

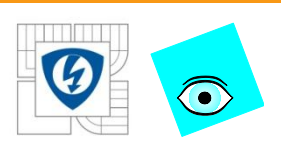
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

Počítačová analýza vícerozměrných dat v oborech přírodních, technických a společenských věd

Prof. RNDr. Milan Meloun, DrSc. (Univerzita Pardubice, Pardubice)

20.-24. června 2011

Tato prezentace je spolufinancována Evropským sociálním fondem a státním rozpočtem České republiky.

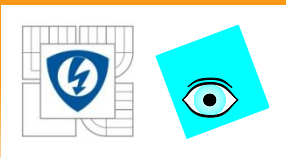


4.7 KLASIFIKACE OBJEKTŮ

25.2.2010

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ



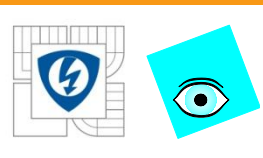


Hledání struktury a vzájemných vazeb v objektech

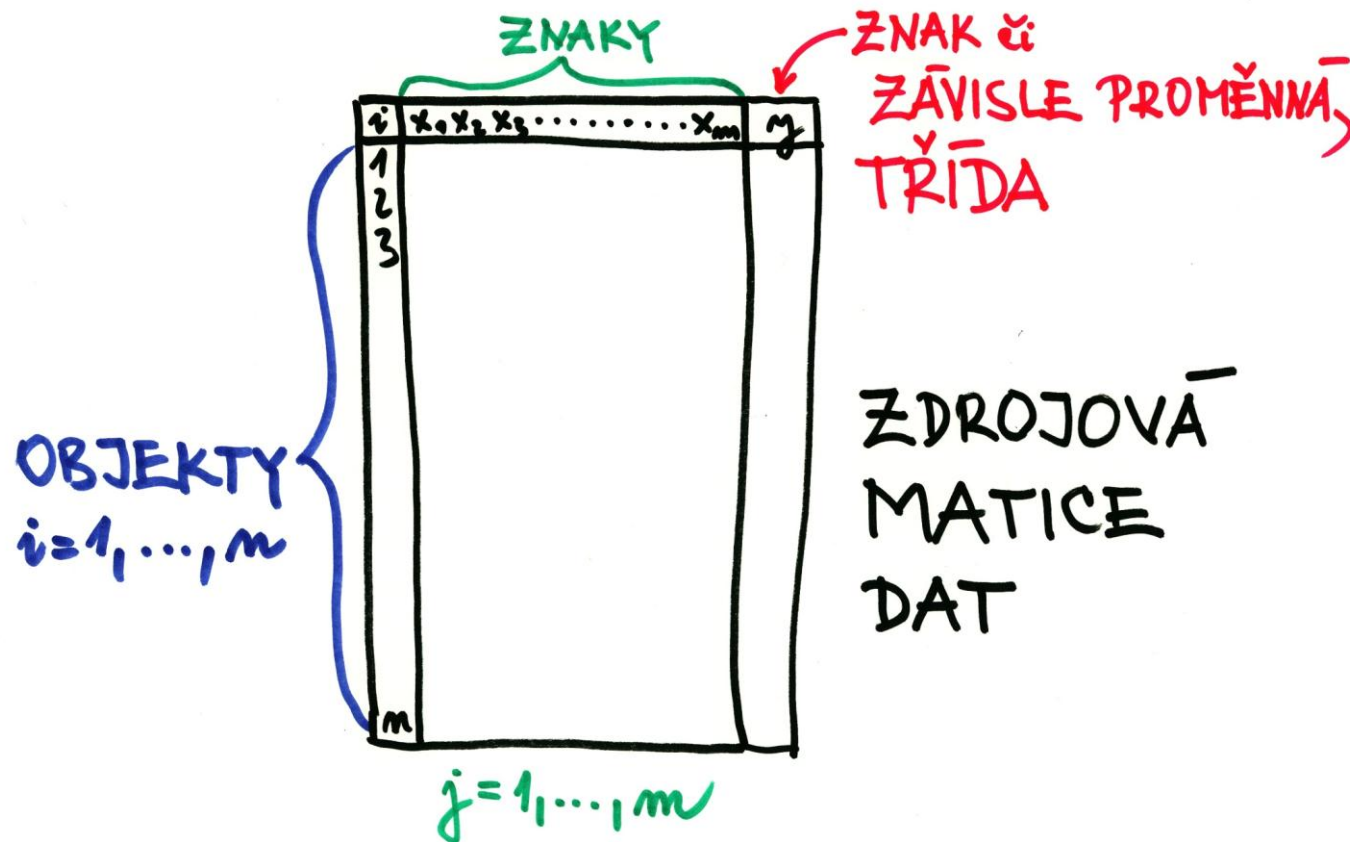
- 1) *Diskriminační analýza DA*: nový objekt se zařadí do již existující třídy.
- 2) *Analýza shluků CLU*: neuspořádanou skupinu objektů lze uspořádat do několika vnitřně sourodých tříd či shluků
- 3) *Vícerozměrné škálování MDS*: hledá strukturu a vazby mezi objekty na základě jejich podobnosti.

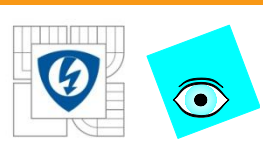
Analyzovaný výběr jsou trénovací data, která obsahují pro každý objekt jak výstup y , tak i hodnoty všech znaků $x, \{y; x_1, x_2, \dots, x_m\}$

Klasifikovaný výběr: na základě analyzovaného výběru sestavit predikční model, který umožní zařazení nových objektů do tříd.



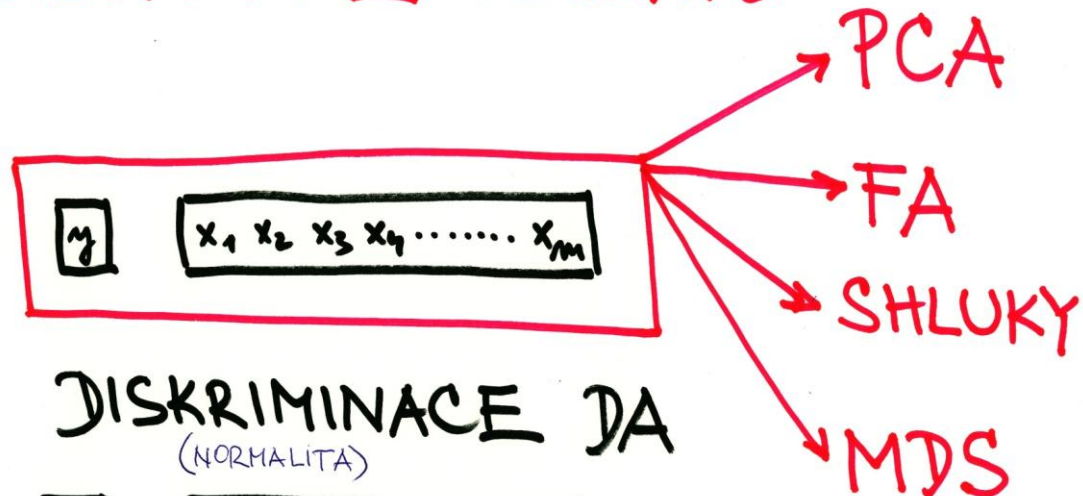
Zdrojová matice dat





Klasifikace objektů

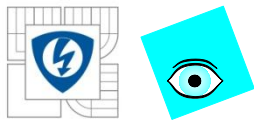
KLASIFIKACE OBJEKTŮ



DISKRIMINACE DA
(NORMALITA)

y $x_1 x_2 x_3 x_4 \dots x_m$
TRÍDA PROMĚNNÉ

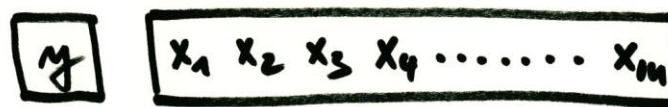
LOGISTICKÁ REGRESE, Logit
(NENORMALITA)



Diskriminační analýza

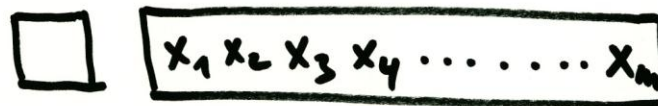
DISKRIMINAČNÍ ANALÝZA

TRÉNOVÁNÍ S UČITELEM (SUPERVISED LEARNING)



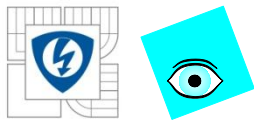
→ MODEL
(FISHEROVA
FUNKCE)

1. TRÉNOVACÍ DATA
(ANALYZOVANÝ SOUBOR)



← VYUŽITÍ
MODELU

2. KLASIFIKACE OBJEKTŮ
(KLASIFIKOVANÝ SOUBOR)

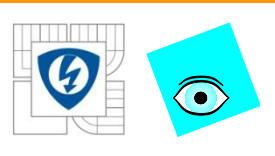


Trénování bez učitele

TRÉNOVÁNÍ BEZ UČITELE (UNSUPERVISED LEARNING)

□ $x_1 x_2 x_3 \dots x_n$

1. obvykle PCA, FA, SHLUKY
2. DA – KLASIFIKAČNÍ SOUBOR



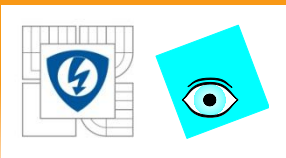
Trénování bez učitele

- 1) **Statistické učení s učitelem (supervised learning):** u dat analyzovaného výběru se sestaví predikční model $y = f(x)$, který umožní predikovat výstup pro nový objekt z dat klasifikovaného výběru x_0 .
- 2) **Statistické učení bez učitele (unsupervised learning):** k dispozici jsou určité znaky pro objekty a nikoliv výstupy. Úlohou je pak pouze stanovit shluky.

Výstupem klasifikace jsou kvalitativní proměnné označující pouze třídy (jako je zdravý, nemocný, nebo nevyhovující, dobrý, vyhovující atd.).

Často jsou výstupem pouze dvě kategorie (binární výstup) kódované jako 0,1, respektive NULA, JEDNA, nebo-1,1.

Když **G** obsahuje pouze dvě třídy (binární proměnné 0 a 1), využívá se náhrady **G** kvantitativním výstupem s tím, že predikce leží v intervalu $[0,1]$ a zařazuje do třídy NULA pro < 0.5 , resp. do třídy JEDNA pro > 0.5 .

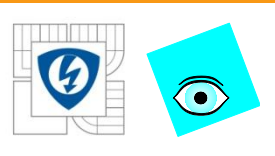


DISKRIMINAČNÍ ANALÝZA

25.2.2010

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ





Hodnotí rozdíly mezi dvěma nebo více skupinami objektů charakterizovaných pomocí několika znaků.

Dělí se na

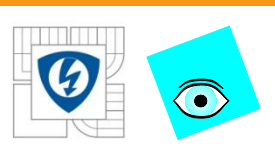
- 1) Techniky, které *vysvětlují rozdíly* mezi předem danými shluky objektů.
- 2) Techniky, u kterých je cílem *klasifikace objektů do skupin*. Jsou porovnávány znaky jednoho objektu (např. charakteristiky sloučenin, vlastnosti pacientů,...) se znaky ostatních.

Příklad: Analyzovaný výběr obsahuje *třídu přežívajících* pacientů a *třídu nepřežívajících*. Dle klinických vyšetření chceme rozhodnout, kam nového pacienta z klasifikovaného výběru zařadit. Potřebujeme znaky k rozlišení dosud nezařazených pacientů do tříd.

Klasická DA (Ronald Fisher, 1936): zkoumání vztahu mezi skupinou p nezávislých znaků (*diskriminátory*) a jednou kvalitativní závisle proměnnou y nabývající hodnotu 0 (objekt je v 1. třídě), nebo hodnotu 1 (objekt je ve 2. třídě).

Třídy jsou zřetelně *odlišené*. Každý objekt patří do jedné ze dvou tříd.

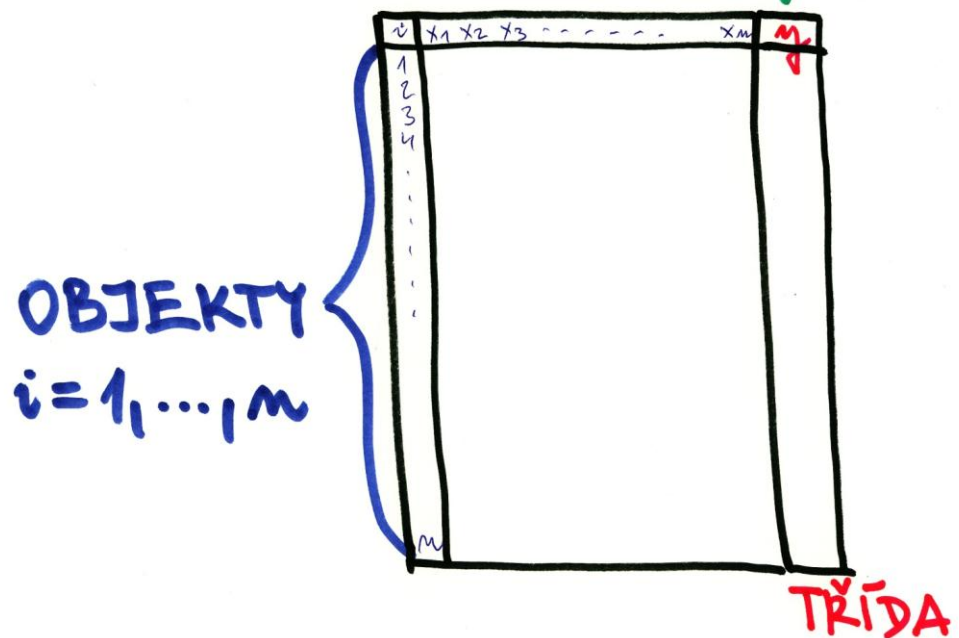
Cílem je nalézt znaky, které úspěšně přispívají do procesu *třídění*

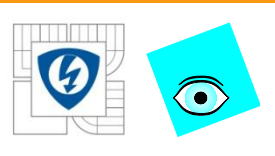


Diskriminační analýza

DISKRIMINAČNÍ ANALÝZA

DISKRIMINÁTOR, $j=1, \dots, m$





Diskriminační analýza

V datech analyzovaného výběru jsou pomocí znaků objekty umístěny do tříd.

V klasifikovaném výběru je třeba nalézt *predikční model* který umístí nové objekty do stávajících tříd.

