiřadíme ke 2. skupině.
6. Účinnost diskriminace posoudíme metodou resubstituce.

Příklad:

V souboru 50 rodin byly zjišťovány tyto údaje:

- zda v posledních dvou letech rodina navštívila jistou rekreační oblast (veličina ID1, nabývá hodnoty 0 pro odpověď „ne“, hodnoty 1 pro odpověď „ano“)
- částka, kterou je rodina ochotná vydat za dovolenou (veličina ID2, nabývá hodnoty 1 pro variantu „malá“, 2 pro variantu „střední“ a 3 pro variantu „velká“)
- roční příjem v tisících dolarů (veličina X_1)
- postoj k cestování (veličina X_2, devítibodová škála, 1 = naprosto odmítavý, 9 = veskrze kladný)
- význam přičítaný rodinné dovolené (veličina X_3, devítibodová škála, 1 = nejnižší, 9 = nejvyšší)
- počet členů rodiny (veličina X_4)
- věk nejstaršího člena rodiny (veličina X_5).

Pro uvedená data proveďte lineární diskriminační analýzu pro dvě skupiny objektů, tj. pro třídění podle ID1.

(Přistoupíme přímo k provedení LDA, protože ověřování předpokladů o datech a testováním hypotéz o shodě variančních matic a shodě vektorů středních hodnot jsme se již zabývali v přednášce o kanonické diskriminační analýze.)

Význam jednotlivých proměnných v modelu

Statistiky – Vícerozměrné průzkumné techniky – Diskriminační analýza – Proměnné - Grupovací ID1 – Seznam nezáv. proměnných X1 až X5 – OK – OK – Výpočet: proměnné v modelu.

N=50	Výsledky diskriminační funkční analýzy (dovolena.sta) Počet prom. v modelu: 5; grupovací: ID1 (2 skup) Wilk. lambda: ,38229 přibliž F (5,44)=14,219 p< ,0000					
	Wilk. Lambda	Parc. Lambda	F na vyj (1,44)	p-hodn.	Toler.	1-toler. R^2
X1	0,627513	0,609207	28,22504	0,000003	0,879866	0,120134
X2	0,388609	0,983729	0,72778	0,398223	0,934715	0,065285
X3	0,400086	0,955507	2,04884	0,159388	0,977164	0,022836
X4	0,382565	0,999270	0,03215	0,858527	0,921303	0,078697
X5	0,439319	0,870177	6,56444	0,013904	0,956782	0,043218

V záhlaví této tabulky je uvedena Wilksova Lambda (na škále od 0 – nejlepší diskriminace do 1 – žádná diskriminace) a její přepočet na testovou statistiku F pro Hotellingův test shody vektorů středních hodnot (14,219) a odpovídající p-hodnota (je blízká 0).
V 1. sloupci (Wilk. Lambda) jsou hodnoty Wilksovy Lambdy při vyřazení dané proměnné z modelu (vyšší hodnoty jsou lepší).
2. sloupec (Parc. Lambda) obsahuje unikátní příspěvky proměnných k diskriminaci.
Ve 3. sloupci jsou přepočty parciálních Lambda na testové statistiky a ve 4. sloupci pak odpovídající p-hodnoty. Podle p-hodnot u jednotlivých proměnných soudíme, že pro diskriminaci jsou významné proměnné X_1 a X_5.
5. sloupec (Tolerance) udává unikátní variabilitu proměnné nevysvětlenou ostatními proměnnými v modelu.
6. sloupec (1-toler., R^2) udává variabilitu proměnné vysvětlenou ostatními proměnnými.

Mahalanobisova vzdálenost v diskriminační analýze

Používá se pro popis vzájemných vzdáleností centroidů jednotlivých skupin.

