

Pokročilé metody analýzy dat v neurovědách



RNDr. Eva Janoušová
doc. RNDr. Ladislav Dušek, Dr.

Jaro 2016

Blok 3

Podobnosti a vzdálenosti ve vícerozměrném prostoru

Osnova

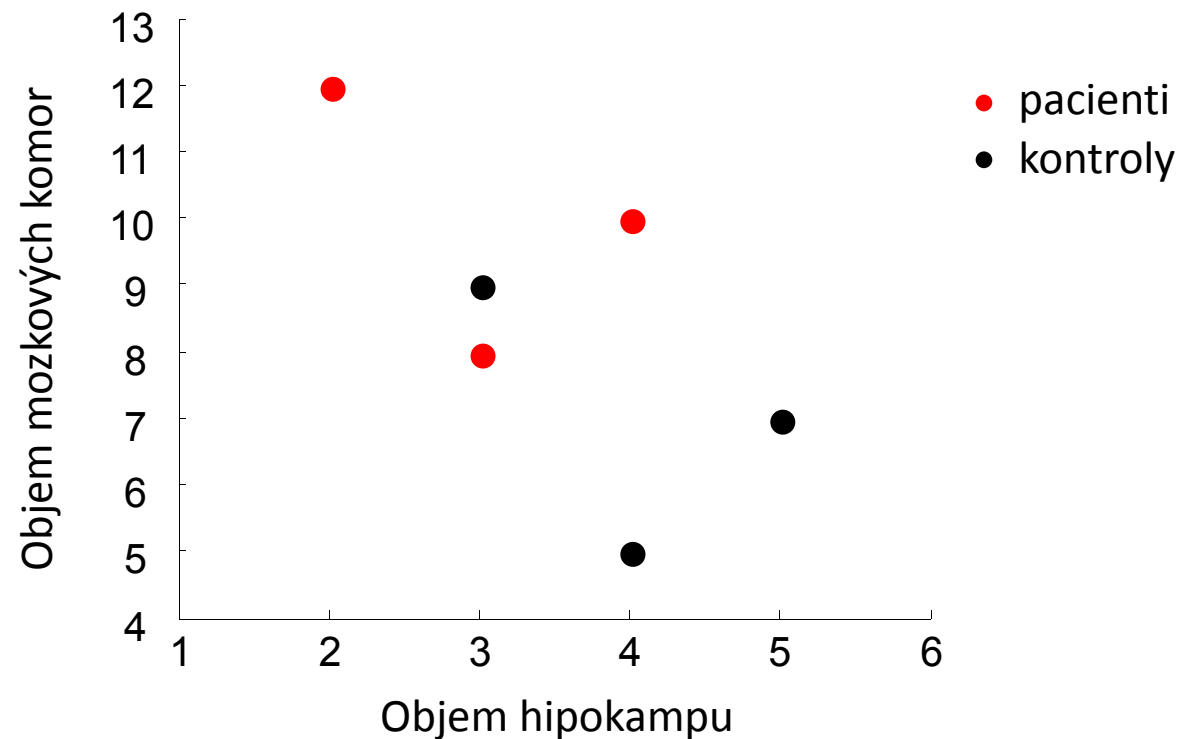
1. Úvod do metrik podobností a vzdáleností
2. Metriky pro určení vzdálenosti mezi dvěma objekty
3. Metriky pro určení podobnosti mezi dvěma objekty
4. Metriky pro určení vzdálenosti mezi dvěma skupinami objektů
5. Asociační matice

Úvod do metrik podobností a vzdáleností

Poznámka

- jednotlivé objekty je možno znázornit pomocí bodu v p -rozměrném prostoru (p je počet proměnných)

$$\mathbf{X}_D = \begin{bmatrix} 2 & 12 \\ 4 & 10 \\ 3 & 8 \end{bmatrix}, \mathbf{X}_H = \begin{bmatrix} 5 & 7 \\ 3 & 9 \\ 4 & 5 \end{bmatrix}$$



Metriky podobnosti vs. metriky vzdálenosti

- **Metriky vzdálenosti** objektu \mathbf{x}_1 od objektu \mathbf{x}_2 – označení: $D(\mathbf{x}_1, \mathbf{x}_2)$
- pozn.: vzdálenost objektu od sebe samého je 0 – tzn. $D(\mathbf{x}_1, \mathbf{x}_1) = 0$
- **Metriky podobnosti** objektu \mathbf{x}_1 od objektu \mathbf{x}_2 – označení: $S(\mathbf{x}_1, \mathbf{x}_2)$
- pozn.: podobnost objektu od sebe samého je maximální hodnota podobnosti pro danou metriku (zpravidla hodnota 1, ale neplatí to vždy)
- Metriky vzdálenosti mohou být různými transformacemi převedeny na metriky podobnosti (a obráceně), např.:

$$S(\mathbf{x}_i, \mathbf{x}_j) = 1 / D(\mathbf{x}_i, \mathbf{x}_j)$$

$$S(\mathbf{x}_i, \mathbf{x}_j) = 1 / (1 + D(\mathbf{x}_i, \mathbf{x}_j))$$

$$S(\mathbf{x}_i, \mathbf{x}_j) = c - D(\mathbf{x}_i, \mathbf{x}_j), \quad c \geq \max D(\mathbf{x}_i, \mathbf{x}_j), \quad \forall i, j$$

Typy měr vzdálenosti (podobnosti)

- podle **typu proměnné** (kvalitativní proměnné, kvantitativní proměnné)
- podle **objektů**, jejichž vztah hodnotíme – obrazy (vektory), množiny obrazů (vektorů)
- **deterministické** (nepravděpodobnosti) vs. **pravděpodobnostní míry**
- výběr konkrétní metriky závisí na:
 - výpočetních nárocích
 - charakteru rozložení dat
 - dosažení optimálních výsledků (klasifikační chyba, ztráta,...)
- obecně bohužel není možné dopředu doporučit vhodnou metriku pro danou situaci
- chybný výběr metriky může vést k chybným závěrům analýzy (stejně jako v klasické statistické analýze výběr nevhodného testu)

Typy metrik a konkrétní příklady

MEZI DVĚMA OBJEKTY

Metriky pro určení **vzdálenosti** mezi 2 objekty s **kvantitativními** proměnnými

Euklidova m., Hammingova (manhattanská) m.,
Minkovského m., Čebyševova m., Mahalanobisova m.,
Canberrská m.

Metriky pro určení **vzdálenosti** mezi 2 objekty s **kvalitativními** proměnnými

Hammingova m.

Metriky pro určení **podobnosti** 2 objektů s **kvantitativními** proměnnými

Skalární součin, m. kosinové podobnosti, Pearsonův
korelační koeficient, Tanimotova m.

Metriky pro určení **podobnosti** 2 objektů s **kvalitativními** proměnnými

Tanimotova m., Jaccardův-Tanimotův a.k., Russelův-
Raovův a.k., Sokalův-Michenerův a.k., Dicův k.,
Rogersův-Tanimotův k., Hamanův k.

MEZI DVĚMA SKUPINAMI OBJEKTŮ

Deterministické metriky pro určení vzdálenosti mezi 2 množinami objektů

Metoda nejbližšího souseda, k nejbližších
sousedů, nejvzdálenějšího souseda, centroidová
metoda, m. průměrné vazby, Wardova metoda

Metriky pro určení vzdálenosti mezi 2 množinami objektů používající jejich pravděpodobnostní charakteristiky

Chernoffova m., Bhattacharyyova m. atd.

Metriky pro určení vzdálenosti mezi dvěma objekty

Typy metrik a konkrétní příklady

MEZI DVĚMA OBJEKTY

Metriky pro určení **vzdálenosti** mezi 2 objekty s **kvantitativními** proměnnými

Euklidova m., Hammingova (manhattanská) m.,
Minkovského m., Čebyševova m., Mahalanobisova m.,
Canberrská m.

Metriky pro určení **vzdálenosti** mezi 2 objekty s **kvalitativními** proměnnými

Hammingova m.

Metriky pro určení **podobnosti** 2 objektů s **kvantitativními** proměnnými

Skalární součin, m. kosinové podobnosti, Pearsonův
korelační koeficient, Tanimotova m.

Metriky pro určení **podobnosti** 2 objektů s **kvalitativními** proměnnými

Tanimotova m., Jaccardův-Tanimotův a.k., Russelův-
Raovův a.k., Sokalův-Michenerův a.k., Dicův k.,
Rogersův-Tanimotův k., Hamanův k.

MEZI DVĚMA SKUPINAMI OBJEKTŮ

Deterministické metriky pro určení vzdálenosti mezi 2 množinami objektů

Metoda nejbližšího souseda, k nejbližších
sousedů, nejvzdálenějšího souseda, centroidová
metoda, m. průměrné vazby, Wardova metoda

Metriky pro určení vzdálenosti mezi 2 množinami objektů používající jejich pravděpodobnostní charakteristiky

Chernoffova m., Bhattacharyyova m. atd.

