Přednáška 4 – Náhodné vektory a regresní analýza

Náhodný vektor $[x, y]$ – sdružené rozdělení

$$\underbrace{f(x,y)}_{\text{sdružené}} = \underbrace{f(x|y)}_{\text{podmíněné}} \cdot \underbrace{f(y)}_{\text{marginální}} = \overbrace{f(y|x)f(x)}^{\text{totéž obráceně}}$$

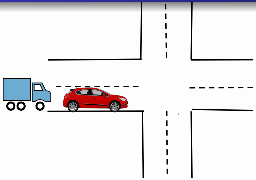
- x, y – spojité nebo diskrétní
- náhodný vektor – vektor N.V. x, y
- vztah mezi x, y

- Sdružené rozdělení $f(x, y)$ popisuje obě veličiny x a y najednou
- Podmíněné rozdělení $f(x|y)$ – chování veličiny x za podmínky znalosti y
- Marginální rozdělení $f(y)$ – informace pouze o veličině y

Výpočet podmíněných rozdělení

$$\underbrace{f(x|y)}_{\text{podmíněné}} = \frac{\overbrace{f(x,y)}^{\text{sdružené}}}{\underbrace{f(y)}_{\text{marginální}}} \qquad \underbrace{f(y|x)}_{\text{podmíněné}} = \frac{\overbrace{f(x,y)}^{\text{sdružené}}}{\underbrace{f(x)}_{\text{marginální}}}$$

Příklad pro diskrétní náhodné veličiny x, y



Data:

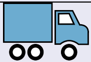
			
	200	250	100
	110	150	190

Křižovatka

$x \in \{$  $= 1,$  $= 2\}$

$y \in \{$  $=1,$  $=2,$  $=3\}$

Sdružené $f(x, y)$, $\sum_{i,j}^{n,m} p_{ij} = 1$

$x \backslash y$			
	0.2	0.25	0.1
	0.11	0.15	0.19

$\Rightarrow$

Marginální $f(x)$, $\sum_{i=1}^{n} p_i = 1$

$f(x)$
0.2+0.25+0.1=0.55
0.11+0.15+0.19=0.45

$\Downarrow$

Marginální $f(y)$, $\sum_{j=1}^{m} p_j = 1$

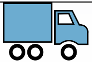
$f(y)$	0.2+0.11=0.31	0.4	0.29

$f(\text{auto}|\text{odbočení})$?

$f(\text{odbočení}|\text{auto})$?

Výpočet podmíněných rozdělení

$f(y|x) = \dfrac{f(x,y)}{f(x)} =$

$x \backslash y$			
	0.2	0.25	0.1
	0.11	0.15	0.19

:

$f(x)$
0.55
0.45

=

f(odbočení|auto)				
	$\frac{0.2}{0.55} = 0.36$	$\frac{0.25}{0.55} = 0.46$	0.18	$\Rightarrow$ 1
	$\frac{0.11}{0.45} = 0.25$	$\frac{0.15}{0.45} = 0.33$	0.42	$\Rightarrow$ 1

$f(x|y) = \dfrac{f(x,y)}{f(y)} =$

f(auto|odbočení)			
	$\frac{0.2}{0.31} = 0.65$	$\frac{0.25}{0.4} = 0.62$	0.34
	$\frac{0.11}{0.31} = 0.35$	$\frac{0.15}{0.4} = 0.38$	0.66
	$\Downarrow$ 1	$\Downarrow$ 1	$\Downarrow$ 1

Příklad pro spojité náhodné veličiny

Sdružená hp: $\quad f(x,y) = 6x^2y, \;$ pro $\; x, y \in (0,\,1)\,,$

Marginální hp: $\quad f(x) = \int_0^1 f(x,y)dy = \int_0^1 6x^2y dy = \left[6x^2\frac{y^2}{2}\right]_0^1 = 3x^2$

Marginální hp: $\quad f(y) = \int_0^1 f(x,y)dx = \int_0^1 6x^2y dx = \left[6\frac{x^3}{3}y\right]_0^1 = 2y$

