

Analýza a klasifikace dat – přednáška 3



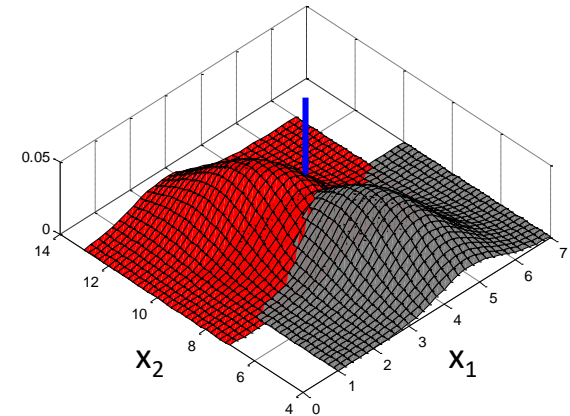
RNDr. Eva Koriťáková, Ph.D.

Podzim 2017

Typy klasifikátorů – podle principu klasifikace

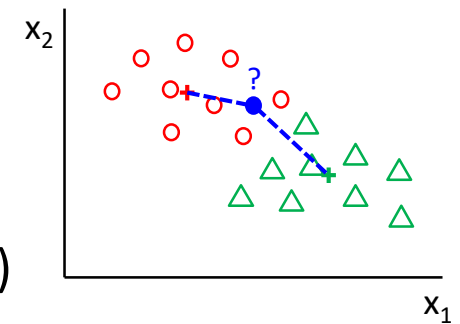
- **klasifikace pomocí diskriminačních funkcí:**

- diskriminační funkce určují míru příslušnosti k dané klasifikační třídě
- pro danou třídu má daná diskriminační funkce nejvyšší hodnotu



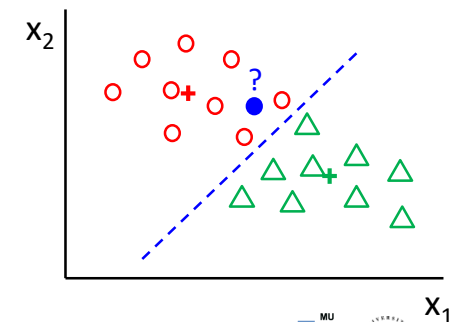
- **klasifikace pomocí vzdálenosti od etalonů klasif. tříd:**

- etalon = reprezentativní objekt(y) klasifikační třídy
- počet etalonů klasif. třídy různý – od jednoho vzorku (např. centroidu) po úplný výčet všech objektů dané třídy (např. u klasif. pomocí metody průměrné vazby)



- **klasifikace pomocí hranic v obrazovém prostoru:**

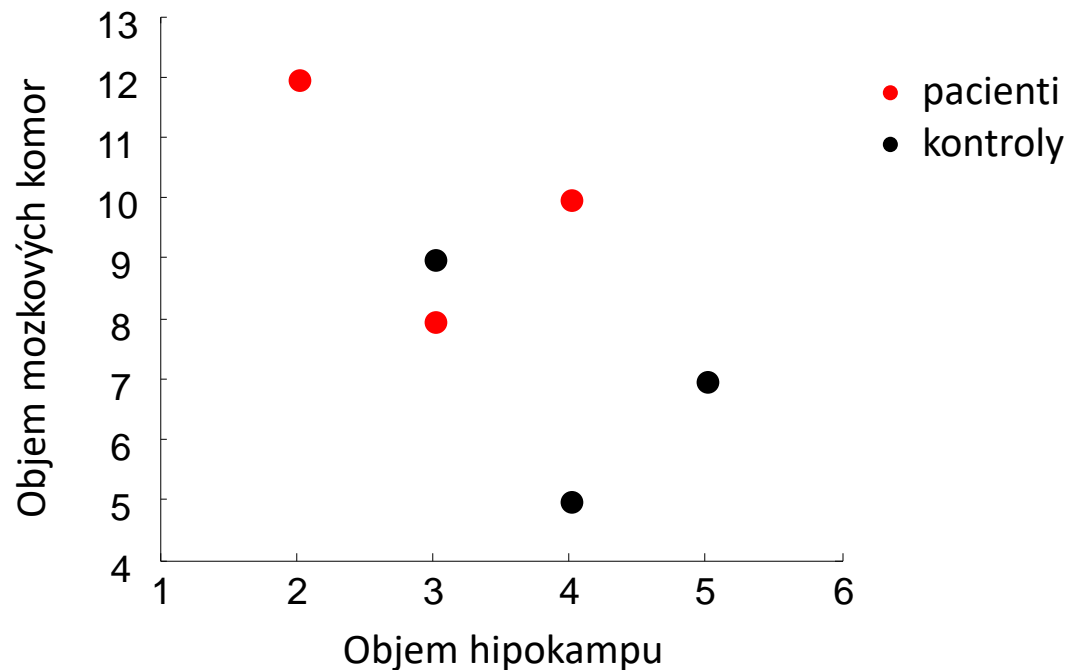
- stanovení hranic (hraničních ploch) oddělujících klasifikační třídy



Poznámka

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

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



Metrika – vzdálenost

Metrika D na X je funkce $D: X \times X \rightarrow \mathbb{R}$, kde \mathbb{R} je množina reálných čísel taková, že:

$$\exists D_0 \in \mathbb{R}: -\infty < D_0 \leq D(x, y) < +\infty, \forall x, y \in X$$

$$D(x, x) = D_0, \forall x \in X$$

a

$$D(x, y) = D(y, x), \forall x, y \in X \text{ (symetrie)}$$

$$D(x, y) = D_0 \text{ když a jen když } x = y \text{ (totožnost)}$$

$$D(x, z) \leq D(x, y) + D(y, z), \forall x, y, z \in X \text{ (\Delta nerovnost)}$$

Prostor X , ve kterém metrika D definována, nazýváme **metrickým prostorem**.

Vzdálenost je hodnota určená podle metriky.

Poznámka: zpravidla $D_0=0$.

Metrika – podobnost

Metrická míra podobnosti S na X je funkce $S: X \times X \rightarrow \mathbb{R}$, taková, že:

$$\exists S_0 \in \mathbb{R}: -\infty < S(x,y) \leq S_0 < +\infty, \forall x,y \in X$$

$$S(x,x) = S_0, \forall x \in X$$

a

$$S(x,y) = S(y,x), \forall x,y \in X \text{ (symetrie)}$$

$$S(x,y) = S_0 \text{ když a jen když } x = y \text{ (totožnost)}$$

$$S(x,y) \cdot S(y,z) \leq [S(x,y) + S(y,z)] \cdot S(x,z), \forall x,y,z \in X$$

Podobnost je hodnota určená podle metrické míry podobnosti.

