Přednáška 4 - náhodné vektory a regresní analýza

Náhodný vektor

V předchozích týdnech jsme se seznámili s dvěma druhy náhodné veličiny - diskrétní a spojitou. Doposud jsme pozorovali pouze jednu náhodnou veličinu. Dnes se podíváme na případ, když máme dvě (nebo více) náhodných veličin. Náhodné veličiny x a y (necháme pro jednoduchost dvě) spolu tvoří

$$\underline{\text{náhodný vektor}} \quad [x, y],$$

při práci s kterým musíme brát v úvahu vztah mezi veličinami x a y.

Náhodný vektor má sdružené rozdělení, které můžeme rozložit na součin podmíněného a marginálního rozdělení:

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

kde

- sdružené rozdělení $f(x, y)$ popisuje obě veličiny x a y najednou.
- podmíněné rozdělení $f(x|y)$ vypráví o chování veličiny x za podmínky znalosti y, například, jsme ji změřili nebo máme její odhad. Veličina, která je za svislítkem, je v podmínce.
- marginální rozdělení $f(y)$ poskytuje informaci o veličině y, kdybychom o veličině x nic nevěděli.

Příklad pro diskrétní náhodné veličiny

Pozorujeme odbočení vozidel na křižovatce. Máme dvě náhodné veličiny: $x \in$ {nákladní auto=1, osobní auto=2}, $y \in$ {doleva=1, doprava=2, rovně=3}. Je dáno sdružené rozdělení $f(x, y)$ ve tvaru tabulky

x \ y	doleva=1	doprava=2	rovně=3
náklaďák= 1	0.2	0.25	0.1
osobák= 2	0.11	0.15	0.19

kde každá pravděpodobnost odpovídá kombinaci hodnot x v řádcích a y ve sloupcích. Např., s pravděpodobností 0.2 $x = 1$ a zároveň $y = 1$, tj náklaďák odbočí doleva, s pravděpodobností 0.25 $x = 1$ a $y = 2$, tj náklaďák odbočí doprava a s 0.1 pojede rovně. Podobně, s pravděpodobností 0.11 osobní auto odbočí doleva a tak dál. Všimněme si, že součet pravděpodobností v celém sdruženém rozdělení se rovná 1.

Pokud máme sdružené rozdělení $f(x, y)$, můžeme z něj spočítat marginální $f(x)$ a $f(y)$ takto:

x \ y	doleva=1	doprava=2	rovně=3
náklaďák= 1	0.2	0.25	0.1
osobák= 2	0.11	0.15	0.19

$\Rightarrow$

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

$\Downarrow$

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

Je vidět, že marginální $f(x)$ je sloupec, kde s pravděpodobností 0.55 to bude náklaďák, aniž bychom něco věděli o tom, kam odbočí, a s pravděpodobností 0.45 to bude osobní auto. Součet pravděpodobností se rovná 1.

Nápodobně, marginální rozdělení $f(y)$ je řádek, obsahující pravděpodobnosti směrů odbočení bez znalosti druhů vozidla. Součet pravděpodobností je zase 1.

Pomocí marginálních rozdělení dokážeme spočítat podmíněné. To by nás zajímalo nejvíc – umožní nám pracovat s x za podmínky, že y jsme už změřili, nebo naopak. Z uvedených vzorců je vidět, že

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

což znamená, že musíme každou pravděpodobnost sdruženého rozdělení ve sloupcích vydělit pravděpodobnostmi marginálního rozdělení $f(y)$ následovně:

x \ y	doleva=1	doprava=2	rovně=3
nákladák= 1	0.2	0.25	0.1
osobák= 2	0.11	0.15	0.19

$=$

x \ y	doleva=1	doprava=2	rovně=3
nákladák= 1	$\frac{0.2}{0.31} = 0.65$	$\frac{0.25}{0.4} = 0.62$	0.34
osobák= 2	$\frac{0.11}{0.31} = 0.35$	$\frac{0.15:}{0.4} = 0.38$	0.66
	$\Downarrow$ 1	$\Downarrow$ 1	$\Downarrow$ 1

$f(y)$	0.31	0.4	0.29

Takovým způsobem vypočtená tabulka vpravo je podmíněné rozdělení $f(x|y)$, které vyjadřuje závislost druhu vozidla na směru odbočení. Např., s pravděpodobností 0.65 to bude náklaďák za podmínky toho, že odbočí doleva a s pravděpodobností 0.35 to bude osobní auto za podmínky odbočení doleva. Všimněme si, že v tomto podmíněném rozdělení součet pravděpodobností se rovná 1 v každém sloupci.