Řešení: v analyzovaném výběru

π_1 značí apriorní pravděpodobnost příslušnosti objektu do první třídy,

$\pi_2 = 1 - \pi_1$ značí pravděpodobnost příslušnosti objektu do druhé třídy.

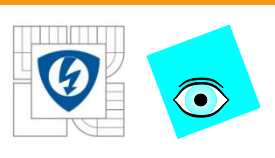
Pomocí Bayesovy věty pak určíme aposteriorní pravděpodobnost příslušnosti k j -té skupině ($j = 1, 2$)

$$P(G = j | \mathbf{x}) = \frac{f_j(\mathbf{x})\pi_j}{\sum_{i=1}^2 f_i(\mathbf{x})\pi_i}, \quad j = 1, 2$$

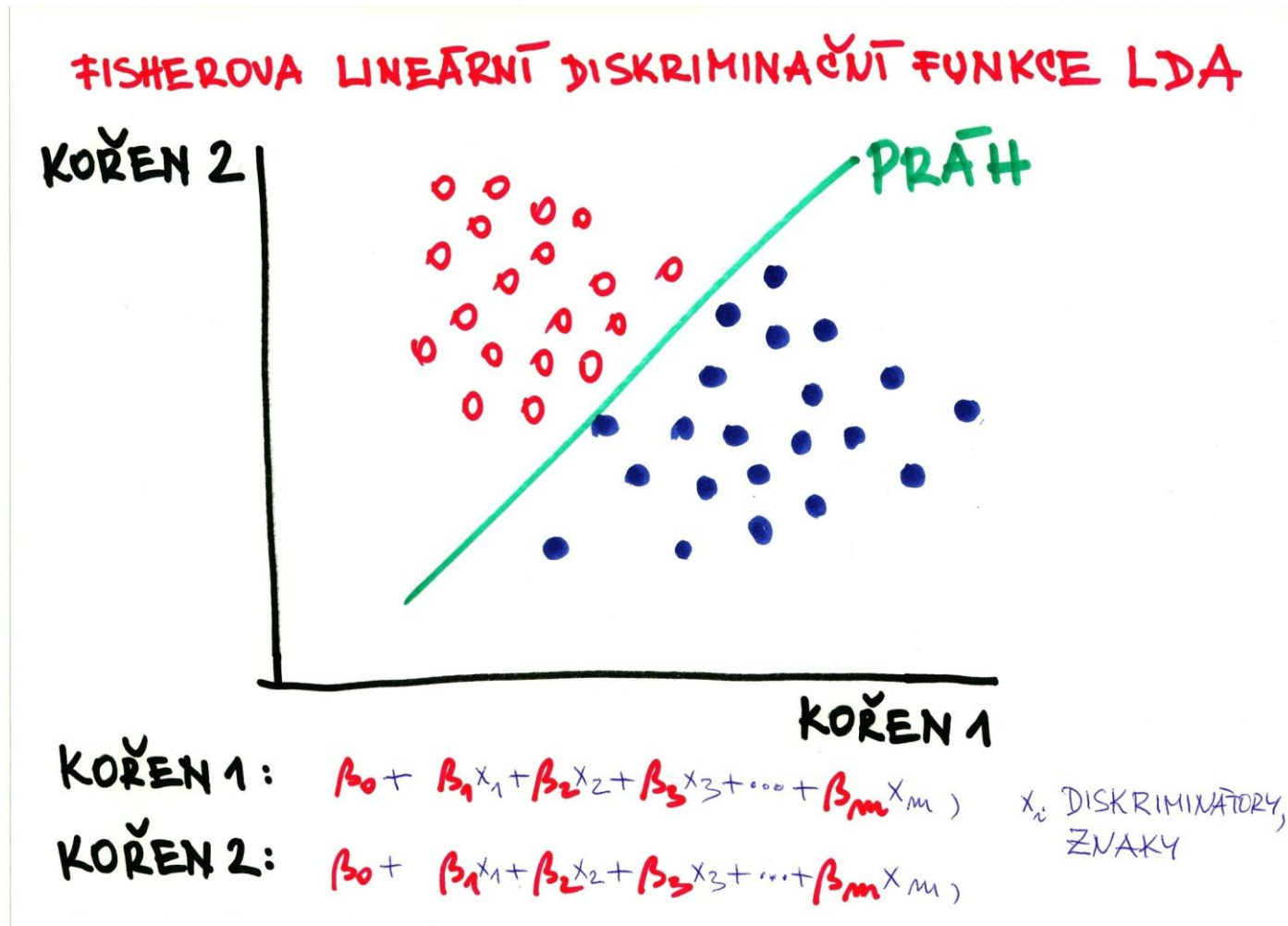
Zápis se dá snadno rozšířit i pro více kategorií.

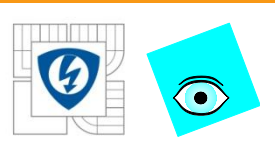
Třídění:

Objekt zařadíme do první skupiny, pokud $\pi_1 f_1(\mathbf{x}) > \pi_2 f_2(\mathbf{x})$. Pro zařazení objektu do první skupiny musí platit, že $\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} > \pi_2 / \pi_1$.



Fischerova lineární diskriminační funkce LDA





Zařazovací pravidla DA

účelem je klasifikace do jedné ze dvou tříd, A nebo B

Klasifikace na základě jediného znaku x (s normálním rozdělením):

ve třídě A jde o rozdělení $N(\mu_A, \sigma_A^2)$,

ve třídě B jde o rozdělení $N(\mu_B, \sigma_B^2)$,

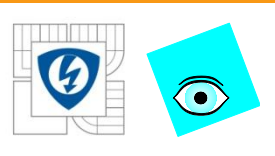
nový objekt má hodnotu jediného znaku x ,

vybrat třídu, pro kterou je x blíže ke střední hodnotě dané třídy μ_A či μ_B .

Určíme *prahový bod* $C = (\mu_A + \mu_B)/2$:

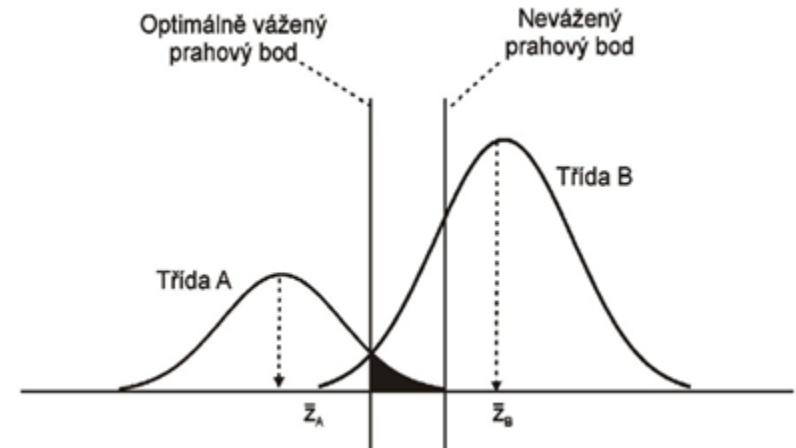
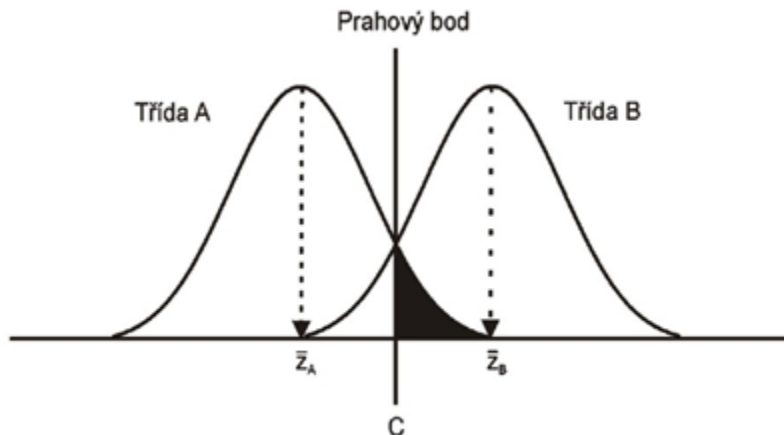
je-li $x < C$, pak se objekt zařadí do první třídy, A,

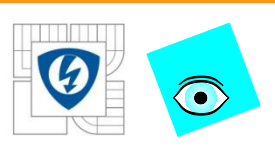
je-li $x \geq C$ pak se objekt zařadí do druhé třídy, B.



Zařazovací pravidla DA

- 1) $\sigma_A^2 = \sigma_B^2$ pravděpodobnost nesprávné klasifikace je šrafovaně. Tato souvisí se vzdáleností obou rozdělení a s jejich rozptýly.
- 2) $\sigma_A^2 < \sigma_B^2$ je pravděpodobnost nesprávné klasifikace pro třídu A větší než pro třídu B. Bude proto třeba penalizovat C s ohledem na nestejný rozptyl.

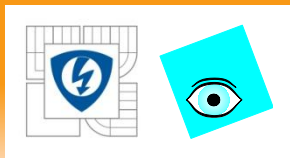




Zařazovací pravidla DA

- 1) Pro **shodné kovarianční matice** $\mathbf{C}_A = \mathbf{C}_B$ a stejné apriorní pravděpodobnosti zařazení do kategorií $\pi_1 = \pi_2$ bude **dělicí čarou** pro obě třídy přímka $a_1 x_1 + a_2 x_2 = C$ jako *lineární diskriminace LDA*.
- 2) Pro **různé kovarianční matice** $\mathbf{C}_A \neq \mathbf{C}_B$ bude **dělicí čára** definovaná polynomem druhého stupně jako *kvadratická diskriminace QDA*.

Zařazování do tříd se takto převede na výpočet hodnoty funkce $Z_i = a_1 x_{1i} + a_2 x_{2i}$ a porovnání Z_i s prahem C .



Lineární LDA a kvadratická QDA diskriminační funkce

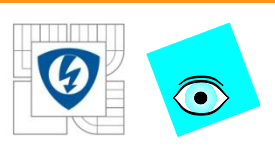
Dle typu hustot pravděpodobnosti pro znaky $f_1(\mathbf{x})$ a $f_2(\mathbf{x})$:

- 1) Pro normální rozdělení, lišící se jen středními hodnotami tříd je *lineární diskriminační analýzu LDA*.
- 2) Pro normální rozdělení, lišící se středními hodnotami a kovariančními maticemi tříd je *kvadratickou diskriminační analýzu QDA*.
- 3) Pro směsi normálních rozdělení je *nelineární diskriminační funkce*.
- 4) Pro neparametrické hustoty rozdělení znaků ve třídách je *flexibilní diskriminační funkce*.

Pro vícerozměrné Gaussovo rozdělení v i -té třídě má odpovídající hustota pravděpodobnosti tvar

$$f_i(\mathbf{x}) = \frac{1}{(2\pi)^{m/2} \det(C_i)^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_i)^T C_i^{-1}(\mathbf{x} - \boldsymbol{\mu}_i)\right]$$

kde $m=2$ je počet znaků.



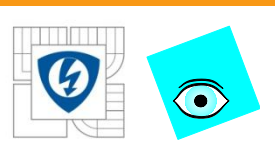
Lineární diskriminační funkce LDA

Pro dvě třídy 1 a 2 vychází LDA z předpokladu, že $\mathbf{C}_1 = \mathbf{C}_2 = \mathbf{C}$.
Logaritmus poměru aposteriorních pravděpodobností je

$$\ln \frac{P(G = 1 | \mathbf{x})}{P(G = 2 | \mathbf{x})} = \ln \left(\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} + \ln \frac{\pi_1}{\pi_2} \right) =$$
$$= \ln \frac{\pi_1}{\pi_2} - 0.5(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \mathbf{C}^{-1} (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) + \mathbf{x}^T \mathbf{C}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$$

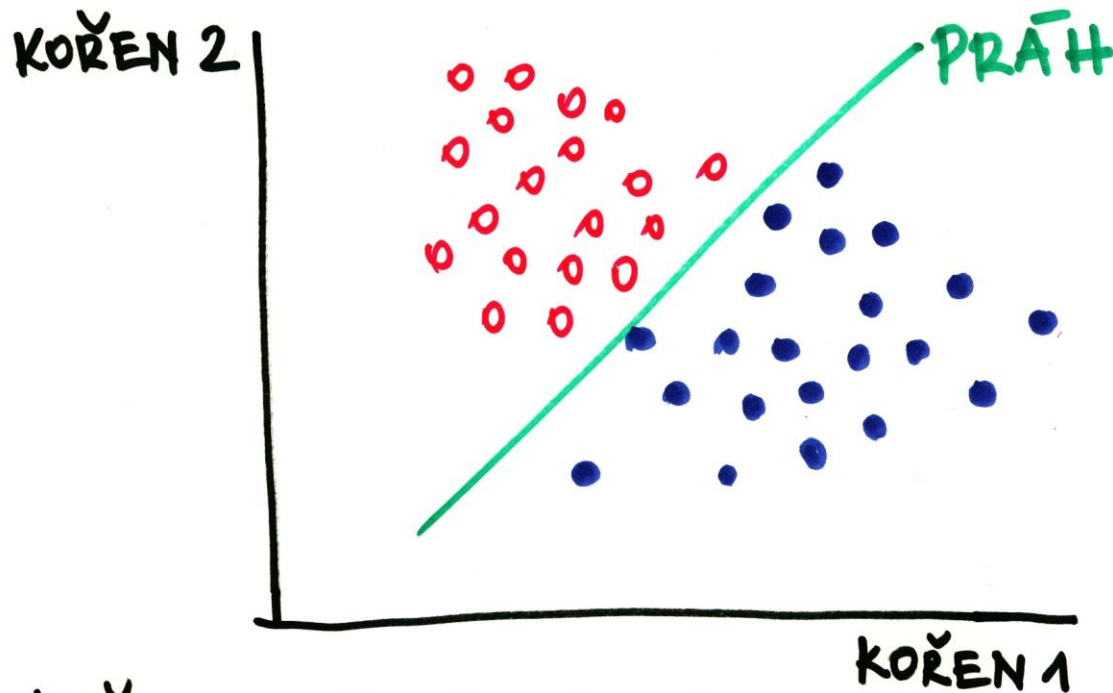
Vzhledem k $\mathbf{x} = (x_1, x_2)^T$ jde o lineární funkci a dělicí funkcí bude přímka. Dělicí funkce je množinou bodů $\mathbf{x}$, pro které platí
 $P(G = 1 | \mathbf{x}) = P(G = 2 | \mathbf{x})$.

Pro více znaků jsou dělicí čáry lineárními úseky procházejícími průsečíky jednotlivých elips konstantní hustoty pro jednotlivé třídy.