Nejpoužívanější metriky pro určení vzdálenosti mezi dvěma obrazy s kvantitativními proměnnými

- Euklidova metrika
- Hammingova (manhattanská) metrika
- Minkovského metrika
- Čebyševova metrika
- Mahalanobisova metrika
- Canberrská metrika

Euklidova metrika

- zřejmě nejpoužívanější metrika s velmi názornou geometrickou interpretací

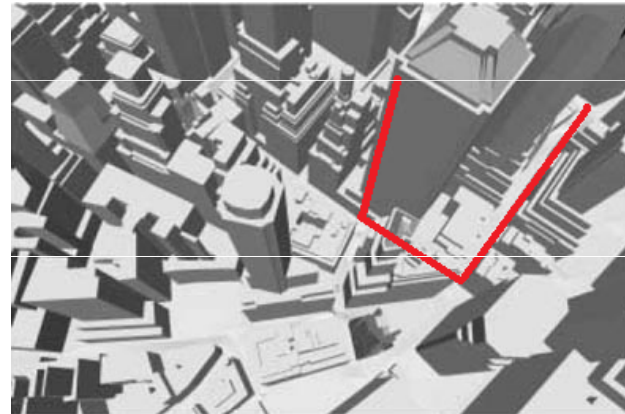
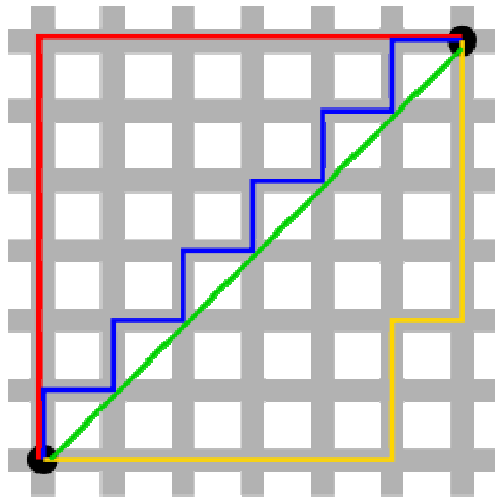
$$D_E(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{\sum_{i=1}^n (x_{1i} - x_{2i})^2}$$

- geometrickým místem bodů s toutéž Euklidovou vzdáleností od daného bodu je povrch hyperkoule (ve dvourozměrném prostoru kružnice)
- dává větší důraz na větší rozdíly mezi souřadnicemi
žádoucí nebo nežádoucí?
- občas se používá čtverec euklidovské vzdálenosti, protože se lépe počítá než euklidovská vzdálenost (není to ale pravá metrika vzdálenosti)

Hammingova (manhattanská) metrika

- v AI názvy: Manhattan distance, city-block distance, taxi driver distance

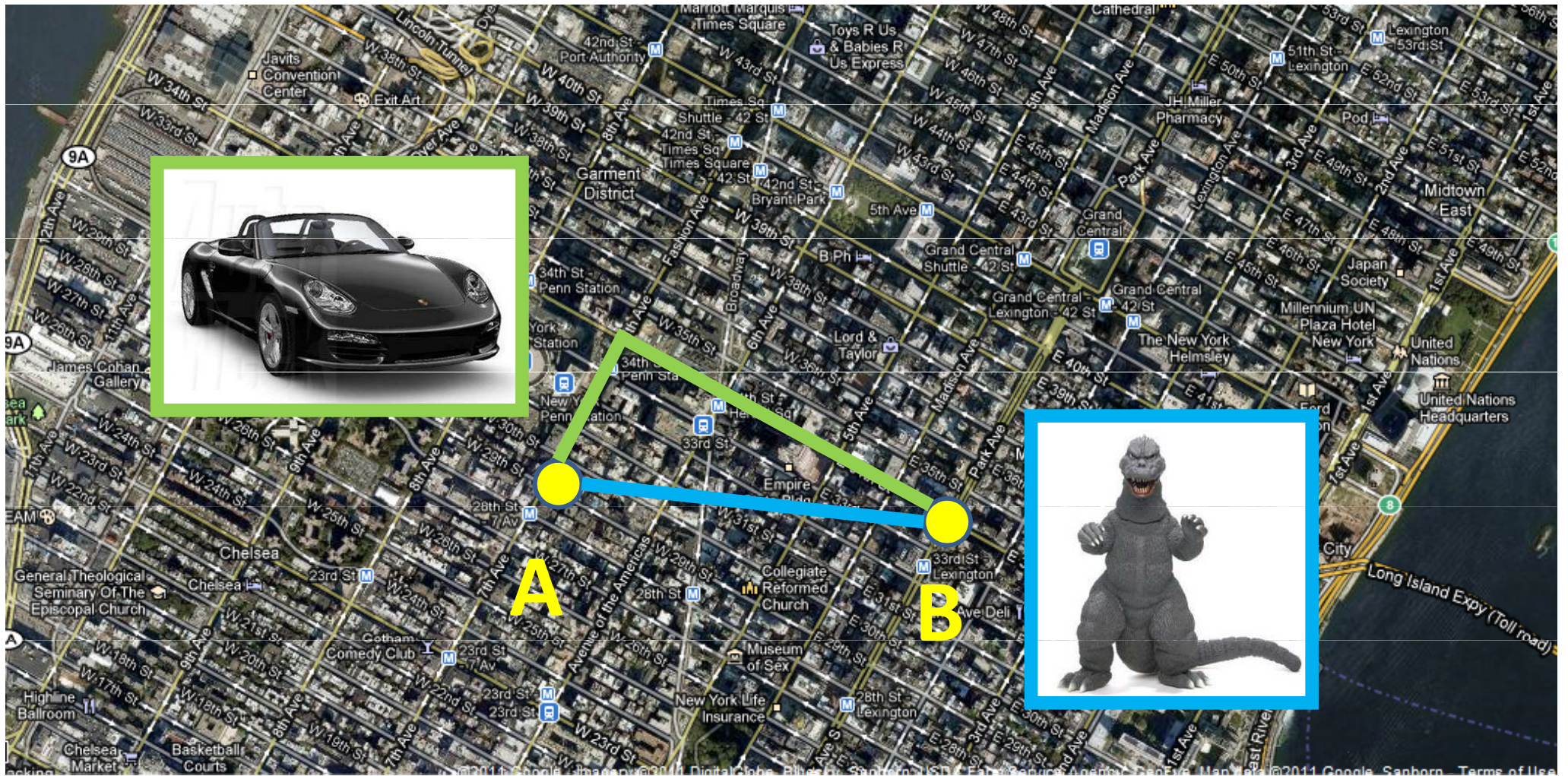
$$D_H(\mathbf{x}_1, \mathbf{x}_2) = \sum_{i=1}^n |x_{1i} - x_{2i}|$$



- nižší výpočetní nároky než Euklidova metrika → použití v úlohách s vysokou výpočetní náročností

