3. Statistika

Statistika nabízí celou řadu teoreticky dobře prozkoumaných a zdůvodněných a léty praxe ověřených metod pro analýzu dat. Pro oblast dobývání znalostí z databází mají význam (ať už přímo jako používané metody nebo nepřímo jako zdroj inspirace)[1]:

- *kontingenční tabulky* – pro zjišťování vztahu mezi dvěma kategoriálními veličinami,
- *regresní analýza* – pro zjišťování funkční závislosti jedné numerické (spojité) veličiny na jiných numerických veličinách,
- *diskriminační analýza* – pro odlišení příkladů (pozorování) patřících do různých tříd,
- *shluková analýza* – pro nalezení skupin (shluků) navzájem si podobných příkladů.

Z dalších metod zmiňme ještě *korelační analýzu* (pro posouzení, zda je mezi dvěma numerickými veličinami lineární závislost), *analýzu rozptylu* (pro posouzení rozdílu mezi průměry z různých výběrů) a *faktorovou analýzu* (pro zjišťování závislosti jedné veličiny na tzv. faktorech vytvořených jako lineární kombinace jiných veličin).

3.1 Kontingenční tabulky

V nejjednodušším případě můžeme kontingenční tabulku použít pro vyhodnocení vztahu mezi dvěma binárními veličinami. Je-li např. první veličina *příjem_klienta* s hodnotami {*vysoký, nízký*} a druhá veličina *úvěr* s hodnotami {*ano, ne*}, bude mít kontingenční tabulka pro *n* pozorování podobu uvedenou v Tab. 1. Takovéto tabulce se také někdy říká *čtyřpolní tabulka*. Jednotlivá "pole" tabulky odpovídají četnostem kombinací hodnot obou veličin v datech. Tedy:

a je počet klientů s vysokým příjmem, kteří získali úvěr,

b je počet klientů s vysokým příjmem, kteří nezískali úvěr,

c je počet klientů s nízkým příjmem, kteří získali úvěr

d je počet klientů s nízkým příjmem, kteří nezískali úvěr.

V tabulce jsou ještě uvedeny řádkové a sloupcové součty (tzv. marginální hodnoty):

a+b=r, c+d=s, a+c=k, b+d=l, a+b+c+d=n

	Úvěr ano	**Úvěr ne**	**Σ**
Vysoký příjem	a	b	r
Nízký příjem	c	d	s
Σ	k	l	n

Tab. 1 (Čtyřpolní) kontingenční tabulka

[1] Vzhledem k tomu, že při dobývání znalostí máme co do činění s množstvím sledovaných veličin, jedná se převážně o mnohorozměrné statistické metody.

V obecném případě může každá z veličin nabývat různého počtu hodnot většího než jedna. Pak má kontingenční tabulka podobu podle Tab. 2, kde

a_{kl} je četnost (frekvence) kombinace $(X=X_k) \wedge (Y=Y_l)$,

$$r_k = \sum_{l=1}^{S} a_{kl}, \quad s_l = \sum_{k=1}^{R} a_{kl}, \quad \text{a} \quad n = \sum_{k=1}^{R} \sum_{l=1}^{S} a_{kl} .$$

	Y_1	Y_2	.....	Y_S	Σ
X_1	a_{11}	a_{12}	.....	a_{1S}	r_1
X_2	a_{21}	a_{22}	.....	a_{2S}	r_2
:	:	:		:	:
:	:	:		:	:
X_R	a_{R1}	a_{R2}	.....	a_{RS}	r_R
Σ	s_1	s_2	.....	s_S	n

Tab. 2 (Obecná) kontingenční tabulka

Pro zjišťování vztahu mezi oběma veličinami se obvykle používá χ^2 test. Tento test je založen na vyhodnocení rozdílu mezi pozorovanými četnostmi jednotlivých kombinací (uvedenými v tabulce) a četnostmi očekávanými při platnosti hypotézy o nezávislosti obou veličin (počítanými z marginálních hodnot). Je-li pozorována četnost

$$a_{kl} ,$$

je jí odpovídající očekávaná četnost (za předpokladu nezávislosti X a Y)

$$e_{kl} = \frac{r_k \, s_l}{n} .$$

Pro kontingenční tabulku spočítáme statistiku χ^2

$$\chi^2 = \sum_{k=1}^{R} \sum_{l=1}^{S} \frac{(a_{kl} - e_{kl})^2}{e_{kl}} = n \times \sum_{k=1}^{R} \sum_{l=1}^{S} \frac{\left(a_{kl} - \frac{r_k \, s_l}{n}\right)^2}{r_k \, s_l} .$$

Při platnosti nulové hypotézy nezávislosti veličin X a Y

$$H_0 : \; P(X=X_k \wedge Y=Y_l) = P(X=X_k) \times P(Y=Y_l) \quad \text{pro všechna } k,l$$

má tato statistika rozdělení χ^2 s *(R-1)(S-1)* stupni volnosti. Je-li hodnota χ^2 statistiky větší nebo rovna hodnotě χ^2 rozdělení [2] s daným počtem stupňů volnosti na zvolené hladině významnosti α

$$\chi^2 \geq \chi^2_{(R-1)(S-1)}(\alpha)$$

[2] Hodnoty rozdělení pro různé počty stupňů volnosti a různé hladiny významnosti jsou uvedeny v každé učebnici statistiky.

zamítneme na této hladině významnosti nulovou hypotézu ve prospěch alternativní hypotézy závislosti obou veličin. Tedy, vytvoříme-li na základě dat čtyřpolní tabulku Tab. 3, spočítáme hodnotu statistiky $\chi^2 = 84.65$ (očekávané četnosti jsou uvedeny v Tab. 4). Vzhledem k tomu, hodnota χ^2 rozdělení s jedním stupněm volnosti je (pro hladinu významnosti $\alpha=0.05$) $\chi^2_{(1)}(0.05) = 3.84$, zamítneme nulovou hypotézu a předpokládáme závislost mezi výší příjmu a poskytnutím úvěru.

