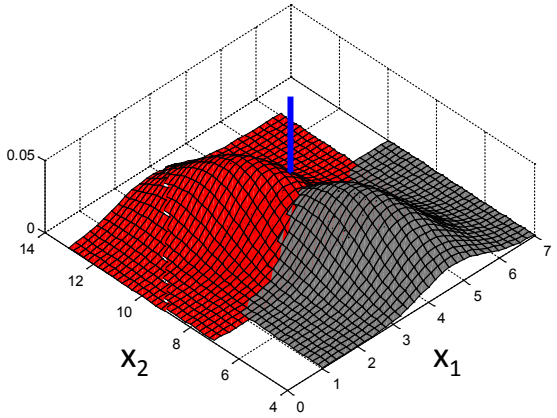
Analýza a klasifikace dat – přednáška 3

RNDr. Eva Janoušová

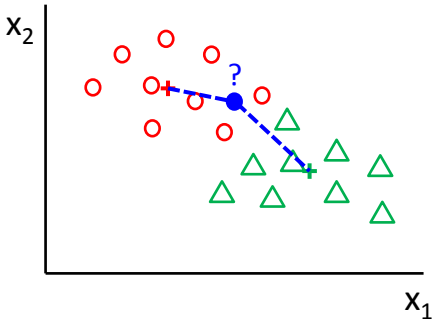
Podzim 2014

Typy klasifikátorů – podle principu klasifikace

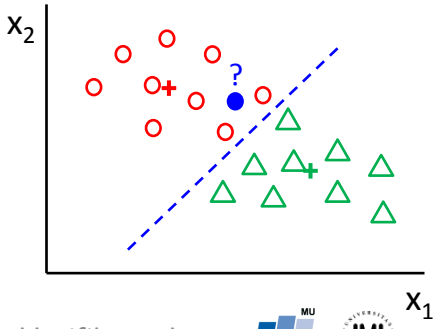
- **klasifikace pomocí diskriminačních funkcí:**
  - diskriminační funkce určují míru příslušnosti k dané klasifikační třídě
  - pro danou třídu má daná diskriminační funkce nejvyšší hodnotu



- **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)



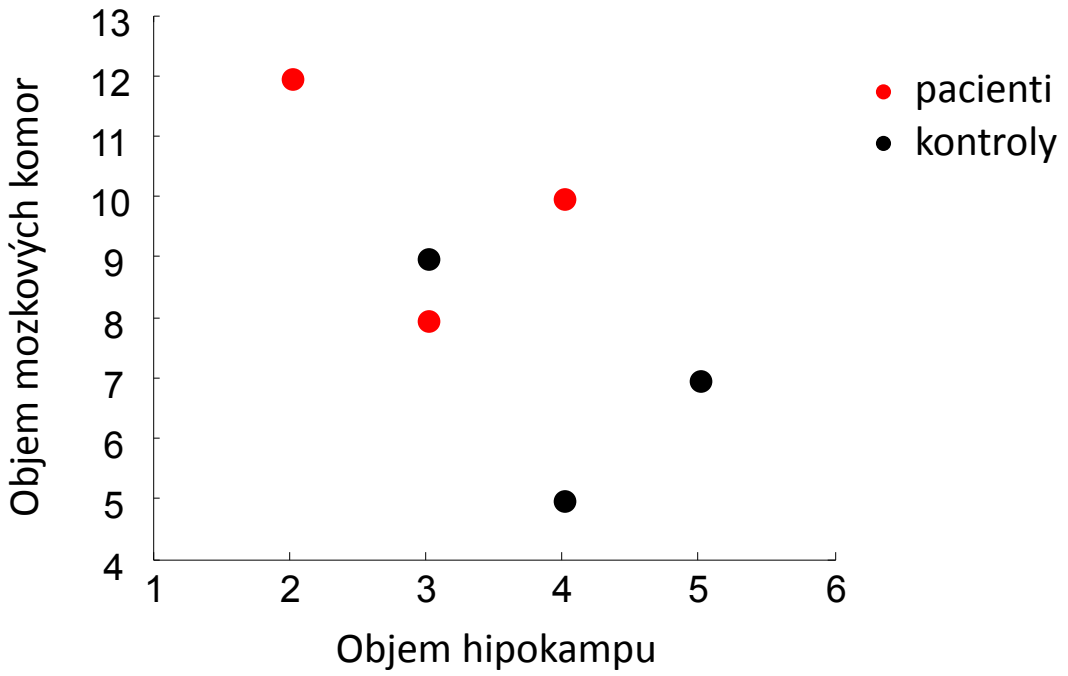
- **klasifikace pomocí hranic v obrazovém prostoru:**
  - stanovení hranic (hraničních ploch) oddělujících klasifikační třídy



Poznámka

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

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



Metrika – vzdálenost

Metrika ρ na X je funkce $\rho: X \times X \rightarrow R$, kde R je množina reálných čísel, taková, že:

$$\exists \rho_0 \in R: -\infty < \rho_0 \leq \rho(x,y) < +\infty, \forall x,y \in X$$

$$\rho(x,x) = \rho_0, \forall x \in X$$

a

$$\rho(x,y) = \rho(y,x), \forall x,y \in X \textit{ (symetrie)}$$

$$\rho(x, y) = \rho_0 \text{ když a jen když } x = y \textit{ (totožnost)}$$

$$\rho(x, z) \leq \rho(x, y) + \rho(y, z), \forall x,y,z \in X \textit{ (Δ nerovnost)}$$

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

Vzdálenost je hodnota určená podle metriky.

