Dobývání znalostí

Doc. RNDr. Iveta Mrázová, CSc.

Katedra teoretické informatiky
Matematicko-fyzikální fakulta
Univerzity Karlovy v Praze

Dobývání znalostí

– Předzpracování dat –

Doc. RNDr. Iveta Mrázová, CSc.

Katedra teoretické informatiky
Matematicko-fyzikální fakulta
Univerzity Karlovy v Praze

Výběr a uspořádání příznaků

Pravděpodobnost chybného rozhodnutí

×

Množství informace obsažené ve vstupních vzorech

- **Příliš velký počet příznaků:**
  - technická realizovatelnost
  - rychlost zpracování
  - nebezpečí přeučení
    - počet proměnných × počet trénovacích vzorů
  - korelace příznaků

Volba informativních příznaků

- **Výběr minimálního počtu příznaků** z předem zvolené množiny příznaků
  - nelze zaručit, že tato množina obsahuje informativní příznaky
  - volba závisí na konkrétní úloze
- **Uspořádání příznaků** v předem zvolené množině příznaků
  - podle množství nesené informace
  - využití např. u sekvenčních klasifikátorů

Karhunen-Loevovův rozvoj (1)

Vlastnosti Karhunen-Loevova rozvoje:

1. Při daném počtu členů rozvoje poskytuje ze všech rozvojů **nejmenší střední kvadratickou odchylku** od původních vzorů
2. Vzory jsou po použití disperzní matice po aproximaci nekorelované

   → **dekorelace příznaků**

Karhunen-Loevovův rozvoj (2)

3. Členy rozvoje **nepřispívají rovnoměrně k aproximaci**
   - Vliv každého z členů na přesnost aproximace se zmenšuje s jeho pořadovým číslem
     → Vliv členů s vysokými indexy bude malý a můžeme je zanedbat (~ vynechat)
4. **Velikost chyby aproximace neovlivňuje strukturu rozvoje**
   - Změna požadavků na chybu aproximace nevyžaduje přepočítávat celý rozvoj
     → Stačí jen přidat či odstranit několik posledních členů

Výhodné zejména u **sekvenčních metod klasifikace**

Karhunen-Loevovův rozvoj (3)

- Volba vhodného zobrazení $V: X^m \rightarrow X^p$ tak, aby vzory z X^p byly nejlepší aproximací původních vzorů z X^m ve smyslu střední kvadratické odchylky

  K vzorů z jedné třídy

  m příznaků

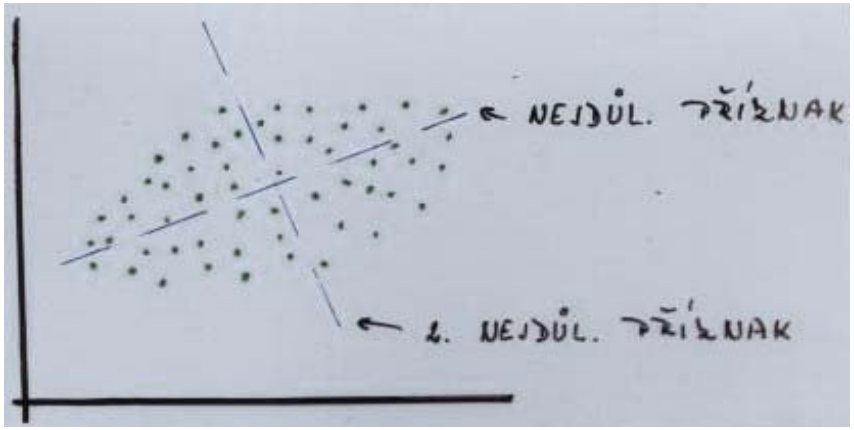
  p ortonormálních vektorů $\mathbf{e}_i$ $(1 \leq i \leq p)$ v X^m $(p \leq m)$

  → Aproximace vektorů $\mathbf{x}_k$ z X^m $(1 \leq k \leq K)$ lineární kombinací vektorů $\mathbf{e}_i$:

  $$\mathbf{y}_k = \sum_{i=1}^{p} c_{ki}\, \mathbf{e}_i$$

  tak, aby kvadrát odchylky $\mathbf{x}_k$ od $\mathbf{y}_k$: $\varepsilon_k^2 = \|\mathbf{x}_k - \mathbf{y}_k\|^2$ byl minimální

Karhunen-Loevovův rozvoj (4)



$\mathbf{v} = (v_1, v_2, \ldots)^T$,

$\mathbf{x} = (x_1, x_2, \ldots)^T$

$\mathbf{y} = \mathbf{v}^T \mathbf{x} = v_1 x_1 + v_2 x_2 + \ldots$

Měřeno m příznaků, z nichž chceme získat p nejdůležitějších příznaků $(1 \leq p << m)$

Matice $\boldsymbol{V}$: $p \times m$

$$\mathbf{V} = \begin{pmatrix} v_{11} & \cdots & v_{1p} \\ \vdots & \ddots & \vdots \\ v_{m1} & \cdots & v_{mp} \end{pmatrix}$$

Výpočet vektoru p nejdůležitějších příznaků:

$$\mathbf{y} = \mathbf{V}^T \mathbf{x}$$

Karhunen-Loevovův rozvoj (5)

Výpočet matice V:

- vycentrovat data:

$$\mu_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$$

- disperzní matice pro trénovací množinu:

$$w_{ij} = w_{ji} = \frac{1}{n}\sum_{k=1}^{n}\left(x_{ki} - \mu_i\right)\left(x_{kj} - \mu_j\right)$$

- vektory definující nejdůležitější příznaky jsou charakteristickými vektory disperzní matice

Karhunen-Loevovův rozvoj (6)