Poznámka: zpravidla $S_0=1$ (ale neplatí to vždy, u některých metrik je maximální hodnota podobnosti jiná než 1)

Metriky podobnosti vs. metriky vzdálenosti

Vzdálenostní míry (míry nepodobnosti) mohou být transformovány na podobnostní míry různými transformacemi, např.:

$$S_{ij} = 1/D_{ij}$$

$$S_{ij} = 1/(1 + D_{ij})$$

$$S_{ij} = c - D_{ij}, c \geq \max D_{ij}, \forall i, j$$

Typy měr vzdálenosti (podobnosti)

- podle **typu proměnné** (kvalitativní proměnné, kvantitativní proměnné)
- podle **počtu objektů**, jejichž vztah hodnotíme – objekty (vektory), množiny objektů (vektorů)
- **deterministické** (nepravděpodobností) vs. **pravděpodobností 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 objekty 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

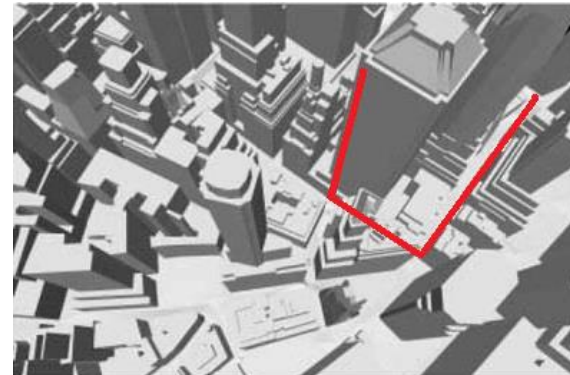
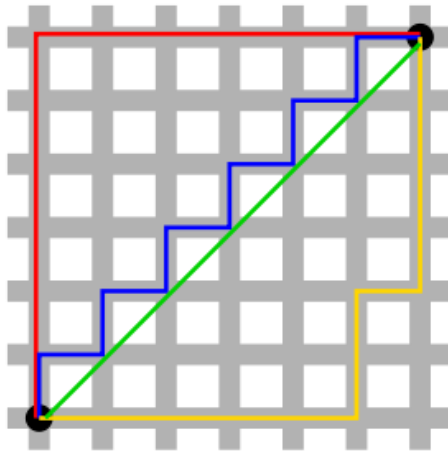
$$D_E(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{\sum_{i=1}^n (x_{1i} - x_{2i})^2}$$

- 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 AJ názvy: Manhattan distance, city-block distance, taxi driver distance

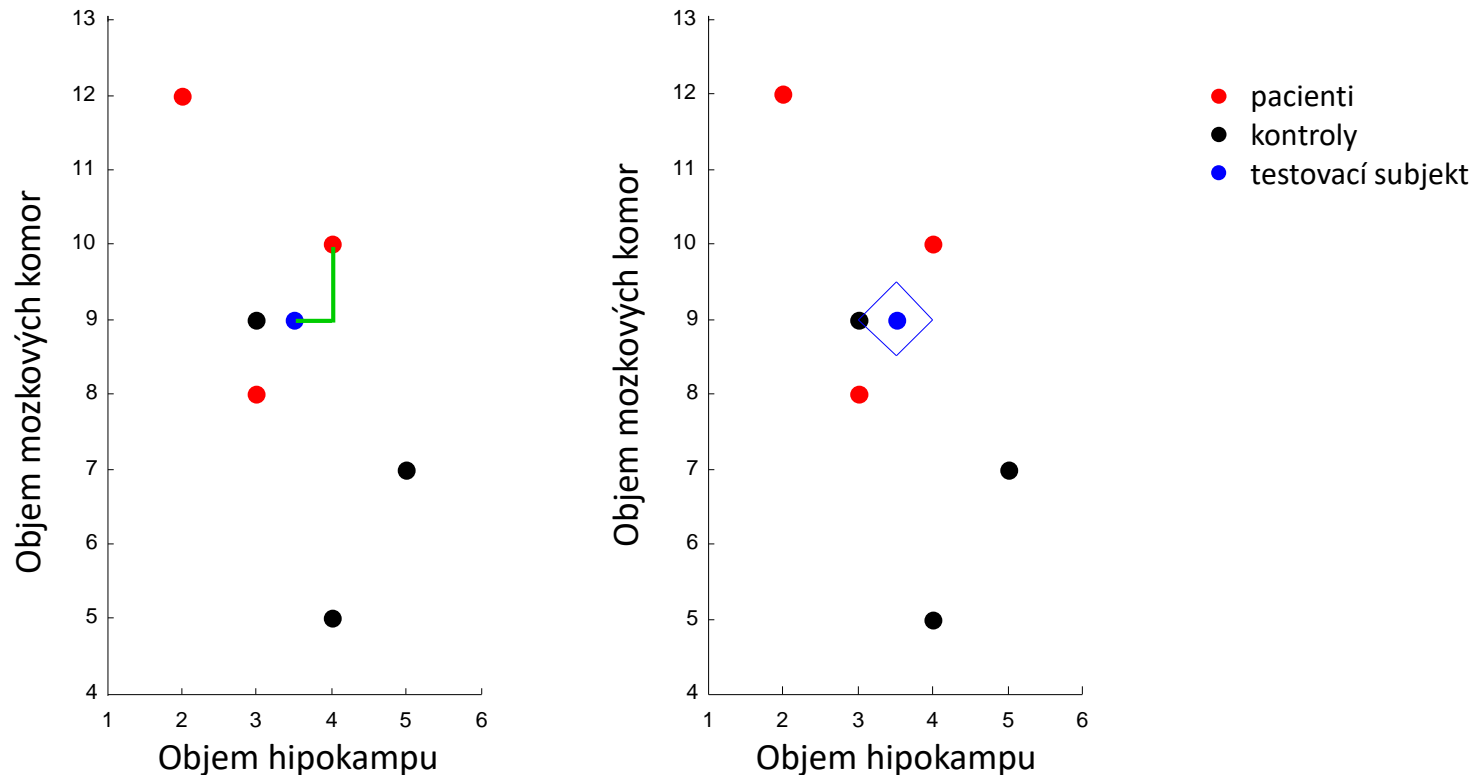
$$D_H(\mathbf{x}_1, \mathbf{x}_2) = \sum_{i=1}^n |x_{1i} - x_{2i}|$$



