Řešení příkladu - klasifikace testovacího subjektu podle minimální vzdálenosti:

Postup:

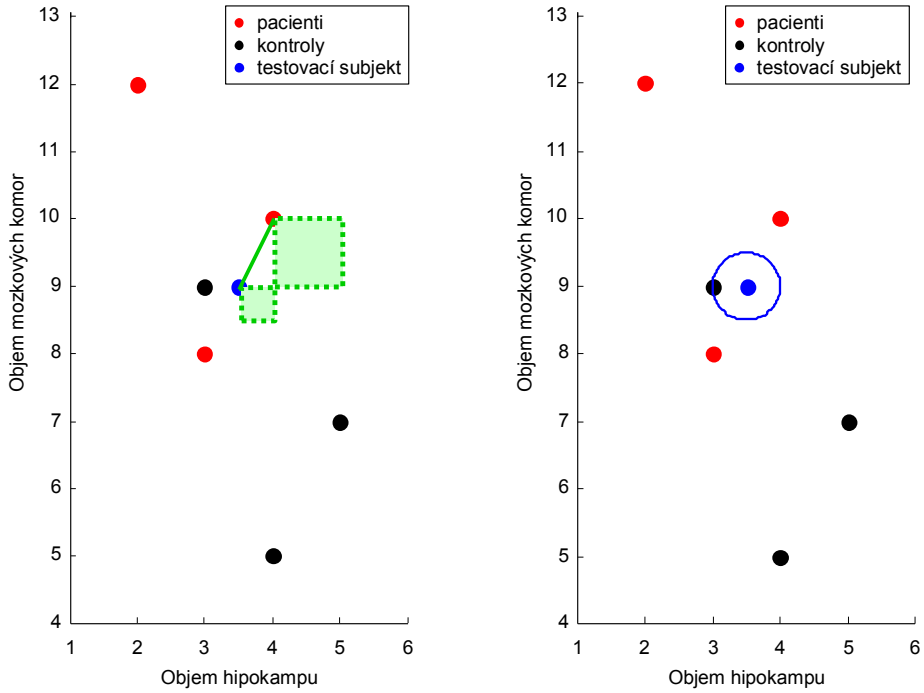
I) zvolení metriky pro výpočet vzdáleností dvou bodů

II) zvolení metriky pro určení vzdálenosti mezi dvěma množinami bodů

Předpoklad: budeme shlukovací algoritmy využívat jako neučící se algoritmy (klasifikátor natrénujeme na celé trénovací množině a pak už pouze klasifikujeme nové subjekty (např. už nepřepočítáváme centroid po zařazení každého nového subjektu či objektu apod.))

1.1 Metoda *k* nejbližších sousedů + Euklidova metrika:

Znázornění výpočtu vzdálenosti dvou bodů pomocí Euklidovy metriky je uvedeno na Obr. 1.



Obr.1: Ilustrace výpočtu vzdálenosti dvou bodů pomocí Euklidovy metriky (vlevo) a znázornění klasifikace podle nejbližšího souseda (vpravo). Modře je vyznačena množina bodů, které mají od testovacího subjektu stejnou vzdálenost.

Výpočet vzdáleností testovacího (nového) subjektu od všech subjektů z obou skupin:

$$d_E(\mathbf{x}_1, \mathbf{x}_0) = \sqrt{(x_{11} - x_{01})^2 + (x_{12} - x_{02})^2} = \sqrt{(2 - 3{,}5)^2 + (12 - 9)^2} = \sqrt{2{,}25 + 9} = 3{,}35$$

$$d_E(\mathbf{x}_2, \mathbf{x}_0) = \sqrt{(x_{21} - x_{01})^2 + (x_{22} - x_{02})^2} = \sqrt{(4 - 3{,}5)^2 + (10 - 9)^2} = \sqrt{0{,}25 + 1} = 1{,}12$$

$$d_E(\mathbf{x}_3, \mathbf{x}_0) = \sqrt{(x_{31} - x_{01})^2 + (x_{32} - x_{02})^2} = \sqrt{(3 - 3{,}5)^2 + (8 - 9)^2} = \sqrt{0{,}25 + 1} = 1{,}12$$

$$d_E(\mathbf{x}_4, \mathbf{x}_0) = \sqrt{(x_{41} - x_{01})^2 + (x_{42} - x_{02})^2} = \sqrt{(5 - 3{,}5)^2 + (7 - 9)^2} = \sqrt{2{,}25 + 4} = 2{,}5$$

$$d_E(\mathbf{x}_5, \mathbf{x}_0) = \sqrt{(x_{51} - x_{01})^2 + (x_{52} - x_{02})^2} = \sqrt{(3 - 3{,}5)^2 + (9 - 9)^2} = \sqrt{0{,}25} = 0{,}5$$

$$d_E(\mathbf{x}_6, \mathbf{x}_0) = \sqrt{(x_{61} - x_{01})^2 + (x_{62} - x_{02})^2} = \sqrt{(4 - 3{,}5)^2 + (5 - 9)^2} = \sqrt{0{,}25 + 16} = 4{,}03$$