	Úvěr ano	**Úvěr ne**	**Σ**
Vysoký příjem	50	0	50
Nízký příjem	30	40	70
Σ	80	40	120

Tab. 3 Data pro výpočet χ^2 testu

Kombinace	**skutečný počet**	**očekávaný počet**	**rozdíl**
Vysoký příjem, úvěr ano	50	33.3	16.7
Vysoký příjem, úvěr ne	0	16.7	-16.7
Nízký příjem, úvěr ano	30	46.7	-16.7
Nízký příjem, úvěr ne	40	13.3	26.7

Tab. 4 Skutečné vs. očekávané počty pozorování

χ^2 test má ale jedno úskalí; lze ho použít pouze v případě, že očekávané četnosti jsou dostatečně veliké. Uvádí se tedy podmínka použitelnosti testu

$$\frac{r_k \; s_l}{n} \geq 5 \text{, pro všechna k, l.}$$

Pro čtyřpolní tabulky lze použít též *Fisherův test*, který naopak dobře funguje pro malé četnosti a netrpí tedy výše uvedeným omezením. Základem tohoto testu je výpočet pravděpodobnosti, že při daných marginálních četnostech *r* a *s* má čtyřpolní tabulka skutečné četnosti a_{kl}.

$$p = \frac{r_1!\; r_2!\; s_1!\; s_2!}{n!\; a_{11}!\; a_{12}!\; a_{21}!\; a_{22}!}$$

Tyto pravděpodobnosti se nasčítají pro různé hodnoty skutečných četností při daných marginálech:

$$\text{Fisher} = \sum_{i=0}^{\min(r_1,s_1)-a_{11}} \frac{r_1!\; r_2!\; s_1!\; s_2!}{n!\; (a_{11}+i)!\; (a_{12}-i)!\; (a_{21}-i)!\; (a_{22}+i)!}\,.$$

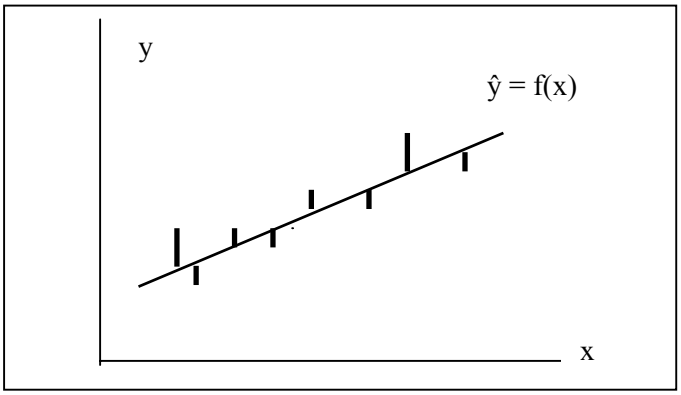
Je-li Fisher $\leq \alpha$, zamítneme nulovou hypotézu o nezávislosti na hladině významnosti α.

3.2 Regresní analýza

Zatímco při korelační analýze nás zajímá pouze to, zda mezi dvěma numerickými veličinami platí lineární závislost [3] v případě *lineární regrese* nás zajímají parametry této závislosti. Řešíme tedy úlohu a*proximace* pozorovaných hodnot daným typem funkce, ovšem s neznámými parametry. V nejjednodušším případě lineární regrese pro dvě veličiny hledáme hodnoty parametrů q_1 a q_0 pro rovnici

$$y = q_1 x + q_0 + \varepsilon.$$

Máme-li k dispozici vhodná pozorování (dvojce hodnot $[x_i, y_i]$) můžeme parametry rovnice spočítat (přesněji řečeno odhadnout) na základě *metody nejmenších čtverců*. Tato metoda minimalizuje rozdíly mezi pozorovanou hodnotou y a očekávanou hodnotou $\hat{y}=f(x)$ spočítanou v tomto případě na základě funkce $q_1 x + q_0$ (viz ilustrační Obr. 1 znázorňující tyto odchylky).



Obr. 1 Metoda nejmenších čtverců

Vzhledem k tomu, že považujeme kladné rozdíly za stejně závažné jako rozdíly záporné, uvažujeme druhou mocninu [4] (kvadrát, čtverec) těchto rozdílů:

$$(y - f(x))^2$$

Hledání minima celkové odchylky pro n pozorování

$$\min \sum_{i=1}^{n} \left(y_i - f(x_i)\right)^2$$

se pak převádí na řešení rovnice

$$\frac{d}{dq} \sum_{i=1}^{n} \left(y_i - f(x_i)\right)^2 = 0$$

ze které pak již lze spočítat (odhadnout) parametry funkce $f(x)$. Pro naši funkci

$$f(x) = q_1 x + q_0$$

[3] Korelační koeficient $\rho \in [-1,1]$ nabývá hodnotu 1 pro přímou úměru obou veličin, hodnotu –1 pro nepřímou úměru obou veličin a hodnotu 0 v případě, že mezi veličinami není lineární závislost. To, že korelační koeficient je nulový ještě neznamená, že mezi veličinami není funkční vztah; příkladem mohou být veličiny x a y, kde $y=|x|$.