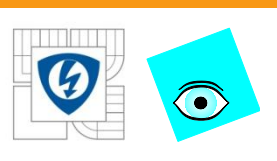
Fisherova lineární diskriminační funkce LDA

FISHEROVA LINEÁRNÍ DISKRIMINAČNÍ FUNKCE LDA

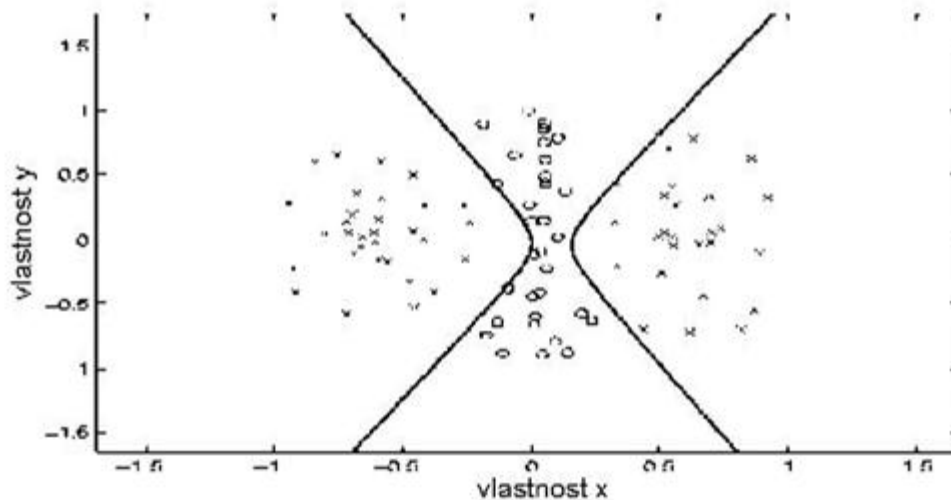


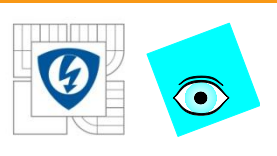
KOŘEN 1: $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \dots + \beta_m x_m$
KOŘEN 2: $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \dots + \beta_m x_m$

x_i DISKRIMINÁTORY,
ZNÁKY



LDA pro dva znaky a tři kategorie: kdy rozdíly jsou pouze ve středních hodnotách $\mu_1 \neq \mu_2$ a shodné rozptyly $\sigma_1^2 = \sigma_2^2$



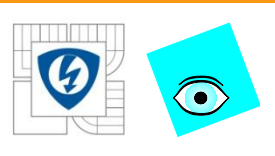


Pravidlo pro zařazení do první třídy

Pravidlo pro zařazení do první třídy je ve tvaru $\mathbf{a}^T \mathbf{x} + \mathbf{b} > 0$, kde $\mathbf{a}^T$ je vektor koeficientů u lineárního členu $\mathbf{a}^T = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \mathbf{C}^{-1}$ a absolutní člen je $\mathbf{b} = -0.5 \mathbf{a}^T (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) - \ln(\pi_2/\pi_1)$

Lineární diskriminační funkce je $L(\mathbf{x}) = \mathbf{a}^T \mathbf{x}$.

Rozdělení funkce je normální a závisí na Mahalanobisově vzdálenosti mezi středními hodnotami v obou třídách



Pravidlo pro zařazení do první třídy

$$D_M^2 = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \mathbf{C}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$$

Pro obě třídy pak platí,

$$E(L(\mathbf{x})) = \frac{D_M^2}{2}, \quad D(L(\mathbf{x})) = D_M^2.$$

Dělicí rovina (přímka pro $m = 2$) je určena rovnicí $L(\mathbf{x}) + \mathbf{b} = \mathbf{a}^T \mathbf{x} + \mathbf{b} = 0$.

Chyba nesprávné klasifikace ω se vyčíslí

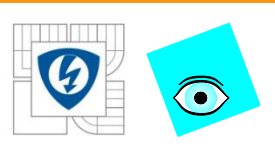
$$\omega = \pi_1 \varphi \left(\frac{\ln \left(\frac{\pi_1}{\pi_2} \right) - \frac{D_M^2}{2}}{D_M} \right) + \pi_2 \varphi \left(\frac{-\ln \left(\frac{\pi_2}{\pi_1} \right) - \frac{D_M^2}{2}}{D_M} \right)$$

kde $\varphi(x)$ je distribuční funkce normovaného normálního rozdělení.

Pro neinformativní apriorní pravděpodobnosti $\pi_1 = \pi_2 = 0.5$ vyjde jednoduchý vztah pro **chybu nesprávné klasifikace** $\omega = \varphi(-D_M/2)$.

Chyba nesprávné klasifikace do první třídy je stejná jako chyba klasifikace do druhé třídy.

Místo lineární diskriminační funkce je možné použít **lineární diskriminační kritérium**, které nevyžaduje stanovení koeficientů $\mathbf{a}$, $\mathbf{b}$.



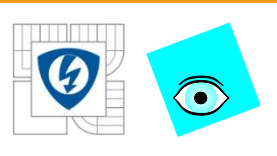
Pravidlo pro zařazení do první třídy

Kritérium $Lk_j(\mathbf{x})$ se počítá pro každou j -tou třídu zvlášť.

Při klasifikaci objektů se pak objekt zařazuje do l -té třídy, pro kterou vyjde $Lk_l(\mathbf{x}) = \max\{Lk_j(\mathbf{x})\}$.

Pro LDA má **lineární diskriminační kritérium** tvar

$$Lk_j(\mathbf{x}) = \mathbf{x}^T \mathbf{C}^{-1} \boldsymbol{\mu}_j - 0.5 \boldsymbol{\mu}_j^T \mathbf{C}^{-1} \boldsymbol{\mu}_j + \ln \pi_j$$



Příklad: použití LDA pro dvě třídy $m = 2$.

Vychází se ze známých matic:

$\mathbf{X}_1$ rozměru $n_1 \times m$ pro třídu 1, $\mathbf{X}_2$ rozměru $n_2 \times m$ pro třídu 2. Jednotlivé objekty v matici $\mathbf{X}$ všech dat se zařadí do tříd podle výstupu $\mathbf{G}$.

Postup:

1) Vyčíslí se výběrové průměry $\bar{x}_1$ a $\bar{x}_2$ a společná kovarianční matice

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

2a) Nejjednodušší je předpoklad $\pi_1 = \pi_2 = 0.5$.

2b) Pokud je výběr informativní a byl pořízen jako celek a pak rozdělen do skupin, je možné použít relativních četností

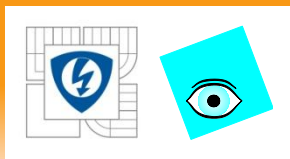
$$\hat{\pi}_1 = \frac{n_1}{n_1 + n_2} \text{ a } \hat{\pi}_2 = \frac{n_2}{n_1 + n_2}.$$

3) Za normality se určí koeficienty Fisherovy lineární diskriminační funkce z odhadů

$$\hat{\mathbf{a}} = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)\mathbf{S} \text{ a } \hat{\mathbf{b}} = -0.5\mathbf{a}^T(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) - \ln\left(\frac{\hat{\pi}_2}{\hat{\pi}_1}\right).$$

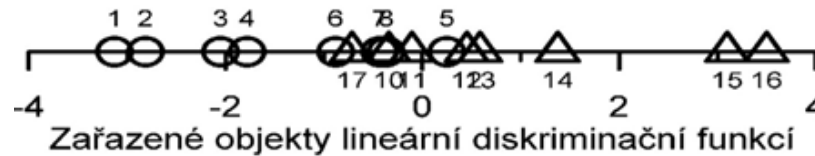
4) Při zařazování nových objektů s hodnotami znaků $\mathbf{j}_{c_0}$ se použije pravidlo, že do první skupiny je objekt klasifikován pokud platí $\hat{\mathbf{a}}\mathbf{x}_0 + \hat{\mathbf{b}} \geq 0$.

5) V opačném případě se klasifikuje do druhé skupiny.



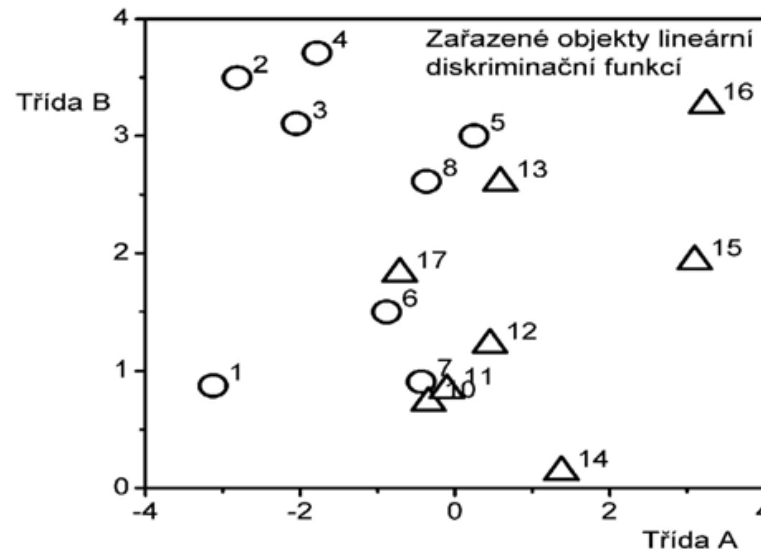
Ukázky

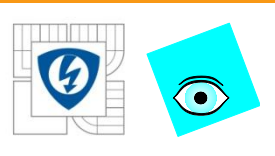
Ukázka 1: zařazování Fisherovou lineární diskriminační funkcí:



Zařazené objekty lineární diskriminační funkcí

Ukázka 2: zařazování Fisherovou lineární diskriminační funkcí:





Příklad 4.22 Třídění lebek Tibetanů lineární diskriminační funkcí

Databáze lebek na pohřebištích v Tibetu svědčí o dvou skupinách lidí: prvních 13 bylo nalezeno v hrobech v Sikkimu a okolí, druhých 15 lebek na bojištích okolo Lhasy. Předpokládejme, že máme data o 2 třídách tibetských lebek.

Data: i index lebky,

x_1 největší délka lebky [mm],

x_2 nej větší horizontální šířka lebky [mm],

x_3 výška lebky [mm],

x_4 výška horní části obličeje [mm],

x_5 šířka obličeje mezi body lícních kostí [mm].

x_1	x_2	x_3	x_4	x_5
190.5	152.5	145	73.5	136.5
...	...	...	...	...
...	...	...	...	...
172.5	132	125.5	63	121
167	130	125.5	69.5	119.5
1825	131	135	68.5	136



Řešení

1) Analyzovaný výběr: na trénovacích datech vytvoříme třídy
První třída (13 lebek z hrobů v Sikkimu) vede ke středním hodnotám
[174.82, 139.35, 132.00, 69.82, 130.35]

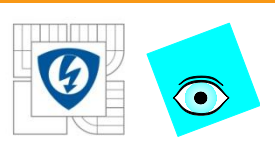
a kovarianční matici S_1

Druhá třída (15 lebek z bojišť v Lhase) vede ke středním hodnotám
[185.73, 138.73, 134.77, 76.47, 137.50]

a kovarianční matici S_2

$$S_1 = \begin{vmatrix} 45.53 & & & & \\ 25.22 & 57.81 & & & \\ 12.39 & 11.88 & 36.09 & & \\ 22.15 & 7.52 & -0.31 & 20.94 & \\ 27.97 & 48.06 & 1.41 & 16.77 & 66.21 \end{vmatrix}$$

$$S_2 = \begin{vmatrix} 74.42 & & & & \\ -9.52 & 37.35 & & & \\ 22.74 & -11.26 & 36.32 & & \\ 17.79 & 0.70 & -10.72 & 15.30 & \\ 11.13 & 9.46 & 7.20 & 8.66 & 17.96 \end{vmatrix}$$



Koeficienty diskriminační funkce $a_1, \dots, a_5$

jsou vyčísleny podle vztahu

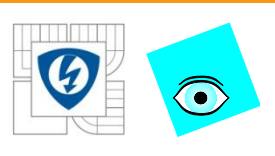
$$\mathbf{a} = \mathbf{S}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) = [-0.09, 0.16, 0.01, -0.18, -0.18]$$

a vedou k *průměrům obou tříd*: $\bar{Z}_1 = -28.71$ a $\bar{Z}_2 = -32.21$.

Optimální prahový bod C ,

dle kterého se budou nezařazené objekty třídit do první nebo druhé třídy, se vyčíslí jako polosuma obou průměrů dle vztahu

$$C = (\bar{Z}_1 + \bar{Z}_2)/2 = (-28.71 + (-32.21))/2 = -30.46.$$



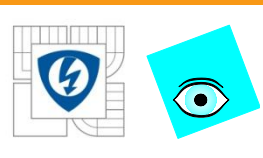
Diskriminace

2) Klasifikovaný výběr: na nových datech provedeme zařazení dosud nezařazených objektů Vezmeme *data pro lebku prvního Tibetana* a pokusíme se ji zařadit do 1. nebo do 2. třídy. Vyčísleme proto pro ni hodnotu *lineární diskriminační funkce*