Seřazení vzdáleností:

$$d_E(\mathbf{x}_5, \mathbf{x}_0) < d_E(\mathbf{x}_2, \mathbf{x}_0) \leq d_E(\mathbf{x}_3, \mathbf{x}_0) < d_E(\mathbf{x}_4, \mathbf{x}_0) < d_E(\mathbf{x}_1, \mathbf{x}_0) < d_E(\mathbf{x}_6, \mathbf{x}_0)$$

pro $k = 1$: nejbližší soused bodu $\mathbf{x}_0$ je bod $\mathbf{x}_5$, protože $d_E(\mathbf{x}_5, \mathbf{x}_0)$ je nejmenší → testovací subjekt bude zařazen do třídy kontrolních subjektů;
lze rovněž zapsat jako: $d_{NN}(\mathrm{D}, \mathbf{x}_0) = \min d_E(\mathbf{x}_i, \mathbf{x}_0) = 1{,}12$, kde $i = 1{,}2{,}3$, a $d_{NN}(\mathrm{H}, \mathbf{x}_0) = \min d_E(\mathbf{x}_i, \mathbf{x}_0) = 0{,}5$, kde $i = 4{,}5{,}6$; protože $d_{NN}(\mathrm{H}, \mathbf{x}_0) < d_{NN}(\mathrm{D}, \mathbf{x}_0)$, testovací subjekt bude zařazen do třídy kontrolních subjektů

pro $k = 2$: nelze rozhodnout

pro $k = 3$: subjekt zařazen do třídy pacientů, protože mezi 3 nejbližšími sousedy jsou 2 pacienti a 1 kontrolní subjekt

pro $k = 4$: nelze rozhodnout

pro $k = 5$: subjekt zařazen do třídy pacientů, protože mezi 5 nejbližšími sousedy jsou 3 pacienti a 2 kontrolní subjekty

pro $k = 6$: nelze rozhodnout

Poznámka: je nutné volit liché k

Poznámka 2: závisí na volbě k, kam subjekt zařadíme (tzn., pro různá k se zařazení může lišit – např. v tomto případě pro $k = 1$ subjekt zařazen do třídy kontrolních subjektů a pro $k = 3$ a $k = 5$ subjekt zařazen do třídy pacientů)

1.2 Metoda průměrné vazby + Euklidova metrika:

$d_{GA}(\mathrm{D}, \mathbf{x}_0) = \frac{d_E(\mathbf{x}_1,\mathbf{x}_0)+d_E(\mathbf{x}_2,\mathbf{x}_0)+d_E(\mathbf{x}_3,\mathbf{x}_0)}{3} = \frac{3{,}35+1{,}12+1{,}12}{3} = 1{,}86$

$d_{GA}(\mathrm{H}, \mathbf{x}_0) = \frac{d_E(\mathbf{x}_4,\mathbf{x}_0)+d_E(\mathbf{x}_5,\mathbf{x}_0)+d_E(\mathbf{x}_6,\mathbf{x}_0)}{3} = \frac{2{,}5+0{,}5+4{,}03}{3} = 2{,}34$

Protože $d_{GA}(\mathrm{D}, \mathbf{x}_0) < d_{GA}(\mathrm{H}, \mathbf{x}_0)$, testovací subjekt bude zařazen do třídy pacientů.

1.3 Centroidová metoda + Euklidova metrika:

$\bar{\mathbf{x}}_D = \left[\frac{1}{n_D}\sum_{i=1}^{n_D} \mathrm{x}_{i1} \quad \frac{1}{n_D}\sum_{i=1}^{n_D} \mathrm{x}_{i2}\right] = \left[\frac{1}{3}(2+4+3) \quad \frac{1}{3}(12+10+8)\right] = [3 \quad 10]$ – centroid pacientů

$\bar{\mathbf{x}}_H = \left[\frac{1}{n_H}\sum_{i=1}^{n_H} \mathrm{x}_{i1} \quad \frac{1}{n_H}\sum_{i=1}^{n_H} \mathrm{x}_{i2}\right] = \left[\frac{1}{3}(5+3+4) \quad \frac{1}{3}(7+9+5)\right] = [4 \quad 7]$ – centroid kontrol

$d_{CE}(\mathrm{D}, \mathbf{x}_0) = d_E(\bar{\mathbf{x}}_D, \mathbf{x}_0) = \sqrt{(3-3{,}5)^2 + (10-9)^2} = \sqrt{0{,}25+1} = 1{,}12$

$d_{CE}(\mathrm{H}, \mathbf{x}_0) = d_E(\bar{\mathbf{x}}_H, \mathbf{x}_0) = \sqrt{(4-3{,}5)^2 + (7-9)^2} = \sqrt{0{,}25+4} = 2{,}06$

Protože $d_{CE}(\mathrm{D}, \mathbf{x}_0) < d_{CE}(\mathrm{H}, \mathbf{x}_0)$, testovací subjekt bude zařazen do třídy pacientů.

Znázornění klasifikace testovacího subjektu pomocí centroidové metody, přičemž vzdálenosti testovacího subjektu od centroidů skupin jsou počítány pomocí Euklidovy metriky, je na Obr. 2.