Poznámka: zpravidla $\rho_0=0$.

Metrika – podobnost

Metrická míra podobnosti s na X je funkce $\rho: X \times X \rightarrow R$, taková, že:

$$\exists s_0 \in R: -\infty < s(x,y) \leq s_0 < +\infty, \forall x,y \in X$$

$$s(x,x) = s_0, \forall x \in X$$

a

$$s(x,y) = s(y,x), \forall x,y \in X \textit{ (symetrie)}$$

$$s(x,y) = s_0 \text{ když a jen když } x = y \textit{ (totožnost)}$$

$$s(x,y).s(y,z) \leq [s(x,y) + s(y,z)].s(x,z), \forall x,y,z \in X$$

Podobnost je hodnota určená podle metrické míry podobnosti.

Metriky podobnosti vs. metriky vzdálenosti

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

$$s_{ij} = 1/\rho_{ij}$$

$$s_{ij} = 1/(1+ \rho_{ij})$$

$$s_{ij} = c - \rho_{ij},\ c \geq \max \rho_{ij},\ \forall i,j$$

Typy měr vzdálenosti (podobnosti)

- dle **typu proměnné** (kvalitativní proměnné, kvantitativní proměnné)
- dle **objektů**, jejichž vztah hodnotíme – obrazy (vektory), množiny obrazů (vektorů), rozdělení
- **deterministické** (nepravděpodobností) vs. **pravděpodobností míry**

- výběr konkrétní metriky závisí na:
  - výpočetních nárocích
  - charakteru rozložení dat
  - dosažení optimálních výsledků (klasifikační chyba, ztráta,...)
- obecně nelze doporučit vhodnou metriku pro danou situaci

Metriky pro určení vzdálenosti mezi dvěma obrazy s kvantitativními proměnnými

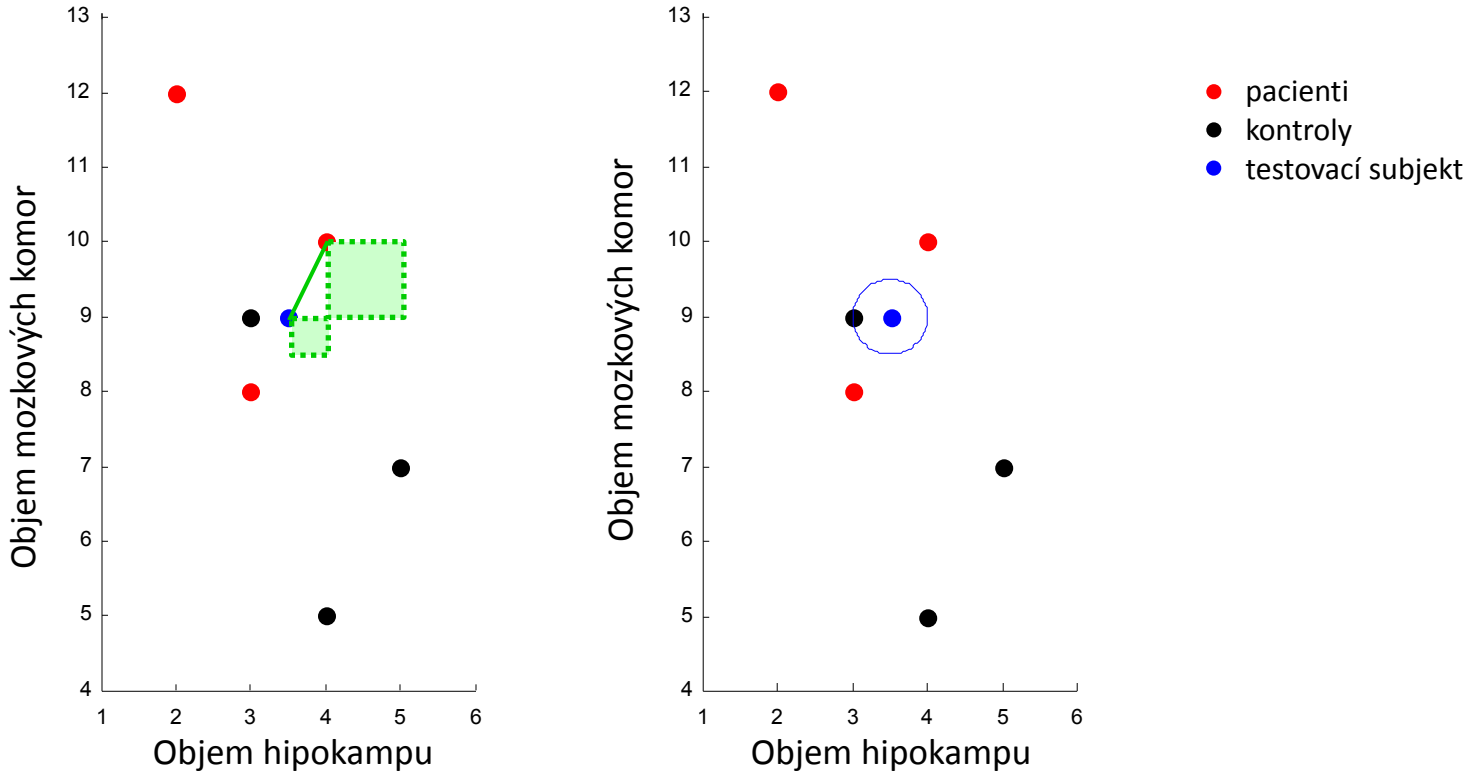
Nejpoužívanější metriky pro určení vzdálenosti mezi dvěma obrazy s kvantitativními proměnnými