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

Hammingova (manhattanská) metrika

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



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

Minkovského metrika

- zobecněním Euklidovy a Hammingovy (manhattanské) metriky

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

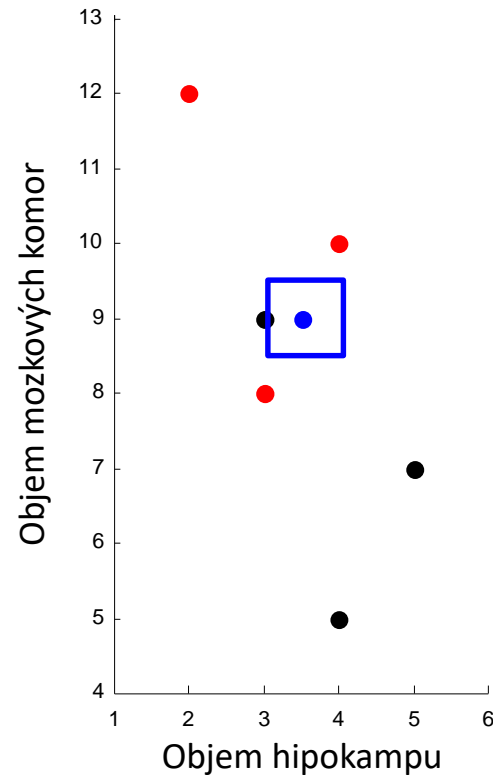
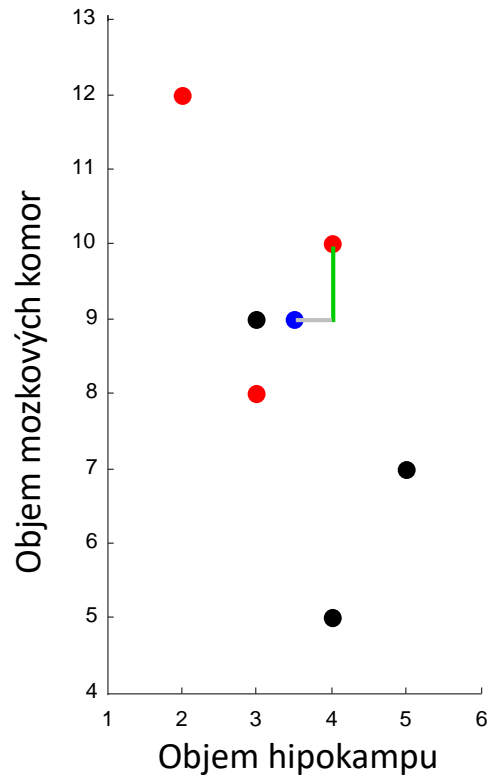
- Euklidova metrika pro $m = 2$, Hammingova (manhattanská) metrika pro $m = 1$
- volba m závisí na tom, jak moc chceme váhovat velké rozdíly mezi proměnnými (čím větší m , tím větší váha na velké rozdíly mezi proměnnými)
- pro $m \rightarrow \infty$ metrika konverguje k **Čebyševově metricě**

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

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

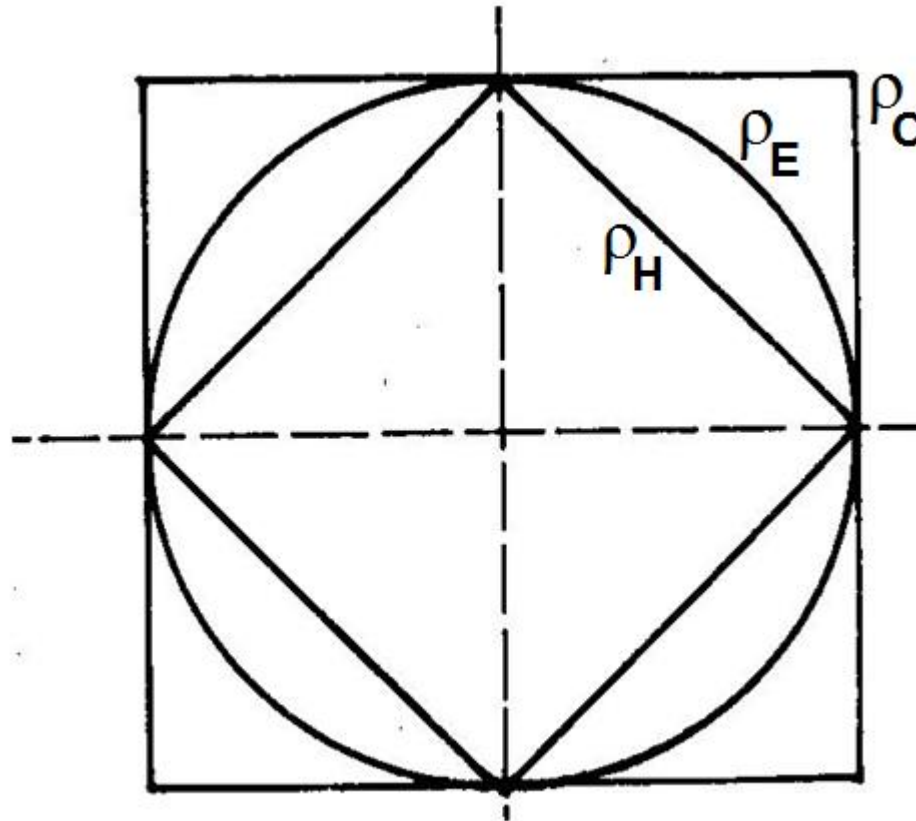
Čebyševova metrika

- 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

Srovnání metrik

- pokud je potřeba použít „euklidovskou“ metriku, ale s nižší výpočetní náročností, používá se v první řadě Hammingova nebo Čebyševova metrika
- případně kombinace obou metrik:

$$D_A(\mathbf{x}_1, \mathbf{x}_2) = \max(2D_H/3; D_C)$$

- geometrickým místem bodů s toutéž vzdáleností je pak ve dvourozměrném prostoru osmiúhelník

Nevýhody metrik

- je zpravidla problematické vytvářet součet rozdílů veličin s různým 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 $\Delta_j = \max_i x_{ij} - \min_i x_{ij}$) 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**: $D_{WM}(\mathbf{x}_1, \mathbf{x}_2) = \left(\sum_{i=1}^n a_i \cdot |x_{1i} - x_{2i}| \right)^2$
 3. začlenění kovarianční matice do výpočtu:
 - začleněním inverze kovarianční matice získáváme **Mahalanobisovu metriku** (což je Euklidova metrika váhovaná inverzí kovarianční matice):
$$D_{MA}(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{(\mathbf{x}_1 - \mathbf{x}_2)^T \cdot \mathbf{S}^{-1} \cdot (\mathbf{x}_1 - \mathbf{x}_2)}$$

Canberrská metrika

- relativizovaná varianta Hammingovy (manhattanské) metriky

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

- vhodná pro proměnné s nezápornými hodnotami
- pokud se vyskytují 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 objekty nacházejí v blízkosti počátku souřadnicové soustavy; naopak méně citlivá na změny hodnot proměnných, pokud jsou tyto hodnoty velké

Příklad 1a

Jsou dány dva vektory $\mathbf{x}_1 = (0,001; 0,001)^T$ a $\mathbf{x}_2 = (0,01; 0,01)^T$. Předpokládejme, že se souřadnice prvního vektoru změní na $\mathbf{x}'_1 = (0,002; 0,001)^T$. Jaká je Hammingova (manhattanská) a canberrská vzdálenost v obou případech a jaká je relativní změna vzdáleností, vyvolaná uvedenou modifikací?