Centroidová metoda s využitím medoidu:

Medoid (odvozen vizuálně – spočítal by se tak, že by se našel nejbližší bod k centroidu u dané skupiny nebo jako bod s nejmenší sumou vzdáleností od ostatních bodů)

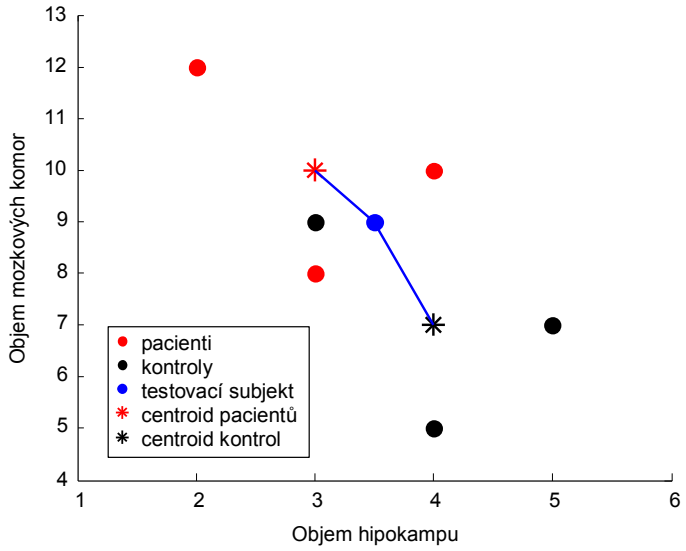
medoid pro pacienty: $\tilde{\mathbf{x}}_D = \mathbf{x}_2 = [4 \quad 10]$

medoid pro kontroly: $\tilde{\mathbf{x}}_H = \mathbf{x}_4 = [5 \quad 7]$

$d_{CEM}(\mathrm{D}, \mathbf{x}_0) = d_E(\tilde{\mathbf{x}}_D, \mathbf{x}_0) = d_E(\mathbf{x}_2, \mathbf{x}_0) = \sqrt{(4-3{,}5)^2 + (10-9)^2} = \sqrt{0{,}25+1} = 1{,}12$

$d_{CEM}(\mathrm{H}, \mathbf{x}_0) = d_E(\tilde{\mathbf{x}}_H, \mathbf{x}_0) = d_E(\mathbf{x}_4, \mathbf{x}_0) = \sqrt{(5-3{,}5)^2 + (7-9)^2} = \sqrt{2{,}25+4} = 2{,}5$

Protože $d_{CEM}(\mathrm{D}, \mathbf{x}_0) < d_{CEM}(\mathrm{H}, \mathbf{x}_0)$, testovací subjekt bude zařazen do třídy pacientů.



Obr. 2: Ilustrace klasifikace testovacího subjektu pomocí centroidové metody, přičemž vzdálenosti testovacího subjektu od centroidů skupin jsou počítány pomocí Euklidovy metriky. Je patrné, že subjekt bude zařazen do třídy pacientů, protože jeho Euklidova vzdálenost od centroidu pacientů je menší než od centroidu kontrol.

2.1 Metoda *k* nejbližších sousedů + Hammingova (manhattanská) metrika:

Znázornění výpočtu vzdálenosti dvou bodů pomocí Hammingovy (manhattanské) metriky je uvedeno na Obr. 3.

Výpočet vzdáleností testovacího (nového) subjektu od všech subjektů z obou skupin:

$$d_H(\mathbf{x}_1,\mathbf{x}_0) = |x_{11} - x_{01}| + |x_{12} - x_{02}| = |2 - 3{,}5| + |12 - 9| = 1{,}5 + 3 = 4{,}5$$
$$d_H(\mathbf{x}_2,\mathbf{x}_0) = |x_{21} - x_{01}| + |x_{22} - x_{02}| = |4 - 3{,}5| + |10 - 9| = 0{,}5 + 1 = 1{,}5$$
$$d_H(\mathbf{x}_3,\mathbf{x}_0) = |x_{31} - x_{01}| + |x_{32} - x_{02}| = |3 - 3{,}5| + |8 - 9| = 0{,}5 + 1 = 1{,}5$$
$$d_H(\mathbf{x}_4,\mathbf{x}_0) = |x_{41} - x_{01}| + |x_{42} - x_{02}| = |5 - 3{,}5| + |7 - 9| = 1{,}5 + 2 = 3{,}5$$
$$d_H(\mathbf{x}_5,\mathbf{x}_0) = |x_{51} - x_{01}| + |x_{52} - x_{02}| = |3 - 3{,}5| + |9 - 9| = 0{,}5 + 0 = 0{,}5$$
$$d_H(\mathbf{x}_6,\mathbf{x}_0) = |x_{61} - x_{01}| + |x_{62} - x_{02}| = |4 - 3{,}5| + |5 - 9| = 0{,}5 + 4 = 4{,}5$$