Statistiky – Vícerozměrné průzkumné techniky – Diskriminační analýza - Proměnné – Grupovací proměnná ID, Seznam nezávislých proměnných X1 až X5 — OK – OK – na záložce Detaily zvolíme Vzdálenosti mezi skupinami. Současně dostaneme i p-hodnoty pro testy hypotéz, že vzdálenosti jsou nulové:

	Mahalanobisovy vzdálenosti^2 (dovolena.sta)	
ID1	návštěva ne	návštěva ano
návštěva ne	0,000000	6,367867
návštěva ano	6,367867	0,000000

	p-hodnot (dovolena.sta)	
ID1	návštěva ne	návštěva ano
návštěva ne		0,000000
návštěva ano	0,000000	

Lze také získat Mahalanobisovy vzdálenosti jednotlivých objektů od centroidů skupin.

Na záložce Klasifikace zvolíme Mahalanobisovy vzdálenosti^2:

	Mahalanobisovy vzdálenosti (dovolena.sta) Nesprávná klasifikace je označena *		
Případ	**Pozorova Klasif.**	návštěva ne p=,58000	návštěva ano p=,42000
1	návštěva ne	9,18363	18,11825
2	návštěva ne	0,88533	10,53314
3	návštěva ne	3,90372	12,30937
4	návštěva ne	5,35649	8,74744
5	návštěva ne	4,41397	11,30806
6	návštěva ne	0,62136	7,62423

Stanovení odhadu Fisherovy lineární diskriminační funkce:

$L(\mathbf{x}) = \mathbf{b}'\mathbf{x} + g$, kde $\mathbf{b}' = (\mathbf{M}_1 - \mathbf{M}_2)'\mathbf{S}^{-1}$, $g = -\frac{1}{2}\,\mathbf{b}'(\mathbf{M}_1 + \mathbf{M}_2) + \ln p_1 - \ln p_2$.

Odhad vektoru středních hodnot v 1. skupině:

	Popisné statistiky (dovolena.sta) Zhrnout podmínku: ID=0	
Proměnná	N platných	Průměr
X1	29	42,84483
X2	29	4,24138
X3	29	4,27586
X4	29	3,72414
X5	29	46,93103

Odhad vektoru středních hodnot ve 2. skupině:

	Popisné statistiky (dovolena.sta) Zhrnout podmínku: ID=1	
Proměnná	N platných	Průměr
X1	21	59,76190
X2	21	5,14286
X3	21	5,76190
X4	21	4,33333
X5	21	53,61905

Odhad společné varianční matice **S**:

	X_1	X_2	X_3	X_4	X_5
X1	63,53	2,37	1,36	2,60	-7,32
X2	2,37	2,75	-0,17	0,17	-2,32
X3	1,36	-0,17	2,70	-0,09	0,22
X4	2,60	0,17	-0,09	1,51	0,11
X5	-7,32	-2,32	0,22	0,11	60,02

Postup v systému STATISTICA :

Statistiky – Vícerozměrné průzkumné techniky – Diskriminační analýza – Proměnné – Grupovací ID, Seznam nezáv. proměnných X1-X5 – OK, zapneme Další možnosti (kroková analýza) – OK – Popisné statistiky – Zobrazit popisné statistiky – Vnitřní kovariance a korelace.

Odhady apriorních pravděpodobností:

$$p_1 = \frac{n_1}{n} = \frac{29}{50} = 0{,}58,\ p_2 = \frac{n_2}{n} = \frac{21}{50} = 0{,}42$$

Po dosazení dostaneme:

$\mathbf{b}' = (\mathbf{M}_1 - \mathbf{M}_2)'\mathbf{S}^{-1} = (-0{,}2865 \;\; -0{,}2556 \;\; -0{,}4169 \;\; 0{,}0736 \;\; -0{,}1527)$

$g = -\frac{1}{2}\,\mathbf{b}'(\mathbf{M}_1 + \mathbf{M}_2) + \ln p_1 - \ln p_2 = 24{,}7666$

$L(\mathbf{x}) = \mathbf{b}'\mathbf{x} + g = -0{,}2685X_1 - 0{,}2556X_2 - 0{,}4169X_3 + 0{,}0736X_4 - 0{,}1527X_5 + 24{,}7666$