$$d_H(\mathbf{x}_1, \mathbf{x}_2) = |0,001 - 0,01| + |0,001 - 0,01| = 0,009 + 0,009 = 0,018$$

$$d_H(\mathbf{x}'_1, \mathbf{x}_2) = |0,002 - 0,01| + |0,001 - 0,01| = 0,008 + 0,009 = 0,017$$

$$d_{CA}(\mathbf{x}_1, \mathbf{x}_2) = \frac{|0,001-0,01|}{|0,001|+|0,01|} + \frac{|0,001-0,01|}{|0,001|+|0,01|} = \frac{0,009}{0,011} + \frac{0,009}{0,011} = 1,6364$$

$$d_{CA}(\mathbf{x}'_1, \mathbf{x}_2) = \frac{|0,002-0,01|}{|0,002|+|0,01|} + \frac{|0,001-0,01|}{|0,001|+|0,01|} = \frac{0,008}{0,012} + \frac{0,009}{0,011} = 1,4849$$

Relativní změny vzdáleností, určující citlivost té které metriky, které jsou způsobeny změnou hodnoty první souřadnice, jsou:

$$\Delta d_H = \frac{|d_H(\mathbf{x}_1, \mathbf{x}_2) - d_H(\mathbf{x}'_1, \mathbf{x}_2)|}{d_H(\mathbf{x}_1, \mathbf{x}_2)} = \frac{|0,018 - 0,017|}{0,018} = \frac{0,001}{0,018} = 0,056$$

$$\Delta d_{CA} = \frac{|d_{CA}(\mathbf{x}_1, \mathbf{x}_2) - d_{CA}(\mathbf{x}'_1, \mathbf{x}_2)|}{d_{CA}(\mathbf{x}_1, \mathbf{x}_2)} = \frac{|1,6364 - 1,4849|}{1,6364} = 0,093$$

Ze získaných výsledků je zřejmé, že relativní změna vzdáleností je v případě canberrské metriky pro toto zadání téměř dvakrát větší.

Příklad 1b

Nyní mějme dány dva vektory $\mathbf{x}_1 = (1000; 1000)^T$ a $\mathbf{x}_2 = (100; 100)^T$ a předpokládejme, že se souřadnice prvního vektoru změní na $\mathbf{x}'_1 = (1002; 1000)^T$. Jaká je Hammingova (manhattanská) a canberrská vzdálenost v obou případech a jaká je relativní změna vzdáleností, vyvolaná uvedenou modifikací?

$$d_H(\mathbf{x}_1, \mathbf{x}_2) = |1000 - 100| + |1000 - 100| = 900 + 900 = 1800$$

$$d_H(\mathbf{x}'_1, \mathbf{x}_2) = |1002 - 100| + |1000 - 100| = 902 + 900 = 1802$$

$$d_{CA}(\mathbf{x}_1, \mathbf{x}_2) = \frac{|1000-100|}{|1000|+|100|} + \frac{|1000-100|}{|1000|+|100|} = \frac{900}{1100} + \frac{900}{1100} = 1,6364$$

$$d_{CA}(\mathbf{x}'_1, \mathbf{x}_2) = \frac{|1002-100|}{|1002|+|100|} + \frac{|1000-100|}{|1000|+|100|} = \frac{902}{1102} + \frac{900}{1100} = 1,6367$$

Relativní změny vzdáleností, určující citlivost té které metriky, které jsou způsobeny změnou hodnoty první souřadnice, jsou:

$$\Delta d_H = \frac{|d_H(\mathbf{x}_1, \mathbf{x}_2) - d_H(\mathbf{x}'_1, \mathbf{x}_2)|}{d_H(\mathbf{x}_1, \mathbf{x}_2)} = \frac{|1800 - 1802|}{1800} = \frac{2}{1800} = 0,0011$$

$$\Delta d_{CA} = \frac{|d_{CA}(\mathbf{x}_1, \mathbf{x}_2) - d_{CA}(\mathbf{x}'_1, \mathbf{x}_2)|}{d_{CA}(\mathbf{x}_1, \mathbf{x}_2)} = \frac{|1,6364 - 1,6367|}{1,6364} = 0,00018$$

Ze získaných výsledků je zřejmé, že citlivost canberrské metriky je v tomto případě řádově nižší.

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
- i když existují doporučení, jak volit obě hodnoty na základě statistických vlastností vektorového prostoru (např. pomocí $H = \frac{\Gamma(n/2)}{D^n \sqrt{\pi^n}}$), výhodnější je volit obě hodnoty 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



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

Hammingova metrika vzdálenosti

- pro $k = 2$, kdy jsou hodnoty obou vektorů binární, se definiční vztah Hammingovy vzdálenosti transformuje na:

$$D_{HQB}(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^p (x_i + y_i - 2x_i y_i)$$

kde třetí člen v závorce kompenzuje případ, kdy jsou hodnoty x_i i y_i rovny jedné a součet prvních členů v závorce je tím pádem roven dvěma, nicméně nastává shoda hodnot, která k celkové vzdálenosti nemůže přispět.

- protože x_i a y_i nabývají hodnot pouze 0 a 1, můžeme také psát:

$$D_{HQB}(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^p (x_i^2 + y_i^2 - 2x_i y_i) = \sum_{i=1}^p (x_i - y_i)^2$$

- díky speciálnímu případu hodnot x_i a y_i je možná i nejjednodušší forma:

$$D_{HQB}(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^p |x_i - y_i|$$

Hammingova metrika vzdálenosti – příklad 2

Určete Hammingovu vzdálenost binárních vektorů

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

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

Podle definičního principu (tzn. počet pozic, ve kterých se oba vektory liší):

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

$$\text{Dle jiného vztahu: } d_{HQB}(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^p (x_i + y_i - 2x_i y_i) =$$

$$= (0+1-2 \cdot 0 \cdot 1) + (1+0-2 \cdot 1 \cdot 0) + (1+0-2 \cdot 1 \cdot 0) + (0+0-2 \cdot 0 \cdot 0) + (1+1-2 \cdot 1 \cdot 1) = 3$$

$$\text{Dle dalšího vztahu: } d_{HQB}(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^p (x_i - y_i)^2 =$$

$$= (0-1)^2 + (1-0)^2 + (1-0)^2 + (0-0)^2 + (1-1)^2 = 1+1+1+0+0 = 3$$

$$\text{Dle posledního vztahu: } d_{HQB}(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^p |x_i - y_i| =$$

$$= |0-1| + |1-0| + |1-0| + |0-0| + |1-1| =$$

$$= 1+1+1+0+0 = 3$$

Hammingova metrika vzdálenosti

V případě bipolárních vektorů, kdy jednotlivé složky vektorů nabývají hodnot +1 a -1, je Hammingova vzdálenost určena vztahem:

$$D_{HQP}(\mathbf{x}, \mathbf{y}) = \frac{\left(p - \sum_{i=1}^p x_i y_i \right)}{2}$$

Příklad 3:

Určete Hammingovu vzdálenost bipolárních vektorů

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

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

Podle definičního principu (tzn. počet pozic, ve kterých se liší): $d_{HQP}(\mathbf{x}, \mathbf{y}) = 2$

Z kontingenční matice (součet prvků mimo hlavní diagonálu): $\mathbf{A}(\mathbf{x}, \mathbf{y}) = \begin{bmatrix} 2 & 2 \\ 0 & 1 \end{bmatrix}$

Pomocí vztahu:

$$d_{HQP}(\mathbf{x}, \mathbf{y}) = \frac{5 - ((1 \cdot 1) + (1 \cdot (-1)) + (1 \cdot 1) + ((-1) \cdot (-1)) + (1 \cdot (-1)))}{2} = \frac{5 - (1 - 1 + 1 + 1 - 1)}{2} = \frac{5 - 1}{2} = 2$$

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

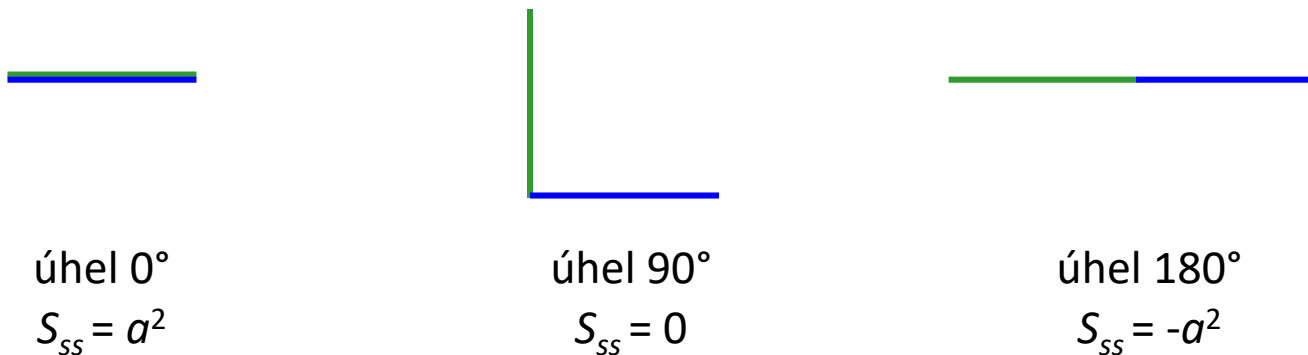
Metriky pro určení vzdálenosti mezi 2 množinami objektů používající jejich pravděpodobnostní charakteristiky

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:



- 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

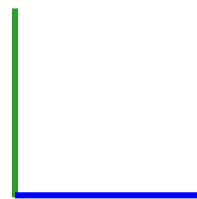
$$S_{\cos}(\mathbf{x}_1, \mathbf{x}_2) = \frac{\mathbf{x}_1^T \cdot \mathbf{x}_2}{\|\mathbf{x}_1\| \cdot \|\mathbf{x}_2\|}$$

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 $S_{\cos}(\mathbf{x}_1, \mathbf{x}_2)$ 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

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

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

\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

Tanimotova metrika podobnosti

$$S_T(\mathbf{x}_1, \mathbf{x}_2) = \frac{\mathbf{x}_1^T \mathbf{x}_2}{\|\mathbf{x}_1\|^2 + \|\mathbf{x}_2\|^2 - \mathbf{x}_1^T \mathbf{x}_2}$$

Přičteme-li a odečteme-li ve jmenovateli výraz $\mathbf{x}_1^T \mathbf{x}_2$ a podělíme-li čitatele i jmenovatele zlomku toutéž hodnotou, dostaneme

$$S_T(\mathbf{x}_1, \mathbf{x}_2) = \frac{1}{1 + \frac{(\mathbf{x}_1 - \mathbf{x}_2)^T (\mathbf{x}_1 - \mathbf{x}_2)}{\mathbf{x}_1^T \mathbf{x}_2}}$$

Tanimotova podobnost vektorů \mathbf{x}_1 a \mathbf{x}_2 je nepřímo úměrná kvadrátu Euklidovy vzdálenosti vektorů \mathbf{x}_1 a \mathbf{x}_2 vztahené k jejich skalárnímu součinu. Pokud skalární součin považujeme za míru korelace obou vektorů, můžeme formulovat výše uvedený vztah tak, že $S_T(\mathbf{x}_1, \mathbf{x}_2)$ je nepřímo úměrná kvadrátu Euklidovy vzdálenosti podělené velikostí jejich korelace, což znamená, že je korelaci, jako míře podobnosti přímo úměrná.

„Bezejmenná“ metrika podobnosti

$$S_C(\mathbf{x}_1, \mathbf{x}_2) = 1 - \frac{D_E(\mathbf{x}_1, \mathbf{x}_2)}{\|\mathbf{x}_1\| + \|\mathbf{x}_2\|}$$

Vzdálenost podle metriky je rovna jedné,

když $\mathbf{x}_1 = \mathbf{x}_2$

a svého minima (tj. $S_C(\mathbf{x}_1, \mathbf{x}_2) = -1$) nabývá,

když $\mathbf{x}_1 = -\mathbf{x}_2$.

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



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

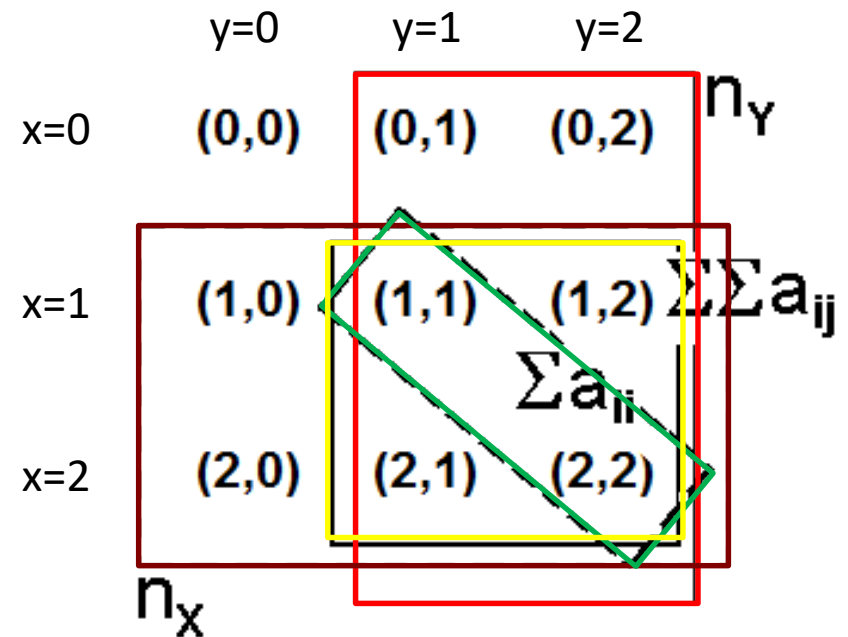
Obečné metriky – Tanimotova metrika

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

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

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

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:

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

Tento vztah se dominantně používá v ekologických studiích.