[4] Obecněji uvažujeme nějakou sudou funkci, tj funkci $f(x)$ takovou, že $f(-x)=f(x)$. Druhé mocnině se dává přednost před absolutní hodnotou, protože druhá mocnina je hladká funkce.

bude

$$\frac{\partial}{\partial q_0}\sum_{i=1}^{n}\left(y_i - (q_1 x_i + q_0)\right)^2 = -2\sum_{i=1}^{n} y_i + 2q_1\sum_{i=1}^{n} x_i + 2q_0 n$$

$$\frac{\partial}{\partial q_1}\sum_{i=1}^{n}\left(y_i - (q_1 x_i + q_0)\right)^2 = -2\sum_{i=1}^{n} x_i y_i + 2q_1\sum_{i=1}^{n} x_i^2 + 2q_0\sum_{i=1}^{n} x_i$$

Položíme-li obě parciální derivace rovny nule, můžeme spočítat parametry regresní přímky

$$q_0 = \frac{(\sum_i y_i)(\sum_i x_i^2) - (\sum_i x_i y_i)(\sum_i x_i)}{n(\sum_i x_i^2) - (\sum_i x_i)^2}$$

$$q_1 = \frac{n(\sum_i x_i y_i) - (\sum_i x_i)(\sum_i y_i)}{n(\sum_i x_i^2) - (\sum_i x_i)^2}$$

Pokud předpokládáme lineární závislost mezi x a y, nalezneme uvedeným postupem optimální parametry rovnice $y = ax + b$. Náš předpoklad ale nemusí platit. Pro posouzení „síly“ lineární závislosti proto můžeme použít korelační koeficient (sčítáme pro pozorování $i=1,...,n$):

$$\rho(x,y) = \frac{S_{xy}}{\sqrt{S_x^2\, S_y^2}}\ ,$$

kde $S_{xy} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{n-1}$ je výběrová kovariance, $S_x^2 = \frac{\sum_i (x_i - \bar{x})^2}{n-1}$ a $S_y^2 = \frac{\sum_i (y_i - \bar{y})^2}{n-1}$ jsou výběrové rozptyly a $\bar{x} = \frac{1}{n}\sum_i x_i$ a $\bar{y} = \frac{1}{n}\sum_i y_i$ jsou průměry.

Pro dobývání znalostí má větší význam mnohorozměrná regrese, ať už lineární nebo nelineární. V případě mnohorozměrné lineární regrese předpokládáme lineární závislost jedné vysvětlované (závislé) veličiny y na více vysvětlujících (nezávislých) veličinách $x_1, x_2, \ldots, x_m$. Pro i-té pozorování tedy předpokládáme

$$y_i = q_0 + q_1 x_{i1} + q_2 x_{i2} + \ldots + q_m x_{im} + \varepsilon_i.$$

V maticovém tvaru

$$\mathbf{y} = \mathbf{X}\mathbf{q},$$

kde

$$\mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \qquad \mathbf{X} = \begin{bmatrix} 1, & x_{11}, & \ldots, & x_{1m} \\ \vdots & \vdots & & \vdots \\ 1, & x_{n1}, & \ldots, & x_{nm} \end{bmatrix}, \qquad \mathbf{q} = \begin{bmatrix} q_0 \\ \vdots \\ q_m \end{bmatrix}.$$

Řešením rovnice $\boldsymbol{y} = \boldsymbol{Xq}$, které získáme metodou nejmenších čtverců, je pak [5]:

$$\mathbf{q} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}\ .$$

[5] Předpokládáme, že hodnoty y jsou nekorelované a že mají stejný rozptyl. Pokud tento předpoklad neplatí, má řešení tvar $\mathbf{q} = (\mathbf{X}^T \mathbf{S}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{S}^{-1} \mathbf{y}$, kde je $\boldsymbol{S}$ kovarianční matice.