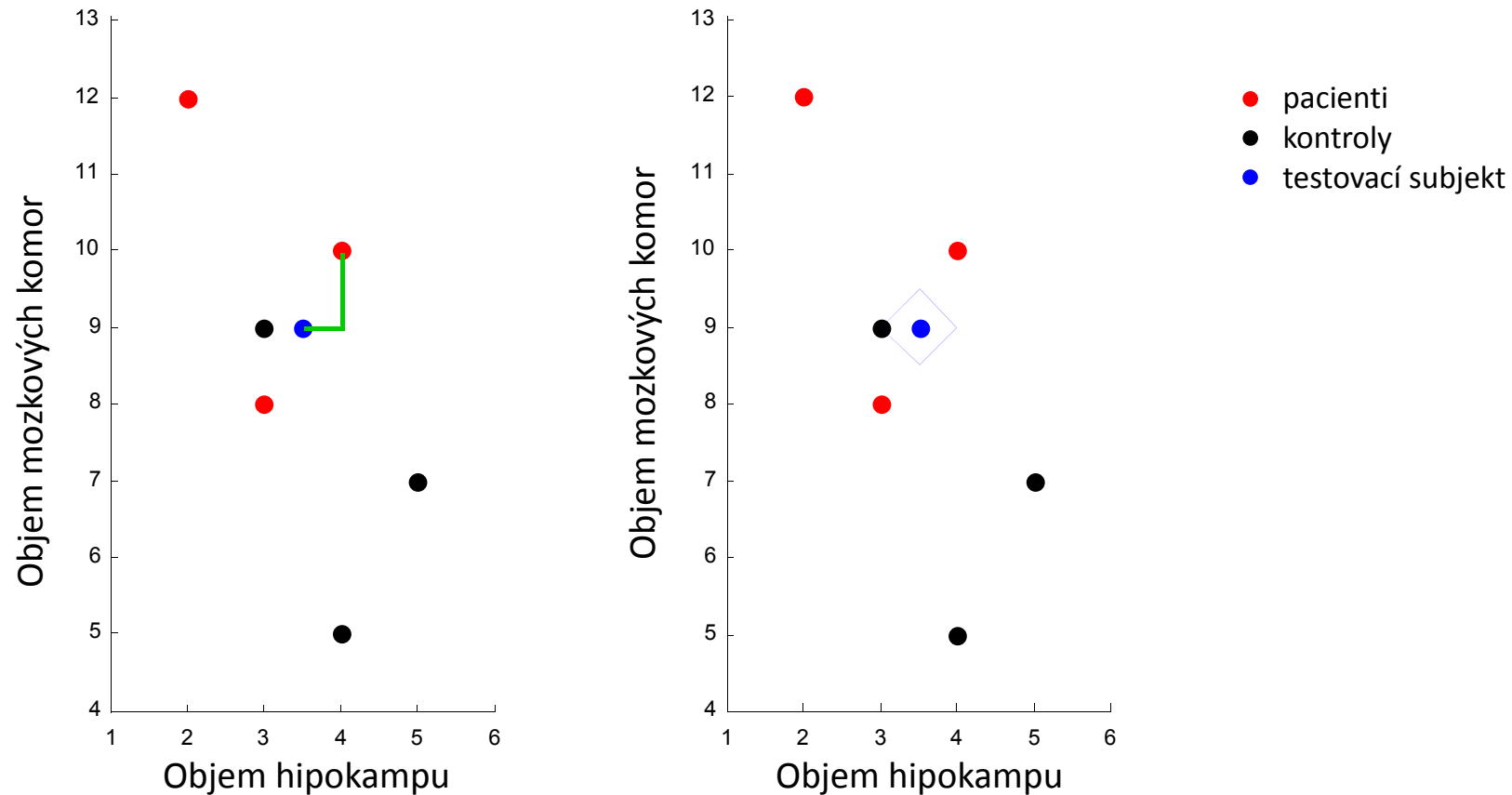
Hammingova (manhattanská) metrika

- srovnání Hammingovy (manhattanské) metriky a Euklidovy metriky



Hammingova (manhattanská) metrika

$$D_H(\mathbf{x}_1, \mathbf{x}_2) = \sum_{i=1}^n |x_{1i} - x_{2i}|$$



- geometrickým místem bodů s toutéž manhattanskou vzdáleností od daného bodu je hyperkrychle (ve dvourozměrném prostoru čtverec)

Minkovského metrika

- zobecněním Euklidovy a Hammingovy (manhattanské) metriky

$$D_M(\mathbf{x}_1, \mathbf{x}_2) = \left(\sum_{i=1}^n |x_{1i} - x_{2i}|^m \right)^{1/m}$$

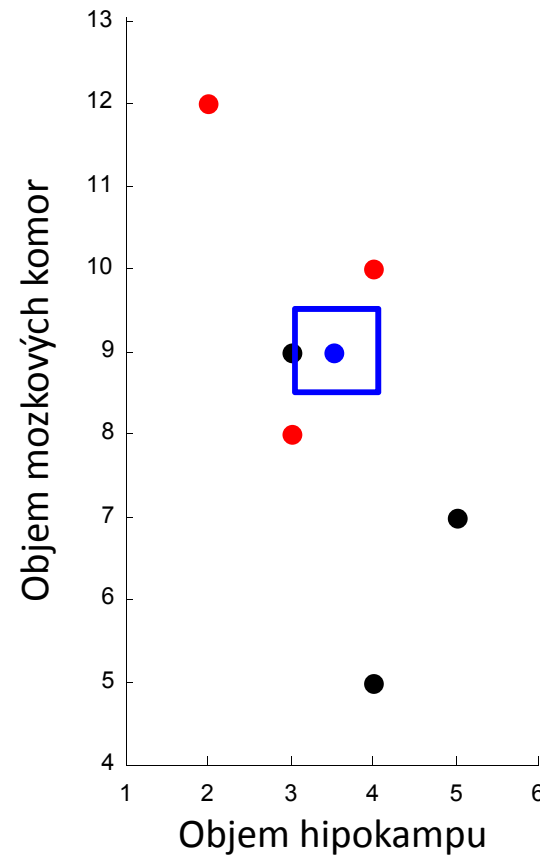
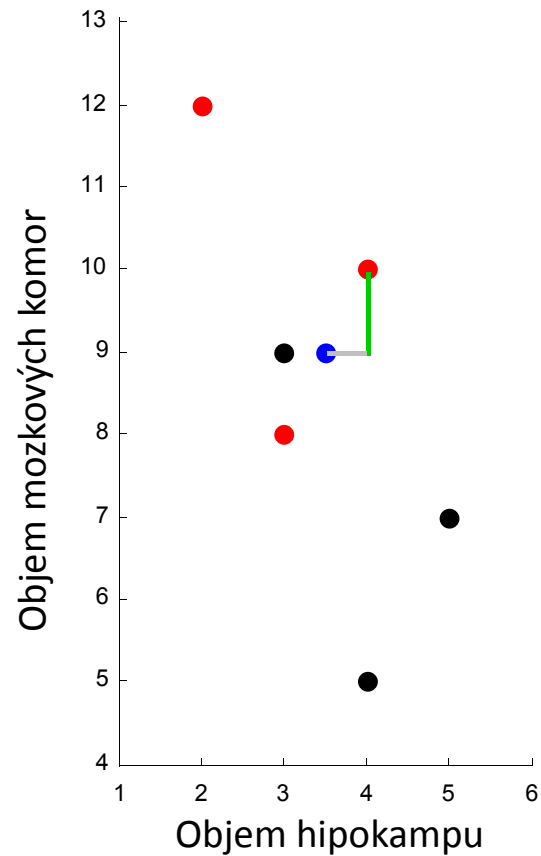
- Euklidova metrika pro $m = 2$, Hammingova (manhattanská) metrika pro $m = 1$
- volba m závisí na tom, jak moc chceme váhovat velké rozdíly mezi proměnnými (čím větší m , tím větší váha na velké rozdíly mezi proměnnými)
- pro $m \rightarrow \infty$ metrika konverguje k **Čebyševově metrice**

$$D_C(\mathbf{x}_1, \mathbf{x}_2) = \lim_{m \rightarrow \infty} D_M(\mathbf{x}_1, \mathbf{x}_2) = \max_{\forall i} |x_{1i} - x_{2i}|$$

Čebyševova metrika

- odvozena z Minkovského metriky pro $m \rightarrow \infty$

$$D_C(\mathbf{x}_1, \mathbf{x}_2) = \max_{\forall i} |x_{1i} - x_{2i}|$$



- pacienti
- kontroly
- testovací subjekt

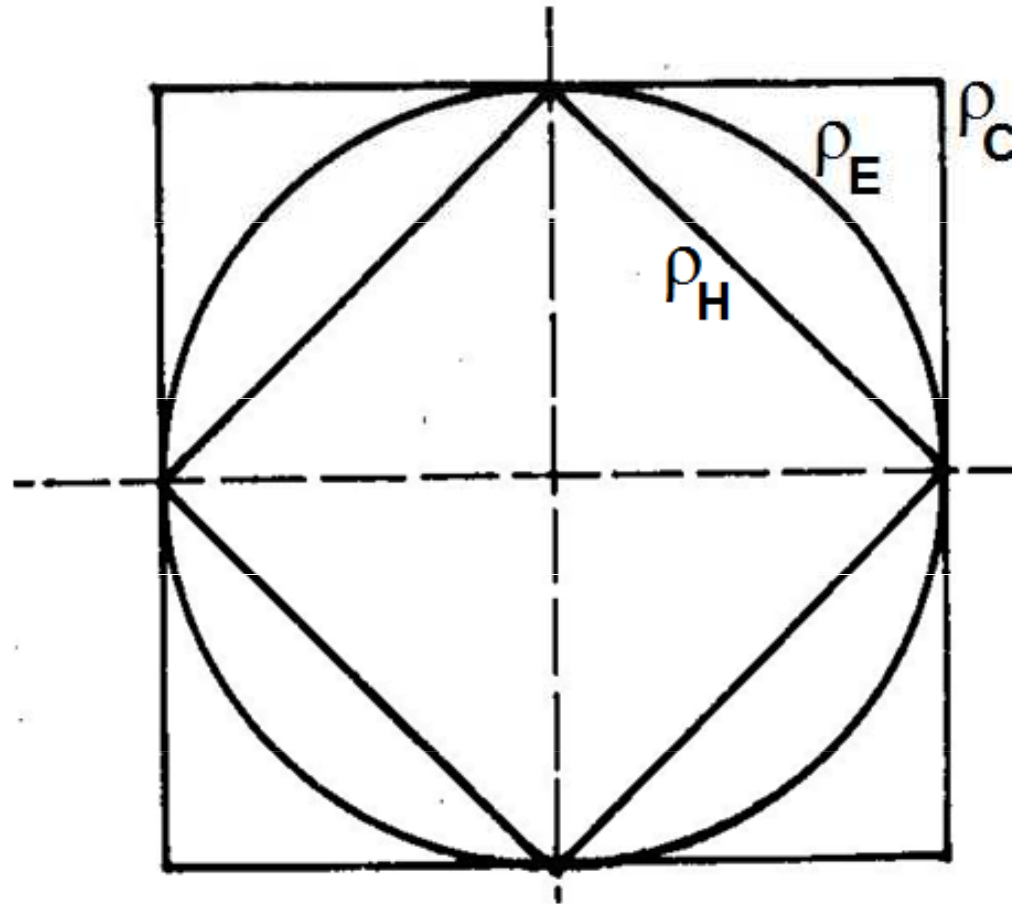
Čebyševova metrika

- odvozena z Minkovského metriky pro $m \rightarrow \infty$

$$D_C(\mathbf{x}_1, \mathbf{x}_2) = \max_{\forall i} |x_{1i} - x_{2i}|$$

- používá se ve výpočetně kriticky náročných případech, kdy je pracnost výpočtu pomocí Euklidovy metriky nepřijatelná
- geometrickým místem bodů s toutéž Čebyševovou vzdáleností od daného bodu je hyperkrychle (ve dvourozměrném prostoru čtverec), ale jinak orientovaná než v případě Hammingovy (manhattanské) vzdálenosti