Postup v systému STATISTICA :

Statistiky – Vícerozměrné průzkumné techniky – Diskriminační analýza - Proměnné – Grupovací proměnná ID, Seznam nezávislých proměnných X1 až X5 — OK – OK – na záložce Klasifikace zvolíme Klasifikační funkce. Dostaneme tabulku tvaru:

	Klasifikační funkce; grupovací : ID (dovolena
Proměnná	návštěva ne p=,58000	návštěva ano p=,42000
X1	0,6369	0,9054
X2	1,7840	2,0395
X3	1,3391	1,7560
X4	1,1866	1,1130
X5	0,9216	1,0743
Konstant	-44,6709	-69,4375

Abychom získali odhad Fisherovy lineární diskriminační funkce, přidáme do této tabulky novou proměnnou a do jejího Dlouhého jména napíšeme =v1-v2

	Klasifikační funkce; grupovací : ID (dovolena		
Proměnná	návštěva ne p=,58000	návštěva ano p=,42000	NProm =v1-v2
X1	0,6369	0,9054	-0,26847
X2	1,7840	2,0395	-0,25557
X3	1,3391	1,7560	-0,41694
X4	1,1866	1,1130	0,073566
X5	0,9216	1,0743	-0,15266
Konstant	-44,6709	-69,4375	24,76658

Klasifikace nového případu

Předpokládejme nyní, že jsme prozkoumali další rodinu, která

má roční příjem $X_1 = 51{,}8$ tisíc dolarů,

k cestování zaujímá postoj ohodnocený $X_2 = 6$ body,

rodinné dovolené přičítá význam ohodnocený $X_3 = 7$ body,

má $X_4 = 4$ členy

a nejstaršímu členovi je $X_5 = 51$ let.

Na základě těchto údajů se pokusíme pomocí Fisherovy lineární diskriminační funkce zařadit tuto rodinu do skupiny rodin, které buď navštěvují nebo nenavštěvují danou rekreační oblast:

$$L(\mathbf{x}) = -0{,}2685X_1 - 0{,}2556X_2 - 0{,}4169X_3 + 0{,}0736X_4 - 0{,}1527X_5 + 24{,}7666 =$$

$$= -0{,}2685*51{,}8 - 0{,}2556*6 - 0{,}4169*7 + 0{,}0736*4 - 0{,}1527*51 + 24{,}7666 = -1{,}0836.$$

Protože $L(\mathbf{x}) < 0$, zařadíme tuto rodinu do skupiny rodin, které navštěvují danou rekreační oblast.

Posouzení účinnosti diskriminace resubstituční metodou:

Na záložce Klasifikace zvolíme Klasifikační matice.

Skup.	Klasifikační matice (dovolena) Řádky: pozorované klasifikace Sloupce: předpovězené klasifikace % správných	návštěva ne p=,58000	návštěva ano p=,42000
návštěva ne	93,10345	27	2
návštěva ano	76,19048	5	16
Celkem	86,00000	32	18

Podíl správně zařazených objektů:

$$\frac{n_{11}+n_{22}}{n}=\frac{27+16}{50}=0{,}86$$

Podíl mylně zařazených objektů:

$$\frac{n_{12}+n_{21}}{n}=\frac{5+2}{50}=0{,}14$$

Pro určení chybně zařazených případů zvolíme na záložce Klasifikace možnost Klasifikace případů. Zjistíme, že v 1. skupině došlo k mylnému zařazení u rodin č. 9 a 10, ve 2. skupině u rodin číslo 30, 33, 36, 43, 45.

Výběr proměnných pro klasifikaci krokovou metodou

Kroková metoda postupně vyhledává nejvhodnější soubor proměnných pro diskriminaci. Používá se buď jako dopředná nebo jako zpětná.