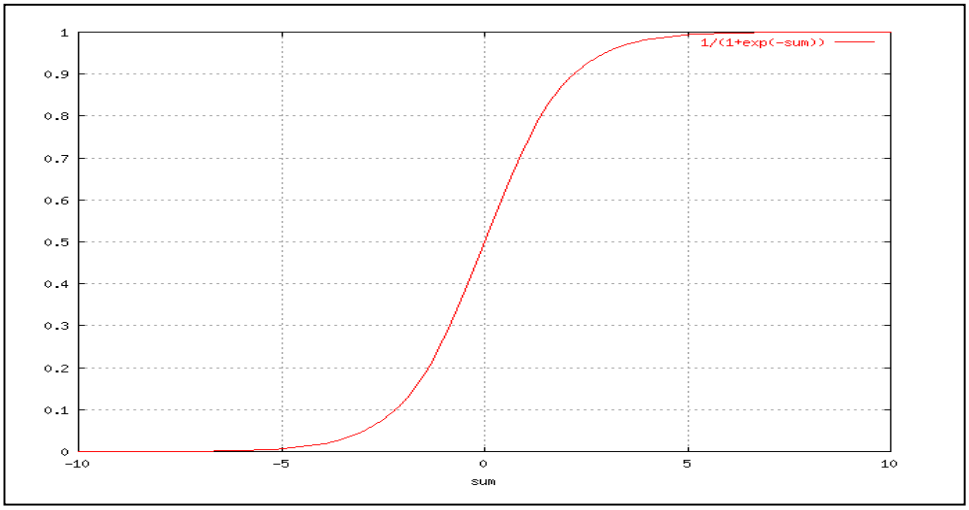
V případě nelineární regrese předpokládáme složitější funkční závislost mezi y a $\boldsymbol{x}$ (kvadratickou, obecně polynomiální, exponenciální,….). Zajímavým případem nelineární regrese je regrese *logistická*. U tohoto modelu předpokládáme, že závislá veličina y je kategoriální (v nejjednodušším případě dvouhodnotová). Při logistické regresi nemodelujeme přímo veličinu y ale pravděpodobnost, že tato veličina má konkrétní hodnotu v závislosti na kombinaci hodnot nazávislých veličin $\boldsymbol{x}$. Vycházíme přitom z pojmu *podmíněné šance* jakožto podílu $P(y|\boldsymbol{x})$ a $(1 - P(y|\boldsymbol{x}))$. V případě, že veličina y nabývá pouze hodnot 0 a 1 (nejjednodušší případ), má logistický model podobu

$$\ln \frac{P(y=1\,|\,x_1, x_2, \dots x_m)}{1 - P(y=1\,|\,x_1, x_2, \dots x_m)} = q_0 + q_1x_1 + q_2x_2 + \dots + q_rx_m \quad ,$$

resp.

$$P(y=1\,|\,x_1, x_2, \dots x_m) = \frac{e^{q_0+\sum_j q_jx_j}}{1+e^{q_0+\sum_j q_jx_j}} = \frac{1}{1+e^{-q_0-\sum_j q_jx_j}}$$

Tento model umožňuje odhadnout šance resp. pravděpodobnosti hodnoty $y=1$ [6]. Průběh funkce $f(sum)=1/(1+e^{-sum})$, kde $sum=q_0+\sum_j q_jx_j$ ukazuje Obr. 2. Této funkci se pro její tvar často říká *sigmoida*.



Obr. 2 Průběh sigmoidální funkce

Odhad parametrů modelu se provádí na základě metody maximální věrohodnosti (maximum likelihood). Při této metodě se maximalizuje součin pravděpodobností $p(y_i|\ x_{i1}, x_{i2}, \dots x_{im})$ [7] přes všechna pozorování $i=1,\dots,n$:

$$L = \prod_{i=1}^{n} P(y_i|\ x_{i1}, x_{i2}, \dots x_{im}) = \prod_{i=1}^{n} \left(\left(\frac{1}{1+e^{-q_0-\sum_j q_jx_j}} \right)^{y_i} \times \left(\frac{1}{1+e^{q_0+\sum_j q_jx_j}} \right)^{1-y_i} \right).$$

[6] Pravděpodobnost $P(y=0|\ x_1, x_2, \dots x_m)$ lze spočítat na základě vztahu $P(y=0|\ x_1, x_2, \dots x_m) = 1 - P(y=1|\ x_1, x_2, \dots x_m)$.

[7] Vzhledem k tomu že y nabývá hodnot 0 nebo 1, je pro každý příklad x_i vždy jeden člen součinu $A^y \times B^{y-1}$ roven 1.

3.3 Diskriminační analýza

Diskriminační analýza je vlastně úloha klasifikace příkladů do předem zadaných tříd. Z pohledu statistického tedy úloha hledání závislosti jedné nominální veličiny (určující příslušnost ke třídě) na dalších v numerických veličinách. Při diskriminační analýze předpokládáme, že ke každé třídě (hodnotě nominální veličiny) c_t, $t=1,\ldots,T$ existuje *(diskriminační) funkce* f_t taková, že

$$f_t(\mathbf{x}) = \max_k f_k(\mathbf{x}), \quad k=1,\ldots,T$$

právě když příklad $\mathbf{x}=[x_1, x_2, \ldots, x_m]$ patří do třídy c_t.

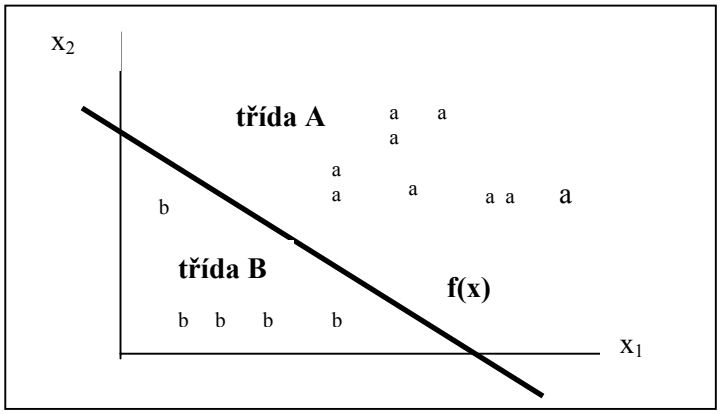
V případě *lineární diskriminační analýzy* mají diskriminační funkce f_t podobu lineární kombinace

$$f_t = q_{0j} + q_{1j}x_1 + q_{2j}x_2 + \ldots + q_{mj}x_m$$

Ukažme si pouze ten nejjednodušší případ, diskriminaci do dvou tříd, kdy místo funkcí f_1 a f_2 můžeme hledat funkci

$$f(\mathbf{x}) = f_1(\mathbf{x}) - f_2(\mathbf{x}).$$

Příklady pak můžeme klasifikovat podle znaménka této funkce. Pokud si představíme příklady jako body v m-rozměrném prostoru veličin, bude funkce $f(\mathbf{x})$ představovat nadrovinu v tomto prostoru, oddělující od sebe příklady obou tříd. Pro $m=2$ je funkce $f(x)$ přímka (Obr. 3)[8].