Pokud bychom potřebovali modelovat odbočení v závislosti na druhu vozidla (například pro průzkum, kde nejlépe postavit novou pobočku supermarketu), musíme použít vzorec obráceně:

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

Výpočet je stejný, jenom počítáme v řádcích.

x \ y	doleva=1	doprava=2	rovně=3
nákladák= 1	0.2	0.25	0.1
osobák= 2	0.11	0.15	0.19

$:$

$f(x)$
0.55
0.45

$=$

x \ y	doleva=1	doprava=2	rovně=3
nákladák= 1	$\frac{0.2}{0.55} = 0.36$	$\frac{0.25}{0.55} = 0.46$	0.18
osobák= 2	$\frac{0.11}{0.45} = 0.25$	$\frac{0.15}{0.45} = 0.33$	0.42

Dokážete říct, co znamená pravděpodobnost 0.46 v podmíněném rozdělení $f(y|x)$ (tabulka dole)?

S pravděpodobností 0.46 vozidlo odbočí doprava za podmínky toho, že je nákladák.
Podobně, s pravděpodobností 0.42 pojede rovně za podmínky, že to je osobní auto.

Nezávislost veličin Pokud platí

$$f(x,y) = f(x)f(y),$$

veličiny x a y jsou nezávislé. Toho si můžeme všimnout ve vztahu $f(x,y) = f(x|y)f(y)$ – pokud odebereme y z podmínky za svislítkem, zůstanou tam pouze dvě marginální rozdělení $f(x)$ a $f(y)$. Pro náš příklad:

$$f(x)f(y) = \begin{bmatrix} 0.55 \\ 0.45 \end{bmatrix} \cdot \begin{bmatrix} 0.31 & 0.4 & 0.29 \end{bmatrix} = \begin{bmatrix} 0.1705 & 0.22 & 0.1595 \\ 0.1395 & 0.18 & 0.1305 \end{bmatrix} \neq \underbrace{\begin{bmatrix} 0.2 & 0.25 & 0.1 \\ 0.11 & 0.15 & 0.19 \end{bmatrix}}_{\text{původní sdružené}}$$

což znamená, že druh vozidla na křižovatce a směr odbočení nejsou nezávislé.

Přiklad pro sdružené, podmíněné a marginální rozdělení pro spojité náhodné veličiny

Základní vztahy pro sdružené, podmíněné a marginální rozdělení jsou obecné a platí i v případě spojitých náhodných veličin x a y. Rozdíl je v tom, že pro výpočet marginální hustoty pravděpodobnosti (hp) je potřeba integrovat sdružené rozdělení podle druhé veličiny. Například,

$$\text{je daná sdružená hp} \qquad f(x,y) = 6x^2y, \text{ pro } x,y \in (0,1),$$

$$\text{marginální hp} \quad f(x) = \int_0^1 6x^2y\,dy = \left[6x^2\frac{y^2}{2}\right]_0^1 = 3x^2, \qquad \text{marginální hp} \quad f(y) = \int_0^1 6x^2y\,dx = \left[6\frac{x^3}{3}y\right]_0^1 = 2y.$$

Dále počítáme obě podmíněné hp podobně, jako v předchozím příkladě:

$$\text{podmíněná hp} \quad f(x|y) = \frac{f(x,y)}{f(y)} = \frac{6x^2y}{2y} = 3x^2, \qquad \text{podmíněná hp} \quad f(y|x) = \frac{f(x,y)}{f(x)} = \frac{6x^2y}{3x^2} = 2y.$$