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

Euklidova metrika

- zřejmě nejpoužívanější metrika s velmi názornou geometrickou interpretací

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



Euklidova metrika

- zřejmě nejpoužívanější metrika s velmi názornou geometrickou interpretací

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

- geometrickým místem bodů s toutéž Euklidovou vzdáleností od daného bodu je hyperkoule (ve dvourozměrném prostoru kruh)
- dává větší důraz na větší rozdíly mezi souřadnicemi

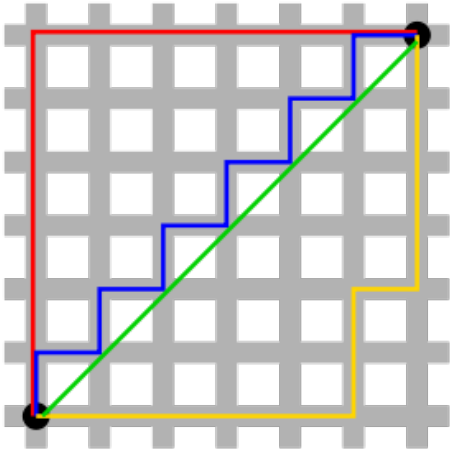
žádoucí nebo nežádoucí?

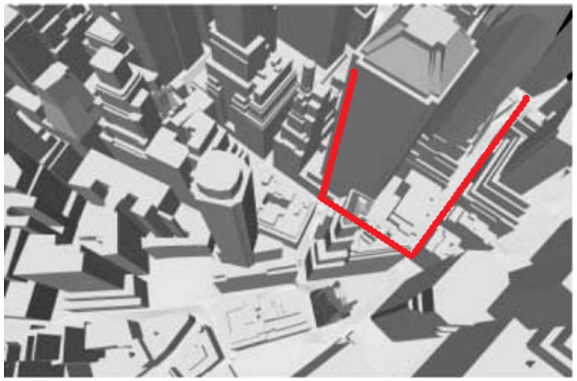
- čtverec euklidovské vzdálenosti (lépe se počítá než euklidovská vzdálenost) je mírou nepodobnosti, ale není metrikou