Další asociační koeficienty I

		\mathbf{x}_j	
		false/0	true/1
\mathbf{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 metrika vzdálenosti
- 'squaredeuclidean' – čtverec Euklidovy metriky vzdálenosti
- 'seuclidean' – standardizovaná Euklidova metrika vzdálenosti
- 'cityblock' – Hammingova (manhattanská) metrika vzdálenosti
- 'minkowski' – Minkovského metrika vzdálenosti
- 'chebychev' – Čebyševova metrika vzdálenosti
- 'mahalanobis' – Mahalanobisova metrika vzdálenosti
- 'cosine' – 1 mínus kosinová podobnost
- 'correlation' – 1 mínus Pearsonův korelační koeficient
- 'spearman' – 1 mínus Spearmanův korelační koeficient
- 'hamming' – Hamminova vzdálenost (pro kvalitativní proměnné)
- 'jaccard' – 1 mínus Jaccardův koeficient
- lze případně nadefinovat i jinou metriku

Výpočet vzdáleností v R

Funkce *dist* na výpočet vzdáleností objektů (či subjektů) s výběrem metrik:

- „euclidean“ – Euklidovska metrika
- „maximum“ – Čebyševova metrika
- „manhattan“ – Hammingova (manhattanská) metrika
- „canberra“ – Canberrská metrika
- „minkowski“ – Minkovského metrika

Metriky pro určení vzdálenosti mezi dvěma skupinami objektů

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) objektů či „vzdáleností“ jednoho objektu z každé skupiny – použitelné při shlukování
- zavedeme funkci, která ke každé dvojici skupin objektů (C_i, C_j) přiřazuje číslo $D(C_i, C_j)$, které podobně jako míry podobnosti či nepodobnosti (metriky) jednotlivých objektů musí splňovat minimálně podmínky:
 - (S1) $D(C_i, C_j) \geq 0$
 - (S2) $D(C_i, C_j) = D(C_j, C_i)$
 - (S3) $D(C_i, C_i) = \max_{i,j} D(C_i, C_j)$ (pro míry podobnosti)
 - (S3') $D(C_i, C_i) = 0$ pro všechna i (pro míry vzdálenosti)

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

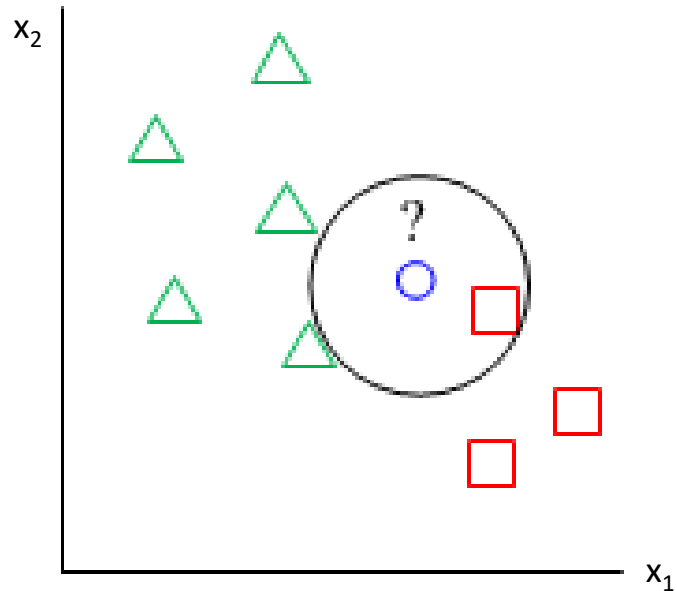
Chernoffova m., Bhattacharyyova m. atd.

Nejpoužívanější metriky pro určení vzdálenosti mezi dvěma množinami objektů

- Metoda nejbližšího souseda
- Metoda k nejbližších sousedů
- Metoda nejvzdálenějšího souseda
- Metoda průměrné vazby
- Wardova metoda

Metoda nejbližšího souseda

- je-li d libovolná míra nepodobnosti (vzdálenosti) dvou objektů a ω_i a ω_j jsou libovolné skupiny objektů, potom metoda nejbližšího souseda definuje mezi skupinami ω_i a ω_j vzdálenost
$$D_{NN}(\omega_i, \omega_j) = \min_{\substack{x_p \in \omega_i \\ x_q \in \omega_j}} d(x_p, x_q)$$



- pacienti
- △ kontroly
- testovací subjekt

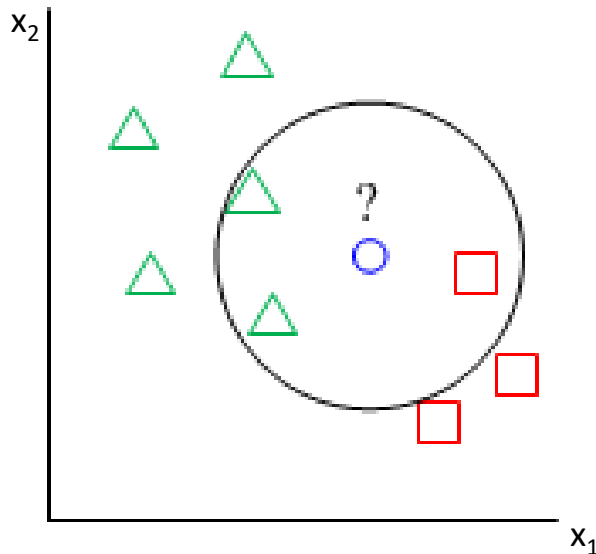
→ testovací subjekt zařadíme do třídy, ze které je jeho nejbližší sused

- výhody a nevýhody použití této metody pro klasifikaci:
 - + žádné předpoklady o rozložení
 - citlivé na odlehlé hodnoty
 - zpravidla nevhodné při nevyvážených počtech objektů ve skupinách

Metoda k nejbližších sousedů

- zobecněním metody nejbližšího souseda
- definována vztahem $D_{NNk}(\omega_i, \omega_j) = \min_{\substack{x_p \in \omega_i \\ x_q \in \omega_j}} \sum_{k} d(x_p, x_q)$, tzn. vzdálenost dvou

shluků je definována součtem nejkratších vzdáleností mezi objekty obou skupin



- pacienti
- △ kontroly
- testovací subjekt

→ testovací subjekt zařadíme do třídy, která převažuje mezi jeho k nejbližšími sousedy

- výhody a nevýhody použití této metody pro klasifikaci:
 - + žádné předpoklady o rozložení
 - + méně citlivé na odlehlé hodnoty
 - zpravidla nevhodné při nevyvážených počtech objektů ve skupinách