Můžeme ukázat, že x a y jsou nezávislé:

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

Kovariance

Kovariance je charakteristika dvou náhodných veličin x a y, která vypovídá o jejích vazbě. Kovariance se počítá

$$\text{pro diskrétní náhodné veličiny} \qquad C[x,y] = \sum_i \sum_j (x_i - E[x])(y_j - E[y])f(x_i,y_j),$$

$$\text{pro spojité náhodné veličiny} \qquad C[x,y] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x - E[x])(y - E[y])f(x,y)\,dx\,dy.$$

Všimněme si, že vzorec výpočtu kovariance je totožný s výpočtem rozptylu

$$\underbrace{D[x] = \sum_{i=1}^{N}(x_i - E[x])^2 f(x_i)}_{\text{pro diskrétní náhodné veličiny}}, \qquad \underbrace{D[x] = \int_{-\infty}^{\infty}(x - E[x])^2 f(x)\,dx}_{\text{pro spojité náhodné veličiny}}.$$

Jediný rozdíl je v tom, že kovarianci počítáme s dvěma veličinami, a rozptyl jenom s jednou. S kovariancí souvisí také pojem kovarianční matice, která pro dvě veličiny x a y má tvar

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

tj na hlavní diagonále se nachází rozptyly veličin a na vedlejší diagonále jsou jejich kovariance.

Je vidět, že pro výpočet kovariance potřebujeme střední hodnoty obou veličin, které spočteme pomocí marginálních rozdělení $f(x)$ a $f(y)$. Pro náš předchozí příklad to je

$$f(x) = \begin{bmatrix} 0.55 \\ 0.45 \end{bmatrix}, \quad E[x] = \sum_{i=1}^{N} x_i f(x_i) = 1 * 0.55 + 2 * 0.45 = 1.45,$$

$$f(y) = \begin{bmatrix} 0.31 & 0.4 & 0.29 \end{bmatrix}, \quad E[y] = \sum_{j=1}^{M} y_j f(y_j) = 1 * 0.31 + 2 * 0.4 + 3 * 0.29 = 1.98.$$

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

$$= (1 - 1.45) * (1 - 1.98) * 0.2 + (1 - 1.45) * (2 - 1.98) * 0.25 + (1 - 1.45) * (3 - 1.98) * 0.1 +$$
$$+ (2 - 1.45) * (1 - 1.98) * 0.11 + (2 - 1.45) * (2 - 1.98) * 0.15 + (2 - 1.45) * (3 - 1.98) * 0.19 = 0.089$$

Platí, že

- $C[x,y] > 0$ znamená přímou závislost veličin, tj hodnoty obou veličin mají společnou tendenci narůstat nebo klesat: $x \Uparrow y \Uparrow$ nebo $x \Downarrow y \Downarrow$
- $C[x,y] < 0$ znamená nepřímou závislost veličin, tedy $x \Uparrow y \Downarrow$ nebo $x \Downarrow y \Uparrow$
- pro nezávislé veličiny $C[x,y] = 0$ (neplatí to obráceně).

Pro náš příklad hodnota kovariance 0.089 naznačuje slabou přímou závislost, kterou můžeme interpretovat takto: nákladák spíš odbočí doleva (např., tam je odbočka na dálnici), osobní auto spíš pojede doprava (např., je tam supermarket) nebo rovně.

Pearsonův korelační koeficient

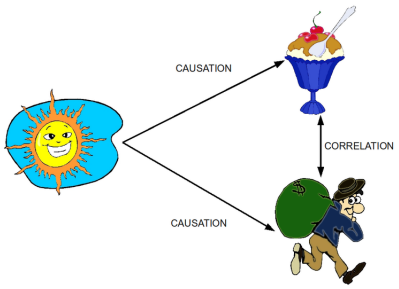
Kovarianci budeme používat pro výpočet Pearsonova korelačního koeficientu

$$\rho_{xy} = \frac{C[x,y]}{\sigma_x \sigma_y}, \quad \rho_{xy} \in \langle -1, 1 \rangle$$

kde σ_x a σ_y jsou směrodatné odchylky obou veličin. Korelační koeficient vyjadřuje míru korelace mezi veličinami na intervalu od -1 do 1. Čím blíž je hodnota korelačního koeficientu k 1, tím vyšší je korelovanost mezi veličinami. Naopak, hodnota korelačního koeficientu blížící se k -1 znamená výraznou nepřímou závislost (antikorelaci) veličin.