Srovnání metrik



ρ_C ... Čebyševova metrika

ρ_E ... Euklidova metrika

ρ_H ... Hammingova (manhattanská) metrika

Canberrská metrika

- relativizovaná varianta Hammingovy (manhattanské) metriky

$$D_{CA}(\mathbf{x}_1, \mathbf{x}_2) = \sum_{i=1}^n \frac{|x_{1i} - x_{2i}|}{|x_{1i}| + |x_{2i}|}$$

- je vhodná pro proměnné s nezápornými hodnotami
- pokud se vyskytnou nulové hodnoty:
 - pokud jsou obě hodnoty x_{1i} a x_{2i} nulové, potom předpokládáme, že hodnota zlomku je nulová
 - je-li jenom jedna hodnota nulová, pak je zlomek roven 1 bez ohledu na velikost druhé hodnoty
 - někdy se nulové hodnoty nahrazují malým kladným číslem (menším než nejmenší naměřené hodnoty)
- velice citlivá na malé změny souřadnic, pokud se oba obrazy nacházejí v blízkosti počátku souřadnicové soustavy; naopak méně citlivá na změny hodnot proměnných, pokud jsou tyto hodnoty velké

Nevýhody metrik

- je nesmyslné vytvářet součet rozdílů veličin s různým fyzikálním rozměrem, a tudíž často s velmi rozdílným rozsahem
- při začlenění korelovaných veličin se zvyšuje jejich vliv na výslednou hodnotu
- řešení:
 1. transformace proměnných:
 - vztažení k nějakému vyrovnávacímu faktoru (střední hodnotě, směrodatné odchylce, rozpětí $\Delta_i = \max_j x_{ij} - \min_j x_{ij}$) či pomocí standardizace $u_{ij} = \frac{x_{ij} - \bar{x}_j}{\sigma_j}$; $i = 1, \dots, n; j = 1, \dots, p$; kde n je počet subjektů a p je počet proměnných
 2. váhování:
 - např. **Minkovského váhovaná metrika**:
$$D_{WM}(\mathbf{x}_1, \mathbf{x}_2) = (\sum_{i=1}^n a_i \cdot |x_{1i} - x_{2i}|^m)^{1/m}$$
 3. začlenění kovarianční matice do výpočtu:
 - začleněním inverze kovarianční matice získáváme **Mahalanobisovu metriku** (což je Euklidova metrika váhovaná inverzí kovarianční matice):
$$D_{MA}(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{(\mathbf{x}_1 - \mathbf{x}_2)^T \cdot \mathbf{S}^{-1} \cdot (\mathbf{x}_1 - \mathbf{x}_2)}$$

Nelineární metrika

$$\rho_N(\mathbf{x}_1, \mathbf{x}_2) = \begin{cases} 0 & \text{když } \rho_E(\mathbf{x}_1, \mathbf{x}_2) < D \\ H & \text{když } \rho_E(\mathbf{x}_1, \mathbf{x}_2) \geq D \end{cases}$$

- kde D je prahová hodnota a H je nějaká konstanta
- obě hodnoty se zpravidla volí na základě expertní analýzy řešeného problému
- ve vztahu může figurovat jakákoliv metrika vzdálenosti, nejen Euklidova metrika

Typy metrik a konkrétní příklady

MEZI DVĚMA OBJEKTY

Metriky pro určení **vzdálenosti** mezi 2 objekty s **kvantitativními** proměnnými

Euklidova m., Hammingova (manhattanská) m.,
Minkovského m., Čebyševova m., Mahalanobisova m.,
Canberrská m.

Metriky pro určení **vzdálenosti** mezi 2 objekty s **kvalitativními** proměnnými

Hammingova m.

Metriky pro určení **podobnosti** 2 objektů s **kvantitativními** proměnnými

Skalární součin, m. kosinové podobnosti, Pearsonův
korelační koeficient, Tanimotova m.

Metriky pro určení **podobnosti** 2 objektů s **kvalitativními** proměnnými

Tanimotova m., Jaccardův-Tanimotův a.k., Russelův-
Raovův a.k., Sokalův-Michenerův a.k., Dicův k.,
Rogersův-Tanimotův k., Hamanův k.

MEZI DVĚMA SKUPINAMI OBJEKTŮ

Deterministické metriky pro určení vzdálenosti mezi 2 množinami objektů

Metoda nejbližšího souseda, k nejbližších
sousedů, nejvzdálenějšího souseda, centroidová
metoda, m. průměrné vazby, Wardova metoda

Metriky pro určení vzdálenosti mezi 2 množinami objektů používající jejich pravděpodobnostní charakteristiky

Chernoffova m., Bhattacharyyova m. atd.

Příklad

Předpokládejme, že množina F obsahuje symboly $\{0, 1, 2\}$, tj. $k = 3$ a vektory \mathbf{x} a \mathbf{y} jsou následující 6-prvkové vektory (tj. $p = 6$):

$$\mathbf{x} = (0, 1, 2, 1, 2, 1)^T$$

$$\mathbf{y} = (1, 0, 2, 1, 0, 1)^T$$

Spočtěte vzdálenost obou vektorů.

Kontingenční matice $\mathbf{A}(\mathbf{x}, \mathbf{y})$ je:

$$\mathbf{A}(\mathbf{x}, \mathbf{y}) = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 2 & 0 \\ 1 & 0 & 1 \end{bmatrix}$$

Součet hodnot všech prvků matice $\mathbf{A}(\mathbf{x}, \mathbf{y})$ je roven délce p obou vektorů, tj. v našem případě:

$$\sum_{i=0}^2 \sum_{j=0}^2 a_{ij} = 6$$

Hammingova metrika vzdálenosti

$$D_{HQ}(\mathbf{x}, \mathbf{y}) = \sum_{i=0}^{k-1} \sum_{\substack{j=0 \\ i \neq j}}^{k-1} a_{ij}$$

- definována počtem pozic, v nichž se oba vektory liší
- tzn. je dána součtem všech prvků matice \mathbf{A} , které leží mimo hlavní diagonálu.

Příklad:

$$\mathbf{x} = (0, 1, 2, 1, 2, 1)^T$$

$$\mathbf{y} = (1, 0, 2, 1, 0, 1)^T$$



liší se ve 3 souřadnicích



$$d_{HQ}(\mathbf{x}, \mathbf{y}) = 3$$

$$\mathbf{A}(\mathbf{x}, \mathbf{y}) = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 2 & 0 \\ 1 & 0 & 1 \end{bmatrix}$$



3 prvky mimo diagonálu



$$d_{HQ}(\mathbf{x}, \mathbf{y}) = 3$$

Metriky pro určení podobnosti mezi dvěma objekty

Typy metrik a konkrétní příklady

MEZI DVĚMA OBJEKTY

Metriky pro určení **vzdálenosti** mezi 2 objekty s **kvantitativními** proměnnými

Euklidova m., Hammingova (manhattanská) m., Minkovského m., Čebyševova m., Mahalanobisova m., Canberrská m.

Metriky pro určení **vzdálenosti** mezi 2 objekty s **kvalitativními** proměnnými

Hammingova m.

Metriky pro určení **podobnosti** 2 objektů s **kvantitativními** proměnnými

Skalární součin, m. kosinové podobnosti, Pearsonův korelační koeficient, Tanimotova m.

Metriky pro určení **podobnosti** 2 objektů s **kvalitativními** proměnnými

Tanimotova m., Jaccardův-Tanimotův a.k., Russelův-Raovův a.k., Sokalův-Michenerův a.k., Dicův k., Rogersův-Tanimotův k., Hamanův k.

MEZI DVĚMA SKUPINAMI OBJEKTŮ

Deterministické metriky pro určení vzdálenosti mezi 2 množinami objektů

Metoda nejbližšího souseda, k nejbližších sousedů, nejvzdálenějšího souseda, centroidová metoda, m. průměrné vazby, Wardova metoda

Metriky pro určení vzdálenosti mezi 2 množinami objektů používající jejich pravděpodobnostní charakteristiky

Chernoffova m., Bhattacharyyova m. atd.