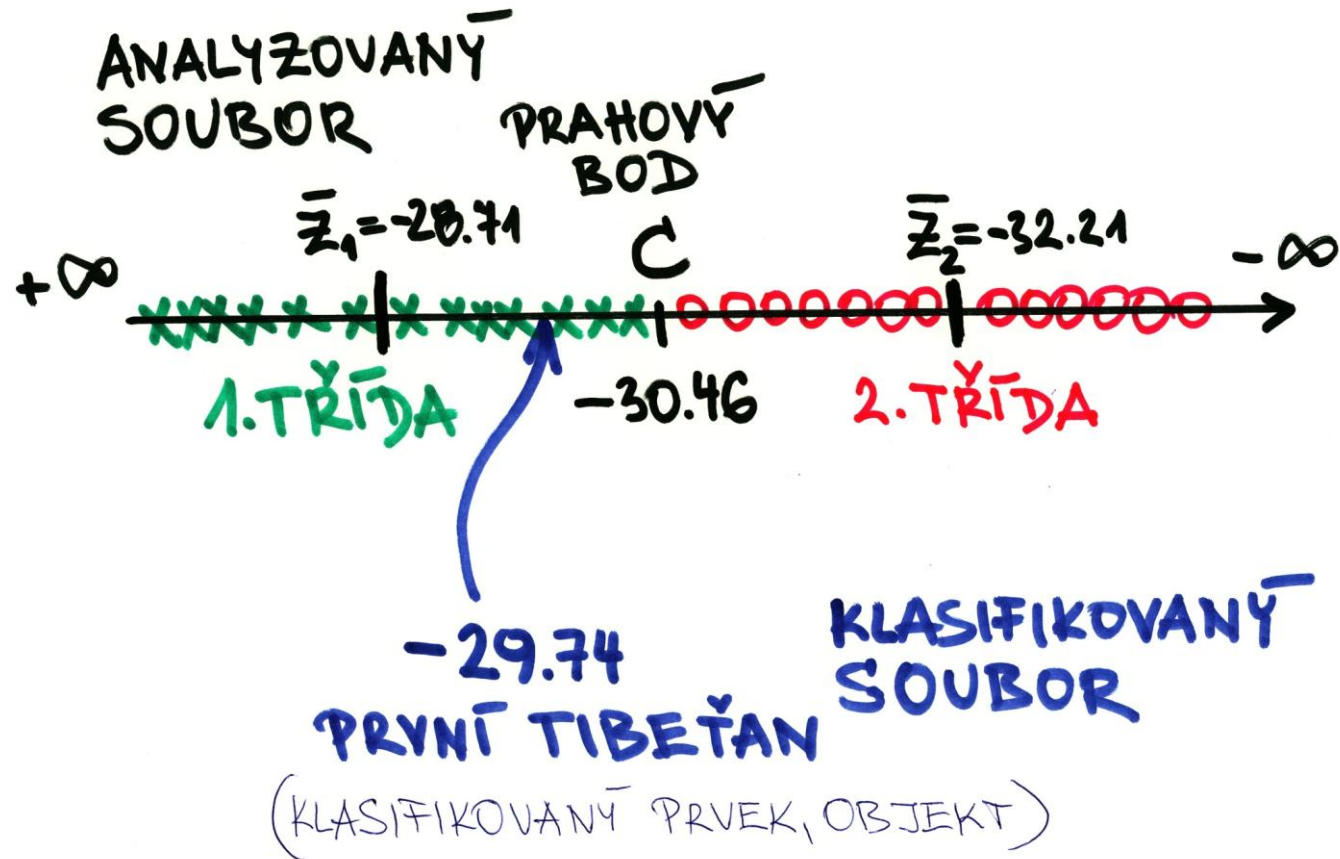
$$Z_1 = -0.09 \times 190.5 + 0.16 \times 152.5 + 0.01 \times 145.0 - 0.18 \times 73.5 - 0.18 \times 136.5$$
$$= -29.74,$$

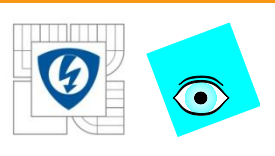
Závěr: Protože lineární diskriminační funkce $Z_1 = -29.74$ je menší než optimální prahový bod

$C = -30.46$, patří lebka prvního Tibetana do první třídy.



Analyzovaný soubor





Kvadratická diskriminační funkce QDA

Pokud nejsou kovarianční matice stejné, vede pravidlo pro zařazení do první skupiny $f_1(x)\pi_1 > f_1(x)\pi_2$ ke kvadratické nerovnosti

$$\mathbf{x}^T \mathbf{G} \mathbf{x} + \mathbf{h}^T \mathbf{x} + C > 0$$

kde matice $\mathbf{G} = 0.5 (\mathbf{C}_2^{-1} - \mathbf{C}_1^{-1})$

vektor $\mathbf{h}^T = \boldsymbol{\mu}_1^T \mathbf{C}_1^{-1} - \boldsymbol{\mu}_2^T \mathbf{C}_2^{-1}$

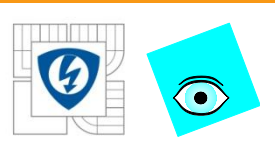
a konstanta $C = 0.5 \ln \frac{\det(\mathbf{C}_2)}{\det(\mathbf{C}_1)} - 0.5(\boldsymbol{\mu}_1^T \mathbf{C}_1^{-1} \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2^T \mathbf{C}_2^{-1} \boldsymbol{\mu}_2) - \ln(\frac{\pi_2}{\pi_1})$

Platí-li pro nové $\mathbf{x}_0$ tato kvadratická nerovnost, zařazuje se objekt do skupiny 1 a v opačném případě do skupiny 2.

Lze také definovat **kvadratické diskriminační kritérium**

$$QK_j(\mathbf{x}) = -0.5 \ln \det(\mathbf{C}_j) - 0.5 (\mathbf{x} - \boldsymbol{\mu}_j)^T \mathbf{C}_j^{-1} (\mathbf{x} - \boldsymbol{\mu}_j) + \ln \pi_j$$

Objekt $\mathbf{x}_0$ se pak zařazuje do třídy, které odpovídá maximální hodnota $QK_j(\mathbf{x}_0)$. Při kvadratické diskriminační analýze se objekty zařazují do tříd podle minima Mahalanobisových vzdáleností od středů tříd $\boldsymbol{\mu}_j$.



Úprava prahového bodu

Volba prahového bodu C poskytuje požadovaný poměr apriorních pravděpodobností π_1 a π_2 .

Optimální volba prahového bodu C je daná vzorcem $C =$

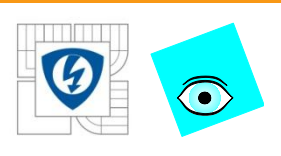
$$\frac{(\bar{Z}_1 + \bar{Z}_2)}{2} + \ln \frac{\pi_1}{\pi_2}$$

a když bude $\pi_1 = \pi_2 = 0.5$, bude prahový bod C roven $C = \frac{(\bar{Z}_1 + \bar{Z}_2)}{2}$

Standardizované koeficienty: hodnoty koeficientů $a_1, a_2, \dots, a_p$ **nejsou** přímo porovnatelné.

Relativní vliv na každou proměnnou v diskriminační funkci získáme ze *standardizovaných diskriminačních koeficientů*.

Standardizované koeficienty se vypočtou vynásobením koeficientů a_i odpovídající směrodatnou odchylkou s_i .

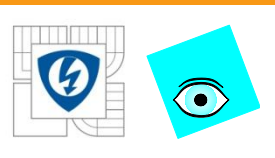


Volba znaků, diskriminátorů

Znaky musí zajistit přesné zařazení objektů do tříd čili diskriminaci.

Principem selekce znaků je zajištění dostatečné separability tříd a maximalizace některé zvolené míry:

- Začneme se všemi znaky. Pak se a vypouštějí se takové znaky, které provedou nedostatečnou separaci.
- V mnoha situacích může být DA užita jako exploratorní pomůcka.
- Nejprve jsou do dat zahrnuty všechny využitelné znaky. Na začátku není známo, které znaky jsou k zařazení objektů do tříd účinné.
- V diskriminační analýze místo testování změny hodnoty čtverce korelačního koeficientu R^2 přidáním nebo odebráním proměnné testujeme změnu Mahalanobisovy vzdálenosti D_M^2 .
- Užívají se stejná testovací kritéria jako při výběru nezávisle proměnných v lineární regresní analýze.
- **Krokový výběr znaků** kombinuje jak jejich přidávání, tak i jejich odstraňování.
- První znak, zahrnutý do modelu ve výběrovém kritériu, má největší přijatelnou hodnotu.
- **Vybírání znaků** se ukončí, když žádné další znaky nesplňují závaděcí nebo odstraňovací kritérium.



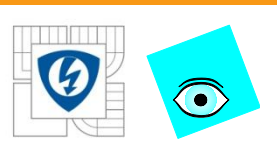
Kritéria k vybírání znaků

Wilkovo kritérium λ když znak poskytuje v diskriminační funkci nejmenší hodnotu Wilkova kritéria λ , je zahrnut do modelu:

- k zavedení nebo odstranění znaku je dovolen jeden krok.
- maximální počet krokuje roven dvojnásobku počtu znaků.
- **tolerance** je mírou stupně lineární asociace mezi znaky. Pro i -tý znak platí vztah $1 - R_i^2$, kde R_i^2 je čtverec vícenásobného korelačního koeficientu, když je uvažován i -tý znak za závisle proměnnou a když je uvažována regresní rovnice mezi tímto i -tým znakem a ostatními znaky.
- **využití tolerance**: malé hodnoty tolerance indikují, že i -tý znak je lineární kombinací ostatních znaků. Znaky s tolerancí menší než 0.001 nejsou do modelu zařazeny.

Testování: významnost změny Wilkova kritéria λ po zavedení znaku do modelu nebo odstranění z modelu je založena na testačním kritériu F .

Na začátku procesu vybírání znaků: tolerance a minimum tolerance jsou položeny rovny 1, protože v modelu zatím nejsou znaky.



F-test významnosti každého znaku

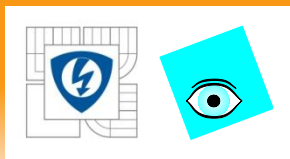
hodnota F pro změnu Wilkova kritéria λ při přidání znaku do modelu se vyčíslí dle

$$F_{zmeny} = \frac{n - g - p}{g - 1} \left(\frac{\frac{1 - \lambda_{p+1}}{\lambda_p}}{\frac{\lambda_{p+1}}{\lambda_p}} \right)$$

kde n je celkový počet objektů, g udává počet tříd, p je počet znaků,

λ_p značí Wilkovo lambda před přidáním a λ_{p+1} je Wilkovo lambda po přidání znaku do modelu.

Testování: Do modelu je zařazen ten znak, který způsobuje nejmenší hodnotu Wilkova kritéria λ .



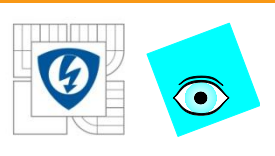
Raovo V kritérium, známé jako Lawley-Hotellingova stopa, je

$$V = (n - g) \sum_{i=1}^m \sum_{j=1}^m w_{ij} \sum_{k=1}^g (\bar{x}_{ik} - \bar{x}_i) (\bar{x}_{jk} - \bar{x}_j)$$

kde p udává počet znaků v modelu, g značí počet tříd, n_k je velikost k -té třídy, $\bar{x}_{ik}$ je střední hodnota i -tého znaku v k -té třídě, $\bar{x}$ je průměr i -tého znaku pro všechny třídy kombinované dohromady a w_{ij}^* představuje prvek inverzní kovarianční matice mezi třídami $\mathbf{B}^{-1}$

Testování: čím větší jsou rozdíly mezi středními hodnotami (průměry) tříd, tím větší je hodnota Raova V .

Může se však stát, že znak po zařazení do modelu sníží hodnotu Raova V .



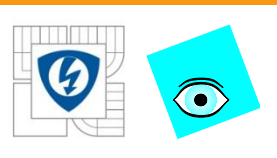
Mahalanobisova vzdálenost $D_{1,2}^2$

je zobecněná míra vzdálenosti mezi dvěma třídami 1 a 2 definovaná vztahem

$$D_{1,2}^2 = (n - g) \sum_{i=1}^m \sum_{j=1}^m w_{ij} (\bar{x}_{i1} - \bar{x}_{i2})(\bar{x}_{j1} - \bar{x}_{j2})$$

kde m udává počet znaků v modelu, $\bar{x}_{i1}$ je průměr i -tého znaku ve třídě 1, w_{ij} je prvek inverzní kovarianční matice $\mathbf{B}^{-1}$.

Testování: kritérium všech párů tříd vyčísleno je jako první. Znak, který měl pro dvě od začátku nejtěsnější třídy, nejmenší hodnotu $D_{1,2}^2$, je zařazen do modelu.



F-testační kritérium mezi třídami

testuje nulovou hypotézu H_0 : *dva vektory středních hodnot tříd objektů jsou stejné.*

Kritérium je postaveno na Mahalanobisově vzdálenosti $D_{1,2}^2$.

Testační kritérium F je definováno vztahem

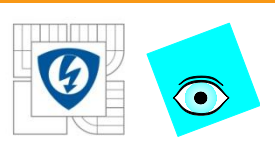
$$F = \frac{(n - 1 - p)n_1n_2}{p(n - 2)(n_1 + n_2)} D_{1,2}^2$$

a může být použito k výběru znaků.

Testování: V každém kroku je zařazen do modelu ten znak, který vykazuje nejvyšší hodnotu kritéria F .

Po zavedení dalších znaků do modelu jsou sledovány změny hodnoty testačního kritéria F_{zmeny} .

Proces přidávání znaků do modelu buď pokračuje, nebo je zastaven terminačním kritériem.



F-test vyšetření přidaného 1 znaku:

Testační kritérium je

$$F = \frac{(n_1 n_2)(n_1 + n_2 - p - 2)(D_{p+1}^2 - D_p^2)}{(n_1 + n_2)(n_1 + n_2 - 2) + (n_1 n_2)D_p^2}$$

které má Fisherovo-Snedecorovo rozdělení s 1 a $n_1 + n_2 - p - 2$ stupni volnosti.

Testování: Vyšetřujeme, zda přidaný znak x_{p+1} zlepší zařazení objektů o znacích $x_1, x_2, \dots, x_p$ významně zvětší hodnotu Mahalanobisovy vzdálenosti D_{p+1}^2 , čili $H_0: D_{p+1}^2 = D_p^2$.

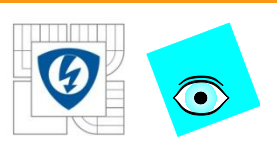
F-test o přidání p znaků současně:

Testační kritérium je

$$F = \frac{(n_1 n_2)(n_1 + n_2 - p - q - 2)(D_{p+q}^2 - D_p^2)}{q(n_1 + n_2)(n_1 + n_2 - 2) + (n_1 n_2)D_p^2}$$

má F -rozdělení s q a $n_1 + n_2 - p - q - 2$ stupni volnosti.

Testování: vyšetřujeme, zda další znaky $x_{p+1}, x_{p+2}, \dots, x_{p+q}$ zlepší predikci, $H_0: D_{p+q}^2 = D_p^2$.

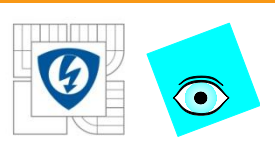


Procento správně zařazených objektů je jedním z indikátorů efektivity diskriminační funkce.

Aktuální diskriminační skóre ve třídách: "Dobrá" diskriminační funkce je taková, která má hodně mezitřídní proměnlivosti vzhledem k proměnlivostem přes všechny ostatní třídy **čili** koeficienty diskriminačních funkcí jsou voleny tak, že poměr sumy čtverců mezi třídami a součtem čtverců přes třídy je co možná největší.

Test H_0 : *V souborech není rozdíl mezi třídními průměry* je založen na Wilkově kritériu λ .

Pro dvě třídy představuje **Wilkovo kritérium λ** poměr sumy čtverců mezi třídami a totální sumy čtverců čili podíl celkového rozptylu v diskriminačním skóre, který je nevysvětlen rozdíly mezi třídami.