Pro nekorelované x a y korelační koeficient $\rho_{xy} = 0$, protože $C[x,y] = 0$. Nicméně, nulový korelační koeficient neznamená, že jsou veličiny x a y nezávislé.

Důležitá poznámka: nenulový korelační koeficient nemusí nutně znamenat, že růst nebo pokles hodnot veličiny x jsou vyvolány chováním veličiny y, nebo naopak. Pro to mezi veličinami musí existovat kauzální vztah. Například, v létě narůstá prodej zmrzliny a kriminalita. Tyto veličiny spolu korelují, ale to neznamená, že vyšší spotřeba zmrzliny vyvolává nárůst kriminality. Obě veličiny obecně souvisí s vyššími letními teplotami a přílivem turistů. V praxi to znamená, že je potřeba pečlivě zvážit, jaké veličiny bereme v úvahu.



(zdroj: wikipedia)

Regresní analýza

Pokud máme naměřená data dvou korelovaných náhodných veličin x a y

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

mohou být vhodná k regresní analýze. Podíváme se na nejčastěji používané regresní metody.

Lineární regrese

Lineární regrese je metoda proložení naměřených hodnot regresní přímkou. Předpokládáme, že závislost mezi veličinami x a y se dá popsat matematickým vztahem

$$y = b_0 + b_1 x,$$

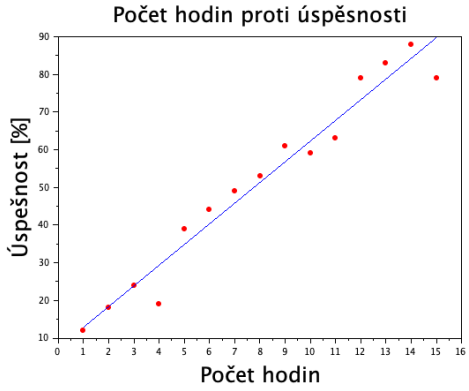
kde b_0 a b_1 jsou regresní koeficienty. V praxi tato závislost nikdy nebude dána takto deterministickým vztahem, proto pro proložení přímky budeme pracovat s naměřenými daty $[x_i, y_i]_{i=1}^{N}$ a počítat s odchylkami e_i v každém i-tem bodě:

$$y_i = b_0 + b_1 x_i + e_i$$

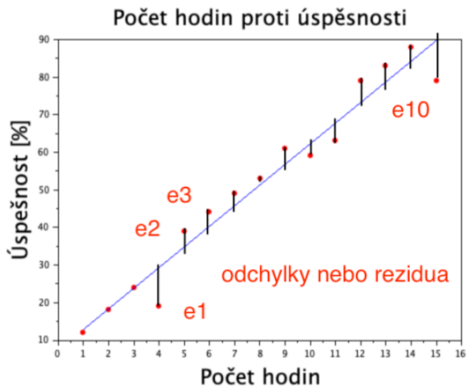
Nejdřív ukážeme, co to znamená graficky. Například, přípravovali jsme se ke zkoušce ze statistiky a dělali jsme záznamy o počtu hodin přípravy a následné úspěšnosti u zkoušky v %. Zajímalo by nás, zda úspěšnost u zkoušky roste lineárně při vyšším počtu hodin přípravy – zkusíme na to použit lineární regresi.
Máme data

$$\begin{aligned} \text{počet hodin přípravy} \quad x &= [1 \quad 2 \quad 3 \quad 4 \quad 5 \quad \ldots], \\ \text{úspěšnost v \%} \quad y &= [12 \quad 18 \quad 30 \quad 49 \quad 79 \quad \ldots]. \end{aligned}$$

Nejdřív si rozmyslíme, která z veličin je závislá. Předpokládáme, že úspěšnost je veličina y závislá na počtu hodin x. Pokud vykreslíme data x a y proti sobě a proložíme přes data přímku (zatím intuitivně), můžeme si všimnout tendence vývoje dat.