Seřazení vzdáleností:

$$d_H(\mathbf{x}_5,\mathbf{x}_0) < d_H(\mathbf{x}_2,\mathbf{x}_0) \leq d_H(\mathbf{x}_3,\mathbf{x}_0) < d_H(\mathbf{x}_4,\mathbf{x}_0) < d_H(\mathbf{x}_1,\mathbf{x}_0) \leq d_E(\mathbf{x}_6,\mathbf{x}_0)$$

pro $k = 1$: nejbližší soused bodu $\mathbf{x}_0$ je bod $\mathbf{x}_5$, protože $d_H(\mathbf{x}_5,\mathbf{x}_0)$ je nejmenší → testovací subjekt bude zařazen do třídy kontrolních subjektů

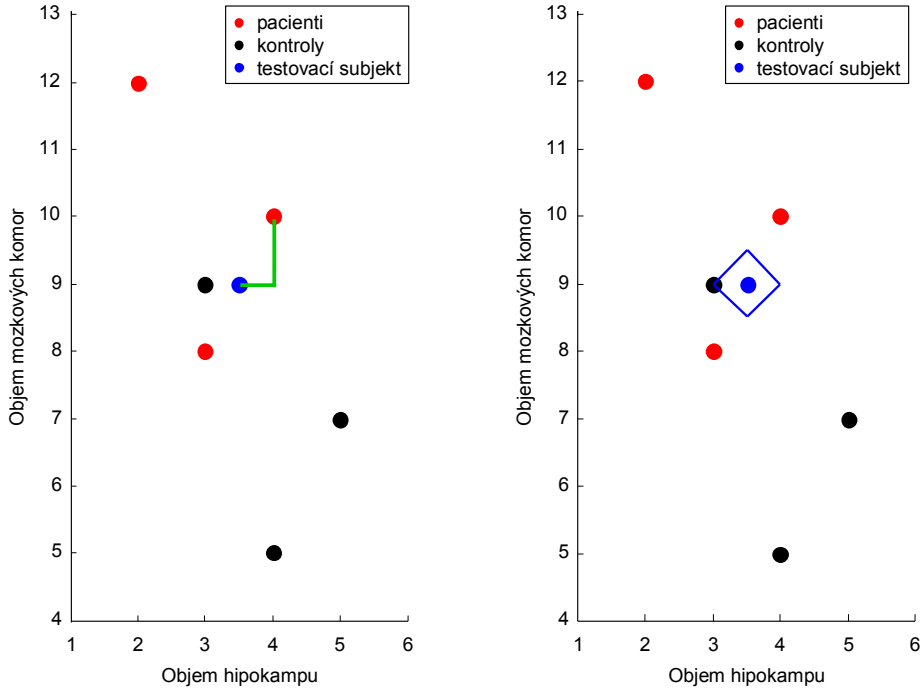
pro $k = 2$: nelze rozhodnout

pro $k = 3$: subjekt zařazen do třídy pacientů, protože mezi 3 nejbližšími sousedy jsou 2 pacienti a 1 kontrolní subjekt

pro $k = 4$: nelze rozhodnout

pro $k = 5$: nelze rozhodnout

pro $k = 6$: nelze rozhodnout



Obr. 3: Ilustrace výpočtu vzdálenosti dvou bodů pomocí Hammingovy (manhattanské) metriky (vlevo) a znázornění klasifikace podle nejbližšího souseda (vpravo). Modře je vyznačena množina bodů, které mají od testovacího subjektu stejnou vzdálenost.

2.2 Metoda průměrné vazby + Hammingova (manhattanská) metrika:

$d_{GA}(\mathrm{D}, \mathbf{x}_0) = \frac{d_H(\mathbf{x}_1,\mathbf{x}_0)+d_H(\mathbf{x}_2,\mathbf{x}_0)+d_H(\mathbf{x}_3,\mathbf{x}_0)}{3} = \frac{4{,}5+1{,}5+1{,}5}{3} = 2{,}5$

$d_{GA}(\mathrm{H}, \mathbf{x}_0) = \frac{d_H(\mathbf{x}_4,\mathbf{x}_0)+d_H(\mathbf{x}_5,\mathbf{x}_0)+d_H(\mathbf{x}_6,\mathbf{x}_0)}{3} = \frac{3{,}5+0{,}5+4{,}5}{3} = 2{,}83$

Protože $d_{GA}(\mathrm{D}, \mathbf{x}_0) < d_{GA}(\mathrm{H}, \mathbf{x}_0)$, testovací subjekt bude zařazen do třídy pacientů.

2.3 Centroidová metoda + Hammingova (manhattanská) metrika:

$d_{CE}(\mathrm{D}, \mathbf{x}_0) = d_H(\bar{\mathbf{x}}_D, \mathbf{x}_0) = |3 - 3{,}5| + |10 - 9| = 0{,}5 + 1 = 1{,}5$

$d_{CE}(\mathrm{H}, \mathbf{x}_0) = d_H(\bar{\mathbf{x}}_H, \mathbf{x}_0) = |4 - 3{,}5| + |7 - 9| = 0{,}5 + 2 = 2{,}5$

Protože $d_{CE}(\mathrm{D}, \mathbf{x}_0) < d_{CE}(\mathrm{H}, \mathbf{x}_0)$, testovací subjekt bude zařazen do třídy pacientů.