Hammingova (manhattanská) metrika

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

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



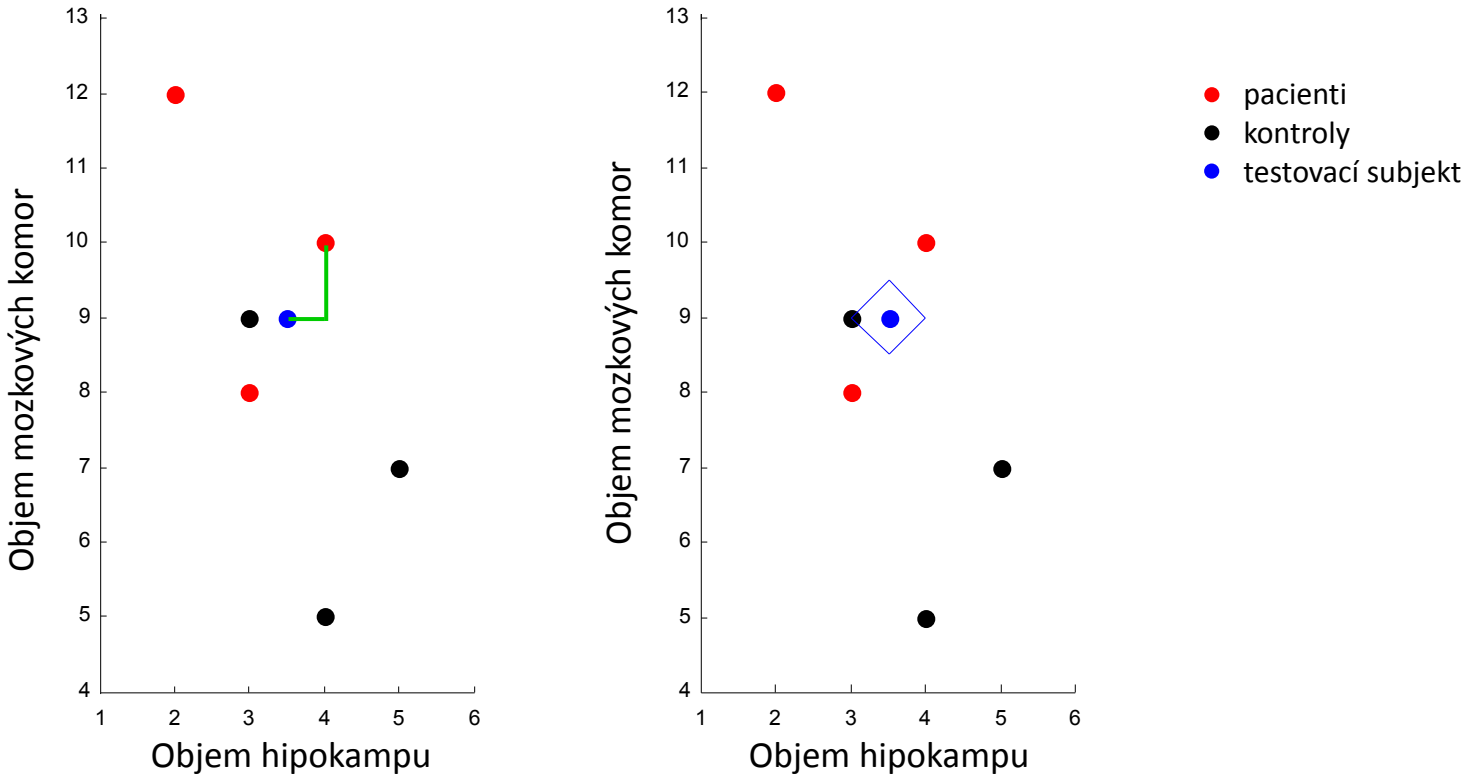


- nižší výpočetní nároky než Euklidova metrika → použití v úlohách s vysokou výpočetní náročností
- geometrickým místem bodů s toutéž manhattanskou vzdáleností od daného bodu je hyperkrychle (ve dvourozměrném prostoru čtverec)

Hammingova (manhattanská) metrika

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

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



Minkovského metrika

- zobecněním Euklidovy a Hammingovy (manhattanské) metriky

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

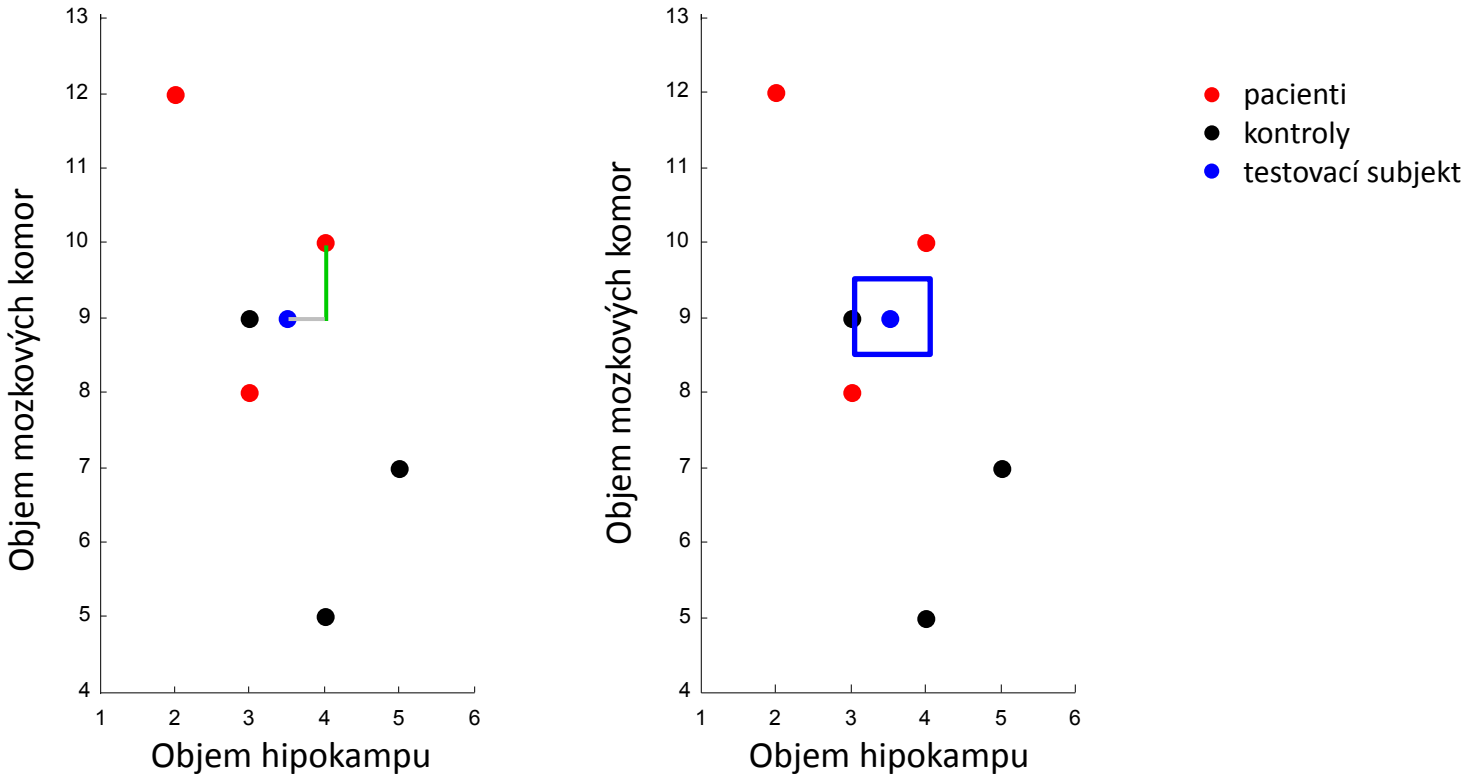
- Euklidova metrika pro $m = 2$, Hammingova (manhattanská) metrika pro $m = 1$
- volba m závisí na tom, jak moc chceme váhovat velké rozdíly mezi proměnnými (čím větší m, tím větší váha na velké rozdíly mezi proměnnými)
- pro $m \to \infty$ metrika konverguje k **Čebyševově metrice**