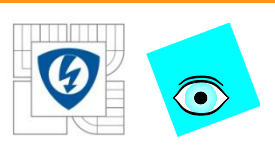


Testování

- a) *Malé hodnoty λ jsou spojeny s diskriminačními funkcemi, které mají hodně proměnlivosti mezi třídami a málo proměnlivosti přes třídy.*
- b) *Hodnota $\lambda = 1$ se objevuje, když průměr diskriminačního skóre je stejný ve všech třídách a když není žádná mezitřídní proměnlivost.*

Veličina λ je transformována do proměnné, která má χ^2 -rozdělení. Poskytuje pouze test, že průměry tříd jsou stejné. Malé rozdíly mohou být statisticky významné, ale nedovolují dobrou diskriminaci mezi třídami.

Jsou-li průměry a kovariance matic shodné, diskriminace není možná.



Kvalita zařazení objektů do tříd

Existují data o k třídách s n_j , $j = 1, \dots, k$, objekty v každé třídě a n představuje celkový počet objektů (např. $n = n_1 + n_2 + n_3 = 150$). Objekty jsou popsány m znaky a stupni volnosti $df1 = k - 1$ a $df2 = n - k$.

Cíl: Diskriminační funkce je váženým průměrem znaků. Problém úlohy spočívá v nalezení vah tak, aby dobře zařazovaly objekty do tříd.

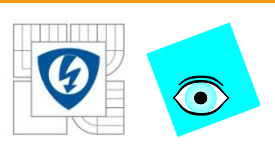
Mírou těsnosti proložení je Wilkovo kritérium

$$\lambda = \prod_{j=1}^m \frac{1}{1 + \lambda_j},$$

kde λ_j je j -té největší vlastní číslo matice $\mathbf{C}^{-1} \mathbf{B}$, m je minimum ze dvou čísel, $k-1$ a m .

Kanonická korelace mezi j -tou diskriminační funkcí a znaky je vztažena k těmto vlastním číslům $r_{cj} = \sqrt{\frac{\lambda_j}{1 + \lambda_j}}$

Korelace mezi znaky a kanonickými proměnnými jsou dány $r_{jk} = \frac{1}{\sqrt{c_{jj}}} \sum_{i=1}^p v_{ik} c_{ji}$



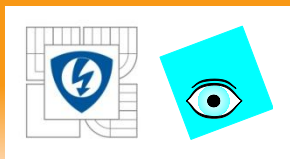
Logistická diskriminace

Nejprve se vyšetřuje normalita a shodnost kovariančních matic:

- 1) *Vícerozměrné normální rozdělení se stejnými kovariančními maticemi*: užije se **klasická LDA** k výpočtu pravděpodobnosti, že objekt je členem jedné ze dvou tříd.
- 2) *Silné nenormalita, přítomnost binárních proměnných*: **logistická diskriminace** k výpočtu pravděpodobnosti, že objekt je členem jedné ze dvou tříd.

$$P(G = 1|\mathbf{x}) = \frac{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p)}{1 + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p)}$$
$$P(G = 0|\mathbf{x}) = \frac{1}{1 + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p)}$$

Po vyčíslení jejich odhadů $b_0, b_1, b_2, \dots, b_p$ se uplatní klasifikační pravidlo zařazení objektu do třídy G_1 platí-li $b_0 + b_1 x_1 + b_2 x_2 + \dots + b_p x_p > 0$ což odpovídá pravděpodobnosti $P(G = 1|\mathbf{x}) > P(G = 0|\mathbf{x})$.



Příklad 4.23 Logistická diskriminace k určení rakoviny prostaty

Rakovina lymfatických uzlin. Rozhodující metodou vyšetření je laparotomie, vyjádřená proměnnou B_{412x6} , kde 0 značí nepřítomnost rakoviny a 1 přítomnost rakoviny. Brownovo vyšetření pěti diskriminantů u 53 pacientů mělo nahradit toto obtížnější vyšetření:

Data: B_{412x1} značí věk pacienta,

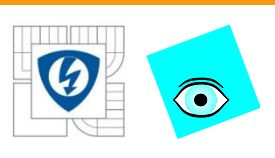
B_{412x2} je hladina sérové kyselý fosfatázy v Kingových-Armstrongových jedn.,

B_{412x3} rentgenového vyšetření: 0 značí nepřítomnost a 1 přítomnost tumoru,

B_{412x4} velikost tumoru rektálně: 0 = malý, 1 = velký,

B_{412x5} závěr patologického bodování z biopsie: 0 méně vážný, 1 velmi vážný.

B_{412x1}	B_{412x2}	B_{412x3}	B_{412x4}	B_{412x5}	B_{412x6}
66	0.48	0	0	0	0
..	..	..	..	..	..
..	..	..	..	..	..
64	0.89	1	1	0	1
68	1.26	1	1	1	1



Řešení:

1) Analyzovaný výběr: odhady parametrů (směrodat. odchylky v závorce) jsou b_0 1.52 (3.56), b_1 0.10 (0.06), b_2 2.64 (1.33), b_3 1.68 (0.80), b_4 2.04 (0.83), b_5 0.35 (0.80).

Dvě třídy: 1. třída (rakovinu má), 2. třída (rakovinu nemá).

Pacient rakovinu lymfatických uzlin má (je zařazen do 1. třídy), je-li splněna nerovnost

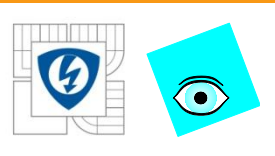
$$1.52 - 0.10 x_1 + 2.64 x_2 + 1.68 x_3 + 2.04 x_4 + 0.35 x_5 > 0.$$

vykazuje-li pacient **kladnou hodnotu** v této nerovnosti, má rakovinu uzlin a je zařazen do 1. třídy, zatímco vykazuje-li **zápornou hodnotu** v této nerovnosti, nemá rakovinu a je zařazen do 2. třídy.

2) Klasifikovaný výběr: dosadíme hodnoty prvního pacienta do této nerovnosti a vyčíslíme

$$1.52 - 0.10 \times 66 + 2.64 \times 0.48 + 1.68 \times 0 + 2.04 \times 0 + 0.35 \times 0 = -3.81.$$

Závěr: Protože **výsledek v nerovnosti -3.81 je záporný**, je pacient zařazen do 2. třídy a **nemá rakovinu** lymfatických uzlin.



Průběh diagnostikování DA

Úlohy diskriminační analýzy se řeší v šesti krocích:

1. krok: Cíle diskriminační analýzy

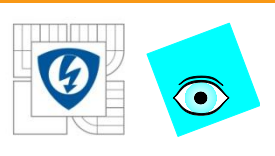
Zajímáme se o správné zařazování objektů do tříd:

- a) Vyšetřujeme, zda existují statisticky významné rozdíly mezi profily průměrného skóre znaků pro více předem definovaných tříd?
- b) Hledáme, který ze znaků se projevuje nejvíce v rozdílových profilech průměrného skóre dvou či více tříd?
- c) Na základě skóre znaků stanovíme postup k zařazování objektů do tříd.
- d) Určíme počet a složení nalezených tříd znaků.

DA pracuje nejlépe, když jde o jedinou závisle proměnnou a několik metrických nezávislých proměnných či znaků.

DA poskytuje ve znacích objektivní vyčíslení rozdílů mezi třídami.

DA je dosti podobná vícerozměrné analýze rozptylu, umožňuje pohled do jednotlivých znaků a definuje rozměry diskriminace mezi třídami.



Průběh diagnostikování DA

DA poskytuje podklad k zařazování objektů do tříd.

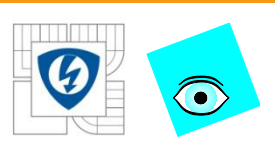
DA určuje diskriminační funkce k zařazování nových objektů do předem nadefinovaných tříd.

2. krok: Formulace úlohy, volba znaků

1) Volby závisle proměnné a několika nezávislých proměnných, znaků:

Závisle proměnná: je většinou nemetrická, kategorická proměnná. Každý objekt může být umístěn pouze do jediné třídy. Obvykle se závisle proměnná týká dvou tříd - dichotomie, např. *dobrý* versus *špatný*, nebo multichotomie, např. *lékař*, *řidič*, *učitel*

Znaky: jsou znaky, metrické proměnné.



Průběh diagnostikování DA

2) Určení velikosti výběru k odhadu diskriminační funkce:

DA je značně citlivá *na poměr velikosti výběru ku počtu znaků*.

Empirická úmluva: poměr 20 objektů najeden znak.

Minimální velikost výběru: 5 objektů najeden znak.

Rozdílná velikost tříd: způsobí problémy při odhadování diskriminační funkce a zařazování objektů, větší třídy mají větší šanci při zařazování.

3) Dělení výběru ke klasifikačním účelům.

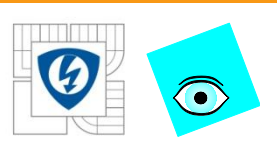
Výběr je často rozdělen (v poměru 50/50 nebo 60/40 nebo 75/25) na:

Analyzovaný výběr k výstavbě diskriminační funkce.

Klasifikovaný výběr k testování diskriminační funkce.

Kategorizované třídy jsou nestejně, pak i velikosti vybraných tříd pro klasifikovaný výběr měly být úměrné jejich podílu v celkovém výběru:

např. výběr 50 mužů a 50 žen, pak bude klasifikovaný výběr 25 mužů a 25 žen, výběr 70 žen a 30 mužů, pak bude klasifikovaný výběr 35 žen a 15 mužů.



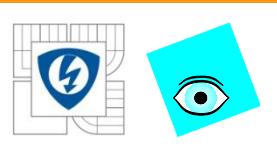
Průběh diagnostikování DA

4) Chybí-li u některého objektu nějaký znak:

DA se omezí na malý počet úplně definovaných objektů, což může způsobit v DA určité problémy.

3. krok: Předpoklady diskriminační analýzy

- Předpokladem DA je *vícerozměrná normalita znaků*, jinak se užije alternativně *logistická regrese*.
- *Nestejné kovarianční matice* mohou negativně ovlivnit klasifikační proces LDA.
- Příliš malý výběr a nestejné kovarianční matice nepříznivě ovlivní statistickou významnost odhadů.
- *Multikolinearita* značí velkou korelaci dvou znaků. To je nevhodné při užití *krokové metody*.
- Klasická DA vychází z předpokladu, že *všechny vztahy jsou lineárního charakteru*.
- *Odlehlé hodnoty* mají rovněž špatný dopad na klasifikační schopnost DA.



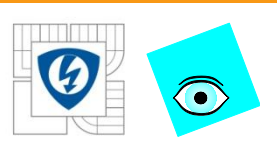
4.krok: Nalezené řešení a dosažená těsnost proložení

Dělení souboru: objekty souboru náhodně rozdělíme do dvou výběrů:

1. *Analyzovaný výběr se použije k vyčíslení diskriminační funkce na základě známých a dostatečně popsanych objektů.*
2. *Klasifikovaný výběr se použije ke klasifikační matici a zařazování nových objektů.*

*Predikční schopnost diskriminační funkce vyjadřuje počet správně klasifikovaných objektů. Diagnostiky každého objektu mohou ukázat na správnost zařazení. **Odvození diskriminační funkce:** přímá metoda a kroková metoda.*

Přímá metoda: diskriminační funkce je vyčíslena na základě všech znaků navzdory jejich rozličné diskriminační síle.



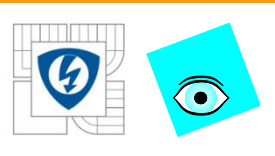
Kritéria posouzení statistické významnosti diskriminace

Wilkovo kritérium λ je poměr sumy čtverců mezi třídami a celkového součtu čtverců:

- Velké $\lambda = 1$ ukazuje, že průměry tříd jsou shodné.
- Malé $\lambda = 0$ ukazuje, že průměry tříd jsou velmi rozdílné a většinu variability lze přičíst rozdílům mezi průměry tříd.

Raovo největší vlastní číslo vyhodnocuje pouze první diskriminační funkci, Mahalanobisova D^2 a Raovo V jsou nejvhodnějšími mírami k posouzení obecné vzdálenosti.

Mahalanobisova vzdálenost D^2 je však nespolehlivá, když počet znaků roste, protože neprovádí žádnou redukci dimenze.

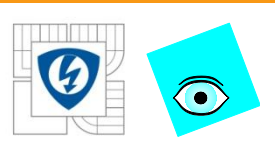


Kritéria těsnosti proložení

a) **Diskriminační Z-skóre pro každý objekt:** vyčíslí se Z-skóre nebo-li Fisherova lineární diskriminační funkce pro k -tý objekt $L(x_k)$.

Vlastnosti Z-skóre:

- 1) Z-skóre je metrická proměnná a nabízí přímé průměry k porovnání objektů.
- 2) Objekty s podobnými Z-skóre jsou podobné ve znacích, které tvoří tuto funkci, na rozdíl od objektů s nepodobnými skóre.
- 3) Standardizovaná verze diskriminační funkce je užitečnější k interpretaci.
- 4) Vysvětlení koeficientů je podobné jako ve vícenásobné lineární regresi.
- 5) Znaky s velkými koeficienty přispívají více do celkové diskriminační funkce.



Kritéria těsnosti proložení

b) Vyhodnocení třídních rozdílů v diskriminačních Z-skóre: určení rozdílu mezi objekty každé třídy v termínech diskriminačního Z-skóre.

Souhrnou mírou třídních rozdílů: je porovnání těžiště třídy čili průměru diskriminačního Z-skóre pro všechny objekty ve třídě.

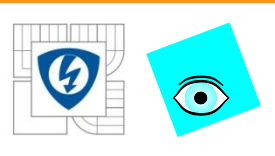
Mírou úspěchu diskriminační analýzy: je schopnost definovat diskriminační funkce, které vedou k významně rozličným třídním těžištím v jednotkách Mahalanobisovy vzdálenosti D^2 .

Rozdíly mezi třídami na každé diskriminační funkci však vizuální inspekce neodhalí.

c) Přiřazení predikční schopnosti tříd:

Závisle proměnná je nemetrická a proto není možné k vyhodnocení predikční schopnosti použít míru R^2 .

U diskriminační analýzy je procento správně klasifikovaných objektů označeno pojmem *hit poměr*, který je analogií vůči regresnímu R^2 .



Kritéria těsnosti proložení

d) Určení prahového bodu:

Při konstrukci klasifikačních matic je třeba určit optimální prahový bod C .