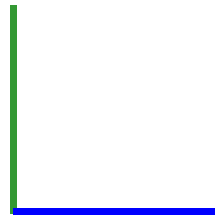
Skalární součin

$$S_{ss}(\mathbf{x}_1, \mathbf{x}_2) = \mathbf{x}_1^T \cdot \mathbf{x}_2 = \sum_{i=1}^n x_{1i} x_{2i}$$

Většinou pro vektory \mathbf{x}_1 a \mathbf{x}_2 o stejné délce (např. a); záleží na úhlu, který svírají:



úhel 0°
 $S_{ss} = a^2$



úhel 90°
 $S_{ss} = 0$



úhel 180°
 $S_{ss} = -a^2$

skalární součin invariantní vůči rotaci – absolutní orientace nepodstatná, důležitý pouze úhel
skalární součin není invariantní vůči lineární transformaci (tzn. závisí na délce vektorů)

odvození metriky vzdálenosti:

$$D_{ss}(\mathbf{x}_1, \mathbf{x}_2) = a^2 - S_{ss}(\mathbf{x}_1, \mathbf{x}_2)$$

Metrika kosinové podobnosti

$$S_{\cos}(\mathbf{x}_1, \mathbf{x}_2) = \frac{\mathbf{x}_1^T \cdot \mathbf{x}_2}{\|\mathbf{x}_1\| \cdot \|\mathbf{x}_2\|}$$

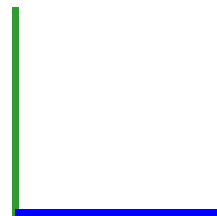
kde $\|\mathbf{x}_i\|$ je norma (délka) vektoru \mathbf{x}_i
= skalární součin vektorů o jednotkové délce

vhodná v případě, pokud je informativní pouze relativní hodnota příznaků

hodnoty $\sigma_{\cos}(\mathbf{x}_1, \mathbf{x}_2)$ jsou rovny kosinu úhlu mezi oběma vektory



úhel 0°
 $S_{\cos} = 1$



úhel 90°
 $S_{\cos} = 0$



úhel 180°
 $S_{\cos} = -1$

Pearsonův korelační koeficient

Pearsonův korelační koeficient

$$S_{PC}(\mathbf{x}_1, \mathbf{x}_2) = \frac{\mathbf{x}_{d1}^T \cdot \mathbf{x}_{d2}}{\|\mathbf{x}_{d1}\| \cdot \|\mathbf{x}_{d2}\|}$$

kde $\mathbf{x}_{di} = (x_{i1} - \bar{x}_i, x_{i2} - \bar{x}_i, \dots, x_{ip} - \bar{x}_i)^T$

\mathbf{x}_{di} jsou tzv. **diferenční vektory**

také nabývá hodnot z intervalu $\langle -1; 1 \rangle$

odvození metriky vzdálenosti:

$$D_{PC}(\mathbf{x}_1, \mathbf{x}_2) = \frac{1 - S_{PC}(\mathbf{x}_1, \mathbf{x}_2)}{2}$$

→ hodnoty se (díky dělení dvěma) vyskytují v intervalu $\langle 0; 1 \rangle$

→ používá se např. při analýze dat genové exprese

Metrika kosinové podobnosti

$$S_{\cos}(\mathbf{x}_1, \mathbf{x}_2) = \frac{\mathbf{x}_1^T \cdot \mathbf{x}_2}{\|\mathbf{x}_1\| \cdot \|\mathbf{x}_2\|}$$

Typy metrik a konkrétní příklady

MEZI DVĚMA OBJEKTY

Metriky pro určení **vzdálenosti** mezi 2 objekty s **kvantitativními** proměnnými

Euklidova m., Hammingova (manhattanská) m., Minkovského m., Čebyševova m., Mahalanobisova m., Canberrská m.

Metriky pro určení **vzdálenosti** mezi 2 objekty s **kvalitativními** proměnnými

Hammingova m.

Metriky pro určení **podobnosti** 2 objektů s **kvantitativními** proměnnými

Skalární součin, m. kosinové podobnosti, Pearsonův korelační koeficient, Tanimotova m.

Metriky pro určení **podobnosti** 2 objektů s **kvalitativními** proměnnými

Tanimotova m., Jaccardův-Tanimotův a.k., Russelův-Raovův a.k., Sokalův-Michenerův a.k., Dicův k., Rogersův-Tanimotův k., Hamanův k.

MEZI DVĚMA SKUPINAMI OBJEKTŮ

Deterministické metriky pro určení vzdálenosti mezi 2 množinami objektů

Metoda nejbližšího souseda, k nejbližších sousedů, nejvzdálenějšího souseda, centroidová metoda, m. průměrné vazby, Wardova metoda

Metriky pro určení vzdálenosti mezi 2 množinami objektů používající jejich pravděpodobnostní charakteristiky

Chernoffova m., Bhattacharyyova m. atd.

Metriky pro určení podobnosti 2 objektů s kvalitativními prom.

1. případy obecné
2. případy s dichotomickými příznaky, pro které je definována celá řada tzv. **asociačních koeficientů**.

(Asociační koeficienty až na výjimky nabývají hodnot z intervalu $\langle 0, 1 \rangle$, hodnoty 1 v případě shody vektorů, 0 pro případ nepodobnosti.)

Obecné metriky – Hammingova metrika podobnosti

$$S_{HQ}(\mathbf{x}, \mathbf{y}) = p - D_{HQ}(\mathbf{x}, \mathbf{y})$$

Příklad:

$$\mathbf{x} = (0, 1, 2, 1, 2, 1)^T$$

$$\mathbf{y} = (1, 0, 2, 1, 0, 1)^T$$



liší se ve 3 souřadnicích



$$d_{HQ}(\mathbf{x}, \mathbf{y}) = 3$$



shoda ve 3 souřadnicích



$$s_{HQ}(\mathbf{x}, \mathbf{y}) = 6 - 3 = 3$$

$$A(\mathbf{x}, \mathbf{y}) = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 2 & 0 \\ 1 & 0 & 1 \end{bmatrix}$$



3 prvky mimo diagonálu



$$d_{HQ}(\mathbf{x}, \mathbf{y}) = 3$$



součet prvků na diagonále roven 3



$$s_{HQ}(\mathbf{x}, \mathbf{y}) = 6 - 3 = 3$$

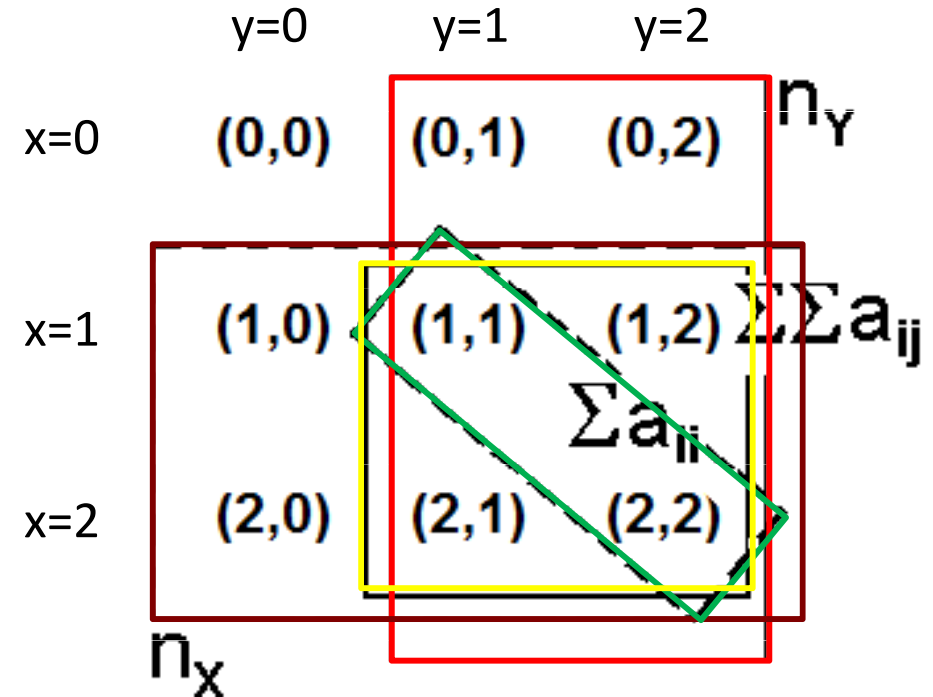
Obecné metriky – Tanimotova metrika

$$S_{TQ}(\mathbf{x}, \mathbf{y}) = \frac{n_{X \cap Y}}{n_X + n_Y - n_{X \cap Y}} = \frac{\sum_{i=1}^{k-1} a_{ii}}{n_x + n_y - \sum_{i=1}^{k-1} \sum_{j=1}^{k-1} a_{ij}}$$

$$n_x = \sum_{i=1}^{k-1} \sum_{j=0}^{k-1} a_{ij}$$

$$n_y = \sum_{i=0}^{k-1} \sum_{j=1}^{k-1} a_{ij}$$

Pro výpočet Tanimotovy podobnosti dvou vektorů s kvalitativními příznaky jsou použity všechny páry složek srovnávaných vektorů, kromě těch, jejichž hodnoty jsou obě nulové.