- Charakteristická čísla odpovídají rozptylu nejdůležitějších příznaků
  - prvním sloupcem matice V bude charakteristický vektor odpovídající největšímu charakteristickému číslu, …
  - další sloupce V se přestanou přidávat poté, co lze další charakteristická čísla vzhledem k jejich velikosti zanedbat

Problém:

- volba odpovídajícího počtu charakteristických čísel (p)
- nelze zaručit optimální volbu p vzhledem ke skutečnému významu jednotlivých příznaků

Karhunen-Loevovův rozvoj (7)

Modifikace:

1. **Centrované nejdůležitější příznaky**

   $\mathbf{y} = \mathbf{V}^T(\mathbf{x} - \boldsymbol{\mu})$, kde $\boldsymbol{\mu} = (\mu_1, \ldots)$ je vektor středních hodnot

2. **Normalizované nejdůležitější příznaky**

   $\mathbf{y} = \mathbf{L}^{-1/2}\,\mathbf{V}^T(\mathbf{x} - \boldsymbol{\mu})$, kde $\mathbf{L}$ je matice $p \times p$, prvky diagonály jsou charakteristická čísla odpovídající sloupcům $\mathbf{V}$, ostatní prvky jsou nulové

3. **Normalizace nejdůležitějších příznaků vzhledem k rozptylům**

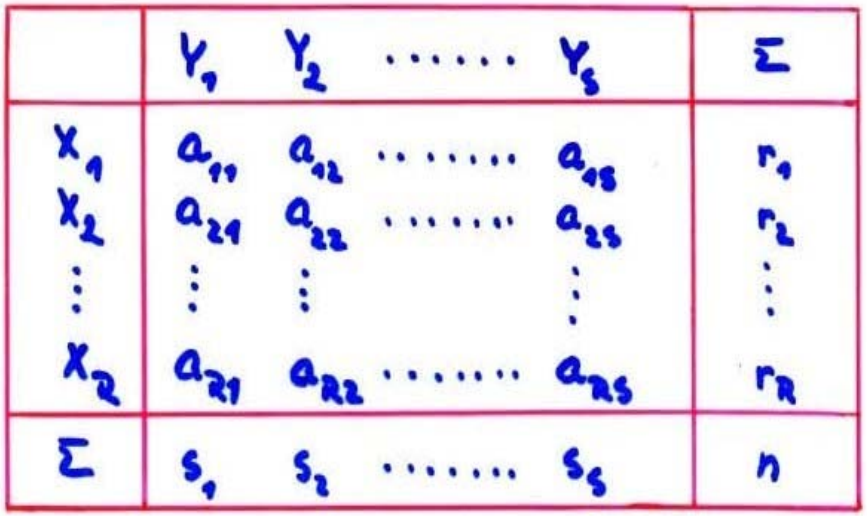
$$w'_{ij} = \frac{w_{ij}}{\sqrt{w_{ii}\, w_{jj}}}$$

Kontingenční tabulky

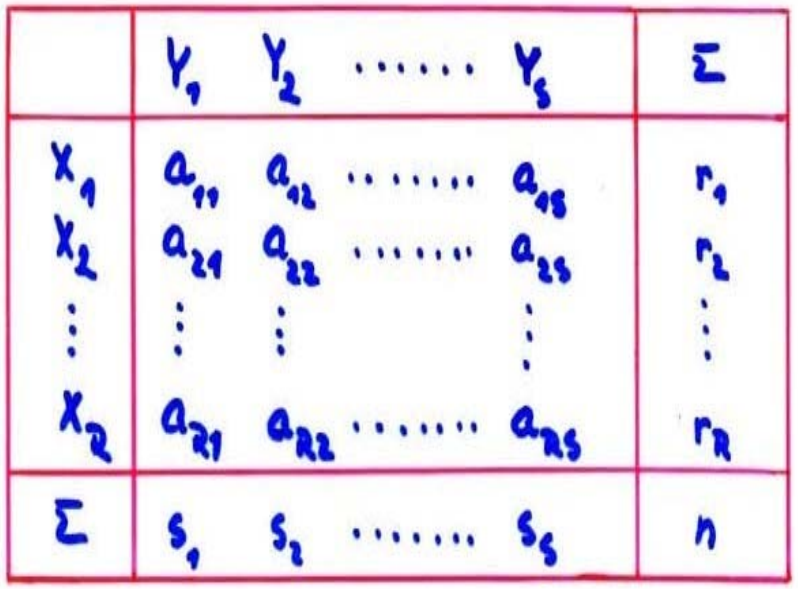
~ vztah mezi dvěma kategoriálními veličinami, např. binárními

Obecná kontingenční tabulka

- Pro ***n*** pozorování s ***R*** hodnotami pro veličinu ***X*** a ***S*** hodnotami pro veličinu ***Y***



Kontingenční tabulky (2)



- a_{kl} … četnost (frekvence) kombinace $(X = X_k) \wedge (Y = Y_l)$
- r_k, s_l … řádkové, sloupcové součty (tzv. marginální hodnoty)
- e_{kl} … očekávaná četnost kombinace $(X = X_k) \wedge (Y = Y_l)$ při nezávislosti X a Y

$$r_k = \sum_{l=1}^{S} a_{kl} \quad ; \quad s_l = \sum_{k=1}^{R} a_{kl} \quad ; \quad n = \sum_{k=1}^{R}\sum_{l=1}^{S} a_{kl} \quad ; \quad e_{kl} = \frac{r_k \cdot s_l}{n}$$

χ^2 - test

- Zjišťování vztahu mezi ***X*** a ***Y***
- Vyhodnocení rozdílu mezi pozorovanými četnostmi jednotlivých kombinací (uvedenými v tabulce) a četnostmi očekávanými při platnosti hypotézy o nezávislosti obou veličin (počítanými z marginálních hodnot)