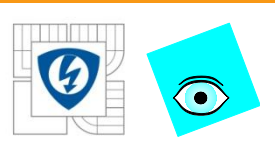
1) **Pro třídy stejné velikosti:** optimální prahový bod C diskriminace bude

$C = \frac{\bar{Z}_1 + \bar{Z}_2}{2}$, kde $\bar{Z}_1$ je těžiště třídy 1 a $\bar{Z}_2$ je těžiště třídy 2.

2) **Pro nestejně třídni velikosti:** je *vážený průměr třídních těžišť* optimálním prahovým bodem $C = \frac{n_1 \bar{Z}_1 + n_2 \bar{Z}_2}{n_1 + n_2}$, kde n_1 je počet prvků ve třídě 1, n_2 je počet prvků ve třídě 2.

Diskriminační skóre Z_n pro klasifikované objekty jsou porovnávána s prahovým bodem C dle schématu

prvek bude zařazen do třídy 1, když je $Z_n < C$, nebo bude zařazen do třídy 2, když je $Z_n > C$.

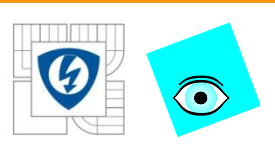


Kritéria těsnosti proložení

Test: *Studentův t-test klasifikační správnosti* podle testačního kritéria

$$t = \frac{p - 0.5}{\sqrt{\left(\frac{0.5(1 - 0.5)}{n}\right)}}$$

kde p je podíl správně zařazených prvků a n je velikost výběru.



Kritéria těsnosti proložení

e) Měření predikční schopnosti diskriminační funkce:

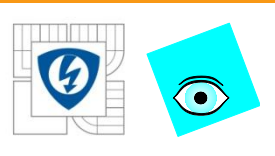
Hit poměr odhalí, jak dobře diskriminační funkce zařazuje objekty. Nejprve se určuje podíl správně klasifikovaných pravděpodobností:

(a) Stejně třídí velikosti: klasifikační apriorní pravděpodobnost π_i , je rovna reciproké hodnotě počtu tříd k , $C_{pr} = 1/k$.

Například: pro dvojtřídní funkci $k = 2$ je pravděpodobnost $C_{pr} = 0.50$, pro trojtřídní $k = 3$ pak $C_{pr} = 0.33$, atd.

(b) Nestejně třídí velikosti: klasifikační apriorní pravděpodobnost π_i , je založena na velikosti největší třídy a nazývá se kritérium maximální pravděpodobnosti. Určuje se vyčíslením procenta celkového výběru, tvořeného největším ze dvou (či více) tříd.

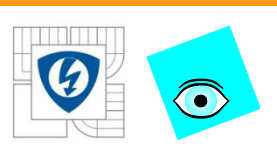
Například: velikosti dvou tříd jsou 65 a 35, a kritérium maximální pravděpodobnost bude $C_{pr} = 65\%$ značící počet správných klasifikací.



Kritéria těsnosti proložení

(c) Kritérium poměrné pravděpodobnosti se užije pro nestejně veliké třídy a klasifikování prvků do dvou či více tříd dle vzorce $\pi_1 = p^2 + (1 - p)^2$, kde p je podíl prvků v 1. třídě a $(1 - p)$ je podíl prvků ve 2. třídě.

Například: když máme třídy 75% a 25%, bude $\pi_1 = 0.75^2 + (1 - 0.75)^2 = 0.625$.



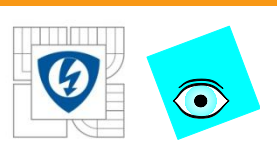
Statistické míry klasifikace spolehlivosti diskriminace:

1) **Pressovo q -kritérium** vyjadřuje míru porovnání počtu správných klasifikací vůči celkové velikosti výběru a počtu tříd. Vyčíslí se vztahem

$$q = \frac{(n - n_s k)^2}{n(k - 1)}$$

kde n je velikost výběru, n_s udává počet objektů správně klasifikovaných a k značí počet tříd,

Testování: vyšetřuje diskriminační sílu klasifikační matice v porovnání s modelem pravděpodobnosti: vypočtené q se porovnává s $\chi^2_{1-\alpha}(1)$ při dané α , a to když q překročí $\chi^2_{1-\alpha}(1)$ klasifikační matice se jeví statisticky lepší než klasifikační pravděpodobnost.



Statistické míry klasifikace spolehlivosti diskriminace:

Vlastnosti testu:

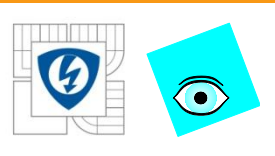
Test je citlivý na velikost výběru, protože velké výběry jsou náchylné snadněji vykazovat statistickou významnost než výběry malé.

I když se totiž velikosti výběru zvyšují, menší klasifikační poměr bude stále významný.

Například, pro $n = 50$, $n_s = 42$ a $k=2$ bude $q = (50 - 42 \times 2)^2 / (50(2 - 1)) = 23.12$ a kritická hodnota pro $\alpha = 0.01$ je $\chi^2_{1-\alpha}(1) = 6.63$.

Závěr: Predikce jsou statisticky významnější než apriorní pravděpodobnost π_i , která uvádí správnou klasifikaci pro 50 %.

Například zvětšíme velikost výběru na 100 a klasifikační poměr zůstane 84 %, Pressovo q se zvýší na 46.24.



Diagnostiky jednotlivých objektů

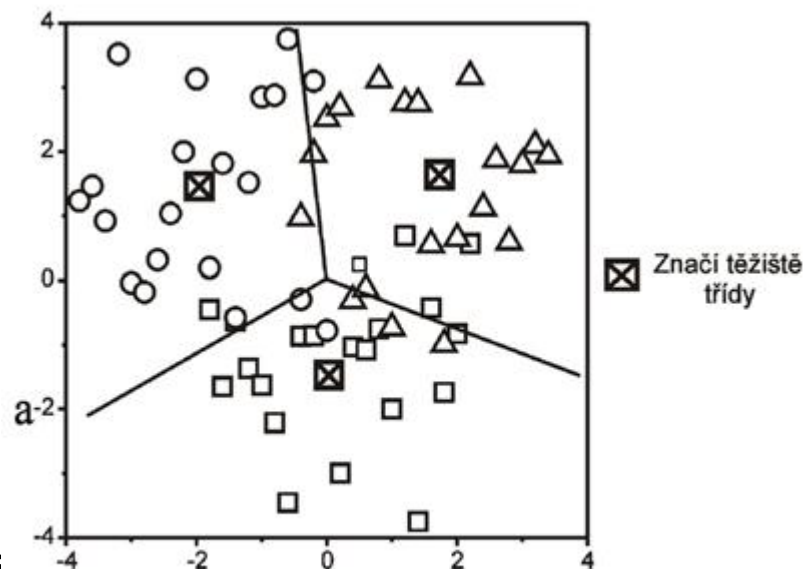
2) Diagnostiky jednotlivých objektů: stanovení těsnosti proložení modelem spočívá na vyšetření predikčních výsledků celé řady objektů.

Hledáme, který objekt byl chybně klasifikován a který nereprezentuje zbývající objekty třídy.

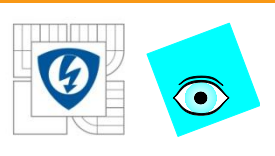
I když znázorníme, které objekty jsou správně a které nesprávně klasifikovány, stále potřebujeme míru podobnosti objektů vůči zbytku třídy.

Podobnost objektu vůči ostatním objektům

třídy: Na základě vyšetření Mahalanobisovy vzdálenosti D^2 mezi posuzovaným objektem a těžištěm třídy. Nachází-li se objekty poblíž těži charakteristiku třídy, ale také na rozptýlení prvků ve třídě.



Obrázek teritoriální mapy.



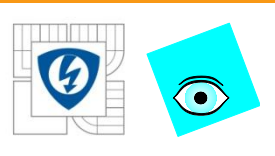
5. krok: Interpretace výsledků

Tři metody určují relativní důležitost znaku:

i) Standardizované diskriminační koeficienty: interpretace diskriminačních funkcí vyšetřuje znaménko a velikost *standardizovaných diskriminačních koeficientů* $\mathbf{a}^T = [a_0, a_1, \dots, a_p]$, které představují relativní příspěvek svého znaku do Fisherovy lineární diskriminační funkce:

Diskuze:

- 1) Znaky s *relativně velkými koeficienty* přispívají více do diskriminační síly diskriminační funkce než znaky s menšími koeficienty.
- 2) Znaménko ukazuje, že znak dává buď kladný, nebo záporný příspěvek.
- 3) Malý koeficient indikuje buď, že odpovídající znak je nevýznamný k určování vztahu nebo je neúplným vztahem, protože je zde vysoký stupeň multikolinearity.
- 4) Problémem je také značná nestabilita diskriminačních koeficientů.



Diskriminační zátěže

2) Diskriminační zátěže (označované také *strukturní korelace*):

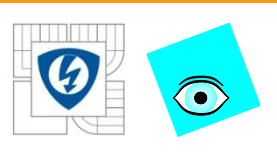
Jsou mírou korelace mezi každým znakem x_i a diskriminační funkcí Z_k .

Vystihují rozptyl, který sdílejí znaky x_i s diskriminační funkcí Z_k

Jsou to faktorové zátěže při posouzení relativního příspěvku každého znaku do diskriminační funkce.

Vyčíslí se:

- 1) Pro každý objekt hodnota diskriminační funkce Z_k a korelační koeficient mezi Z_k a znakem x_i .
- 2) Samostatnou *korelační matici* lze vyčíslit pro každou třídu.
- 3) Výsledky lze dále kombinovat k vyčíslení *společné korelační matice* mezi všemi třídami.
- 4) Můžeme také uvažovat všechny objekty dohromady a vyčíslit *celkovou korelační matici*. Často se stává, že korelace mezi znaky ovlivní velikosti i znaménka svých koeficientů.



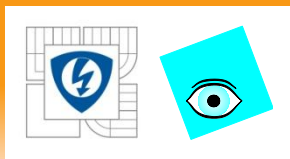
Parciální F-hodnoty

3) Parciální F-hodnoty:

U krokové metody je k dispozici způsob vysvětlování relativní diskriminační síly znaků prostřednictvím parciálních F -hodnot.

Vyšetřuje se absolutní hodnota významné F -hodnoty: velká F -hodnota indikuje velkou diskriminační sílu.

Vysvětlení dvou a i více diskriminačních funkcí: jak přispívá každý znak x_i do funkce Z_k .



Index vlivu je relativní mírou diskriminační síly každého znaku x_i .

- obsahuje oba příspěvky proměnné k diskriminační funkci Z_k (její diskriminační zátěže) a relativní příspěvek diskriminační funkce do celkového řešení.
- interpretace je však omezena na vykreslení relativní polohy (jako je třeba pořadí hodnoty) každého znaku. Jeho absolutní hodnota nemá žádný skutečný smysl.

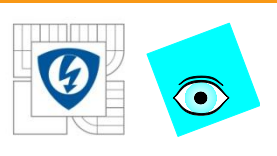
Vyčíslí se:

1) *Výpočet indexu vlivu pro každou významnou diskriminační funkci*: vypočte se míra relativního vlastního čísla pro každou diskriminační funkci dle

$$\lambda_{j,R} = \frac{\lambda_j}{\sum_{j=1}^m \lambda_j} . \text{ Index vlivu každého znaku v diskriminační funkci je vyčíslen dle}$$

Index vlivu i -tého znaku v j -té funkci = (Deskriptivní zátěž_{ij})² $\lambda_{j,R}$

2) *Výpočet indexu vlivu přes všechny významné funkce*: je vyčíslen složený index vlivu jako suma hodnot indexů vlivu všech významných diskriminačních funkcí. Index vlivu *představuje* celkový diskriminační efekt znaku x_i přes všechny významné diskriminační funkce.



6. krok: Ověření výsledků

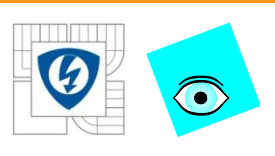
Konečné stadium DA se týká potvrzení diskriminačních výsledků. DA má tendenci přeceňovat hit poměr. Metoda dělení do skupin (cross-validation) je velmi užitečná.

Rozdělení výběru:

1. Soubor je náhodně rozdělen *na analyzovaný výběr a na klasifikovaný výběr*.
2. Místo dělení na *analyzovaný a klasifikovaný* můžeme zcela náhodně rozdělit soubor několikrát.

Testujeme potvrzení diskriminační funkce pomocí *klasifikační matice a hit poměru*.

Když *najdeme znaky*, které mají největší vliv na diskriminaci mezi třídami, je dalším krokem profilování charakteristiky tříd, založené na třídních průměrech.



6. krok: Ověření výsledků

Ve Fisherově úloze o rozměrech okvětních lístků u 150 kosatců se analyzují květy tří základních tříd: (1) *Iris setosa*, (2) *Iris versicolor*, (3) *Iris virginica*. Květy kosatců jsou popsány čtyřmi znaky: délkou kališních lístků v mm anglicky *lsepal* a jejich šířkou *wsepal*, dále délkou korunních plátků v mm *lpetal* a jejich šířkou *wpetal*.

Každý objekt je popsán $p = 4$ znaky, a to *SepalLength*, *SepalWidth*, *PetalLength*, *PetalWidth*.