Význam jednotlivých proměnných pro diskriminaci se k každém kroku zkoumá pomocí zaváděcího a odstraňovacího kritéria.
Vybírání proměnných či jejich odstraňování skončí, když žádné další proměnné nesplňují zaváděcí nebo odstraňovací kritérium.

Upozornění: Před zařazením j-té proměnné do modelu se stanoví její tolerance $1-R_j^2$ (R_j^2 je čtverec vícenásobného koeficientu korelace, tj. koeficientu, který měří těsnost lineární závislosti veličiny X_j na ostatních veličinách). Tolerance je implicitně nastavená na 0,01.

Příklad: Použijte krokovou dopřednou (a poté zpětnou) metodu pro zařazování rodin do dvou skupin.

Řešení:

Statistika – Vícerozměrné průzkumné techniky – Diskriminační analýza – Proměnné - Grupovací ID1 – Seznam nezáv. proměnných X1 až X5 – OK – zaškrtneme Další možnosti (kroková analýza) – OK – Metoda – zvolíme kroková dopředná. Na záložce Detaily můžeme změnit Možnosti kroku (ponecháme implicitní nastavení) a také pomocí tlačítka Výsledky můžeme zvolit, zda chceme zobrazovat výsledky po každém kroku nebo chceme pouze shrnutí (ponecháme shrnutí) – OK.

Zvolíme-li tlačítko Výpočet: proměnné v modelu, dostaneme tabulku

N=50	Výsledky diskriminační funkční analýzy (dovolena.sta) krok 3, poč. prom. v modelu: 3; grupovací: ID1 (2 skup) Wilk. lambda: ,38880 přibliž F (3,46)=24,104 p< ,0000					
	Wilk. Lambda	Parc. Lambda	F na vyj (1,46)	p-hodn.	Toler.	1-toler. R^2
X1	0,719493	0,540386	39,12429	0,000000	0,974791	0,025209
X5	0,441811	0,880024	6,27128	0,015879	0,985042	0,014958
X3	0,405987	0,957678	2,03285	0,160683	0,988398	0,011602

Vidíme, že algoritmus skončil po třech krocích a vybral proměnné X_1, X_5 a X_3.

Zvolíme-li tlačítko Proměnné neobsažené v modelu, zjistíme, že jde o proměnné X_2 a X_4.
Na záložce Klasifikace vybereme Klasifikační funkce. Dostaneme lineární diskriminační skóry pro 1. a 2. skupinu objektů. Do vzniklé tabulky přidáme novou proměnnou L, do jejíhož Dlouhého jména napíšeme =v1-v2 a tím získáme odhad Fisherovy lineární diskriminační funkce:

	Klasifikační funkce; grupovací : ID1 (dovolena.sta)		
Proměnná	návštěva ne p=,58000	návštěva ano p=,42000	L =v1-v2
X1	0,7504	1,0247	-0,2742808
X5	0,8693	1,0128	-0,1434212
X3	1,1355	1,5365	-0,4009242
Konstant	-39,4479	-63,0649	23,6170025

Vidíme, že $L(\mathbf{x}) = -0{,}2743*X1 - 0{,}1434*X5 - 0{,}4009*X3 + 23{,}617$

Klasifikační matice je stejná jako v případě diskriminace podle všech proměnných a chybně zařazené případy jsou také stejné.

	Klasifikační matice (dovolena.sta) Řádky: pozorované klasifikace Sloupce: předpovězené klasifikace		
Skup.	% správnýc	návštěva ne p=,58000	návštěva ano p=,42000
návštěva ne	93,10345	27	2
návštěva ano	76,19048	5	16
Celkem	86,00000	32	18

Použijeme-li krokovou zpětnou metodu, je vybrána pouze proměnná X_1 a účinnost diskriminace poklesne na 80 %.

Porovnání s náhodnou klasifikací

Kdybychom zařazovali rodiny do skupin náhodně, pouze s ohledem na apriorní pravděpodobnosti π_1, π_2, tak bychom s pravděpodobností π_1 našli rodinu patřící do 1. skupiny, avšak s pravděpodobností π_2 bychom ji mylně zařadili do 2. skupiny. Naopak s pravděpodobností π_2 najdeme rodinu patřící do 2. skupiny, kterou s pravděpodobností π_1 mylně zařadíme do 1. skupiny.