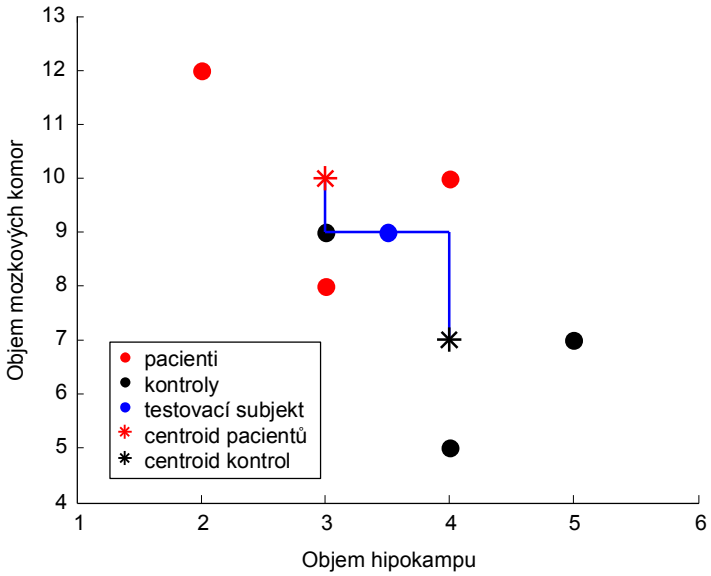
Znázornění klasifikace testovacího subjektu pomocí centroidové metody, přičemž vzdálenosti testovacího subjektu od centroidů skupin jsou počítány pomocí Hammingovy (manhattanské) metriky, je na Obr. 4.

Centroidová metoda s využitím medoidu:

$d_{CEM}(\mathrm{D}, \mathbf{x}_0) = d_H(\tilde{\mathbf{x}}_D, \mathbf{x}_0) = d_H(\mathbf{x}_2, \mathbf{x}_0) = |4 - 3{,}5| + |10 - 9| = 0{,}5 + 1 = 1{,}5$

$d_{CEM}(\mathrm{H}, \mathbf{x}_0) = d_H(\tilde{\mathbf{x}}_H, \mathbf{x}_0) = d_H(\mathbf{x}_4, \mathbf{x}_0) = |5 - 3{,}5| + |7 - 9| = 1{,}5 + 2 = 3{,}5$

Protože $d_{CEM}(\mathrm{D}, \mathbf{x}_0) < d_{CEM}(\mathrm{H}, \mathbf{x}_0)$, testovací subjekt bude zařazen do třídy pacientů.



Obr. 4: Ilustrace klasifikace testovacího subjektu pomocí centroidové metody, přičemž vzdálenosti testovacího subjektu od centroidů skupin jsou počítány pomocí Hammingovy (manhattanské) metriky. Je patrné, že subjekt bude zařazen do třídy pacientů, protože jeho Hammingova (manhattanská) vzdálenost od centroidu pacientů je menší než od centroidu kontrol.

3.1 Metoda *k* nejbližších sousedů + Čebyševova metrika:

Výpočet vzdáleností testovacího (nového) subjektu od všech subjektů z obou skupin:

$$d_C(\mathbf{x}_1, \mathbf{x}_0) = \max(|x_{11} - x_{01}|; |x_{12} - x_{02}|) = \max(|2 - 3{,}5|; |12 - 9|) = \max(1{,}5; 3) = 3$$
$$d_C(\mathbf{x}_2, \mathbf{x}_0) = \max(|x_{21} - x_{01}|; |x_{22} - x_{02}|) = \max(|4 - 3{,}5|; |10 - 9|) = \max(0{,}5; 1) = 1$$
$$d_C(\mathbf{x}_3, \mathbf{x}_0) = \max(|x_{31} - x_{01}|; |x_{32} - x_{02}|) = \max(|3 - 3{,}5|; |8 - 9|) = \max(0{,}5; 1) = 1$$
$$d_C(\mathbf{x}_4, \mathbf{x}_0) = \max(|x_{41} - x_{01}|; |x_{42} - x_{02}|) = \max(|5 - 3{,}5|; |7 - 9|) = \max(1{,}5; 2) = 2$$
$$d_C(\mathbf{x}_5, \mathbf{x}_0) = \max(|x_{51} - x_{01}|; |x_{52} - x_{02}|) = \max(|3 - 3{,}5|; |9 - 9|) = \max(0{,}5; 0) = 0{,}5$$
$$d_C(\mathbf{x}_6, \mathbf{x}_0) = \max(|x_{61} - x_{01}|; |x_{62} - x_{02}|) = \max(|4 - 3{,}5|; |5 - 9|) = \max(0{,}5; 4) = 4$$