Obr. 3 Rozdělovací přímka

Pokud chceme zaručit optimalitu klasifikace ve smyslu minimální chyby, musíme použít jako diskriminační funkce podmíněné (aposteriorní) pravděpodobnosti zařazení pozorování $\mathbf{x}$ do třídy c_t.

$$f_t(\mathbf{x}) = P(c_t \mid \mathbf{x}) = \frac{P(\mathbf{x} \mid c_t)\, P(c_t)}{\sum_k P(\mathbf{x} \mid c_k)\, P(c_k)}.$$

Ve variantě pro dvě třídy lze opět použít rozdíl aposteriorních pravděpodobností

$$f(\mathbf{x}) = f_1(\mathbf{x}) - f_2(\mathbf{x}) = P(\mathbf{x} \mid c_1)\, P(c_1) - P(\mathbf{x} \mid c_2)\, P(c_2).$$

[8] V případě R diskriminačních funkcí lze získat (R*(R-1))/2 nadrovin.

Pokud se uvedené pravděpodobnosti $P(\mathbf{x}|c_k)\ P(c_k)$ řídí normálním rozdělením, lze odvodit podobu diskriminační funkce

$$f(\mathbf{x}) = \frac{1}{2}\mathbf{X}^T(\mathbf{S}_1^{-1} - \mathbf{S}_2^{-1})\mathbf{X}^T + (\boldsymbol{\mu}_1^T\mathbf{S}_1^{-1} - \boldsymbol{\mu}_2^T\mathbf{S}_2^{-1})\mathbf{X} + \frac{1}{2}\ln\frac{|\mathbf{S}_2|}{|\mathbf{S}_1|} - \frac{1}{2}(\boldsymbol{\mu}_1^T\mathbf{S}^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2^T\mathbf{S}^{-2}\boldsymbol{\mu}_2) - \ln\frac{P(c_1)}{P(c_2)},$$

která je obecně kvadratická.

Pokud však předpokládáme, že rozdělení mají stejné kovarianční matice $\boldsymbol{S}_1 = \boldsymbol{S}_2 = \boldsymbol{S}$, získáme lineární diskriminační funkci

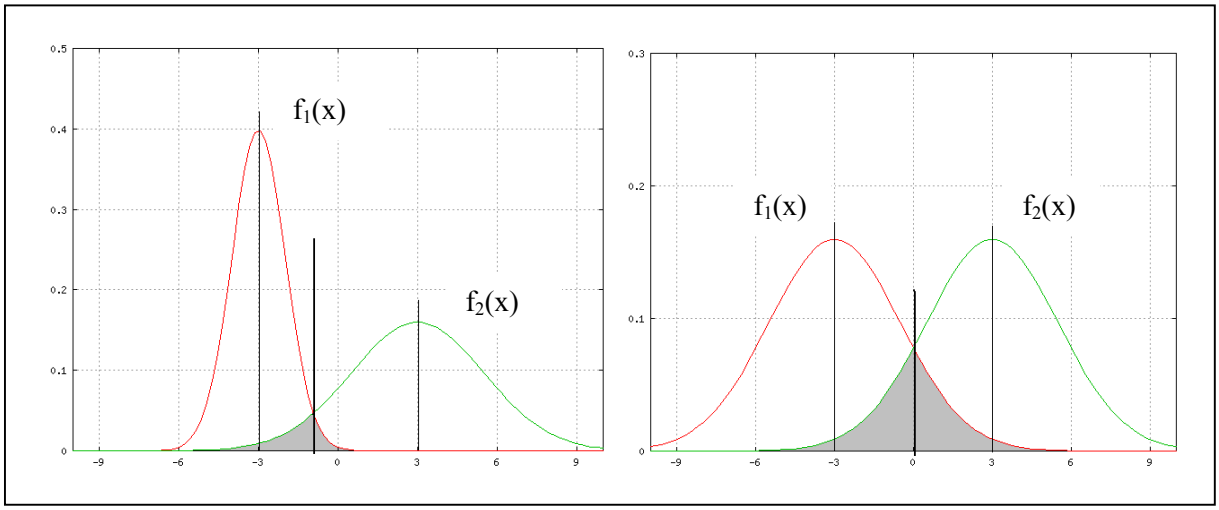
$$f(\mathbf{x}) = (\boldsymbol{\mu}_1^T - \boldsymbol{\mu}_2^T)\,\mathbf{S}^{-1}\,\mathbf{X} - \frac{1}{2}(\boldsymbol{\mu}_1^T - \boldsymbol{\mu}_2^T)\,\mathbf{S}^{-1}\,(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) - \ln\frac{P(c_1)}{P(c_2)}.$$

Pokud navíc předpokládáme, že kovarianční matice jsou jednotkové a obě třídy jsou stejně pravděpodobné, zjednoduší se diskriminační funkce do tvaru

$$f(\mathbf{x}) = (\boldsymbol{\mu}_1^T - \boldsymbol{\mu}_2^T)\,\mathbf{X} - \frac{1}{2}(\boldsymbol{\mu}_1^T - \boldsymbol{\mu}_2^T)\,(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2).$$

Hledání diskriminační funkce se tedy za předpokladu normality "redukuje" na odhad středních hodnot $\boldsymbol{\mu}_i$ (na základě výběrových průměrů) a kovariančních matic $\boldsymbol{S}_i$ (na základě výběrových rozptylů).