Podmíněná hp: $\quad f(x|y) = \frac{f(x,y)}{f(y)} = \frac{6x^2y}{2y} = 3x^2$

Podmíněná hp: $\quad f(y|x) = \frac{f(x,y)}{f(x)} = \frac{6x^2y}{3x^2} = 2y$

Pokud $f(x,y) = f(x)f(y)$, veličiny jsou nezávislé

$$f(x)f(y) = 3x^2 2y = 6x^2y = f(x,y)$$

Porovnejte: $f(x,y) = f(y|\cancel{x})f(x) = f(y)f(x)$

Kovariance

– charakteristika dvou náhodných veličin x a y, která vypovídá o jejích vazbě

Kovariance diskrétních náhodných veličin

$$C[x,y] = \sum_i \sum_j (x_i - E[x])(y_j - E[y]) f(x_i, y_j)$$

Kovariance spojitých náhodných veličin

$$C[x,y] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - E[x])(y - E[y]) f(x,y)\, dx\, dy$$

- Přímá závislost veličin
  $C[x,y] > 0$
- Nepřímá závislost veličin
  $C[x,y] < 0$
- Nezávislé veličiny
  $C[x,y] = 0$

Kovarianční matice

$$cov[x,y] = \begin{bmatrix} D[x] & C[x,y] \\ C[x,y] & D[y] \end{bmatrix}$$

Korelace – míra vztahu mezi dvěma veličinami (síla a směr)

- Korelační koeficient

Pearsonův korelační koeficient

$$\rho_p = \frac{C[x, y]}{\sigma_x \sigma_y}$$

lineární závislost, normální data

Spearmanův koeficient pořadové korelace

$$\rho_s = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n(n^2 - 1)}, \quad d_i - \text{rozdíl pořadí}$$

lineární/nelineární závislost, data bez normality

- $\rho \in (-1, 1)$
- $0 < \rho \leq 1$ – přímá závislost (pozitivní korelace)
- $-1 \leq \rho < 0$ – nepřímá závislost (negativní korelace)
- pro nekorelované veličiny $\rho = 0$

Poznámka

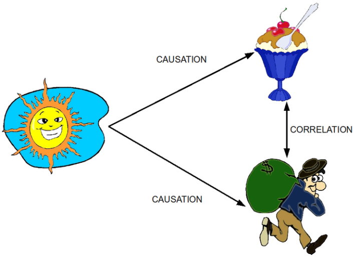
- Korelace $\neq$ kauzalita – musí existovat kauzální vztah

Příklad:

x – prodej zmrzliny, y – kriminalita

$\rho_{xy} > 0$ nebo $r_s > 0$

- vliv – vyšší letní teploty, příliv turistů
- pečlivý výběr veličin



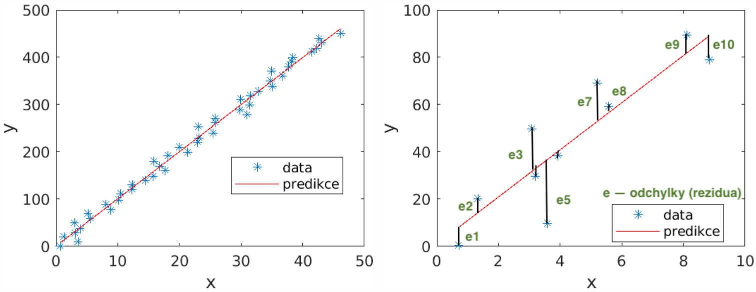
zdroj: wikipedia

Regresní analýza – popis vztahu mezi korelovanými x a y

Lineární regrese – metoda proložení naměřených hodnot přímkou

$y = b_0 + b_1 x, \qquad b_0,\ b_1$ – regresní koeficienty

Příklad:  y – spotřeba paliva, x – poloha plynového pedálu



- přímka neprochází všemi body – odchylky (rezidua) e
- minimální odchylky
- optimální regresní přímka – metoda nejmenších čtverců

Metoda nejmenších čtverců – b_0, b_1

Pro data $[x_i, y_i]_{i=1}^N$

- $e_i = y_i - b_0 - b_1 x_i$
- $\sum_{i=1}^N e_i^2 \to \min$

Odhad regresních koeficientů

$$\hat{b}_1 = \frac{\sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^N (x_i - \bar{x})^2}, \qquad \hat{b}_0 = \bar{y} - \hat{b}_1 \bar{x}$$

Předpověď (predikce)