Seřazení vzdáleností:

$$d_C(\mathbf{x}_5, \mathbf{x}_0) < d_C(\mathbf{x}_2, \mathbf{x}_0) \leq d_C(\mathbf{x}_3, \mathbf{x}_0) < d_C(\mathbf{x}_4, \mathbf{x}_0) < d_C(\mathbf{x}_1, \mathbf{x}_0) < d_C(\mathbf{x}_6, \mathbf{x}_0)$$

pro $k = 1$: nejbližší soused bodu $\mathbf{x}_0$ je bod $\mathbf{x}_5$, protože $d_C(\mathbf{x}_5, \mathbf{x}_0)$ je nejmenší → testovací subjekt bude zařazen do třídy kontrolních subjektů

pro $k = 2$: nelze rozhodnout

pro $k = 3$: subjekt zařazen do třídy pacientů, protože mezi 3 nejbližšími sousedy jsou 2 pacienti a 1 kontrolní subjekt

pro $k = 4$: nelze rozhodnout

pro $k = 5$: subjekt zařazen do třídy pacientů, protože mezi 5 nejbližšími sousedy jsou 3 pacienti a 2 kontrolní subjekty

pro $k = 6$: nelze rozhodnout

3.2 Metoda průměrné vazby + Čebyševova metrika:

$$d_{GA}(\mathrm{D}, \mathbf{x}_0) = \frac{d_C(\mathbf{x}_1, \mathbf{x}_0) + d_C(\mathbf{x}_2, \mathbf{x}_0) + d_C(\mathbf{x}_3, \mathbf{x}_0)}{3} = \frac{3+1+1}{3} = 1{,}67$$

$d_{GA}(\mathrm{H}, \mathbf{x}_0) = \frac{d_C(\mathbf{x}_4,\mathbf{x}_0)+d_C(\mathbf{x}_5,\mathbf{x}_0)+d_C(\mathbf{x}_6,\mathbf{x}_0)}{3} = \frac{2+0{,}5+4}{3} = 2{,}17$

Protože $d_{GA}(\mathrm{D}, \mathbf{x}_0) < d_{GA}(\mathrm{H}, \mathbf{x}_0)$, testovací subjekt bude zařazen do třídy pacientů.

3.3 Centroidová metoda + Čebyševova metrika:

$d_{CE}(\mathrm{D}, \mathbf{x}_0) = d_C(\bar{\mathbf{x}}_D, \mathbf{x}_0) = \max(|3-3{,}5|; |10-9|) = \max(0{,}5; 1) = 1$

$d_{CE}(\mathrm{H}, \mathbf{x}_0) = d_C(\bar{\mathbf{x}}_H, \mathbf{x}_0) = \max(|4-3{,}5|; |7-9|) = \max(0{,}5; 2) = 2$

Protože $d_{CE}(\mathrm{D}, \mathbf{x}_0) < d_{CE}(\mathrm{H}, \mathbf{x}_0)$, testovací subjekt bude zařazen do třídy pacientů.

Centroidová metoda s využitím medoidu:

$d_{CEM}(\mathrm{D}, \mathbf{x}_0) = d_C(\tilde{\mathbf{x}}_D, \mathbf{x}_0) = d_C(\mathbf{x}_2, \mathbf{x}_0) = \max(|4-3{,}5|; |10-9|) = \max(0{,}5; 1) = 1$

$d_{CEM}(\mathrm{H}, \mathbf{x}_0) = d_C(\tilde{\mathbf{x}}_H, \mathbf{x}_0) = d_C(\mathbf{x}_4, \mathbf{x}_0) = \max(|5-3{,}5|; |7-9|) = \max(1{,}5; 2) = 2$

Protože $d_{CEM}(\mathrm{D}, \mathbf{x}_0) < d_{CEM}(\mathrm{H}, \mathbf{x}_0)$, testovací subjekt bude zařazen do třídy pacientů.

4.1 Metoda *k* nejbližších sousedů + Canberrská metrika:

Výpočet vzdáleností testovacího (nového) subjektu od všech subjektů z obou skupin:

$d_{CA}(\mathbf{x}_1, \mathbf{x}_0) = \frac{|x_{11}-x_{01}|}{|x_{11}|+|x_{01}|} + \frac{|x_{12}-x_{02}|}{|x_{12}|+|x_{02}|} = \frac{|2-3{,}5|}{|2|+|3{,}5|} + \frac{|12-9|}{|12|+|9|} = \frac{1{,}5}{5{,}5} + \frac{3}{21} = 0{,}42$

$d_{CA}(\mathbf{x}_2, \mathbf{x}_0) = \frac{|x_{21}-x_{01}|}{|x_{21}|+|x_{01}|} + \frac{|x_{22}-x_{02}|}{|x_{22}|+|x_{02}|} = \frac{|4-3{,}5|}{|4|+|3{,}5|} + \frac{|10-9|}{|10|+|9|} = \frac{0{,}5}{7{,}5} + \frac{1}{19} = 0{,}12$

$d_{CA}(\mathbf{x}_3, \mathbf{x}_0) = \frac{|x_{31}-x_{01}|}{|x_{31}|+|x_{01}|} + \frac{|x_{32}-x_{02}|}{|x_{32}|+|x_{02}|} = \frac{|3-3{,}5|}{|3|+|3{,}5|} + \frac{|8-9|}{|8|+|9|} = \frac{0{,}5}{6{,}5} + \frac{1}{17} = 0{,}14$

$d_{CA}(\mathbf{x}_4, \mathbf{x}_0) = \frac{|x_{41}-x_{01}|}{|x_{41}|+|x_{01}|} + \frac{|x_{42}-x_{02}|}{|x_{42}|+|x_{02}|} = \frac{|5-3{,}5|}{|5|+|3{,}5|} + \frac{|7-9|}{|7|+|9|} = \frac{1{,}5}{8{,}5} + \frac{2}{16} = 0{,}30$