Obecné metriky – Tanimotova metrika – příklad

Určete hodnoty Tanimotových podobností $s_{TQ}(\mathbf{x}, \mathbf{x})$, $s_{TQ}(\mathbf{x}, \mathbf{y})$ a $s_{TQ}(\mathbf{x}, \mathbf{z})$, když:

$$\mathbf{x} = (0, 1, 2, 1, 2, 1)^T \text{ a}$$

$$\mathbf{y} = (1, 0, 2, 1, 0, 1)^T \text{ a}$$

$$\mathbf{z} = (2, 0, 0, 0, 0, 2)^T.$$

Ze zadání je množina symbolů $F = \{0, 1, 2\}$, $k = 3$, $p = 6$.

Kontingenční tabulky jsou:

$$\mathbf{A}(\mathbf{x}, \mathbf{x}) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 2 \end{bmatrix}$$

$$\mathbf{A}(\mathbf{x}, \mathbf{y}) = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 2 & 0 \\ 1 & 0 & 1 \end{bmatrix}$$

$$\mathbf{A}(\mathbf{x}, \mathbf{z}) = \begin{bmatrix} 0 & 0 & 1 \\ 2 & 0 & 1 \\ 2 & 0 & 0 \end{bmatrix}$$

$$s_{TQ}(\mathbf{x}, \mathbf{x}) = \frac{5}{5+5-5} = 1$$

$$s_{TQ}(\mathbf{x}, \mathbf{y}) = \frac{3}{5+4-3} = 0,5$$

$$s_{TQ}(\mathbf{x}, \mathbf{z}) = \frac{0}{5+2-1} = 0$$

Další obecné metriky

- definovány pomocí různých prvků kontingenční matice $\mathbf{A}(\mathbf{x}, \mathbf{y})$
- některé z nich používají pouze počet shodných pozic v obou vektorech (ovšem s nenulovými hodnotami):

$$S_1(\mathbf{x}, \mathbf{y}) = \frac{\sum_{i=1}^{k-1} a_{ii}}{p}$$

$$S_2(\mathbf{x}, \mathbf{y}) = \frac{\sum_{i=1}^{k-1} a_{ii}}{p - a_{00}}$$

- některé z nich používají i shodu s nulovými hodnotami:

$$S_3(\mathbf{x}, \mathbf{y}) = \frac{\sum_{i=0}^{k-1} a_{ii}}{p}$$

Asociační koeficienty

| | | x_j | |
|-------|---------|---------|--------|
| | | false/0 | true/1 |
| x_i | false/0 | D | C |
| | true/1 | B | A |

- A** - u obou objektů sledovaný jev nastal (obě odpovídající si proměnné mají hodnotu true, resp.1) – **pozitivní shoda**;
- B** - u objektu x_i jev nastal ($x_{ik} = \underline{\text{true}}$), zatímco u objektu x_j nikoliv ($x_{jk} = \underline{\text{false}}$, resp.0);
- C** - u objektu x_i jev nenastal ($x_{ik} = \underline{\text{false}}$), zatímco u objektu x_j ano ($x_{jk} = \underline{\text{true}}$);
- D** - sledovaný jev nenastal ani u jednoho z objektů (obě odpovídající si proměnné mají hodnotu false, resp. 0) – **negativní shoda**.

Při výpočtu podobnosti dvou objektů sledujeme, kolikrát pro všechny souřadnice obou vektorů x_i a x_j nastaly případy shody či neshody:

- **A+D** určuje celkový počet shod
- **B+C** celkový počet neshod
- **A+B+C+D** = p (tj. celk. počet souřadnic obou vektorů – tzn. počet proměnných)

Jaccardův – Tanimotův asociační koeficient

$$S_{JT}(\mathbf{x}, \mathbf{y}) = \frac{A}{A + B + C}$$

| | | \mathbf{x}_j | |
|----------------|---------|----------------|--------|
| | | false/0 | true/1 |
| \mathbf{x}_i | false/0 | D | C |
| | true/1 | B | A |

což je díky zjednodušení i dichotomická varianta metriky podle vztahu:

$$S_{TQ}(\mathbf{x}, \mathbf{y}) = \frac{\sum_{i=1}^{k-1} a_{ii}}{n_x + n_y - \sum_{i=1}^{k-1} \sum_{i=1}^{k-1} a_{ij}}$$

Tento vztah se dominantně používá v ekologických studiích.

Další asociační koeficienty I

| | | | |
|-------|---------|---------|--------|
| | | x_j | |
| | | false/0 | true/1 |
| x_i | false/0 | D | C |
| | true/1 | B | A |

Russelův – Raoův asociační koeficient

$$S_{RR}(\mathbf{x}, \mathbf{y}) = \frac{A}{A + B + C + D}$$

dichotomická varianta
metriky:

$$S_1(\mathbf{x}, \mathbf{y}) = \frac{\sum_{i=1}^{k-1} a_{ii}}{p}$$

Sokalův – Michenerův asociační koeficient

$$S_{SM}(\mathbf{x}, \mathbf{y}) = \frac{A + D}{A + B + C + D}$$

dichotomická varianta
metriky:

$$S_3(\mathbf{x}, \mathbf{y}) = \frac{\sum_{i=0}^{k-1} a_{ii}}{p}$$

Další asociační koeficienty II

| | | | |
|-------|---------|---------|--------|
| | | x_j | |
| | | false/0 | true/1 |
| x_i | false/0 | D | C |
| | true/1 | B | A |

Diceův (Czekanowského) asociační koeficient

$$S_{DC}(\mathbf{x}, \mathbf{y}) = \frac{2A}{2A + B + C} = \frac{2A}{(A + B) + (A + C)}$$

V případě Jaccardova a Diceova koeficientu pokud nastane úplná negativní shoda (tzn. $A = B = C = 0$), pak často: $S_{JT}(\mathbf{x}, \mathbf{y}) = S_{DC}(\mathbf{x}, \mathbf{y}) = 1$.

Rogersův – Tanimotův asociační koeficient

$$S_{RT}(\mathbf{x}, \mathbf{y}) = \frac{A + D}{A + D + 2 \cdot (B + C)} = \frac{A + D}{(B + C) + (A + B + C + D)}$$

Hamanův asociační koeficient

$$S_{HA}(\mathbf{x}, \mathbf{y}) = \frac{A + D - (B + C)}{A + B + C + D}$$

nabývá na rozdíl od všech dříve uvedených koeficientů hodnot z intervalu $\langle -1, 1 \rangle$. Hodnoty -1, pokud se příznaky pouze neshodují; hodnoty 0, když je počet shod a neshod v rovnováze; +1 v případě úplné shody všech příznaků

Asociační koeficienty – poznámka

| | | x_j | |
|-------|---------|---------|--------|
| | | false/0 | true/1 |
| x_i | false/0 | D | C |
| | true/1 | B | A |

Na základě četností A až D lze pro případ binárních příznaků vytvářet i zajímavé vztahy pro již dříve uvedené míry:

Hammingova metrika $D_H(\mathbf{x}, \mathbf{y}) = B + C$

Euklidova metrika $D_H(\mathbf{x}, \mathbf{y}) = \sqrt{B + C}$

Pearsonův korelační koeficient

$$S_{PC}(\mathbf{x}, \mathbf{y}) = \frac{A \cdot D - B \cdot C}{\sqrt{(A + B) \cdot (C + D) \cdot (A + C) \cdot (B + D)}}$$

Výpočet vzdáleností z asociačních koeficientů

Z asociačních koeficientů, které vyjadřují míru podobnosti, lze jednoduše odvodit i míry nepodobnosti (vzdálenosti) pomocí:

$$D_X(\mathbf{x}, \mathbf{y}) = 1 - S_X(\mathbf{x}, \mathbf{y})$$

Výpočet vzdáleností v Matlabu

Funkce:

- pdist (vzdálenost mezi páry objektů matice X či páry proměnných matice X^T)
- pdist2 (vzdálenost mezi maticemi X a Y)

Výběr metrik vzdáleností u obou těchto funkcí:

- 'euclidean' – Euklidova vzdálenost
- 'squaredeuclidean' – čtverec Euklidovy vzdálenosti
- 'seuclidean' – standardizovaná Euklidova vzdálenost
- 'cityblock' – Hammingova (manhattanská) vzdálenost
- 'minkowski' – Minkovského vzdálenost
- 'chebychev' – Čebyševova vzdálenost
- 'mahalanobis' – Mahalanobisova vzdálenost
- 'cosine' – 1 minus kosinová podobnost
- 'correlation' – 1 minus Pearsonův korelační koeficient
- 'spearman' – 1 minus Spearmanův korelační koeficient
- 'hamming' – Hamminova vzdálenost (pro kvalitativní proměnné)
- 'jaccard' – 1 minus Jaccardův koeficient
- lze případně nadefinovat i jinou metriku

Metriky pro určení vzdálenosti mezi dvěma skupinami objektů

Typy metrik a konkrétní příklady

MEZI DVĚMA OBJEKTY

Metriky pro určení **vzdálenosti** mezi 2 objekty s **kvantitativními** proměnnými

Euklidova m., Hammingova (manhattanská) m., Minkovského m., Čebyševova m., Mahalanobisova m., Canberrská m.