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

χ^2 – test (2)

- Při platnosti nulové hypotézy nezávislosti veličin ***X*** a ***Y***:

  H_0: $P(X = X_k \wedge Y = Y_l) = P(X = X_k)\, P(Y = Y_l)$; $\forall k,l$

  má χ^2 $(R - 1) . (S - 1)$ stupňů volnosti

- Je-li hodnota χ^2 - statistiky $\geq$ hodnotě χ^2 - rozdělení s příslušným počtem stupňů volnosti na zvolené hladině významnosti α : $\chi^2 \geq \chi^2_{(R-1)(S-1)}$

  zamítne se nulová hypotéza

  => alternativní hypotéza závislosti

χ^2 – test (3)

Příklad: čtyřpolní kontingenční tabulka

PŘÍJEM \ ÚVĚR	ÚVĚR ANO	ÚVĚR NE	Σ
VYSOKÝ PŘÍJEM	50	0	50
NÍZKÝ PŘÍJEM	30	40	70
Σ	80	40	120

χ^2 – test (4)

Příklad (pokračování)**:**

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	23,3	16,7

- Hodnota statistiky χ^2 : 42.857
- Hodnota rozdělení χ^2 s 1 stupněm volnosti je pro hladinu významnosti $\alpha = 0.05$: $\chi^2_{(1)}(0.05) = 3.84$

=> závislost mezi výší příjmu a poskytnutím úvěru

Fisherův test

- χ^2 – test lze použít jen v případě dostatečně velkých četností – pro $(r_k . s_l) / n \geq 5 \quad \forall k, l$
- pro čtyřpolní tabulky lze použít Fisherův test (použitelný pro nízké četnosti)
- Výpočet pravděpodobnosti, že při daných marginálních četnostech r a s má čtyřpolní tabulka skutečné četnosti a_{kl}:

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

Fisherův test (2)

- Pravděpodobnosti p se nasčítají pro různé hodnoty skutečných četností při daných marginálech (předpokládá $a_{11} = \min_{k,l} a_{kl}$):

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

- Je-li $P \leq \alpha$, zamítne se nulová hypotéza o nezávislosti na hladině významnosti α

Regresní analýza

~ určit, jaký vztah má proměnná Y k jedné anebo vícero jiným proměnným $X_1,\ldots,X_n$

Důvody využití:

1. Nákladné měření výstupů => hledáme predikci výstupu na základě snadno získaných vstupů
2. Hodnoty vstupů jsou k dispozici dříve než výstup => potřebujeme pracovat s odhadem výstupu
3. Řízené vstupní hodnoty mohou pomoci správně odhadnout chování odpovídajících výstupů
4. Může existovat kauzální spojitost mezi vstupy a výstupy => tento vztah chceme najít

Regresní analýza (2)

- **Korelační analýza**

  Platí mezi dvěma numerickými veličinami lineární závislost?

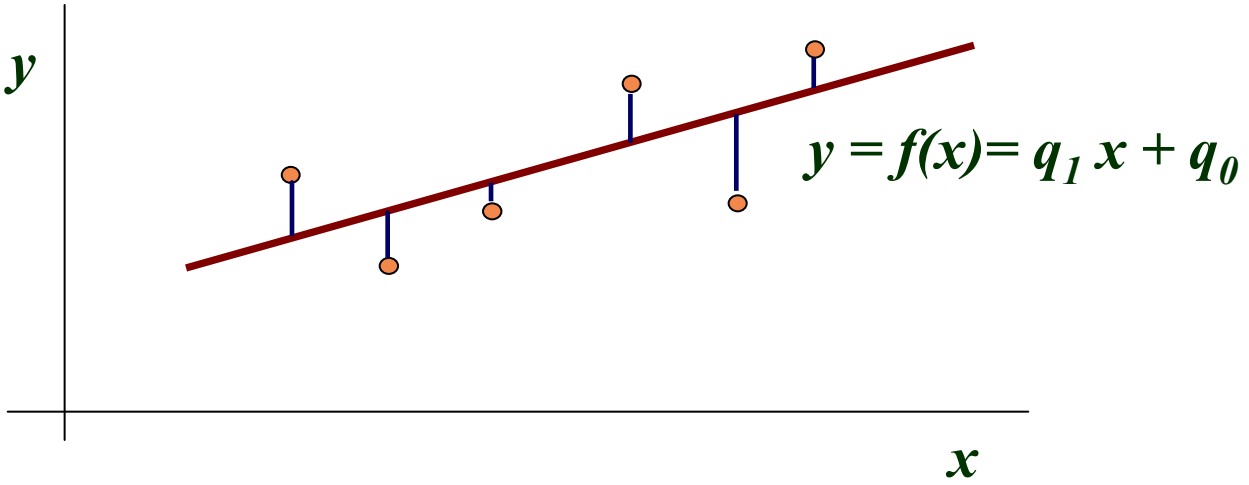
- **Lineární regrese**

  Jaké parametry má lineární závislost mezi dvěma numerickými veličinami?