Celková pravděpodobnost mylné klasifikace je tedy: $\pi_1\pi_2 + \pi_2\pi_1 = 2\pi_1(1- \pi_1)$.

Nahradíme-li apriorní pravděpodobnosti π_1, π_2 jejich odhady p_1, p_2 , dostaneme odhad celkové pravděpodobnosti mylné klasifikace $2p_1(1- p_1) = 2 \cdot \frac{29}{50} \cdot \frac{21}{50} = 0{,}4872$.

Použitím diskriminační analýzy jsme tedy dosáhli výrazného zlepšení, pravděpodobnost mylné klasifikace klesla na 0,14.

Klasifikace pomoci LDA pro $r \geq 3$ **skupin**

Opět předpokládáme, že ve všech r skupinách se vektory pozorování řídí p-rozměrným normálním rozložením, varianční matice jednotlivých skupin jsou shodné a vztahy mezi sledovanými p proměnnými jsou přibližně lineární.

Lineární diskriminační skór pro h-tou skupinu (Andersonova diskriminační statistika) má tvar:

$$\lambda_h(\mathbf{x}) = \boldsymbol{\mu}_h' \boldsymbol{\Sigma}^{-1} \mathbf{x} - \frac{1}{2} \boldsymbol{\mu}_h' \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_h + \ln \pi_h, \; h = 1, \ldots, r$$

Její odhad získáme dosazením $\mathbf{M}_h$, $\mathbf{S}$ a p_h:

$$L_h(\mathbf{x}) = \mathbf{M}_h' \mathbf{S}^{-1} \mathbf{x} - \frac{1}{2} \mathbf{M}_h' \mathbf{S}^{-1} \mathbf{M}_h + \ln p_h$$

Objekt neznámého původu, jehož vektor pozorování je $\mathbf{x}$, bude zařazen do skupiny s nejvyšší hodnotou $L_h(\mathbf{x})$.

Příklad: Soubor rodin nyní roztřiďte do tří skupin podle proměnné ID2, tj. podle toho, jak velkou částku je rodina ochotna vydat z dovolenou (varianty „malá“, „střední“, „velká“).

Řešení: Předběžné analýzy již byly provedeny, přistoupíme proto přímo k LDA pro tři skupiny objektů.
Při zadávání proměnných zvolíme jako grupovací proměnnou ID2. Zvolíme-li Výpočet: proměnné v modelu, dostaneme tabulku:

	Výsledky diskriminační funkční analýzy (dovolena.sta) Počet prom. v modelu: 5; grupovací: ID2 (3 skup) Wilk. lambda: ,26322 přibliž F (10,86)=8,1626 p< ,0000					
N=50	Wilk. Lambda	Parc. Lambda	F na vyj (2,43)	p-hodn.	Toler.	1-toler. R^2
X1	0,602832	0,436636	27,74006	0,000000	0,805704	0,194297
X2	0,289522	0,909148	2,14852	0,129016	0,959666	0,040334
X3	0,270302	0,973794	0,57859	0,564991	0,899531	0,100469
X4	0,269947	0,975075	0,54960	0,581183	0,883696	0,116304
X5	0,319480	0,823896	4,59552	0,015533	0,948842	0,051158

V záhlaví této tabulky je uvedena testová statistika pro Wilksův test shody vektorů středních hodnot (8,1626) a odpovídající p-hodnota (je blízká 0).
Podle p-hodnot u jednotlivých proměnných soudíme, že pro diskriminaci jsou významné proměnné X_1 a X_5.