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

Čebyševova metrika

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

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



Čebyševova metrika

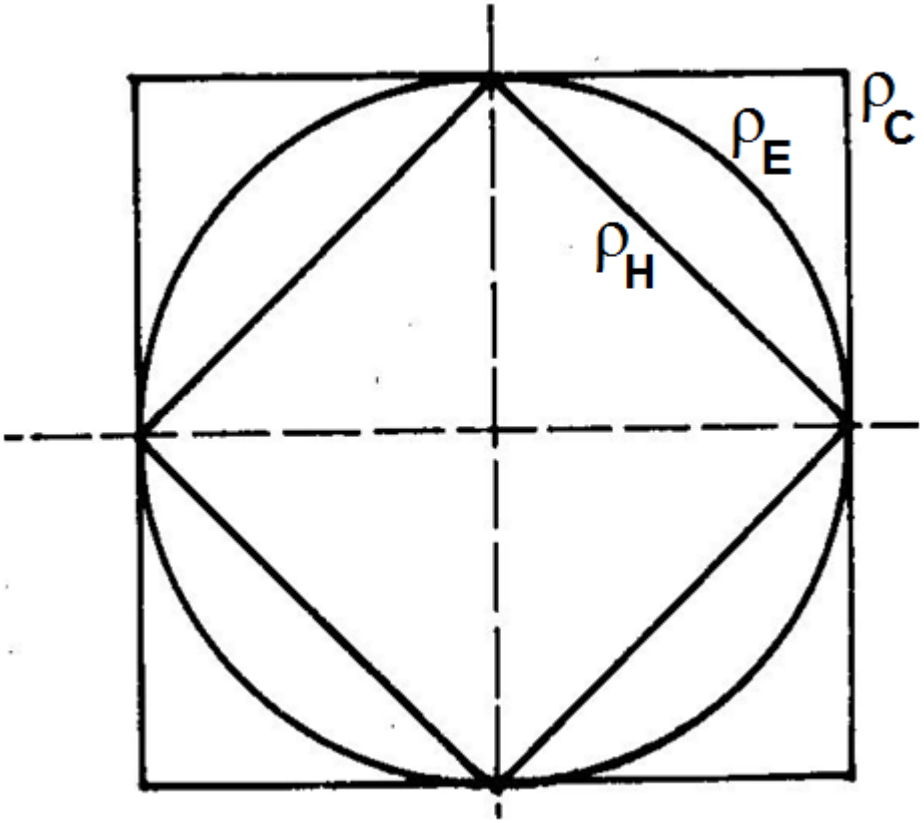
- odvozena z Minkovského metriky pro $m \to \infty$

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

- 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

- lepším přiblížením je kombinace obou metrik

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

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

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 Δ_i = $\max_j x_{ij}$ - $\min_j x_{ij}$) či pomocí standardizace $u_{ij} = \frac{x_{ij} - \bar{x}_j}{\sigma_j}$; $i = 1, \ldots, n; j = 1, \ldots, p$; 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. začleněním 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

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

- je 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 obrazy 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:

$$\Delta d_H = \frac{|d_H(\mathbf{x}_1,\mathbf{x}_2) - d_H(\mathbf{x}'_1,\mathbf{x}_2)|}{d_H(\mathbf{x}_1,\mathbf{x}_2)} = \frac{|0{,}018-0{,}017|}{0{,}018} = \frac{0{,}001}{0{,}018} = 0{,}056$$
$$\Delta d_H = \frac{|d_{CA}(\mathbf{x}_1,\mathbf{x}_2) - d_{CA}(\mathbf{x}'_1,\mathbf{x}_2)|}{d_{CA}(\mathbf{x}_1,\mathbf{x}_2)} = \frac{|1{,}6364-1{,}4849|}{1{,}6364} = 0{,}093$$

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:

$$\Delta d_H = \frac{|d_H(\mathbf{x}_1,\mathbf{x}_2) - d_H(\mathbf{x}'_1,\mathbf{x}_2)|}{d_H(\mathbf{x}_1,\mathbf{x}_2)} = \frac{|1800-1802|}{1800} = \frac{2}{1800} = 0{,}0011$$
$$\Delta d_H = \frac{|d_{CA}(\mathbf{x}_1,\mathbf{x}_2) - d_{CA}(\mathbf{x}'_1,\mathbf{x}_2)|}{d_{CA}(\mathbf{x}_1,\mathbf{x}_2)} = \frac{|1{,}6364-1{,}6367|}{1{,}6364} = 0{,}00018$$

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 (viz vzorec níže), výhodnější je volit obě hodnoty na základě expertní analýzy řešeného problému

$$H = \frac{\Gamma(n/2)}{D^n \sqrt{\pi^n}}$$

- ve vztahu může figurovat jakákoliv metrika vzdálenosti, nejen Euklidova metrika