  → Aproximace pozorovaných hodnot $[x_i, y_i]$; $i = 1, \ldots$ pomocí $y = q_1 x + q_0 + \varepsilon$

Regresní analýza (3)

→ **metodou nejmenších čtverců** (minimalizace rozdílů mezi skutečnou a očekávanou hodnotou)



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

Regresní analýza (4)

→ **metodou nejmenších čtverců** (minimalizace rozdílů mezi skutečnou a očekávanou hodnotou)

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

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

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

→ **obě parciální derivace by měly být rovné nule**

Regresní analýza (5)

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

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

- Pro lineární závislost x a y nalezneme optimální parametry q_0, q_1 vztahu $y = q_1 x + q_0$

Regresní analýza (6)

- Korelační koeficient:

  Posouzení „míry“ lineární závislosti lineární závislost

$$\rho(x,y)=\frac{S_{xy}}{\sqrt{S_x^2 S_y^2}} \quad ; \quad \bar{x}=\frac{\sum_i x_i}{n} \quad ; \quad \bar{y}=\frac{\sum_i y_i}{n}$$

  výběrová kovariance:

$$S_{xy}=\frac{1}{(n-1)}\sum_i (x_i-\bar{x})(y_i-\bar{y})$$

  výběrové rozptyly:

$$S_x^2=\frac{1}{(n-1)}\sum_i (x_i-\bar{x})^2$$

$$S_y^2=\frac{1}{(n-1)}\sum_i (y_i-\bar{y})^2$$

Regresní analýza (7)

Mnohorozměrná regrese:

- **Lineární**

  předpokládáme lineární závislost vysvětlované (závislé) veličiny y na vícero vysvětlujících (nezávislých) veličinách $x_1, x_2, \ldots, x_m$

  → předpoklad pro i – té pozorování:

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

Regresní analýza (7)

Mnohorozměrná regrese:

- **Lineární** (pokračování)

→ maticový zápis: $\vec{y} = X\,\vec{q}\;;\;\; \vec{y} = (y_1, \ldots, y_n)^T\;;\;\; \vec{q} = (q_0, \cdots, q_m)^T$

$$X = \begin{bmatrix} 1 & x_{11} & \ldots & x_{+m} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \ldots & x_{nm} \end{bmatrix}$$

→ řešení $\vec{y} = X\,\vec{q}$ metodou nejmenších čtverců:

$$\vec{q} = \left(X^T X\right)^{-1} X^T \vec{y}$$

Regresní analýza (8)

Mnohorozměrná regrese (pokračování):

- **Nelineární**

  předpokládáme složitější funkční závislost mezi y a $\vec{x}$

  - kvadratickou, exponenciální, …

  - **logistická regrese** (případ nelineární regrese)

    - předpokládáme, že závislá veličina y je kategoriální, např. dvouhodnotová
    - modelujeme pravděpodobnost, že y má konkrétní hodnotu v závislosti na kombinaci hodnot nezávislých veličin $\vec{x}$
    - podmíněná šance: $P(y \mid \vec{x})/(1-P(y \mid \vec{x}))$

Regresní analýza (9)

Mnohorozměrná regrese (pokračování):

- **logistická regrese** (pokračování)
  - Pro ***y*** s hodnotami pouze ***1*** , resp. ***0*** :

$$\ln \frac{P(y=1 \mid x_1, x_2, \ldots, x_m)}{1 - P(y=1 \mid x_1, x_2, \ldots, x_m)} = q_0 + q_1 x_1 + \ldots + q_m x_m$$

resp.

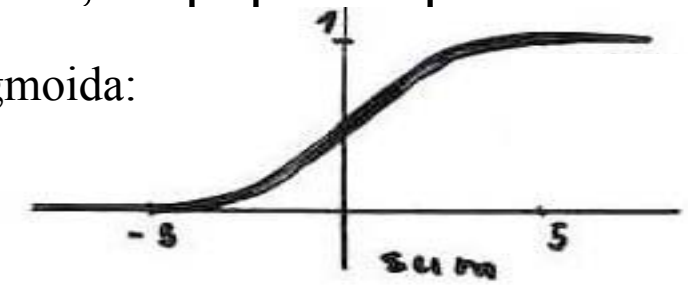
$$P(y=1 \mid x_1, x_2, \ldots, x_m) = \frac{e^{q_0 + \sum_j q_j x_j}}{1 + e^{q_0 + \sum_j q_j x_j}} = \frac{1}{1 + e^{-q_0 - \sum_j q_j x_j}}$$

Regresní analýza (10)

Mnohorozměrná regrese (pokračování):

- **logistická regrese** (pokračování)
  - Odhad šance, resp. pravděpodobnosti hodnoty ***y = 1*** :

Sigmoida:



$$1/(1+\exp(-sum))$$

$$sum = q_0 + \sum_j q_j x_j$$

  - Odhad parametrů modelu **metodou maximální věrohodnosti** (maximalizace ***L***):

$$L = \prod_{i=1}^{n} P(y_i = 1 \mid x_{i,1}, x_{i,2}, \ldots, x_{i,m}) = \prod_{i=1}^{n} \left[\left(\frac{1}{1+e^{-q_0 - \sum_j q_j x_{i,j}}} \right)^{y_i} \cdot \left(\frac{1}{1+e^{q_0 + \sum_j q_j x_{i,j}}} \right)^{1-y_i} \right]$$

Regresní analýza
alternativní odvození (1)

Regresní rovnice:

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n$$

⇨ pro jednotlivé vzory

$$y_j = \alpha + \beta_1 x_{1j} + \beta_2 x_{2j} + \cdots + \beta_n x_{nj} + \varepsilon_j$$

regresní odchylka pro vzor j

Regresní analýza
alternativní odvození (2)

Lineární regrese pro jednu vstupní proměnnou:

- vzory $(x_1, y_1), \ldots, (x_n, y_n); x_i \in X, y_i \in Y$
- regresní rovnice $Y = \alpha + \beta X$

regresní koeficienty

- metoda nejmenších čtverců pro volbu regresních koeficientů
- kvadratická odchylka

$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - y'_i)^2 = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2$$

Regresní analýza
alternativní odvození (3)

$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - y'_i)^2 = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2$$