Metriky pro určení **vzdálenosti** mezi 2 objekty s **kvalitativními** proměnnými

Hammingova m.

Metriky pro určení **podobnosti** 2 objektů s **kvantitativními** proměnnými

Skalární součin, m. kosinové podobnosti, Pearsonův korelační koeficient, Tanimotova m.

Metriky pro určení **podobnosti** 2 objektů s **kvalitativními** proměnnými

Tanimotova m., Jaccardův-Tanimotův a.k., Russelův-Raovův a.k., Sokalův-Michenerův a.k., Dicův k., Rogersův-Tanimotův k., Hamanův k.

MEZI DVĚMA SKUPINAMI OBJEKTŮ

Deterministické metriky pro určení vzdálenosti mezi 2 množinami objektů

Metoda nejbližšího souseda, k nejbližších sousedů, nejvzdálenějšího souseda, centroidová metoda, m. průměrné vazby, Wardova metoda

Metriky pro určení vzdálenosti mezi 2 množinami objektů používající jejich pravděpodobnostní charakteristiky

Chernoffova m., Bhattacharyyova m. atd.

Vzdálenost mezi skupinami objektů

- vzdálenost mezi skupinami dána:
 - „vzdáleností“ jednoho objektu s jedním či více objekty jedné skupiny (třídy) – použitelné při klasifikaci
 - „vzdáleností“ skupin (třídy, shluku) obrazů či „vzdáleností“ jednoho obrazu z každé skupiny – použitelné při shlukování
- jednotlivé deterministické metriky pro určení vzdálenosti mezi dvěma množinami objektů si probereme v rámci shlukové analýzy na příští přednášce

Typy metrik a konkrétní příklady

MEZI DVĚMA OBJEKTY

Metriky pro určení vzdálenosti mezi 2 objekty s kvantitativními proměnnými

Euklidova m., Hammingova (manhattanská) m., Minkovského m., Čebyševova m., Mahalanobisova m., Canberrská m.

Metriky pro určení vzdálenosti mezi 2 objekty s kvalitativními proměnnými

Hammingova m.

Metriky pro určení podobnosti 2 objektů s kvantitativními proměnnými

Skalární součin, m. kosinové podobnosti, Pearsonův korelační koeficient, Tanimotova m.

Metriky pro určení podobnosti 2 objektů s kvalitativními proměnnými

Tanimotova m., Jaccardův-Tanimotův a.k., Russelův-Raovův a.k., Sokalův-Michenerův a.k., Dicův k., Rogersův-Tanimotův k., Hamanův k.

MEZI DVĚMA SKUPINAMI OBJEKTŮ

Deterministické metriky pro určení vzdálenosti mezi 2 množinami objektů

Metoda nejbližšího souseda, k nejbližších sousedů, nejvzdálenějšího souseda, centroidová metoda, m. průměrné vazby, Wardova metoda

Metriky pro určení vzdálenosti mezi 2 množinami objektů používající jejich pravděpodobnostní charakteristiky

Chernoffova m., Bhattacharyyova m. atd.

Metriky založené na pstních charakteristikách

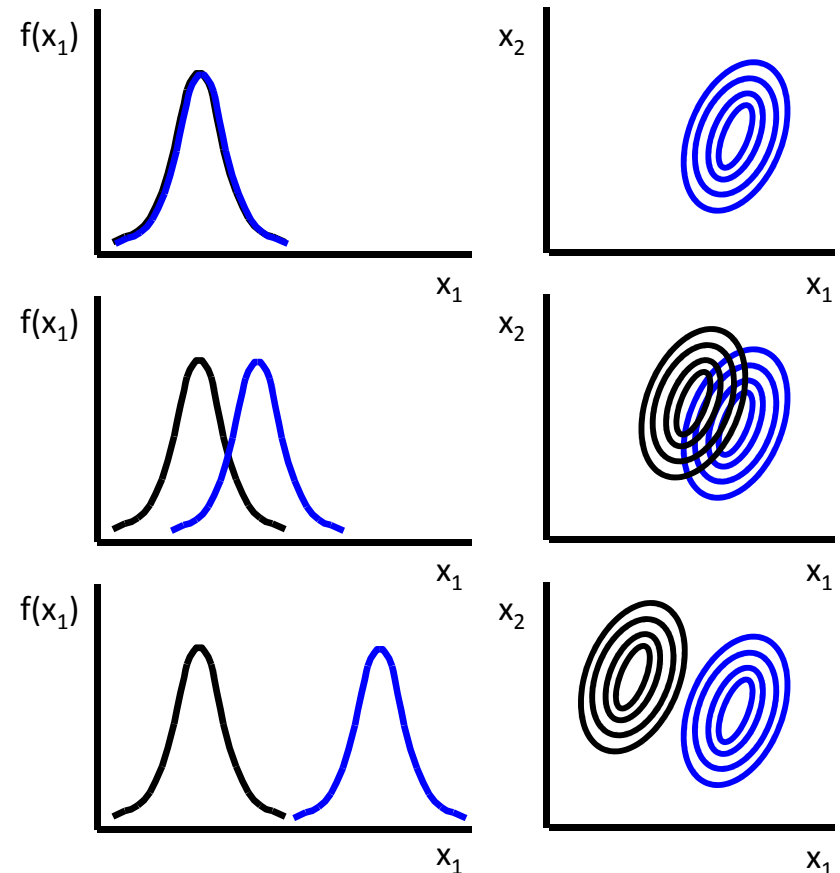
Základní myšlenkou je využití **pravděpodobnosti způsobené chyby při klasifikaci** (tzn. zařazení objektu do skupiny). Čím více se hustoty pravděpodobnosti výskytu obrazů \mathbf{x} v jednotlivých množinách překrývají, tím je větší pravděpodobnost chyby.

Tzn. tyto metriky splňují následující vlastnosti:

1. $J = 0$, pokud jsou hustoty pravděpodobnosti obou množin identické, tj. když $p(\mathbf{x} | \omega_1) = p(\mathbf{x} | \omega_2)$

2. $J > 0$

3. J nabývá maxima, pokud jsou obě množiny disjunktní, tj. když
$$\int_{-\infty}^{\infty} p(\mathbf{x} | \omega_1) \cdot p(\mathbf{x} | \omega_2) d\mathbf{x} = 0$$