$\hat{y}_i = \hat{b}_0 + \hat{b}_1 x_i$

$\hat{y}_i$ – hodnoty na přímce

Metoda nejmenších čtverců v maticovém tvaru

$$\underbrace{\begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_N \end{bmatrix}}_{Y} = \underbrace{\begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \dots & \dots \\ 1 & x_N \end{bmatrix}}_{X} \underbrace{\begin{bmatrix} b_0 \\ b_1 \end{bmatrix}}_{b} + \underbrace{\begin{bmatrix} e_1 \\ e_2 \\ \dots \\ e_N \end{bmatrix}}_{E}, \qquad Y = Xb + E$$

Odhad koeficientů: $$\underbrace{\begin{bmatrix} \hat{b}_0 \\ \hat{b}_1 \end{bmatrix}}_{\hat{b}} = (X'X)^{-1} X' Y$$

Vícenásobná lineární regrese – více nezávislých veličin x

$$y = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + \ldots + b_n x_n$$

$x_1, x_2, \ldots, x_n$ – vzájemně nezávislé

Příklad:

x_1 – poloha plynového pedálu

x_2 – rychlost

x_3 – sklon vozovky

y – spotřeba paliva

Odhad regresních koeficientů – maticový tvar

$$\underbrace{\begin{bmatrix} y_1 \\ y_2 \\ \ldots \\ y_N \end{bmatrix}}_{Y} = \underbrace{\begin{bmatrix} 1 & x_{1;1} & x_{2;1} & x_{3;1} \\ 1 & x_{1;2} & x_{2;2} & x_{3;2} \\ \ldots & \ldots & \ldots & \ldots \\ 1 & x_{1;N} & x_{2;N} & x_{3;N} \end{bmatrix}}_{X} \underbrace{\begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ b_3 \end{bmatrix}}_{b} + \underbrace{\begin{bmatrix} e_1 \\ e_2 \\ \ldots \\ e_N \end{bmatrix}}_{E}$$

Predikce

$$\hat{Y} = X\hat{b}$$

Nelineární regresní metody – polynomiální regrese

Polynomiální regrese – proložení dat křivkou

$$y = b_0 + b_1 x + b_2 x^2 + b_3 x^3 + \ldots + b_n x^n$$

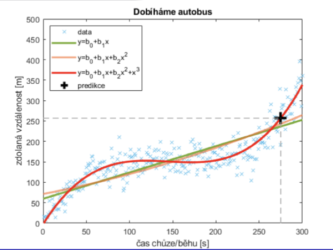
Odhad regresních koeficientů – maticový tvar

$$\underbrace{\begin{bmatrix} y_1 \\ y_2 \\ \ldots \\ y_N \end{bmatrix}}_{Y} = \underbrace{\begin{bmatrix} 1 & x_1 & x_1^2 & x_1^3 \\ 1 & x_2 & x_2^2 & x_2^3 \\ \ldots & \ldots & \ldots & \ldots \\ 1 & x_N & x_N^2 & x_N^3 \end{bmatrix}}_{X} \underbrace{\begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ b_3 \end{bmatrix}}_{b} + \underbrace{\begin{bmatrix} e_1 \\ e_2 \\ \ldots \\ e_N \end{bmatrix}}_{E}$$

Predikce

$$\hat{Y} = X\hat{b}$$

Příklad: x – čas chůze/běhu, y – vzdálenost



Nelineární regresní metody – logistická regrese

Logistická regrese – klasifikace dat

$$\ln\left(\frac{P_1}{P_0}\right) = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_n x_n$$

P_1, P_0 – pravděpodobnost tříd pro klasifikaci

Příklad: úsek silnice

$y \in \{0, 1\}$ – dopravní prostředek (auto, MHD)

x_1 – náklady na cestu (spojitá)

x_2 – příjem (spojitá)

x_3 – dostupnost cyklostezek (diskrétní)

x_4 – vzdálenost od zastávky (spojitá)

Odhad a predikce v inverzní formě

Metoda maximální věrohodnosti:

Věrohodnostní funkce

$$L_t(\theta) = \prod_{t=1}^{N} \frac{e^{y_t z_t}}{1+e^{z_t}}$$

$$\hat{b} = \arg\max_b \ln L_t(b)$$

Predikce třídy:

$$P(y=1) = \frac{e^{x\hat{b}}}{1+e^{x\hat{b}}}$$