Deterministické metriky pro určení vzdálenosti mezi dvěma množinami obrazů

Podobnost a vzdálenost mezi třídami

- podobnost mezi třídami dána:
  - „podobností“ jednoho obrazu s jedním či více obrazy jedné třídy (skupin, shluků) – použitelné při klasifikaci
  - „podobností“ skupin obrazů či „podobností“ jednoho obrazu z každé skupiny – použitelné při shlukování

- zavedeme funkci, která ke každé dvojici skupin obrazů (C_i, C_j) přiřazuje číslo D(C_i, C_j), které podobně jako míry podobnosti či nepodobnosti (metriky) jednotlivých obrazů 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)

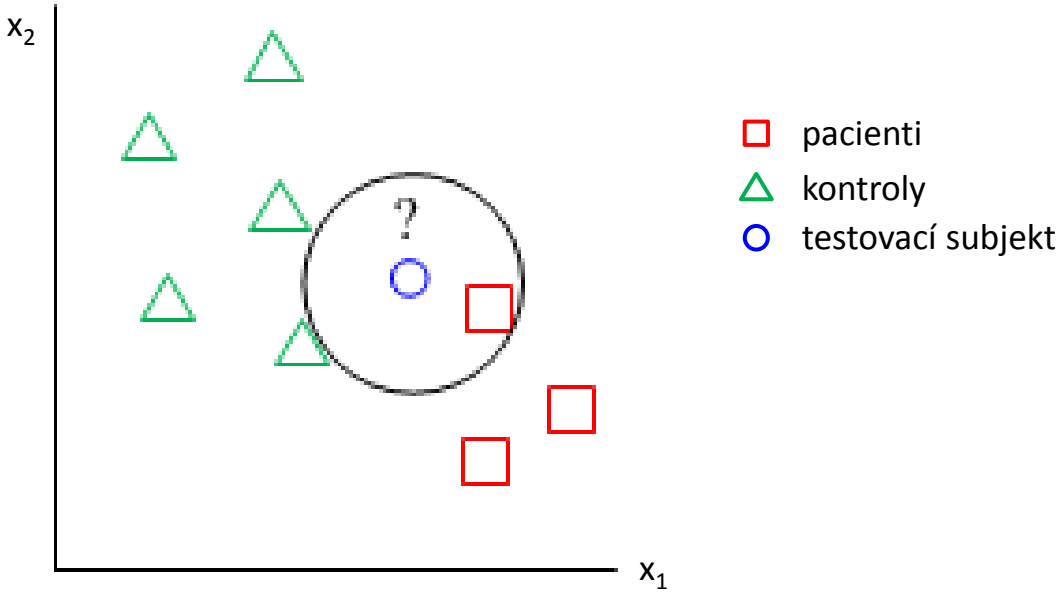
  (S3') $D(C_i, C_i) = 0$ pro všechna i (pro míry vzdálenosti)

Nejpoužívanější metriky pro určení vzdálenosti mezi dvěma množinami obrazů

- 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 obrazů a C_i a C_j jsou libovolné skupiny obrazů, potom metoda nejbližšího souseda definuje mezi skupinami C_i a C_j vzdálenost $D_{NN}(C_i, C_i) = \min_{\substack{x_p \in C_i \\ x_q \in C_j}} d(x_p, x_q)$

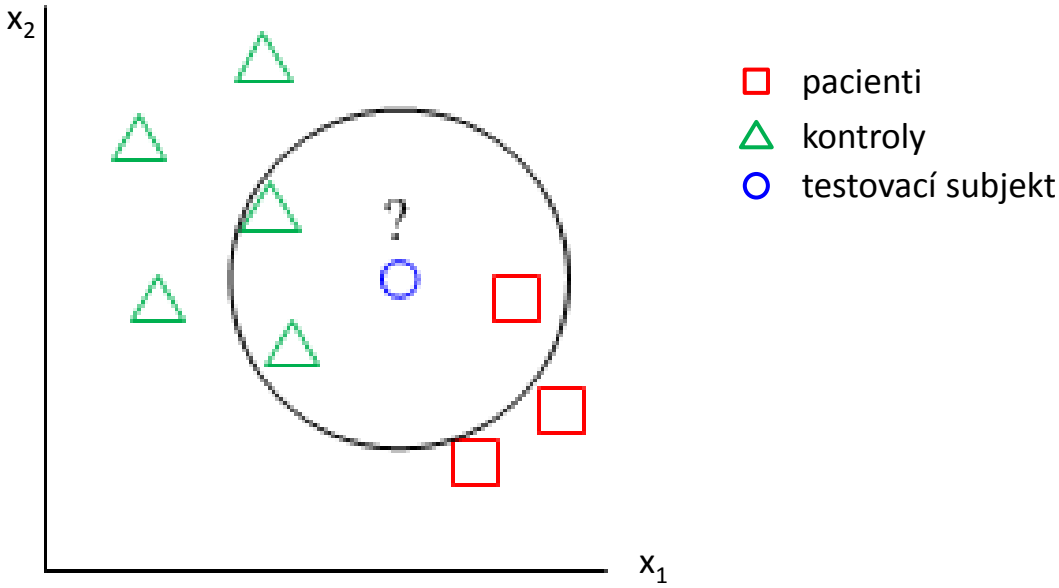


- pozn. (při použití metody nejbližšího souseda pro shlukování): Při použití této metody se mohou vyskytovat v jednom shluku často i poměrně vzdálené obrazy. Tzn. metoda nejbližšího souseda může generovat shluky protáhlého tvaru.