Data: jsou v pořadí proměnných *E418a*, *E418b*, *E418c*, *E418d*:



Data

5.1 3.5 1.4 0.2
 4.9 3.0 1.4 0.2
 4.7 3.2 1.3 0.2
 4.6 3.1 1.5 0.2
 5.0 3.6 1.4 0.2
 5.4 3.9 1.7 0.4
 4.6 3.4 1.4 0.3
 5.0 3.4 1.5 0.2
 4.4 2.9 1.4 0.2
 4.9 3.1 1.5 0.1
 5.4 3.7 1.5 0.2
 4.8 3.4 1.6 0.2
 4.8 3.0 1.4 0.1
 4.3 3.0 1.1 0.1
 5.8 4.0 1.2 0.2
 5.7 4.4 1.5 0.4

5.4 3.9 1.3 0.4
 5.1 3.5 1.4 0.3
 5.7 3.8 1.7 0.3
 5.1 3.8 1.5 0.3
 5.4 3.4 1.7 0.2
 5.1 3.7 1.5 0.4
 4.6 3.6 1.0 0.2
 5.1 3.3 1.7 0.5
 4.8 3.4 1.9 0.2
 5.0 3.0 1.6 0.2
 5.0 3.4 1.6 0.4
 5.2 3.5 1.5 0.2
 5.2 3.4 1.4 0.2
 4.7 3.2 1.6 0.2
 4.8 3.1 1.6 0.2
 5.4 3.4 1.5 0.4

5.2 4.1 1.5 0.1
 5.5 4.2 1.4 0.2
 4.9 3.1 1.5 0.2
 5.0 3.2 1.2 0.2
 5.5 3.5 1.3 0.2
 4.9 3.6 1.4 0.1
 4.4 3.0 1.3 0.2
 5.1 3.4 1.5 0.2
 5.0 3.5 1.3 0.3
 4.5 2.3 1.3 0.3
 4.4 3.2 1.3 0.2
 5.0 3.5 1.6 0.6
 5.1 3.8 1.9 0.4
 4.8 3.0 1.4 0.3
 5.1 3.8 1.6 0.2
 4.6 3.2 1.4 0.2

5.3 3.7 1.5 0.2
 5.0 3.3 1.4 0.2
 7.0 3.2 4.7 1.4
 6.4 3.2 4.5 1.5
 6.9 3.1 4.9 1.5
 5.5 2.3 4.0 1.3
 6.5 2.8 4.6 1.5
 5.7 2.8 4.5 1.3
 6.3 3.3 4.7 1.6
 4.9 2.4 3.3 1.0
 6.6 2.9 4.6 1.3
 5.2 2.7 3.9 1.4
 5.0 2.0 3.5 1.0
 5.9 3.0 4.2 1.5
 6.0 2.2 4.0 1.0
 6.1 2.9 4.7 1.4



Data

5.6	2.9	3.6	1.3	6.7	3.1	4.7	1.5	6.7	2.5	5.8	1.8	7.4	2.8	6.1	1.9
6.7	3.1	4.4	1.4	6.3	2.3	4.4	1.3	7.2	3.6	6.1	2.5	7.9	3.8	6.4	2.0
5.6	3.0	4.5	1.5	5.6	3.0	4.1	1.3	6.5	3.2	5.1	2.0	6.4	2.8	5.6	2.2
5.8	2.7	4.1	1.0	5.5	2.5	4.0	1.3	6.4	2.7	5.3	1.9	6.3	2.8	5.1	1.5
6.2	2.2	4.5	1.5	5.5	2.6	4.4	1.2	6.8	3.0	5.5	2.1	6.1	2.6	5.6	1.4
5.6	2.5	3.9	1.1	6.1	3.0	4.6	1.4	5.7	2.5	5.0	2.0	7.7	3.0	6.1	2.3
5.9	3.2	4.8	1.8	5.8	2.6	4.0	1.2	5.8	2.8	5.1	2.4	6.3	3.4	5.6	2.4
6.1	2.8	4.0	1.3	5.0	2.3	3.3	1.0	6.4	3.2	5.3	2.3	6.4	3.1	5.5	1.8
6.3	2.5	4.9	1.5	5.6	2.7	4.2	1.3	6.5	3.0	5.5	1.8	6.0	3.0	4.8	1.8
6.1	2.8	4.7	1.2	5.7	3.0	4.2	1.2	7.7	3.8	6.7	2.2	6.9	3.1	5.4	2.1
6.4	2.9	4.3	1.3	5.7	2.9	4.2	1.3	7.7	2.6	6.9	2.3	6.7	3.1	5.6	2.4
6.6	3.0	4.4	1.4	6.2	2.9	4.3	1.3	6.0	2.2	5.0	1.5	6.9	3.1	5.1	2.3
6.8	2.8	4.8	1.4	5.1	2.5	3.0	1.1	6.9	3.2	5.7	2.3	5.8	2.7	5.1	1.9
6.7	3.0	5.0	1.7	5.7	2.8	4.1	1.3	5.6	2.8	4.9	2.0	6.8	3.2	5.9	2.3
6.0	2.9	4.5	1.5	6.3	3.3	6.0	2.5	7.7	2.8	6.7	2.0	6.7	3.3	5.7	2.5
5.7	2.6	3.5	1.0	5.8	2.7	5.1	1.9	6.3	2.7	4.9	1.8	6.7	3.0	5.2	2.3
5.5	2.4	3.8	1.1	7.1	3.0	5.9	2.1	6.7	3.3	5.7	2.1	6.3	2.5	5.0	1.9
5.5	2.4	3.7	1.0	6.3	2.9	5.6	1.8	7.2	3.2	6.0	1.8	6.5	3.0	5.2	2.0
5.8	2.7	3.9	1.2	6.5	3.0	5.8	2.2	6.2	2.8	4.8	1.8	6.2	3.4	5.4	2.3
6.0	2.7	5.1	1.6	7.6	3.0	6.6	2.1	6.1	3.0	4.9	1.8	5.9	3.0	5.1	1.8
5.4	3.0	4.5	1.5	4.9	2.5	4.5	1.7	6.4	2.8	5.6	2.1				
6.0	3.4	4.5	1.6	7.3	2.9	6.3	1.8	7.2	3.0	5.8	1.6				



Kosatce (Iris)



Iris Setosa

Wild Iris - Iris setosa



Iris Versicolor



Iris Virginica

Kosatce (Iris)



Iris Setosa



Iris Versicolor



Iris Versicolor



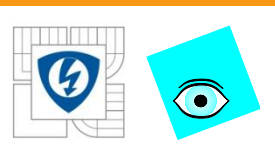
Iris Virginica



Iris Virginica



Iris Virginica



Řešení

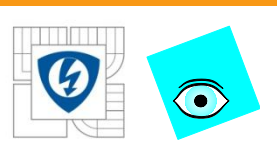
Řešení: zdrojová matice 150 řádků a 4 sloupce neobsahuje žádné chybějící prvky.

1. Znaky

1,1 Aritmetický průměr [mm] objektů ve třídě Setosa G_1 , Versicolor G_2 , Virginica G_3 .

Znak	Setosa G_1	Versicolor G_2	Virginica G_3	Celkově $G_1+G_2+G_3$
SepalLength	50.06	59.36	65.88	58.43333
SepalWidth	34.28	27.7	29.74	30.57333
PetalLength	14.62	42.6	55.52	37.58
PetalWidth	2.46	13.26	20.26	11.99333
Počet	50	50	50	150

Tabulka obsahuje průměry každého znaku pro každou třídu kosatců G_1 , G_2 a G_3 .

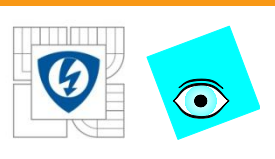


Směrodatné odchylky

1.2 Směrodatné odchylky [mm] objektů ve třídě Setosa G_1 ; Versicolor G_2 ,
Virginica G_3 .

Znak	Setosa G_1	Versicolor G_2	Virginica G_3	Celkově $G_1+G_2+G_3$
SepalLength	3.524897	5.161712	6.358796	8.280662
SepalWidth	3.790644	3.137983	3.224966	4.358663
PetalLength	1.73664	4.69911	5.518947	17.65298
PetalWidth	1.053856	1.977527	2.7465	7.622377
Počet	50	50	50	150

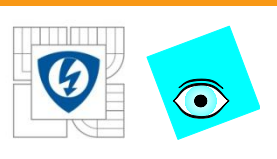
Lineární diskriminační analýza předpokládá, že kovarianční matice jsou stejné pro každou třídu. Tabulka posoudí předpoklad, zda jsou směrodatné odchylky v třídách zhruba stejné.



1.3 Celkové korelace a kovariance znaků.

Korelace jsou v dolní levé části, kovariance jsou v pravé horní části matice. Rozptyly znaků jsou na diagonále matice.

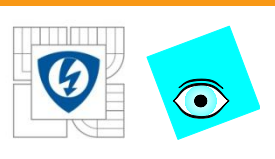
Znak	SepalLength	SepalWidth	PetalLength	PetalWidth
SepalLength	68.56935	-4.243401	127.4315	51.62707
SepalWidth	-0.117570	18.99794	-32.96564	-12.16394
PetalLength	0.871754	-0.428440	311.6278	129.5609
PetalWidth	0.817941	-0.366126	0.962865	58.10063



1.4 Mezitřídní korelace a kovariance znaků.

Tabulka obsahuje korelace a kovariance, vytvořené za použití průměrů tříd objektů místo jednotlivých objektů. Korelace jsou v dolní levé části, mezitřídní kovariance jsou na diagonále matice a v horní pravé části matice.

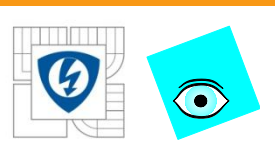
Znak	SepalLength	SepalWidth	PetalLength	PetalWidth
SepalLength	3160.607	-997.6334	8262.42	3563.967
SepalWidth	-0.745075	567.2466	-2861.98	-1146.633
PetalLength	0.994135	-0.812838	21855.14	9338.7
PetalWidth	0.999768	-0.759258	0.996232	4020.667



1.5 Vnitrotřídní korelace a kovariance znaků.

V datech byly třídní průměry objektů odečteny. Korelace jsou v dolní levé části, vnitrotřídní kovariance jsou na diagonále a v pravé horní části matice.

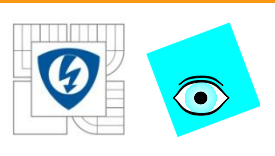
Znak	SepalLength	SepalWidth	PetalLengthPetalWidth
SepalLength	26.50082	9.272109	16.751433.840136
SepalWidth	0.530236	11.53878	5.5243543.27102
PetalLength	0.756164	0.377916	18.518784.266531
PetalWidth	0.364506	0.470535	0.4844594.188163



1.6 Vyšetření vlivu jednotlivých znaků.

Vliv jednotlivých znaků na diskriminační analýzu: *spočtená hladina významnosti při odstranění dotyčného znaku* je vypočtená u F -testu při odstranění dotyčného znaku. Znak je důležitý, je-li tato hodnota menší než uživatelem zadaná hladina významnosti $\alpha = 0.05$. Zde jsou všechny čtyři znaky statisticky významné. *Wilk λ pro dotyčný samotný znak* značí hodnotu za použití jediného znaku, který je důležitý, je-li tato hodnota menší než uživatelem zadaná hladina významnosti $\alpha = 0.05$.

Při odstranění dotyčného znaku			Pro dotyčný samotný znak		R^2		
Diskrimin.	Wilk. λ	F -test	Spočt. α	Wilk. λ	F -test	Spočt. α	Ostatní X
SepalLength	0.938463	4.720	0.010329	0.381294	119.26	0.000000	0.858612
SepalWidth	0.766480	21.94	0.000000	0.599217	49.16	0.000000	0.524007
PetalLength	0.669206	35.59	0.000000	0.058628	1180.2	0.000000	0.968012
PetalWidth	0.743001	24.90	0.000000	0.071117	960.01	0.000000	0.937850



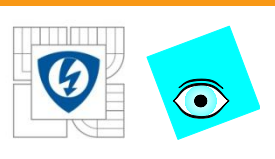
1.7 Automatická volba účinných znaků

Úkolem v diskriminační analýze je výběr znaků. Z velké skupiny možných znaků je třeba vybrat menší výběr těch nejlepších, maximálně 8 proměnných. *Automaticky výběr znaků* se plní krokově - nejprve se nalezne nejlepší znak a potom druhý nejlepší. Postupný krokový proces přidávání nejlepšího zbývajcího znaku s ověřením, zda by jeden aktivní znak mohl být odebrán, pokračuje, dokud je ještě nějaký znak k dispozici. % změny v λ značí procento snížení v hodnotě λ , které je výsledkem tohoto kroku. Wilkovo kritérium λ je analogické $(1 - R^2)$ ve vícenásobné regresi. Vhodnější model sníží Wilkovo kritérium λ .



Data

Krok	Činnost v kroku	Diskrimin.	% změny lambda	v F-test	Spočtená hladina α	Wilkovo lambda
0	Žádné					1.000000
1	Zavedeno	PetalLength	94.14	1180.16	0.000000	0.058628
2	Zavedeno	SepalWidth	37.09	43.04	0.000000	0.036884
3	Zavedeno	PetalWidth	32.29	34.57	0.000000	0.024976
4	Zavedeno	SepalLength	6.15	4.72	0.010329	0.023439
..	.				...	
Detail ve 4. kroku automatického volby znaku:						
Status		Diskrimin.	% změny lambda	v F-test	Spočtená hladina α	R^2 na ostatních X
Uvnitř		SepalLength	6.15	4.72	0.010329	0.858612
Uvnitř		SepalWidth	23.35	21.94	0.000000	0.524007
Uvnitř		PetalLength	33.08	35.59	0.000000	0.968012
Uvnitř		PetalWidth	25.70	24.90	0.000000	0.937850
Celkové Wilkovo $\lambda = 0.023439$						

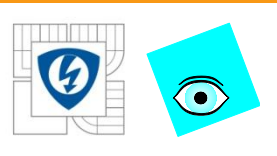


2. Diskriminační funkce.

2.1 Odhady diskriminačních koeficientů $a_0, a_1, \dots, a_p$ diskriminační funkce Z_k pro třídy G_1 Setosa Versicolor G_2 , Virginica G_3 .

Znak	Setosa G_1	Versicolor G_2	Virginica G_3
Absolutní člen	-85.20985	-71.754	-103.2697
SepalLength	2.354417	1.569821	1.244585
SepalWidth	2.358787	0.707251	0.3685279
PetalLength	-1.643064	0.5211451	1.276654
PetalWidth	-1.739841	0.6434229	2.107911

Tabulka obsahuje celkem tři diskriminační funkce, jednu pro každou třídu ve sloupci. Tato diskriminační funkce je určena k zařazování nových, dosud nezařazených objektů do tříd čili ke klasifikaci, a proto se v této formě nazývá *klasifikační funkce*. Klasifikační funkce *pro první třídu* G_1 : - 5.20985 + 2.35442*SepalLength + 2.35879*SepalWidth - 1.64306*PetalLength - 1.73984*PetalWidth. Tyto tři funkce mohou být použity k zařazení nových kosatců do navržených tříd.



Odhady regresních parametrů

2.2 Odhady regresních parametrů $a_0, a_1, \dots, a_p$ lineárního regresního modelu pro každou třídu Setosa G_1 Versicolor G_2 Virginica G_3

Znak	Setosa G_1	Versicolor G_2	Virginica G_3
Absolutní člen	0.1182229	1.577059	-0.6952819
SepalLength	6.602977E-03	-2.015369E-03	-4.587608E-03
SepalWidth	2.428479E-02	-4.456162E-02	2.027684E-02
PetalLength	-2.246571E-02	2.206692E-02	3.987911E-04
PetalWidth	-5.747273E-03	-4.943066E-02	5.517793E-02

Tabulka obsahuje regresní parametry $a_0, a_1, \dots, a_p$ lineárního regresního modelu. Zařazení objektů do třídy se provede tak, že se u každého objektu vybere vždy třída s nejvyšší hodnotou diskriminačního skóre.