$d_{CA}(\mathbf{x}_5, \mathbf{x}_0) = \frac{|x_{51}-x_{01}|}{|x_{51}|+|x_{01}|} + \frac{|x_{52}-x_{02}|}{|x_{52}|+|x_{02}|} = \frac{|3-3{,}5|}{|3|+|3{,}5|} + \frac{|9-9|}{|9|+|9|} = \frac{0{,}5}{6{,}5} + \frac{0}{18} = 0{,}08$

$d_{CA}(\mathbf{x}_6, \mathbf{x}_0) = \frac{|x_{61}-x_{01}|}{|x_{61}|+|x_{01}|} + \frac{|x_{62}-x_{02}|}{|x_{62}|+|x_{02}|} = \frac{|4-3{,}5|}{|4|+|3{,}5|} + \frac{|5-9|}{|5|+|9|} = \frac{0{,}5}{7{,}5} + \frac{4}{14} = 0{,}35$

Seřazení vzdáleností:

$$d_{CA}(\mathbf{x}_5, \mathbf{x}_0) < d_{CA}(\mathbf{x}_2, \mathbf{x}_0) < d_{CA}(\mathbf{x}_3, \mathbf{x}_0) < d_{CA}(\mathbf{x}_4, \mathbf{x}_0) < d_{CA}(\mathbf{x}_6, \mathbf{x}_0) < d_{CA}(\mathbf{x}_1, \mathbf{x}_0)$$

pro $k = 1$: nejbližší soused bodu $\mathbf{x}_0$ je bod $\mathbf{x}_5$, protože $d_{CA}(\mathbf{x}_5, \mathbf{x}_0)$ je nejmenší → testovací subjekt bude zařazen do třídy kontrolních subjektů

pro $k = 2$: nelze rozhodnout

pro $k = 3$: subjekt zařazen do třídy pacientů, protože mezi 3 nejbližšími sousedy jsou 2 pacienti a 1 kontrolní subjekt

pro $k = 4$: nelze rozhodnout

pro $k = 5$: subjekt zařazen do třídy kontrolních subjektů, protože mezi 5 nejbližšími sousedy jsou 2 pacienti a 3 kontrolní subjekty

pro $k = 6$: nelze rozhodnout

4.2 Metoda průměrné vazby + Canberrská metrika:

$d_{GA}(\mathrm{D}, \mathbf{x}_0) = \frac{d_{CA}(\mathbf{x}_1,\mathbf{x}_0)+d_{CA}(\mathbf{x}_2,\mathbf{x}_0)+d_{CA}(\mathbf{x}_3,\mathbf{x}_0)}{3} = \frac{0{,}42+0{,}12+0{,}14}{3} = 0{,}23$

$d_{GA}(\mathrm{H}, \mathbf{x}_0) = \frac{d_{CA}(\mathbf{x}_4,\mathbf{x}_0)+d_{CA}(\mathbf{x}_5,\mathbf{x}_0)+d_{CA}(\mathbf{x}_6,\mathbf{x}_0)}{3} = \frac{0{,}30+0{,}08+0{,}35}{3} = 0{,}24$

Protože $d_{GA}(\mathrm{D}, \mathbf{x}_0) < d_{GA}(\mathrm{H}, \mathbf{x}_0)$, testovací subjekt bude zařazen do třídy pacientů.

4.3 Centroidová metoda + Canberrská metrika:

$$d_{CE}(\mathrm{D}, \mathbf{x}_0) = d_{CA}(\bar{\mathbf{x}}_D, \mathbf{x}_0) = \frac{|3-3{,}5|}{|3|+|3{,}5|} + \frac{|10-9|}{|10|+|9|} = \frac{0{,}5}{6{,}5} + \frac{1}{19} = 0{,}13$$

$$d_{CE}(\mathrm{H}, \mathbf{x}_0) = d_{CA}(\bar{\mathbf{x}}_H, \mathbf{x}_0) = \frac{|4-3{,}5|}{|4|+|3{,}5|} + \frac{|7-9|}{|7|+|9|} = \frac{0{,}5}{7{,}5} + \frac{2}{16} = 0{,}19$$

Protože $d_{CE}(\mathrm{D}, \mathbf{x}_0) < d_{CE}(\mathrm{H}, \mathbf{x}_0)$, testovací subjekt bude zařazen do třídy pacientů.

Centroidová metoda s využitím medoidu:

$$d_{CEM}(\mathrm{D}, \mathbf{x}_0) = d_{CA}(\tilde{\mathbf{x}}_D, \mathbf{x}_0) = d_{CA}(\mathbf{x}_2, \mathbf{x}_0) = \frac{|4-3{,}5|}{|4|+|3{,}5|} + \frac{|10-9|}{|10|+|9|} = \frac{0{,}5}{7{,}5} + \frac{1}{19} = 0{,}12$$

$$d_{CEM}(\mathrm{H}, \mathbf{x}_0) = d_{CA}(\tilde{\mathbf{x}}_H, \mathbf{x}_0) = d_{CA}(\mathbf{x}_4, \mathbf{x}_0) = \frac{|5-3{,}5|}{|5|+|3{,}5|} + \frac{|7-9|}{|7|+|9|} = \frac{1{,}5}{8{,}5} + \frac{2}{16} = 0{,}30$$

Protože $d_{CEM}(\mathrm{D}, \mathbf{x}_0) < d_{CEM}(\mathrm{H}, \mathbf{x}_0)$, testovací subjekt bude zařazen do třídy pacientů.