Metoda *k* nejbližších sousedů

- zobecněním metody nebližšího souseda
- definována vztahem $D_{NNk}(C_i, C_i) = \min_{\substack{x_p \in C_i \\ x_q \in C_j}} \sum^{k} d(x_p, x_q)$, tzn. vzdálenost dvou

  shluků je definována součtem nejkratších vzdáleností mezi obrazy obou skupin



- pozn. (při použití metody *k* nejbližších sousedů pro shlukování): Při shlukování částečně potlačuje generování řetězcových struktur.

Metoda nejvzdálenějšího souseda

- opačný princip než metoda nejbližšího souseda: $D_{FN}(C_i, C_i) = \max_{\substack{x_p \in C_i \\ x_q \in C_j}} d(x_p, x_q)$

- pozn. (při použití metody nejvzdálenějšího souseda pro shlukování): Generování protáhlých struktur tato metoda potlačuje, naopak vede ke tvorbě nevelkých kompaktních shluků.

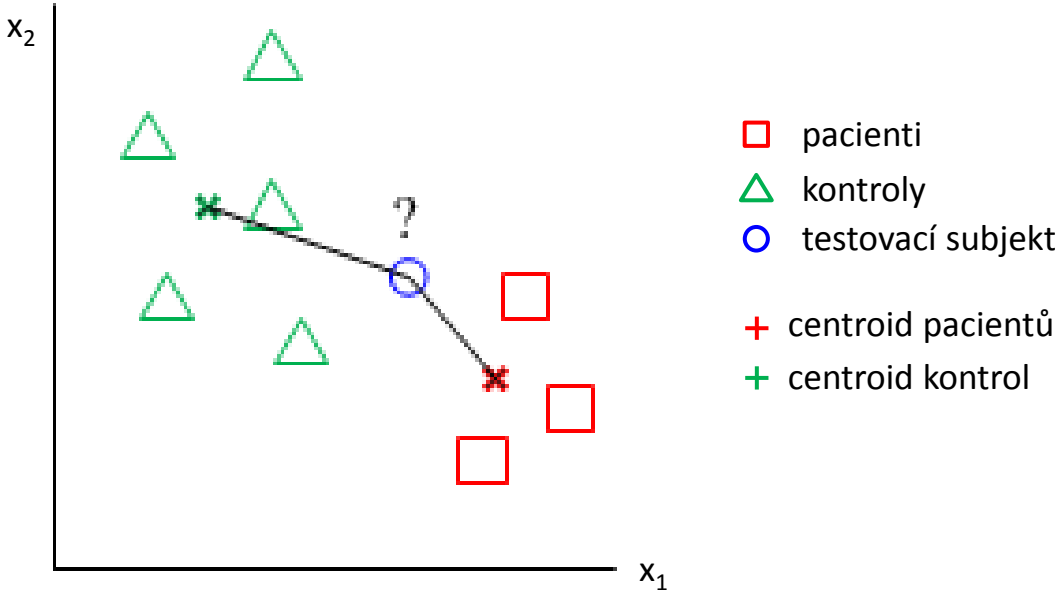
- pozn. 2: pro klasifikaci je obtížně použitelná