Obr. 4 ukazuje pro jednorozměrný případ diskriminační funkci (bod označený silnou čarou) v situaci, kdy předpokládáme normální rozdělení pravděpodobností $P(x|c_k)\ P(c_k)$ s různými rozptyly (vlevo) a stejnými rozptyly (vpravo). Pro stejné rozptyly tedy můžeme diskriminovat pouze na základě odhadů středních hodnot.



Obr. 4 Diskriminace podle minimální chyby, jednorozměrné rozdělení

3.4 Shluková analýza

Shluková analýza hledá odpověď na otázku, zda lze pozorované příklady rozdělit do skupin (shluků) vzájemně si blízkých příkladů. Vychází se tedy z toho, že umíme měřit vzdálenost mezi příklady.

Předpokládejme, že každý příklad je charakterizován m numerickými veličinami. Vzdálenost mezi dvěma příklady $\mathbf{x_1} = [x_{11},\ldots,x_{1m}]$ a $\mathbf{x_2} = [x_{21},\ldots,x_{2m}]$ lze vyjádřit různými mírami. Uveďme zde např.:

- Hammingovu vzdálenost

$$d_H(\mathbf{x_1},\mathbf{x_2}) = \sum_{j=1}^{m} |x_{1j} - x_{2j}|$$

- Eukleidovskou vzdálenost

$$d_E(\mathbf{x_1},\mathbf{x_2}) = \sqrt{\sum_{j=1}^{m} (x_{1j} - x_{2j})^2}$$

- Čebyševovu vzdálenost

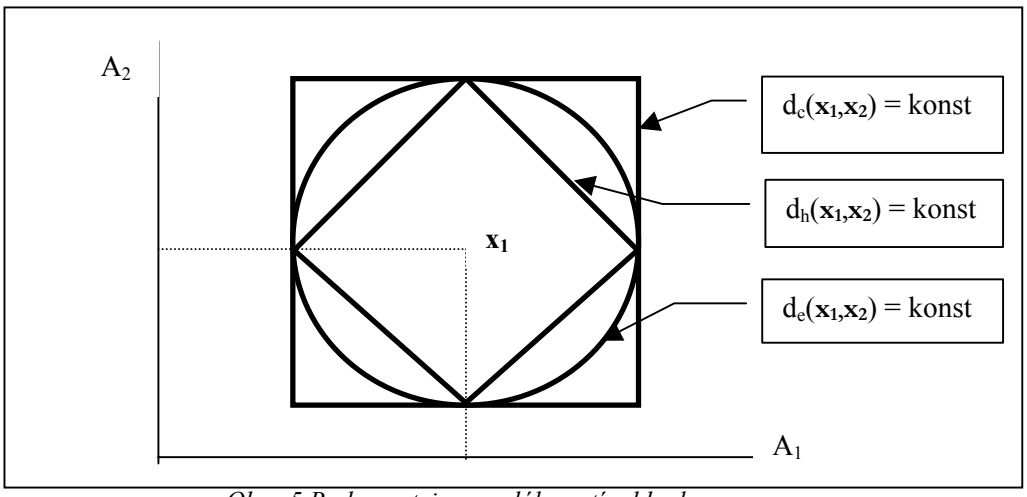
$$d_C(\mathbf{x_1},\mathbf{x_2}) = \max_j |x_{1j} - x_{2j}|$$

Ve všech výše uvedených případech se jedná o speciální příklady Minkovského metriky

$$L^{(z)}(\mathbf{x_1},\mathbf{x_2}) = \sqrt[z]{\sum_{j=1}^{m} (x_{1j} - x_{2j})^z}$$

$$d_H(\mathbf{x_1},\mathbf{x_2}) = L^{(1)}(\mathbf{x_1},\mathbf{x_2}), \quad d_E(\mathbf{x_1},\mathbf{x_2}) = L^{(2)}(\mathbf{x_1},\mathbf{x_2}), \quad d_C(\mathbf{x_1},\mathbf{x_2}) = \lim_{z\to\infty} L^{(z)}(\mathbf{x_1},\mathbf{x_2}),$$

Rozdíl mezi $d_H(\mathbf{x_1},\mathbf{x_2})$, $d_E(\mathbf{x_1},\mathbf{x_2})$ a $d_C(\mathbf{x_1},\mathbf{x_2})$ ukazuje Obr. 5. Zde jsou (pro jednotlivé míry) znázorněny body $\mathbf{x_2}$, které mají stejnou vzdálenost od bodu $\mathbf{x_1}$: kruh pro eukleidovskou vzdálenost, čtverec rovnoběžný s osami pro Čebyševovu vzdálenost a čtverec „na špičce“ pro Hammingovu vzdálenost [Hebák, Hustopecký, 1987].