- Derivace kvadratické odchylky podle α a β:

$$\frac{\partial(SSE)}{\partial \alpha} = -2\sum_{i=1}^{n} (y_i - \alpha - \beta x_i)$$

$$\frac{\partial(SSE)}{\partial \beta} = -2\sum_{i=1}^{n} ((y_i - \alpha - \beta x_i) x_i)$$

- Minimalizace celkové chyby (derivace by měly být rovné 0)

Regresní analýza
alternativní odvození (4)

$$\frac{\partial(SSE)}{\partial\alpha} = -2\sum_{i=1}^{n}(y_i - \alpha - \beta x_i) \qquad \frac{\partial(SSE)}{\partial\beta} = -2\sum_{i=1}^{n}((y_i - \alpha - \beta x_i)x_i)$$

$$n\alpha + \beta\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$

$$\alpha\sum_{i=1}^{n} x_i + \beta\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i$$

$$\alpha + \beta\,\bar{x} = \bar{y} \qquad \text{tedy} \qquad \alpha = \bar{y} - \beta\,\bar{x}$$

$$\left(\bar{y} - \beta\,\bar{x}\right)\sum_{i=1}^{n} x_i + \beta\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i$$

Regresní analýza
alternativní odvození (5)

$$\alpha = \bar{y} - \beta\,\bar{x} \qquad \left(\bar{y} - \beta\,\bar{x}\right)\sum_{i=1}^{n} x_i + \beta\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i$$

$$\beta\left(\sum_{i=1}^{n} x_i^2 - \bar{x}\sum_{i=1}^{n} x_i\right) = \sum_{i=1}^{n} x_i y_i - \bar{y}\sum_{i=1}^{n} x_i$$

$$\beta\sum_{i=1}^{n} x_i\left(x_i - \bar{x}\right) = \sum_{i=1}^{n} x_i\left(y_i - \bar{y}\right) \quad \text{tedy} \qquad \beta = \frac{\sum_{i=1}^{n} x_i\left(y_i - \bar{y}\right)}{\sum_{i=1}^{n} x_i\left(x_i - \bar{x}\right)}$$

- Predikce y pomocí $y = \alpha + \beta x$

Regresní analýza
alternativní odvození (6)

$$\beta = \frac{\sum_{i=1}^{n} x_i \left(y_i - \bar{y}\right)}{\sum_{i=1}^{n} x_i \left(x_i - \bar{x}\right)}$$

- úpravou dostaneme

$$\beta = \frac{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2}$$

$$\alpha = \bar{y} - \beta \bar{x}$$

Regresní analýza
alternativní odvození (7)

Vícerozměrná lineární regrese:

- proměnná Y se modeluje jako lineární funkce vícero predikčních proměnných

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n$$

- maticové vyjádření

$Y = \beta X$ X … rozšířená matice vstupních vzorů

Y … matice výstupů

$\beta = (\beta_0, \beta_1, \ldots, \beta_n); \quad \beta_0 = \alpha$

- kvadratická odchylka $SSE = (Y - \beta X)^T \cdot (Y - \beta X)$

Regresní analýza
alternativní odvození (8)

- optimalizační krok (LMS)

$$\frac{\partial(SSE)}{\partial\beta} = \frac{\partial\left((Y-\beta X)^T(Y-\beta X)\right)}{\partial\beta} = 0$$

$$\Rightarrow \quad \left(X^T \cdot X\right)\beta = X^T \cdot Y$$

- vyjádření regresních koeficientů

$$\beta = \left(X^T \cdot X\right)^{-1}\left(X^T \cdot Y\right)$$

- Vysoké výpočetní nároky při řešení složitých úloh z praxe

  $\Rightarrow$ aproximativní řešení

Diskriminační analýza

~ Klasifikace příkladů do předem zadaných tříd

→ hledání závislosti jedné nominální veličiny (určující příslušnost ke třídě) na dalších ***m*** numerických veličinách

- Předpokládáme, že ke každé třídě (~ hodnotě nominální veličiny) c_t ; $t = 1, \ldots, T$ existuje (**diskriminační**) **funkce** f_t ;

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

$$\Leftrightarrow \; \vec{x} = (x_1, x_2, \ldots, x_m) \text{ patří k } c_t$$

Diskriminační analýza (2)

Lineární diskriminační analýza:

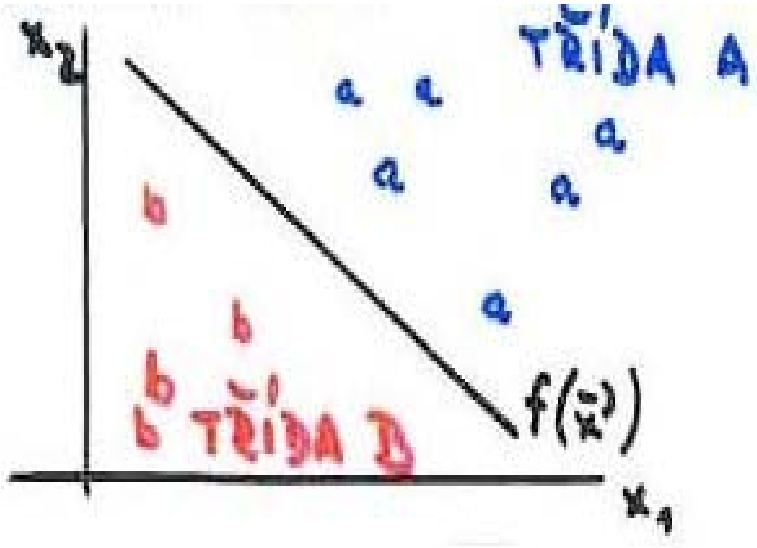
$$f_t = q_{0t} + q_{1t} x_1 + q_{2t} x_2 + \ldots + q_{mt} x_m$$