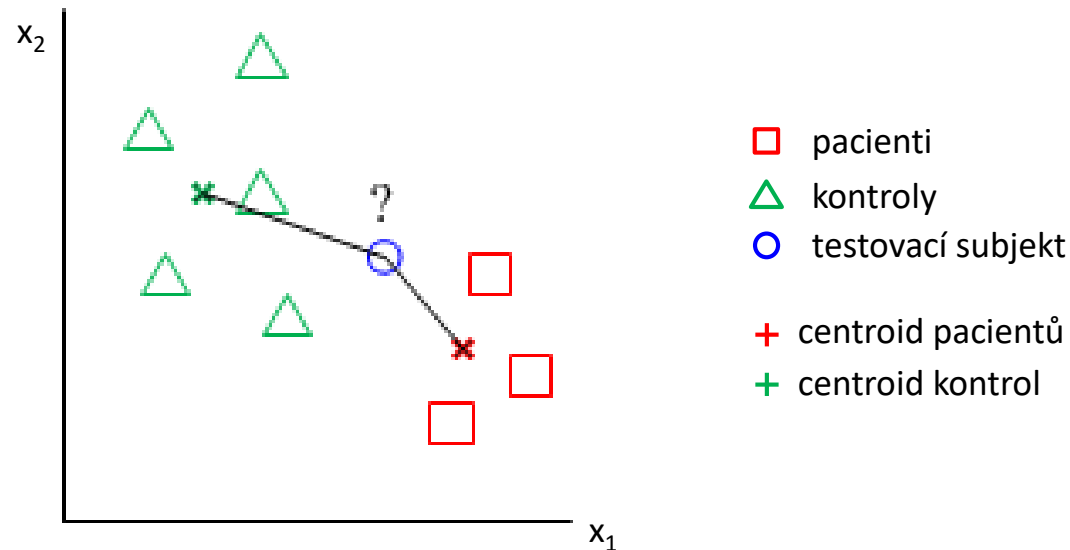
Metoda nejvzdálenějšího souseda

- opačný princip než metoda nejbližšího souseda: $D_{\text{FN}}(C_i, C_j) = \max_{\substack{x_p \in C_i \\ x_q \in C_j}} d(x_p, x_q)$
- pozn.: pro klasifikaci je obtížně použitelná
- pozn. 2: je možné zobecnění i pro více nejvzdálenějších sousedů

$$D_{\text{FNk}}(C_i, C_j) = \max_{\substack{x_p \in C_i \\ x_q \in C_j}} \sum^k d(x_p, x_q),$$

Centroidová metoda

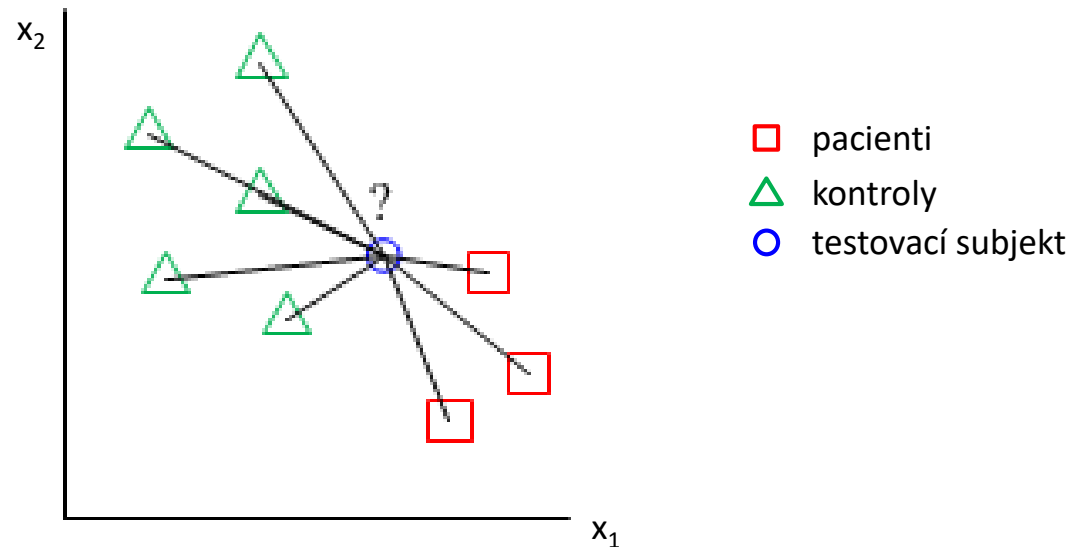
- vychází z výpočtu centroidů pro jednotlivé skupiny
- při klasifikaci: zařazení objektu do skupiny s nejbližším centroidem



- výhody a nevýhody použití této metody pro klasifikaci:
 - + žádné předpoklady o rozložení (pokud je však centroid počítán jako vícerozměrný průměr, může nenormalita působit trochu problémy)
 - + méně citlivé na odlehlé hodnoty než metoda nejbližšího souseda
 - + nebývá problém při nevyvážených počtech objektů ve skupinách

Metoda průměrné vazby

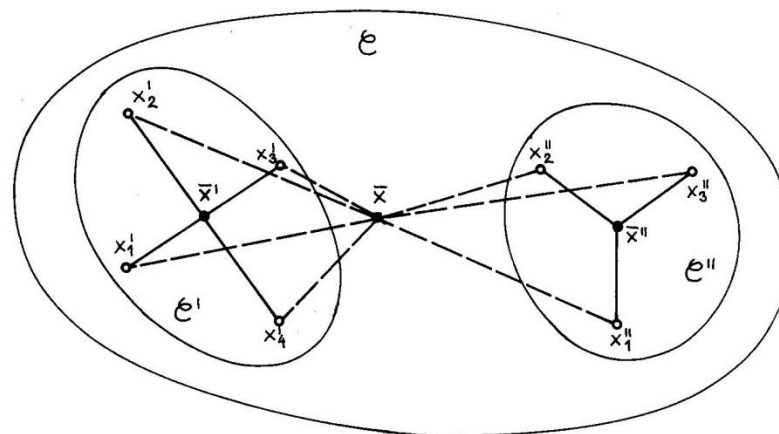
- vzdálenost dvou tříd je průměrná vzdálenost mezi všemi objekty těchto tříd
- při klasifikaci: zařazení subjektu do skupiny s nejmenší průměrnou vzdáleností od všech objektů dané skupiny



- výhody a nevýhody použití této metody pro klasifikaci:
 - + žádné předpoklady o rozložení
 - + méně citlivé na odlehlé hodnoty než metoda nejbližšího souseda
 - + nebývá problém při nevyvážených počtech objektů ve skupinách
 - časově náročnější než centroidová metoda při větším počtu objektů

Wardova metoda

- vzdálenost mezi třídami (shluky) je definována přírůstkou součtu čtverců odchylek mezi těžištěm a objekty shluku vytvořeného z obou uvažovaných shluků C_i a C_j oproti součtu čtverců odchylek mezi objekty a těžišti v obou shlucích C_i a C_j .
- pozn. (při použití Wardovy metody pro shlukování): Metoda má tendenci vytvářet shluky zhruba stejné velikosti, tedy odstraňovat shluky malé, resp. velké.
- pozn. 2: pro klasifikaci se používá zřídka