Na obrázku je vidět, že přímka neprochází všemi body - ve většině z nich se tvoří odchylky (nebo rezidua) e_i.



Metoda nejmenších čtverců

Optimální regresní přímka prochází body tak, aby součet čtverců odchylek byl minimální. K jejímu nalezení se používá metoda nejmenších čtverců pro odvození vzorců výpočtu regresních koeficientů b_0 a b_1 následovně:

$$\begin{aligned}
\text{z rovnice regrese vyjádříme odchylky:} \quad e_i &= y_i - b_0 - b_1 x_i, \\
\text{minimalizujeme v každém bodě:} \quad \sum_{i=1}^{N} e_i^2 &= \sum_{i=1}^{N} (y_i - b_0 - b_1 x_i)^2 \quad \Rightarrow \quad \min, \\
\text{po derivaci podle } b_0 \text{ a } b_1 \text{ a minimalizaci, odhad regresních koeficientů je:} \quad \hat{b}_1 &= \frac{\sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N} (x_i - \bar{x})^2}, \\
\hat{b}_0 &= \bar{y} - \hat{b}_1 x.
\end{aligned}$$

Dosadíme do rovnice regrese odhady $\hat{b}_1$, $\hat{b}_0$ a data x a můžeme proložit data regresní přímkou:

$$\underbrace{\hat{y}_i}_{\text{hodnoty na přímce}} = \overbrace{\hat{b}_0 + \hat{b}_1}^{\text{odhad}} \underbrace{x_i}_{\text{data}}.$$

To znamená, že hodnoty $\hat{y}_i$, které jsme dostali po dosazení odhadů a dat x do rovnice regrese jsou hodnoty regresní přímky. Po dosazení hodnoty x_i, které není naměřená, výsledkem je předpověď (predikce) $\hat{y}_i$ pro tuto hodnotu.

Metoda nejmenších čtverců v maticovém tvaru

Regresní koeficienty můžeme odhadnout jiným způsobem v maticovém tvaru (výsledky budou totožné). Zapíšeme celou soustavu rovnic pro všechny naměřené hodnoty v maticovém tvaru:

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

Pro tato označení spočteme vektor odhadů koeficientů takto:

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

Lineární vícenásobná regrese

Lineární vícenásobná regrese se vyznačuje větším počtem nezávislých veličin x, tj její rovnice obecně je

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

kde $x_1, x_2, \ldots, x_n$ jsou vzájemně nezávislé. Např., x_1 je počet hodin domácí přípravy, x_2 – počet cvičení, která jsme navštívili, x_3 – počet testů, atd.

Odhad regresních koeficientů se provádí podle maticového způsobu. Následná predikce veličiny se počítá

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

Exponenciální regrese

Exponenciální regrese patří k nelineárním regresním metodám. Závislost mezi veličinami y a x je popsána rovnicí

$$y = b_0 \exp\{b_1 x\}.$$

Tento typ regrese vyžaduje linearizaci pomocí logaritmování

$$\ln(y) = \underbrace{\ln(b_0)}_{B_0} + b_1 x,$$

kde $\ln(y)$ a x se používají jako data. Odhad regresních koeficientů B_o a b_1 se provádí v maticovém tvaru. Pro predikci přepočítáme

$$b_0 = \exp\{B_0\},$$
$$\hat{y} = \exp\{\ln(b_0) + b_1 x\}.$$

Polynomiální regrese

Polynomiální regrese má tvar

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

Odhad regresních koeficientů se provádí podle maticového způsobu.

Dynamická regrese

Dynamická regrese je specifickou regresní metodou, kde veličina y je závislá na své předchozí hodnotě, tj.,

$$y_i = b_0 + b_1 y_{i-1} + \ldots + e_i.$$

Poznámka: pro doplnění teorie na webu jsou k dispozici skripta ze statistiky.