Diskriminace do dvou tříd

- Místo funkcí f_1 a f_2 můžeme hledat funkci

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

- Příklady se klasifikují podle znaménka $f(\vec{x})$



Diskriminační analýza (3)

Optimální klasifikace ve smyslu minimální chyby

→ diskriminační funkce ≈ podmíněné (aposteriorní) pravděpodobnosti zařazení pozorování $\vec{x}$ do třídy c_t

$$f_t(\vec{x}) = P(c_t \mid \vec{x}) = \frac{P(\vec{x} \mid c_t) \cdot P(c_t)}{\sum_k P(\vec{x} \mid c_k) P(c_k)}$$

→ pro dvě třídy:

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

$$= P(\vec{x} \mid c_1) \cdot P(c_1) - P(\vec{x} \mid c_2) \cdot P(c_2)$$

Diskriminační analýza (4)

→ normální rozdělení (kvadratická diskriminační funkce):

$$f(\vec{x}) = \frac{1}{2} X^T \left(S_1^{-1} - S_2^{-1}\right) X^T + \left(\vec{\mu}_1^T S_1^{-1} - \vec{\mu}_2^T S_2^{-1}\right) X +$$

$$+ \frac{1}{2} \ln \frac{|S_2|}{|S_1|} - \frac{1}{2}\left(\vec{\mu}_1^T S_1^{-1} \vec{\mu}_1 - \vec{\mu}_2^T S_2^{-1} \vec{\mu}_2\right) - \ln \frac{P(C_1)}{P(C_2)}$$

→ stejné kovarianční matice, $\boldsymbol{S_1 = S_2 = S}$:

(lineární diskriminační funkce)

$$f(\vec{x}) = \left(\vec{\mu}_1^T - \vec{\mu}_2^T\right) S^{-1} X - \frac{1}{2}\left(\vec{\mu}_1^T - \vec{\mu}_2^T\right) S^{-1} \cdot$$

$$\cdot \left(\vec{\mu}_1 - \vec{\mu}_2\right) - \ln \frac{P(C_1)}{P(C_2)}$$

Diskriminační analýza (5)

→ jednotkové kovarianční matice, obě třídy stejně pravděpodobné

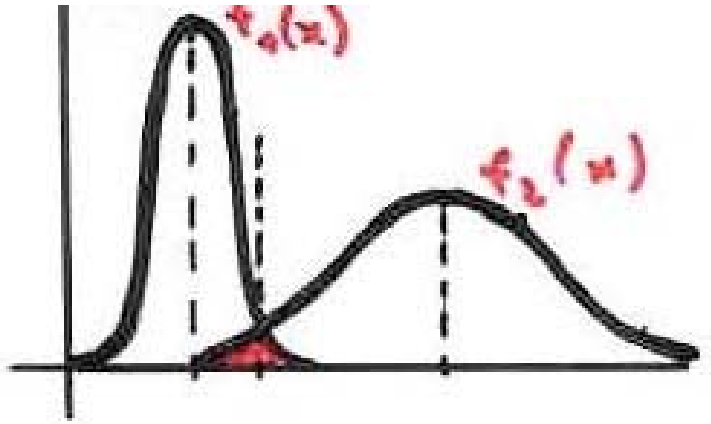
$$f(\vec{x}) = \left(\vec{\mu}_1^T - \vec{\mu}_2^T\right) X - \frac{1}{2}\left(\vec{\mu}_1^T - \vec{\mu}_2^T\right)\left(\vec{\mu}_1 - \vec{\mu}_2\right)$$

→ pro normální rozložení se hledání diskriminační funkce „redukuje“ na odhad středních hodnot $\vec{\mu}_i$ na základě výběrových průměrů a kovariančních matic $\boldsymbol{S_i}$ (z výběrových rozptylů)

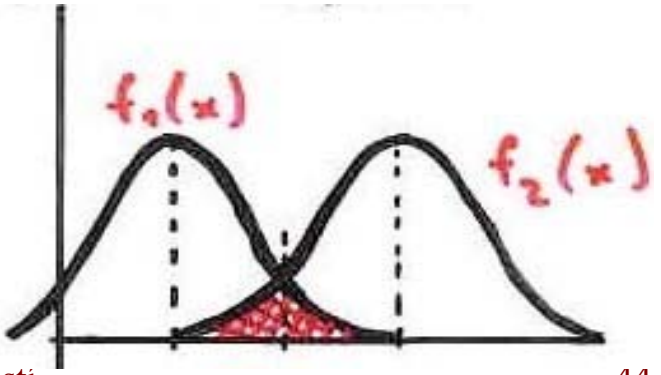
Diskriminační analýza (6)

Příklady:

- Normální rozdělení pravděpodobností $P(x|c_k)\ P(c_k)$ s různými rozptyly



- Normální rozdělení se stejnými rozptyl
  → diskriminace jen podle odhadů středních hodnot



Shluková analýza

- **Lze pozorované vzory rozdělit do skupin (shluků) vzájemně si blízkých vzorů?**

  → **Předpoklad:** umíme měřit vzdálenost mezi vzory

- Každý vzor je charakterizován ***m*** numerickými veličinami
- Vzdálenost mezi dvěma vzory:

$$\vec{x}_1 = (x_{11}, \ldots, x_{1m}) \quad \text{a} \quad \vec{x}_2 = (x_{21}, \ldots, x_{2m})$$

Shluková analýza (2)