- pozn. 3: je možné zobecnění i pro více nejvzdálenějších sousedů

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

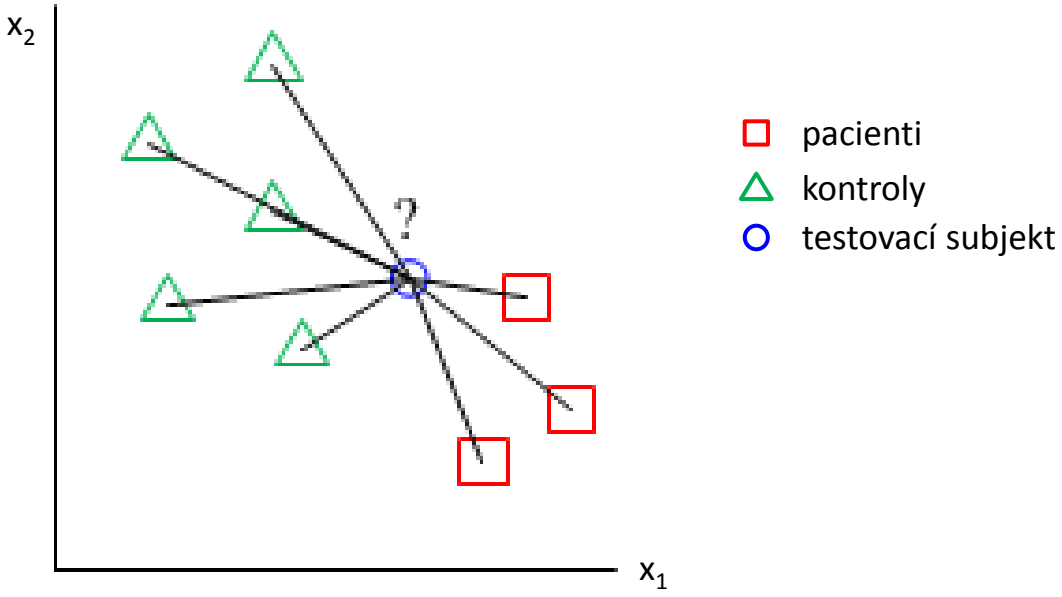
Centroidová metoda

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



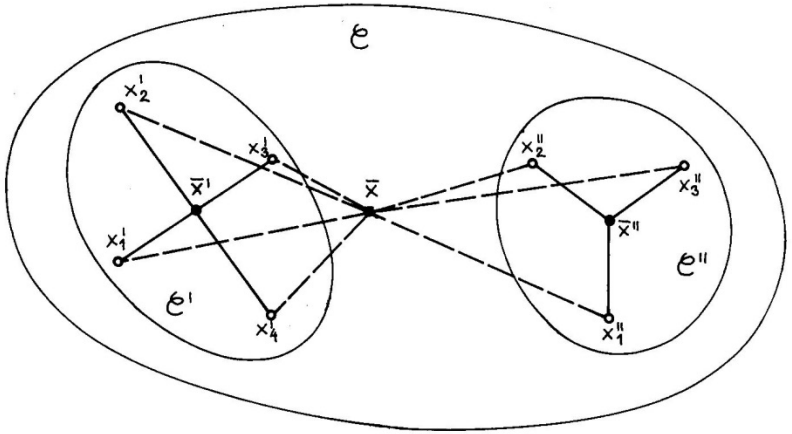
Metoda průměrné vazby

- vzdálenost dvou tříd C_i a C_j je průměrná vzdálenost mezi všemi obrazy tříd C_i a C_j
- při klasifikaci: zařazení subjektu do skupiny s nejmenší průměrnou vzdálenosti od všech obrazů dané skupiny



Wardova metoda

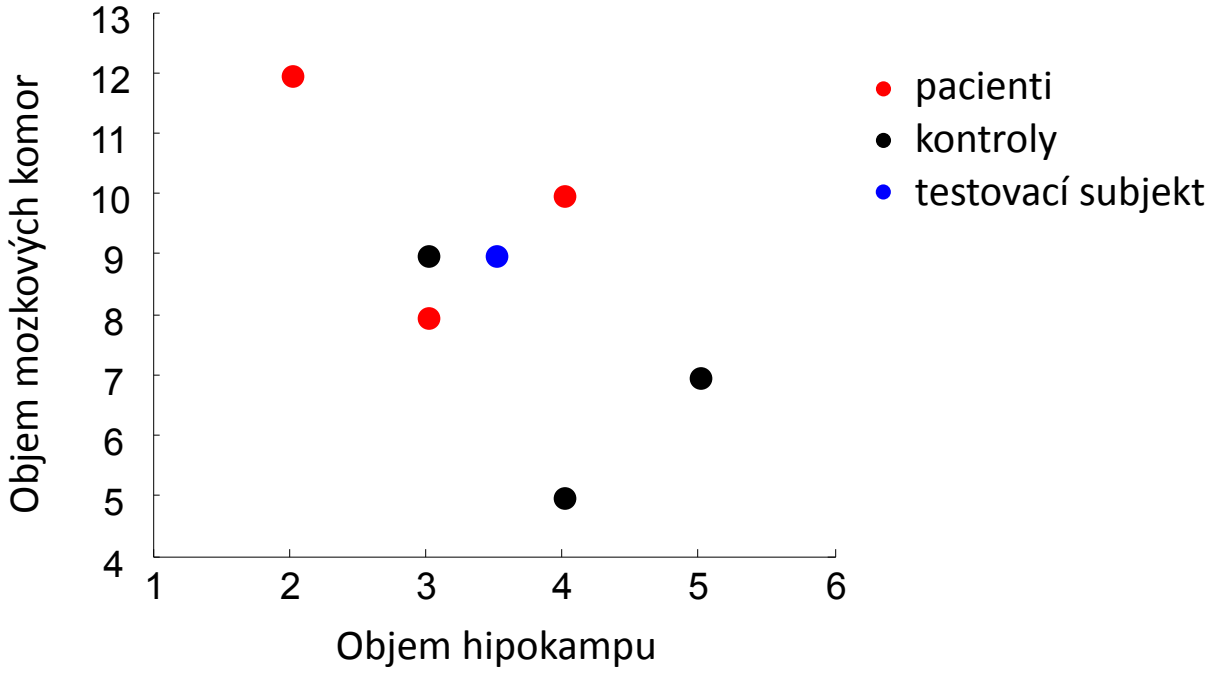
- vzdálenost mezi třídami (shluky) je definována přírůstkem součtu čtverců odchylek mezi těžištěm a obrazy shluku vytvořeného z obou uvažovaných shluků C_i a C_j oproti součtu čtverců odchylek mezi obrazy a těžišti v obou shlucích C_i a C_j.
- pozn. (při použití Wardovy metody pro shlukování): Metoda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



Příklad 2

Bylo provedeno měření objemu hipokampu a mozkových komor (v cm^3) 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.



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“

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