Na záložce Klasifikace zvolíme Klasifikační funkce:

Proměnná	Klasifikační funkce; grupovací : ID2 (dovolena.sta)		
	malá p=,24000	střední p=,48000	velká p=,28000
X1	0,5525	0,8026	1,0981
X2	2,3285	2,4727	3,1155
X3	0,6466	0,3530	0,3648
X4	0,7459	0,4926	0,1242
X5	0,8874	0,7754	0,9120
Konstant	-42,2581	-45,1663	-70,7708

Zde jsou uvedeny koeficienty pro odhady Andersonových diskriminačních skórů pro 1., 2. a 3. skupinu:

$L_1(\mathbf{x}) = 0{,}5525*X1 + 2{,}3285*X2 + 0{,}6466*X3 + 0{,}7459*X4 + 0{,}8874*X5 - 42{,}2581$
$L_2(\mathbf{x}) = 0{,}8026*X1 + 2{,}4727*X2 + 0{,}3530*X3 + 0{,}4926*X4 + 0{,}7754*X5 - 45{,}1663$
$L_3(\mathbf{x}) = 1{,}0981*X1 + 3{,}1155*X2 + 0{,}3648*X3 + 0{,}1242*X4 + 0{,}9120*X5 - 70{,}7708$

Klasifikační matice:

Skup.	Klasifikační matice (dovolena.sta) Řádky: pozorované klasifikace Sloupce: předpovězené klasifikace % správnýc	malá p=,24000	střední p=,48000	velká p=,28000
malá	66,66666	8	4	0
střední	91,66666	1	22	1
velká	78,57143	0	3	11
Celkem	82,00000	9	29	12

Správně zařazeno bylo $\frac{8+22+11}{50}\cdot 100\% = 82\%$ případů, chybně 18 % případů.

V 1. skupině rodin byly chybně zařazeny případy 8, 10, 19, 20 ($\frac{4}{12} = 33{,}3\%$), ve 2. skupině případy 4, 47 ($\frac{2}{24} = 8{,}3\%$) a ve 3. skupině případy 24, 34, 43 ($\frac{3}{14} = 21{,}4\%$)

Zařazení nového případu

Nyní podle těchto skórů zařadíme do jedné ze tří skupin rodinu, která

má roční příjem X_1 = 51,8 tisíc dolarů,

k cestování zaujímá postoj ohodnocený X_2 = 6 body,

rodinné dovolené přičítá význam ohodnocený X_3 = 7 body,

má X_4 = 4 členy

a nejstaršímu členovi je X_5 = 51 let.

Otevřeme nový datový soubor s osmi proměnnými a jedním případem. Do prvních pěti proměnných napíšeme zadané hodnoty a do Dlouhých jmen posledních tří proměnných napíšeme vyjádření pro odhady diskriminačních skórů.

	1 X1	2 X2	3 X3	4 X4	5 X5	6 L1	7 L2	8 L3
1	51,8	6	7	4	51	53,0996	55,23138	54,36618

Největší hodnotu má skór ve 2. skupině, tedy zkoumaná rodina vydá za dovolenou střední částku.

Dále v LDA použijeme pro výběr proměnných krokovou metodu.

Výsledky pro krokovou dopřednou metodu

Proměnné obsažené v modelu

N=50	Wilk. Lambda	Parc. Lambda	F na vyj (2,45)	p-hodn.	Toler.	1-toler. R^2
	Výsledky diskriminační funkční analýzy (dovolena.sta) krok 3, poč. prom. v modelu: 3; grupovací: ID2 (3 skup) Wilk. lambda: ,27663 přibliž F (6,90)=13,519 p< ,0000					
X1	0,652311	0,424084	30,55552	0,000000	0,984948	0,015052
X5	0,338537	0,817147	5,03482	0,010635	0,953070	0,046930
X2	0,303098	0,912692	2,15236	0,128024	0,967370	0,032630

Klasifikační funkce

Proměnná	malá p=,24000	střední p=,48000	velká p=,28000
	Klasifikační funkce; grupovací : ID2 (dovolena.sta)		
X1	0,6401	0,8551	1,1311
X5	0,8991	0,7824	0,9163
X2	2,3409	2,4846	3,1046
Konstant	-41,3768	-44,8553	-70,5840

Klasifikační matice

Skup.	% správnýc	malá p=,24000	střední p=,48000	velká p=,28000
	Klasifikační matice (dovolena.sta) Řádky: pozorované klasifikace Sloupce: předpovězené klasifikace			
malá	75,00000	9	3	0
střední	83,33334	3	20	1
velká	78,57143	0	3	11
Celkem	80,00000	12	26	12

Úspěšnost klasifikace poklesla z 82 % na 80 %.