1. Hammingova vzdálenost: $d_H(\vec{x}_1, \vec{x}_2) = \sum_{j=1}^{m} |x_{1j} - x_{2j}|$

2. Eukleidovská vzdálenost: $d_E(\vec{x}_1, \vec{x}_2) = \sqrt{\sum_{j=1}^{m} (x_{1j} - x_{2j})^2}$

3. Čebyševova vzdálenost: $d_C(\vec{x}_1, \vec{x}_2) = \max_j |x_{1j} - x_{2j}|$

4. Minkovského metrika (1.-3. jsou jejím speciálním případem):

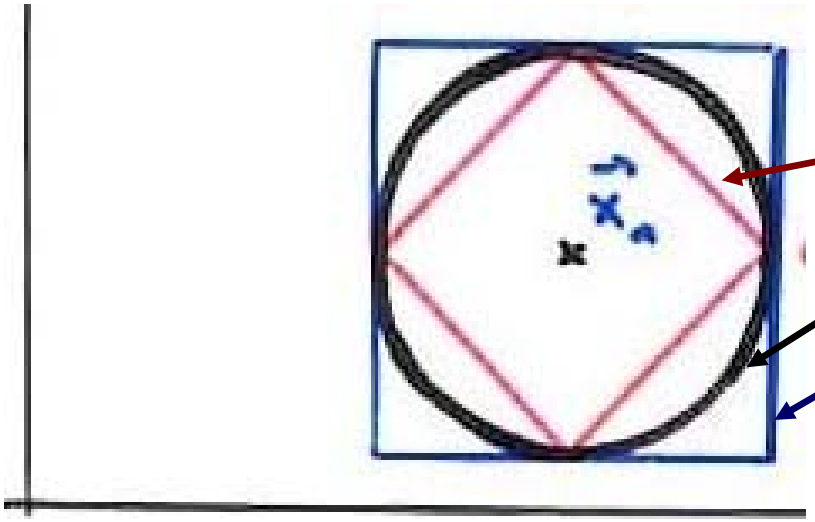
$$L^{(z)}(\vec{x}_1, \vec{x}_2) = \sqrt[z]{\sum_{j=1}^{m} (x_{1j} - x_{2j})^z}$$

Shluková analýza (3)

- $d_H(\vec{x}_1, \vec{x}_2) = L^{(1)}(\vec{x}_1, \vec{x}_2)$

  $d_E(\vec{x}_1, \vec{x}_2) = L^{(2)}(\vec{x}_1, \vec{x}_2)$

  $d_C(\vec{x}_1, \vec{x}_2) = \lim_{z \to \infty} L^{(z)}(\vec{x}_1, \vec{x}_2)$



$d_H(\vec{x}_1, \vec{x}_2) = const$.

$d_E(\vec{x}_1, \vec{x}_2) = const$.

$d_C(\vec{x}_1, \vec{x}_2) = const$.

Shluková analýza (4)

- Volba míry vzdálenosti závisí na měřítku veličin

  → veličiny normovat (~ dělit průměrem, směrodatnou odchylkou, rozpětím (max – min), …)

- předpokládáme stejný rozptyl u všech veličin
- Různý rozptyl veličin → **Mahalanobisova vzdálenost**

$$d_{M^2}(\vec{x}_1, \vec{x}_2) = (\vec{x}_1 - \vec{x}_2)^T S^{-1} (\vec{x}_1 - \vec{x}_2)$$

Shluková analýza (5)

Vzdálenost mezi dvěma shluky U a V :

- Metodou nejbližšího souseda
  - ~ minimum ze vzdáleností mezi jejich prvky

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

- Metodou nejvzdálenějšího souseda
  - ~ maximum ze vzdáleností mezi jejich prvky

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

Shluková analýza (6)

Vzdálenost mezi dvěma shluky *U* a *V* (pokračování):

- Metodou průměrné vzdálenosti ~ průměr ze vzdáleností mezi vzory; ***(n_U*** ~ počet vzorů ve shluku ***U ; n_V*** ~počet vzorů ve shluku ***V)***

$$D(U,V) = \frac{1}{n_U n_V} \sum_{k=1}^{n_U} \sum_{l=1}^{n_V} d(\vec{x}_k, \vec{x}_l) \quad ; \; \vec{x}_k \in U, \; \vec{x}_l \in V$$

- Centroidní metodou ~ vzdálenost mezi středy shluků; ($\overline{\vec{u}}$ ~ střed shluku ***U;*** $\overline{\vec{v}}$ ~ střed shluku ***V)***

$$D\left(U, V\right) = d\left(\overline{\vec{u}}, \overline{\vec{v}}\right)$$

Shluková analýza (7)

Centroid ~ střed shluku

- „Prototyp“ reprezentující daný shluk
- Jeden shluk může být reprezentován i vícero centroidy
  - V závislosti na tvaru shluku a zvolené metrice pro výpočet vzdálenosti
- Shlukování metodou ***k***-středů

Shluková analýza (8)

Shlukování metodou *k*-středů:

1. Náhodně zvol rozklad do k shluků
2. Urči centroidy pro všechny shluky v aktuálním rozkladu
3. Pro každý vzor $\vec{x}$
   1. Urči vzdálenosti $d(\vec{x}, c_k)$ ($1 \le k \le K$; c_k ~ centroid k-tého shluku)
   2. Nechť $d(\vec{x}_1, \vec{c}_l) = \min_k d(\vec{x}_1, \vec{c}_k)$
   3. Není-li $\vec{x}$ součástí shluku l (k jehož centroidu $\vec{c}_l$ má nejblíž), přesuň $\vec{x}$ do shluku 1
4. Došlo-li k nějakému přesunu, potom jdi na 2, jinak KONEC

Shluková analýza (9)

Shlukování metodou *k*-středů:

- Varianty algoritmu:
  - Při počátečním rozkladu prohlásit prvních ***k*** vzorů za centroidy (odpadne Krok 2)
  - Aktualizace centroidů po každém přesunu (v cyklu Kroku 3)
- Shluky jsou následně reprezentovány svými centroidy