Asociační matice

Asociační matice – Q mode analýza

NxP MATICE

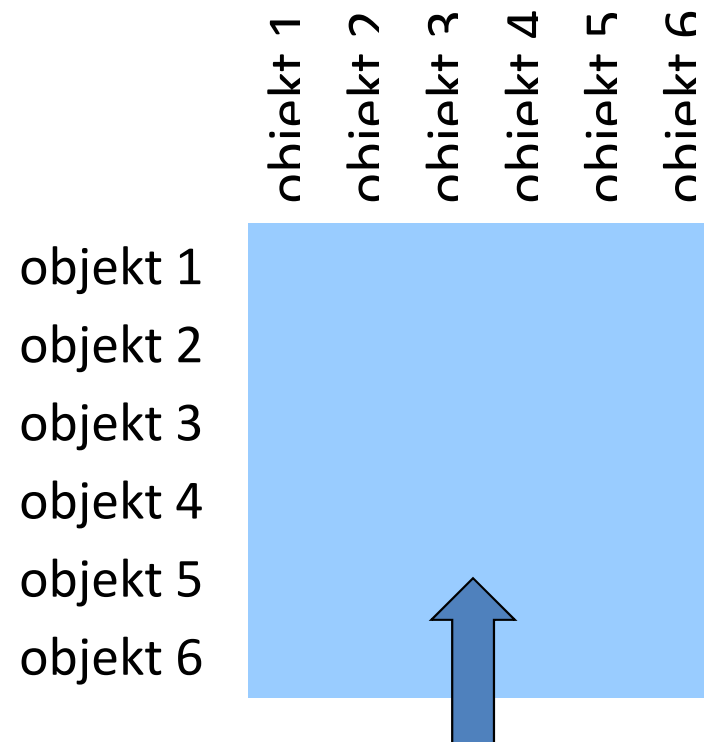


Hodnoty proměnných pro jednotlivé objekty

Výpočet metriky
podobností/
vzdáleností



ASOCIAČNÍ MATICE



Vzdálenost, podobnost, korelace,
kovariance mezi objekty

Asociační matice – R mode analýza

NxP MATICE

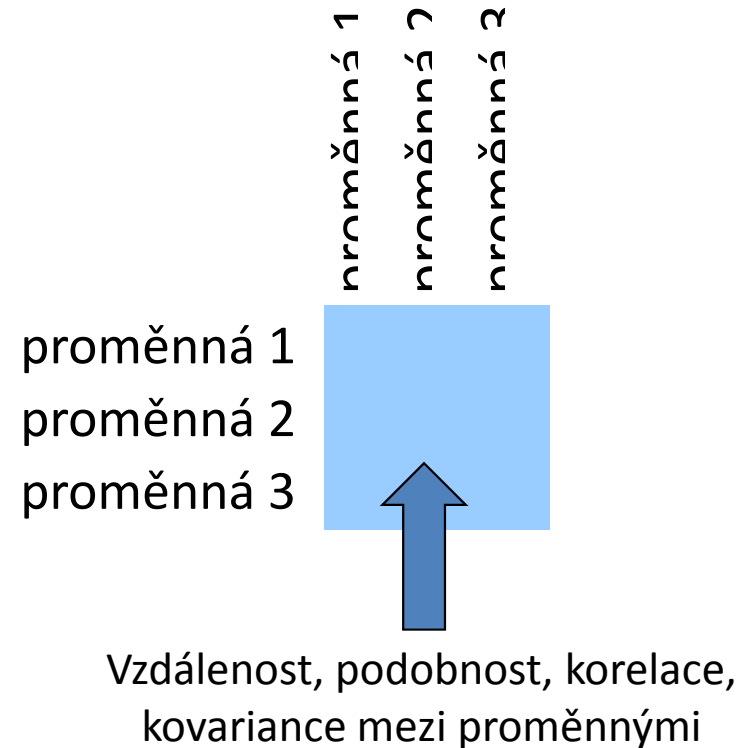


Hodnoty proměnných pro jednotlivé objekty

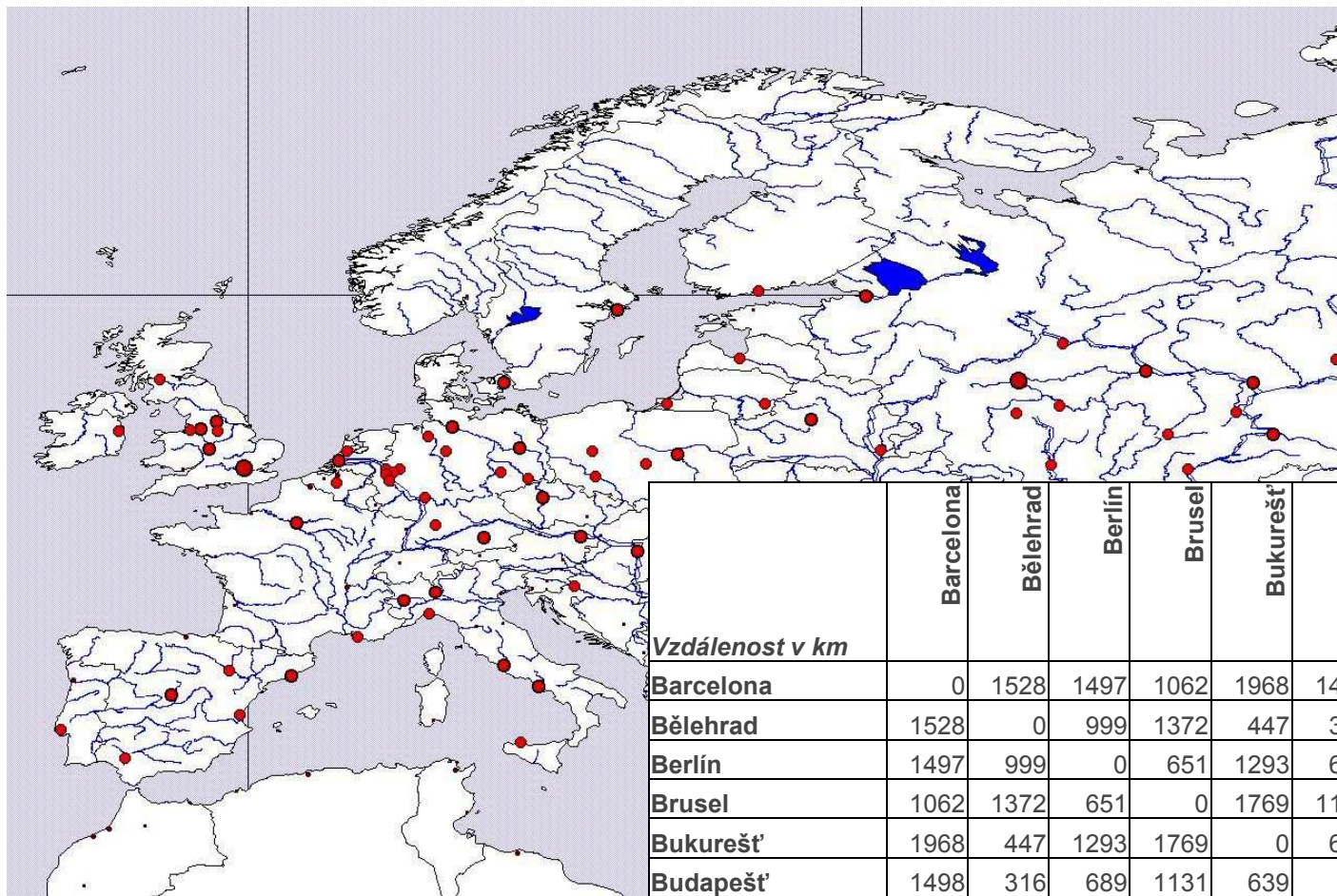
Výpočet metriky
podobností/
vzdáleností



ASOCIAČNÍ MATICE



Asociační matice – ukázka



Vzdálenost měst v mapě není ničím jiným než maticí vzdálenosti v 2D prostoru

| | Barcelona | Bělehrad | Berlín | Brusel | Bukurešť | Budapešť | Kodaň | Dublin | Hamburg | Istanbul | Kiev | Londýn | Madrid |
|------------------------|-----------|----------|--------|--------|----------|----------|-------|--------|---------|----------|------|--------|--------|
| <i>Vzdálenost v km</i> | | | | | | | | | | | | | |
| Barcelona | 0 | 1528 | 1497 | 1062 | 1968 | 1498 | 1757 | 1469 | 1471 | 2230 | 2391 | 1137 | 504 |
| Bělehrad | 1528 | 0 | 999 | 1372 | 447 | 316 | 1327 | 2145 | 1229 | 809 | 976 | 1688 | 2026 |
| Berlín | 1497 | 999 | 0 | 651 | 1293 | 689 | 354 | 1315 | 254 | 1735 | 1204 | 929 | 1867 |
| Brusel | 1062 | 1372 | 651 | 0 | 1769 | 1131 | 766 | 773 | 489 | 2178 | 1836 | 318 | 1314 |
| Bukurešť | 1968 | 447 | 1293 | 1769 | 0 | 639 | 1571 | 2534 | 1544 | 445 | 744 | 2088 | 2469 |
| Budapešť | 1498 | 316 | 689 | 1131 | 639 | 0 | 1011 | 1894 | 927 | 1064 | 894 | 1450 | 1975 |
| Kodaň | 1757 | 1327 | 354 | 766 | 1571 | 1011 | 0 | 1238 | 287 | 2017 | 1326 | 955 | 2071 |
| Dublin | 1469 | 2145 | 1315 | 773 | 2534 | 1894 | 1238 | 0 | 1073 | 2950 | 2513 | 462 | 1449 |
| Hamburg | 1471 | 1229 | 254 | 489 | 1544 | 927 | 287 | 1073 | 0 | 1983 | 1440 | 720 | 1785 |
| Istanbul | 2230 | 809 | 1735 | 2178 | 445 | 1064 | 2017 | 2950 | 1983 | 0 | 1052 | 2496 | 2734 |
| Kiev | 2391 | 976 | 1204 | 1836 | 744 | 894 | 1326 | 2513 | 1440 | 1052 | 0 | 2131 | 2859 |
| Londýn | 1137 | 1688 | 929 | 318 | 2088 | 1450 | 955 | 462 | 720 | 2496 | 2131 | 0 | 1263 |
| Madrid | 504 | 2026 | 1867 | 1314 | 2469 | 1975 | 2071 | 1449 | 1785 | 2734 | 2859 | 1263 | 0 |

Asociační matice – shrnutí

- Typická asociační matice je čtvercová matice
- Typická asociační matice je symetrická kolem diagonály
 - Ve speciálních případech existují i asymetrické asociační matice
- Diagonála obsahuje:
 - 0 (v případě vzdáleností)
 - identitu objektu se sebou samým (v případě podobnosti, obvykle 1 nebo 100%)
- Asociační matice může být spočtena mezi objekty (Q mode analýza) nebo mezi proměnnými (R mode analýza)
- Asociační matice mohou být jak vstupem do vícerozměrných analýz, tak vstupem pro klasické jednorozměrné statistické výpočty, kdy základní jednotkou není jeden objekt, ale podobnost/vzdálenost dvojice objektů

Poděkování

Příprava výukových materiálů předmětu „DSAN02 Pokročilé metody analýzy dat v neurovědách“ byla finančně podporována prostředky projektu FRMU č. MUNI/FR/0260/2014 „Pokročilé metody analýzy dat v neurovědách jako nový předmět na LF MU“