Kosatce (Iris)



Iris Setosa

Wild Iris - Iris setosa

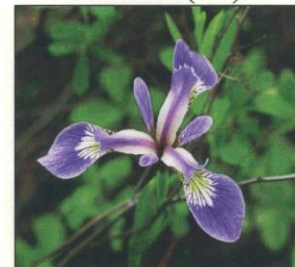


Iris Versicolor



Iris Virginica

Kosatce (Iris)



Iris Setosa



Iris Versicolor



Iris Versicolor



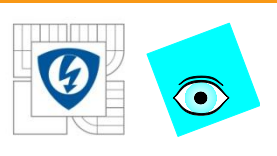
Iris Virginica



Iris Virginica



Iris Virginica



3. Klasifikace objektů.

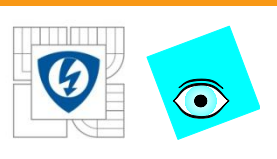
3.1 Zařazení (klasifikace) objektů klasifikační funkcí do tříd tvoří klasifikační matici kosatců ve třídách Setosa G_1 Versicolor G_2 , Virginica G_3

Nalezená predikovaná třída

Daná třída	Setosa G_x	Versicolor G_2	Virginica G_3	Celkově $G_1+G_2+G_3$
Setosa	50	0	0	50
Versicolor	0	34	16	50
Virginica	0	7	43	50
Celkově	50	41	59	150

Redukce v klasifikační správnosti v důsledku proměnných X — 77.0 %.

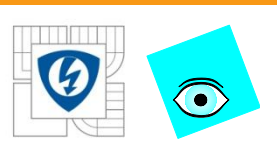
Klasifikační matice ukazuje Jak diskriminační funkce zařazují objekty do tří tříd. Bylo-li dosaženo perfektního zařazení, obdržíme v matici mimo diagonálu nuly. *Redukce v klasifikační správnosti* obsahuje procento snížení v zařazovací správnosti, dosažené diskriminačními funkcemi vůči očekávanému zařazení dle původní závisle proměnné.



3.2 Přehled chybně zařazených objektů v řádcích do tříd Setosa Gu Versicolor G2, Virginica (73 dle klasifikačních funkcí proti původní závisle proměnné.

V řádku je hodnota pravděpodobnosti (v procentech), že objekt se nachází v dané třídě kosatců. Hodnota blízko 100 % značí, že objekt patří do dotyčné třídy.

Objekt	Procento zařazení do jednotlivé třídy				
	Daná třída	Nalezená třída	Setosa G_1	Versicolor G_2	Virginica G_3
5	Virginica	Versicolo	-1.8	58.6	43.1
9	Versicolo	Virginica	10.3	20.2	69.5
22	Versicolo	Virginica	18.8	22.6	58.6
28	Versicolo	Virginica	22.1	35.5	42.4
29	Versicolo	Virginica	22.1	27.4	50.6
38	Versicolo	Virginica	10.6	38.3	51.1
45	Virginica	Versicolo	-31.4	66.4	65.0
57	Virginica	Versicolo	-18.6	83.9	34.7
62	Versicolo	Virginica	24.4	34.0	41.6
66	Versicolo	Virginica	11.9	37.9	50.2
70	Versicolo	Virginica	12.1	41.5	46.3
78	Virginica	Versicolo	-7.3	58.4	48.9
91	Virginica	Versicolo	-16.1	83.8	32.3
95	Versicolo	Virginica	23.7	14.3	62.0
106	Versicolo	Virginica	20.7	30.7	48.7
111	Virginica	Versicolo	-21.4	63.8	57.6
112	Virginica	Versicolo	-23.9	71.8	52.1
114	Versicolo	Virginica	17.1	35.6	47.2
117	Versicolo	Virginica	22.1	38.9	39.0
130	Versicolo	Virginica	30.9	32.4	36.8
131	Versicolo	Virginica	14.0	39.6	46.4
142	Versicolo	Virginica	21.4	38.6	40.0
148	Versicolo	Virginica	6.8	36.8	56.4



3.3 Zařazení všech objektů diskriminačními funkcemi do tříd

Setosa G_1 Versicolor G_2 , Virginica G_3 .

Procento zařazení do jednotlivé třídy

Objekt	Daná třída	Nalezená třída	Setosa G_1	Versicolor G_2	Virginica G_3
1	Setosa	Setosa	92.4	21.6	-14.0
2	Virginica	Virginica	-16.4	34.9	81.5
3	Versicolo	Versicolo	10.8	47.2	42.0
...	...	...	...	...	...
150	Setosa	Setosa	101.8	5.4	-7.2

Tabulka obsahuje pro každý objekt skutečnou třídu kosatců, dále predikovanou třídu kosatců a procento pravděpodobnosti zařazení do jednotlivé třídy kosatců.



Kosatce (Iris)



Iris Setosa

Wild Iris - Iris setosa

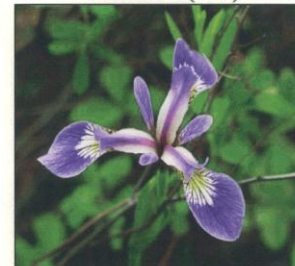


Iris Versicolor



Iris Virginica

Kosatce (Iris)



Iris Setosa



Iris Versicolor



Iris Versicolor



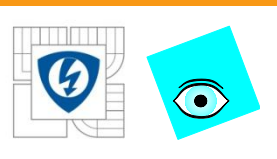
Iris Virginica



Iris Virginica



Iris Virginica



4. Klasifikace objektů kanonickou korelační analýzou.

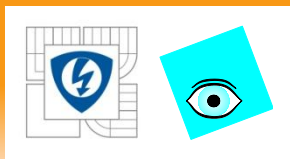
4.1 Analýza kanonických proměnných:

$C^1 B$	Ind.	Total	Kanon.	Kanon.	Čítatel	Jmenov.	Spočtená	Wilk. λ
Fn vlast.číslo	%	%	korel.	korel2	F-test	SVSV	kritérium	
1 32.191929	99.1	99.1	0.9848	0.9699	199.1	8.0288.00.000000.023439		
2 0.285391	0.9	100.0	0.4712	0.2220	13.8	3.0145.00.000000.777973		

F -test testuje, zda tato funkce a další jsou statisticky významné.

Tabulka obsahuje zařazování objektů kanonickou korelační analýzou.

U kanonické korelační analýzy jsou dva typy znaků: první skupina obsahuje znaky a druhá závisle proměnné.



4.2 Odhady koeficientů u kanonických proměnných

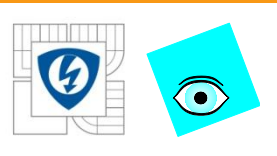
Obsahuje kanonické koeficienty k výpočtu kanonického skóre.

	1. kanonická	2. kanonická
Znak	proměnná	proměnná
Absolutní člen	-2.105106	6.661473
SepalLength	-0.082938	-0.002410
SepalWidth	-0.153447	-0.216452
PetalLength	0.220121	0.093192
PetalWidth	0.281046	-0.283919

4.3 Kanonické proměnné u třídních průměrů

Tabulka obsahuje výsledky kanonických koeficientů k výpočtu kanonické funkce pro průměry, těžiště u každé třídy.

	Kanonická	Kanonická
Třída kosatců	funkce 1	funkce 2
Setosa	-7.6076	-0.215133
Versicolor	1.82505	0.7278996
Virginica	5.78255	-0.5127666



4.4 Standardizované kanonické koeficienty u kanonických proměnných

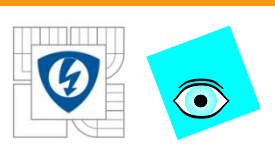
Tabulka obsahuje standardizované kanonické koeficienty.

Znak	1. kanonická proměnná	2. kanonická proměnná
SepalLength	-0.426955	-0.012408
SepalWidth	-0.521242	-0.735261
PetalLength	0.947257	0.401038
PetalWidth	0.575161	-0.581040

4.5 Korelace původních a kanonických pro

Tabulka obsahuje zátěže původních znaků a kanonických proměnných. Každá hodnota představuje korelaci mezi kanonickou proměnnou a znakiem.

Znak	1. kanonická proměnná	2. kanonická proměnná
SepalLength	0.222596	-0.310812
SepalWidth	-0.119012	-0.863681
PetalLength	0.706065	-0.167701
PetalWidth	0.633178	-0.737242



5. Skóre k zařazování objektů do tříd.

5.1 Lineární diskriminační skóre všech objektů k zařazení do tříd G_1 , G_2 a G_3

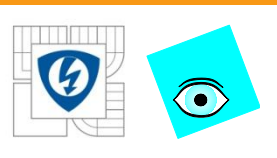
Tabulka obsahuje hodnoty vyčíslených lineárních diskriminačních skóre pro všech 150 kosatců.

Objekt	Daná třída	Nalezená predikovaná třída		
		Skóre třídy G_1	Skóre třídy G_2	Skóre třídy G_3
		Setosa	Versicolor	Virginica
1	Setosa	83.86837	38.65921	-6.790054
2	Virginica	1.230765	91.857	104.5692
..				
150	Setosa	98.72371	46.71882	-0.3055334

5.2 Regresní skóre všech objektů k zařazení do tříd G_1 , G_2 a G_3

Tabulka obsahuje vyčíslená predikovaná skóre, založená na regresních

Objekt	Daná třída	Nalezená predikovaná třída		
		Skóre třídy G_1	Skóre třídy G_2	Skóre třídy G_3
		Setosa	Versicolor	Virginica
1	Setosa	0.923755	0.215832	-0.139588
2	Virginica	-0.163732	0.348623	0.815109
3	Versicolo	0.107759	0.471953	0.420288
..				
..				
150	Setosa	1.018238	0.053607	-0.071844



5. Skóre k zařazování objektů do tříd.

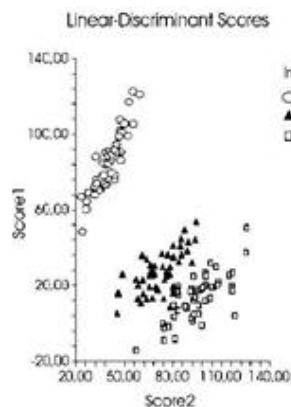
5.2 Kanonická skóre všech objektů k zařazení do dvou tříd

Tabulka obsahuje skóre kanonických proměnných pro každý řádek u všech 150 kosatců.

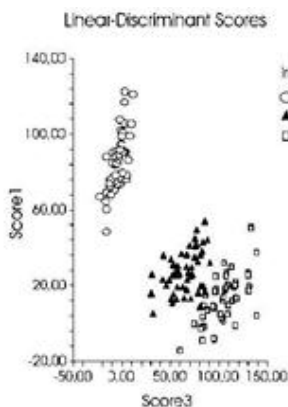
Objekt	Daná třída	Nalezená predikovaná třída	
		Skóre třídy G_1	Skóre třídy G_2
1	Setosa	-7.671967	0.134894
2	Virginica	6.800150	-0.580895
3	Versicolo	2.548678	0.472205
..			
..			
150	Setosa	-8.314449	-0.644953



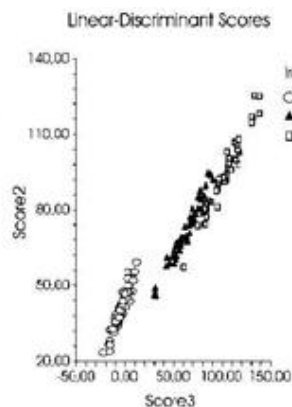
6. Klasifikační grafy. Užijeme grafy: (a) lineárních diskriminačních skóre, (b) regresních skóre nebo (c) kanonických skóre.



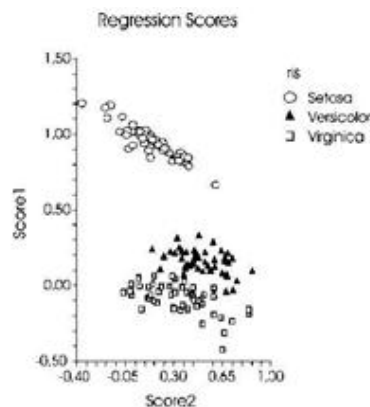
Obr. 7.9a Graf lineárního diskriminačního skóre 1 a 2 u 4 znaků 150 kosáčků zdrojové matice dat *Iris* (STATISTICA).



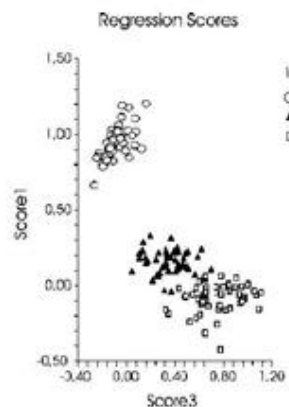
Obr. 7.9b Graf lineárního diskriminačního skóre 1 a 3 u 4 znaků 150 kosáčků zdrojové matice dat *Iris* (STATISTICA).



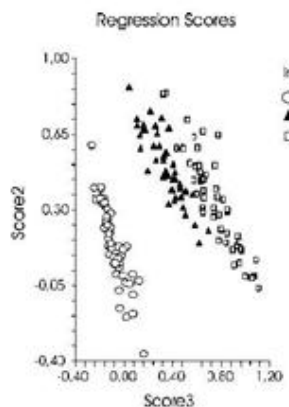
Obr. 7.9c Graf lineárního diskriminačního skóre 2 a 3 u 4 znaků 150 kosáčků zdrojové matice dat *Iris* (STATISTICA).



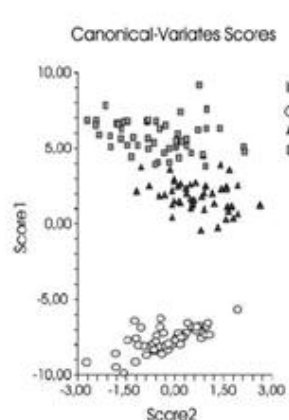
Obr. 7.9d Graf regresního skóre 1 a 2 u 4 znaků 150 kosáčků zdrojové matice dat *Iris* (STATISTICA).



Obr. 7.9e Graf regresního skóre 1 a 3 u 4 znaků 150 kosáčků zdrojové matice dat *Iris* (STATISTICA).



Obr. 7.9f Graf regresního skóre 2 a 3 u 4 znaků 150 kosáčků zdrojové matice dat *Iris* (STATISTICA).



Obr. 7.9g Graf kanonických proměnných 1 a 2 u 4 znaků 150 kosáčků zdrojové matice dat *Iris* (STATISTICA).