Obr. 5 Body se stejnou vzdáleností od bodu u

Výše uvedené míry vzdálenosti závisí na měřítku veličin. Proto je třeba veličiny normovat. Konkrétní hodnota se obvykle dělí nějakou jinou hodnotou: průměrem, směrodatnou odchylkou, nebo rozpětím (max-min). Navíc předpokládají stejný rozptyl u všech veličin. V případě různého rozptylu se doporučuje použít Mahalanobisovu vzdálenost, která je zobecněním vzdálenosti eukleidovské.

$$d_M^2(\mathbf{x}_1,\mathbf{x}_2) = (\mathbf{x}_1 - \mathbf{x}_2)^T \mathbf{S}^{-1} (\mathbf{x}_1 - \mathbf{x}_2)$$

Z používaných shlukové analýzy metod zmiňme:

- hierarchické shlukování
- metoda *K*-středů (*K*-means clustering)

Při hierarchickém shlukování se obvykle postupuje metodou „zdola nahoru“. Začíná se tedy v situaci, kdy každý příklad tvoří jeden samostatný shluk. Postupně se pak jednotlivé shluky spojují až skončíme s jedním shlukem obsahujícím všechny příklady (Obr. 6).

Algoritmus hierarchického shlukování

Inicializace
1. urči vzájemné vzdálenosti mezi všemi příklady
2. zařaď každý příklad do samostatného shluku

hlavní cyklus
1. dokud je více než jeden shluk
   1.1. najdi dva navzájem nejbližší shluky a spoj je
   1.2. spočítej pro tento nový shluk jeho vzdálenost od ostatních shluků

Obr. 6 Algoritmus hierarchického shlukování

Vzdálenost mezi shluky lze stanovit různým způsobem:

- *metoda nejbližšího souseda* - vzdálenost mezi shluky ***U*** a ***V*** je dána minimem ze vzdálenosti mezi jejich příklady

$$D(\mathbf{U},\mathbf{V}) = \min_{k,l} d(\mathbf{x}_k, \mathbf{x}_l), \quad \mathbf{x}_k \in \mathbf{U}, \mathbf{x}_l \in \mathbf{V}$$

- *metoda nejvzdálenějšího souseda* - vzdálenost mezi shluky ***U*** a ***V*** je dána maximem ze vzdálenosti mezi jejich příklady

$$D(\mathbf{U},\mathbf{V}) = \max_{k,l} d(\mathbf{x}_k, \mathbf{x}_l), \quad \mathbf{x}_k \in \mathbf{U}, \mathbf{x}_l \in \mathbf{V}$$

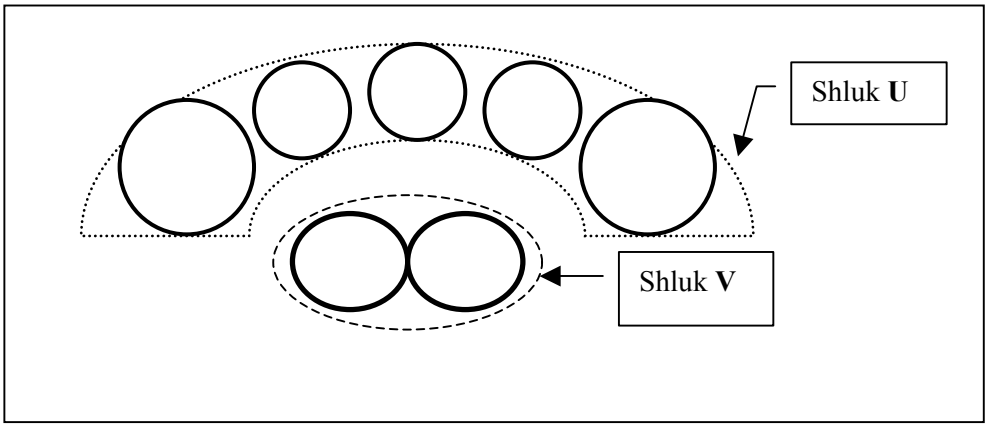
- *metoda průměrné vzdálenosti* - vzdálenost mezi shluky ***U*** a ***V*** je dána průměrem ze vzdálenosti mezi jejich příklady (n_U je počet příkladů ve shluku **U** a n_V je počet příkladů ve shluku **V**)

$$D(\mathbf{U},\mathbf{V}) = \frac{1}{n_U\, n_V} \times \sum_{k=1}^{n_u} \sum_{l=1}^{n_V} d(x_k,x_l)$$

- *centroidní metoda* - vzdálenost mezi shluky ***U*** a ***V*** je dána vzdáleností mezi středy shluků.