Shluková analýza (10)

Algoritmus hierarchického shlukování:

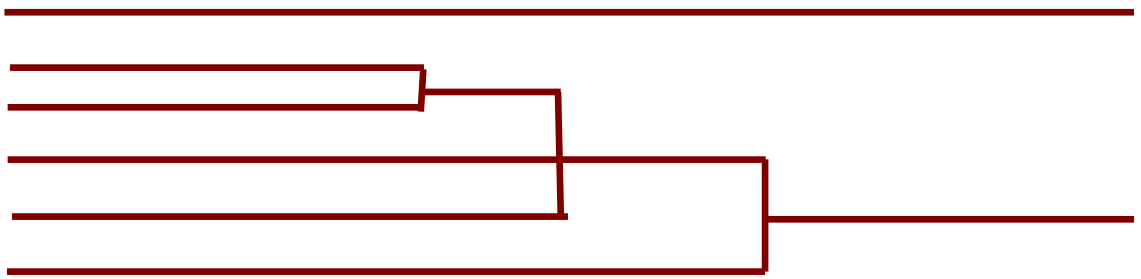
~ metodou „zdola nahoru“

- Inicializace:
  1. Urči vzájemné vzdálenosti mezi všemi vzory
  2. Zařaď každý vzor do samostatného shluku
- Hlavní cyklus:
  1. Dokud je více než jeden shluk
     1. Najdi dva navzájem nejbližší shluky a spoj je
     2. Spočítej pro tento nový shluk jeho vzdálenost od ostatních shluků

Shluková analýza (11)

Algoritmus hierarchického shlukování (pokračování):

- **dendrogram**
  - ~ ukazuje (zleva doprava) postupné spojování shluků
  - ~ optimální počet shluků není předem znám



Vektorová kvantizace: Algoritmus LVQ

Krok 1: Inicializace všech váhových vektorů $\vec{w}_j(0)$

Inicializace parametru učení $\mu(0)$ a nastavení $k = 0$

Krok 2: Otestuj ukončovací podmínku:

IF FALSE => CONTINUE

IF TRUE => QUIT

Krok 3: Pro každý trénovací vzor $\vec{x}_i$ proveď Kroky 4 a 5

Krok 4: Urči index váhového vektoru ($j = q$) tak, aby

$$\min_j \left\| \vec{x}_i - \vec{w}_j(k) \right\|_2^2$$

(použij euklidovskou vzdálenost, $\vec{w}_q(k)$ je váhový vektor s minimální vzdáleností

Vektorová kvantizace: Algoritmus LVQ (2)

Krok 5: Aktualizuj příslušný váhový vektor $\vec{w}_q(k)$ podle:

$$IF \quad C_{\vec{w}_q} = C_{\vec{x}_i} \Rightarrow \vec{w}_q(k+1) = \vec{w}_q(k) + \mu(k)\left[\vec{x}_i - \vec{w}_q(k)\right]$$

$$IF \quad C_{\vec{w}_q} \neq C_{\vec{x}_i} \Rightarrow \vec{w}_q(k+1) = \vec{w}_q(k) - \mu(k)\left[\vec{x}_i - \vec{w}_q(k)\right]$$

Krok 6: Nastav $k \leftarrow k+1$

Sniž parametr učení, např. podle:

$$\mu(k) = \mu(k-1)/(k+1) \qquad (k > 0)$$

Přejdi ke Kroku 2

Vektorová kvantizace: Algoritmus LVQ (3)

MATLAB: Funkce pro LVQ1

```
Function W = lvq1(X,CX,m,mu,maxiter)
% W = lvq1(X,CX,m,mu,maxiter) počítá váhovou matici
%     pro vektorovou kvantizaci LVQ1
% X:  je matice vstupů (každý sloupec odpovídá
%     vstupnímu vektoru
% CX: je řádkový vektor „skalárních“ tříd
%     odpovídajících sloupcovým vektorům z X
% m:  počet různých tříd
% mu: počáteční parametr učení
% maxiter:  maximální počet iterací
N = size(X,2);
```

Vektorová kvantizace: Algoritmus LVQ (4)

MATLAB: **Funkce pro LVQ1** (pokračování)

```
% inicializace váhových vektorů podle prvních m vektorů
% z trénovací množiny (musí obsahovat vzory ze všech tříd)
W = X(:,1:m);
CW = CX(1:m);          % třídy pro váhové vektory
snorm = zeros(1,m);
niter = 1;
while niter <= maxiter
  if niter == 1
     for i = m+1:N
        for j = 1:m
           snorm(1,j) = norm(X(:,i) - W(:,j))^2;
        end
        [mind,index] = min(snorm);
```

Vektorová kvantizace: Algoritmus LVQ (5)

MATLAB: **Funkce pro LVQ1** (pokračování)

```
        if CX(i) == CW(index)
           W(:,index) = W(:,index) +
                                mu*(X(:,i)-W(:,index));
         else
           W(:,index) = W(:,index) -
                                mu*(X(:,i)-W(:,index));
        end
      end
    else
       for i = 1:N
          for j = 1:m
             snorm(1,j) = norm(X(:,i)-W(:,j))^2;
          end
          mind,index] = min(snorm);
```

Vektorová kvantizace: Algoritmus LVQ (6)

MATLAB: **Funkce pro LVQ1** (pokračování)

```
            if CX(i) == CW(index)
              W(:,index) = W(:,index) + (mu/niter)*
                                        (X(:,i)-W(:,index));
             else
                W(:,index) = W(:,index) - (mu/niter)*
                                          (X(:,i)-W(:,index));
            end
        end
    end
    niter = niter + 1;
end
```