Výsledky pro krokovou zpětnou metodu

Proměnné obsažené v modelu

	Výsledky diskriminační funkční analýzy (dovolena.sta) krok 4, poč. prom. v modelu: 1; grupovací: ID2 (3 skup) Wilk. lambda: ,36521 přibliž F (2,47)=40,846 p< ,0000					
N=50	Wilk. Lambda	Parc. Lambda	F na vyj (2,47)	p-hodn.	Toler.	1-toler. R^2
X1	1,000000	0,365211	40,84639	0,000000	1,000000	0,00

Klasifikační funkce

	Klasifikační funkce; grupovací : ID2 (dovolena.sta)		
Proměnná	malá p=,24000	střední p=,48000	velká p=,28000
X1	0,7506	0,9498	1,2413
Konstant	-15,7327	-23,6411	-40,3976

Klasifikační matice

	Klasifikační matice (dovolena.sta) Řádky: pozorované klasifikace Sloupce: předpovězené klasifikace			
Skup.	% správnýc	malá p=,24000	střední p=,48000	velká p=,28000
malá	83,3333	10	2	0
střední	100,0000	0	24	0
velká	78,5714	1	2	11
Celkem	90,0000	11	28	11

Je-li ke klasifikaci rodin do skupin použita pouze proměnná X_1, je úspěšnost klasifikace nejvyšší, a to 90 %.
Aplikujeme-li toto klasifikační pravidlo na rodinu s vektorem pozorování (51,8 6 7 4 51)', dostaneme výsledek

	1 X1	2 X2	3 X3	4 X4	5 X5	6 L1	7 L2	8 L3
1	51,8	6	7	4	51	23,14838	25,55854	23,90174

Mahalanobisovy vzdálenosti mezi skupinami a jejich statistická významnost

V systému STATISTICA lze vypočítat kvadrát Mahalanobisovy vzdálenosti mezi všemi dvojicemi skupin a získat p-hodnotu pro test hypotézy, že tyto vzdálenosti jsou nulové.
V panelu Diskriminační analýza vybereme záložku Detaily a poté Vzdálenosti mezi skupinami.

(Uvedené výsledky jsou pro případ, kdy k diskriminaci použijeme všechny proměnné)

	Mahalanobisovy vzdálenosti^2 (dovolena.sta)		
ID2	malá	střední	velká
malá	0,00000	3,227044	14,35858
střední	3,22704	0,000000	6,14948
velká	14,35858	6,149479	0,00000

	p-hodnot (dovolena.sta)		
ID2	malá	střední	velká
malá		0,001585	0,000000
střední	0,001585		0,000002
velká	0,000000	0,000002	

Všechny tři dvojice skupin se liší na hladině významnosti 0,05, nejvíce pak skupina 1 a 3.

Poznámka o klasifikaci objektů pomocí umělých neuronových sítích

Diskriminaci objektů je možno provádět také pomocí neuronových sítí. Ty nekladou žádné předběžné požadavky na data (normalita, homogenita variančních matic, linearita vztahů). Použití neuronových sítí v systému STATISTICA ukážeme na datovém souboru dovolena.sta, a to jak pro klasifikaci do dvou skupin, tak do tří skupin.
Statistiky – Automatizované neuronové sítě – Nová analýza – Klasifikace – OK – Proměnné – Kategorická cílová proměnná: ID1, Spojité prediktory: X1 až X5 – OK. Na záložce Vzorkování zadáme velikost trénovací množiny 100 % (kvůli porovnání výsledků s výsledky kanonické DA nebo lineární DA). Velikosti zbylých dvou množin jsou pak 0 %. Následně zvolíme tlačítko Trénovat. Zjistíme, že všechny sítě poskytly trénovací výkon 100 %. Pomocí tlačítka Výběr aktivních sítí vybereme např. síť s indexem 1 – OK. Na záložce Detaily vybereme Matice záměn:

	ID1 (Souhrn klasifikací) (dovolena) Vzorky: Trénovací			
		ID1-návštěva ano	ID1-návštěva ne	ID1-Všechny
1.MLP 5-9-2	Celkem	21,0000	29,0000	50,0000
	Správné	21,0000	29,0000	50,0000
	Chybné	0,0000	0,0000	0,0000
	Správné (%)	100,0000	100,0000	100,0000
	Chybné (%)	0,0000	0,0000	0,0000

Vidíme, že všechny rodiny byly správně klasifikovány, což je lepší výsledek než poskytla LDA.

Na záložce Vlastní predikce můžeme zadat vektory pozorování objektů s neznámou příslušností ke skupině. Použijeme údaje o rodině, jejíž vektor pozorování je 51,8 6 7 4 51.
Získáme tabulku Vlastní predikce:

	Tabulka s uživatelskými predikcemi (dovolena)					
Případy	1.ID1_(t)	X1	X2	X3	X4	X5
1	návštěva ano	51,80000	6,000000	7,000000	4,000000	51,00000

Neuronová síť zařadila tuto rodinu do skupiny rodin, které danou oblast navštěvují.

Stejný postup zopakujeme pro klasifikaci rodin do tří skupin podle proměnné ID2.
První čtyři sítě mají trénovací výkon 100 %, pátá 98 %. Vybereme první síť.

Matice záměn (tj. klasifikační matice):

	ID1 (Souhrn klasifikací) (dovolena) Vzorky: Trénovací			
		ID1-návštěva ano	ID1-návštěva ne	ID1-Všechny
1.MLP 5-9-2	Celkem	21,0000	29,0000	50,0000
	Správné	21,0000	29,0000	50,0000
	Chybné	0,0000	0,0000	0,0000
	Správné (%)	100,0000	100,0000	100,0000
	Chybné (%)	0,0000	0,0000	0,0000

Neuronová síť opět dosáhla lepšího výsledku než LDA.

Poznámka o kvadratické diskriminační analýze

Kvadratické diskriminační analýza se používá v situacích, kdy p-rozměrné vektory pozorování objektů v daných r skupinách pocházejí z normálních rozložení, která mají rozdílné varianční matice.
Při klasifikaci objektů se používají kvadratické diskriminační skóry

$$Q_h(\mathbf{x}) = -\frac{1}{2}\ln(\det \mathbf{\Sigma}_h) - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_h)' \mathbf{\Sigma}_h^{-1}(\mathbf{x}-\boldsymbol{\mu}_h) + \ln \pi_h,\ h = 1, \ldots, r,$$

přičemž v praxi neznámé parametry $\boldsymbol{\mu}_h$, $\mathbf{\Sigma}_h$ a π_h nahradíme jejich odhady $\mathbf{M}_h$, $\mathbf{S}_h$ a p_h. Tím získáme odhad kvadratického diskriminačního skóru

$$\hat{Q}_h(\mathbf{x}) = -\frac{1}{2}\ln(\det \mathbf{S}_h) - \frac{1}{2}(\mathbf{x}-\mathbf{M}_h)' \mathbf{S}_h^{-1}(\mathbf{x}-\mathbf{M}_h) + \ln p_h$$

Objekt s neznámou příslušností, jehož vektor pozorování je **x**, zařadíme do té skupiny, pro niž je $\hat{Q}_h(\mathbf{x})$ maximální.

QDA je velmi citlivá na porušení předpokladu normality. V systému STATISTICA není implementována.