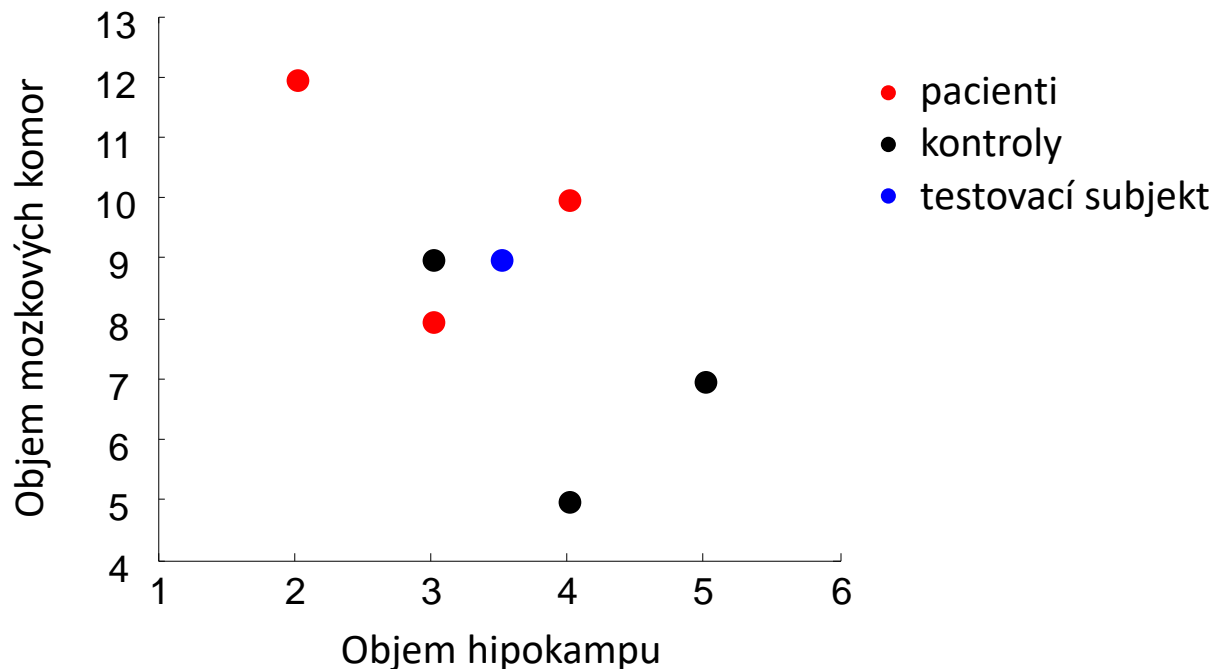


Příklad 2

Bylo provedeno měření objemu hipokampu a mozkových komor (v cm³) u

3 pacientů se schizofrenií a 3 kontrol: $\mathbf{X}_D = \begin{bmatrix} 2 & 12 \\ 4 & 10 \\ 3 & 8 \end{bmatrix}$, $\mathbf{X}_H = \begin{bmatrix} 5 & 7 \\ 3 & 9 \\ 4 & 5 \end{bmatrix}$.

Určete, zda testovací subjekt $\mathbf{x} = [3,5 \quad 9]$ patří do skupiny pacientů či kontrolních subjektů pomocí různých metod klasifikace podle minimální vzdálenosti.



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 psních charakteristikách

Klasifikační třídy (množiny objektů se společnými charakteristikami) nemusí být definovány jen výčtem objektů, ale i vymezením obecnějších vlastností:

- definicí hranic oddělujících část obrazového prostoru náležející dané klasifikační třídě
- diskriminační funkcí
- **pravděpodobnostními charakteristikami výskytu objektů v dané třídě**
- atd.

Metriky založené na pstních charakteristikách

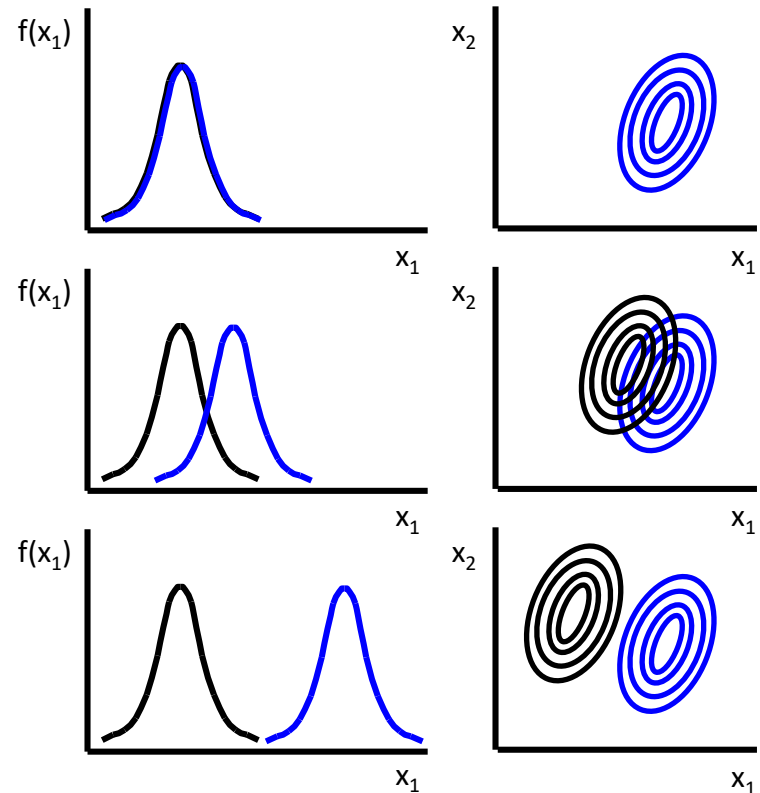
Základní myšlenkou je využití **pravděpodobnosti způsobené chybou při klasifikaci** (tzn. zařazení objektu do skupiny). Čím více se hustoty pravděpodobnosti výskytu objektů \mathbf{x} v jednotlivých množinách překrývají, tím je větší pravděpodobnost chyby.

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
$$\int_{-\infty}^{\infty} p(\mathbf{x}|\omega_1) \cdot p(\mathbf{x}|\omega_2) d\mathbf{x} = 0$$



(Jak vidíme, není mezi vlastnostmi pravděpodobnostních metrik uvedena trojúhelníková nerovnost, jejíž splnění by se zajišťovalo obtížně.)

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.

Příprava nových učebních materiálů pro obor Matematická biologie

je podporována projektem OPVK

č. CZ.1.07/2.2.00/28.0043

„Interdisciplinární rozvoj studijního
oboru Matematická biologie“



MINISTERSTVO ŠKOLSTVÍ,
MLÁDEŽE A TĚLOVÝCHOVY



OP Vzdělávání
pro konkurenceschopnost

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