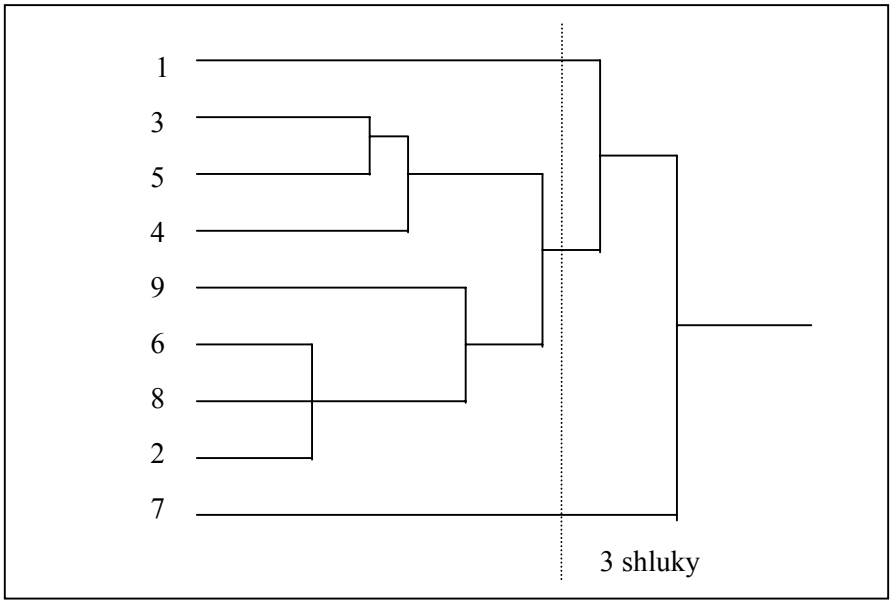
$$D(\mathbf{U},\mathbf{V}) = d(\bar{\mathbf{u}}, \bar{\mathbf{v}}), \quad \bar{\mathbf{u}} \text{ je střed shluku } \boldsymbol{U} \text{ a } \bar{\mathbf{v}} \text{ je střed shluku } \boldsymbol{V}$$

Centroidy (středy shluků), zmíněné v předcházejícím výčtu, představují jakési prototypy reprezentující jednotlivé shluky [9]. Nemusí ale platit, že ke každému shluku patří právě jeden centroid. V závislosti na tvaru shluku a zvolené míře pro výpočet vzdálenosti může být jeden shluk reprezentován více centroidy (na Obr. 7 znázorněnými jako tučné kružnice). [10]



Obr. 7 Více centroidů pro jeden shluk

Proces hierarchického shlukování bývá zachycen v podobě tzv. dendrogramu. Ten ukazuje (zleva doprava) postupné spojování shluků počínaje očíslovanými příklady. Optimální počet shluků zde není předem znám, odvodíme ho rozborem výsledků – tak, že někde dendrogram „rozřízneme“ (Obr. 8).



Obr. 8 Dendrogram

[9] V nejjednodušším případě můžeme centroidy chápat jako příklady, které nabývají průměrných hodnot jednotlivých veličin v rámci daného shluku. Takto chápané centroidy zaručují optimalitu klasifikace v situaci, kdy aposteriorní pravděpodobnosti zařazení příkladů do tříd se řídí normálním rozdělením se stejnou (jednotkovou) kovarianční maticí a stejnou pravděpodobností jednotlivých tříd – více viz odstavec diskriminační analýza.

[10] Všimněme si toho, že dané shluky nejsou lineárně separabilní. Lineární diskriminační analýza tedy nedokáže bezchybně od sebe odlišit příklady obou shluků.

Při shlukování metodou *K*-středů předpokládáme, že víme do kolika shluků je možno příklady rozdělit. Počet shluků se tedy během výpočtu nemění, mění se pouze zařazení příkladů k těmto shlukům. Proto je tato metoda méně výpočetně náročná než hierarchické shlukování (a tudíž vhodnější pro větší datové soubory). Příslušný algoritmus je uveden na Obr. 9.

Metoda *K* -středů

1. náhodně zvol rozklad do *K* shluků
2. urči centroidy pro všechny shluky v aktuálním rozkladu
3. pro každý příklad $\boldsymbol{x}$
   3.1. urči vzdálenosti $d(\mathbf{x},\mathbf{c}_k)$, $k=1,\ldots,K$ kde $\boldsymbol{c_k}$ je centroid *k*-tého shluku
   3.2. nechť $d(\mathbf{x},\mathbf{c}_l) = \min_k d(\mathbf{x},\mathbf{c}_k)$
   3.3. není-li $\boldsymbol{x}$ součástí shluku *l* (k jehož centroidu $\boldsymbol{c_l}$ má nejblíže) přesuň $\boldsymbol{x}$ do shluku *l*
4. došlo-li k nějakému přesunu potom jdi na 2 jinak konec

Obr. 9 Algoritmus shlukování metodou K-středů

Uvedený algoritmus může mít určité varianty:

- místo počátečního rozkladu lze za centroidy prohlásit prvních *K* příkladů; odpadne tak krok 2 při prvním průchodu daty,
- přepočet centroidů lze provádět po každém přesunu (tedy v cyklu v kroku 3).

Výsledné shluky jsou při použití metody *K*-středů reprezentovány svými centroidy. Tuto reprezentaci lze snadno použít pro zařazování nových příkladů. Příklad bude (ve shodě s krokem 3.3 algoritmu) zařazen do shluku, k jehož centroidu má nejblíže.

Literatrura:

[Anděl, 1978] Anděl,J.: Matematická statistika, Praha, SNTL/ALFA 1978.

[Havránek, 1993] Havránek,T.: Statistika pro biologické a lékařské vědy. Academia, 1993, ISBN 80-200-0080-1.

[Hebák, Hustopecký, 1987] Hebák,P. – Hustopecký,J.: Vícerozměrné statistické metody s aplikacemi. SNTL, 1987.

[Kotek a kol.,1980] Kotek,Z. - Chalupa,V. - Brůha,I. - Jelínek,J.: Adaptivní a učící se systémy. SNTL, Praha, 1980.

[Michie a kol., 1994] Michie,D. – Spiegelhalter,D.J. – Taylor,C.: Machine learning, neural and statistical classification. Ellis Horwood, 1994, ISBN 0-13-106360-X.