5.1 Metoda *k* nejbližších sousedů + Mahalanobisova metrika:

Nejprve je potřeba vypočítat výběrové kovarianční matice pro třídu pacientů a kontrol, tzn. $\mathbf{S}_D = \begin{bmatrix} 1 & -1 \\ -1 & 4 \end{bmatrix}$ a $\mathbf{S}_H = \begin{bmatrix} 1 & -1 \\ -1 & 4 \end{bmatrix}$ (výpočet výběrových kovariančních matic lze nalézt ve Cvičení 1) a jejich inverzi její inverzi $\mathbf{S}_D^{-1} = \begin{bmatrix} 4/3 & 1/3 \\ 1/3 & 1/3 \end{bmatrix}$ a $\mathbf{S}_H^{-1} = \begin{bmatrix} 4/3 & 1/3 \\ 1/3 & 1/3 \end{bmatrix}$.

Výpočet vzdáleností testovacího (nového) subjektu od všech subjektů z obou skupin:

$$d_{MA}(\mathbf{x}_1, \mathbf{x}_0) = \sqrt{(\mathbf{x}_1 - \mathbf{x}_0)^T \cdot \mathbf{S}_D^{-1} \cdot (\mathbf{x}_1 - \mathbf{x}_0)} = \sqrt{\begin{bmatrix} 2-3{,}5 & 12-9 \end{bmatrix} \cdot \begin{bmatrix} 4/3 & 1/3 \\ 1/3 & 1/3 \end{bmatrix} \cdot \begin{bmatrix} 2-3{,}5 \\ 12-9 \end{bmatrix}} = 1{,}73$$

$$d_{MA}(\mathbf{x}_2, \mathbf{x}_0) = \sqrt{(\mathbf{x}_2 - \mathbf{x}_0)^T \cdot \mathbf{S}_D^{-1} \cdot (\mathbf{x}_2 - \mathbf{x}_0)} = \sqrt{\begin{bmatrix} 4-3{,}5 & 10-9 \end{bmatrix} \cdot \begin{bmatrix} 4/3 & 1/3 \\ 1/3 & 1/3 \end{bmatrix} \cdot \begin{bmatrix} 4-3{,}5 \\ 10-9 \end{bmatrix}} = 1$$

$$d_{MA}(\mathbf{x}_3, \mathbf{x}_0) = \sqrt{(\mathbf{x}_3 - \mathbf{x}_0)^T \cdot \mathbf{S}_D^{-1} \cdot (\mathbf{x}_3 - \mathbf{x}_0)} = \sqrt{\begin{bmatrix} 3-3{,}5 & 8-9 \end{bmatrix} \cdot \begin{bmatrix} 4/3 & 1/3 \\ 1/3 & 1/3 \end{bmatrix} \cdot \begin{bmatrix} 3-3{,}5 \\ 8-9 \end{bmatrix}} = 1$$

$$d_{MA}(\mathbf{x}_4, \mathbf{x}_0) = \sqrt{(\mathbf{x}_4 - \mathbf{x}_0)^T \cdot \mathbf{S}_H^{-1} \cdot (\mathbf{x}_4 - \mathbf{x}_0)} = \sqrt{\begin{bmatrix} 5-3{,}5 & 7-9 \end{bmatrix} \cdot \begin{bmatrix} 4/3 & 1/3 \\ 1/3 & 1/3 \end{bmatrix} \cdot \begin{bmatrix} 5-3{,}5 \\ 7-9 \end{bmatrix}} = 1{,}53$$

$$d_{MA}(\mathbf{x}_5, \mathbf{x}_0) = \sqrt{(\mathbf{x}_5 - \mathbf{x}_0)^T \cdot \mathbf{S}_H^{-1} \cdot (\mathbf{x}_5 - \mathbf{x}_0)} = \sqrt{\begin{bmatrix} 3-3{,}5 & 9-9 \end{bmatrix} \cdot \begin{bmatrix} 4/3 & 1/3 \\ 1/3 & 1/3 \end{bmatrix} \cdot \begin{bmatrix} 3-3{,}5 \\ 9-9 \end{bmatrix}} = 0{,}58$$

$$d_{MA}(\mathbf{x}_6, \mathbf{x}_0) = \sqrt{(\mathbf{x}_6 - \mathbf{x}_0)^T \cdot \mathbf{S}_H^{-1} \cdot (\mathbf{x}_6 - \mathbf{x}_0)} = \sqrt{\begin{bmatrix} 4-3{,}5 & 5-9 \end{bmatrix} \cdot \begin{bmatrix} 4/3 & 1/3 \\ 1/3 & 1/3 \end{bmatrix} \cdot \begin{bmatrix} 4-3{,}5 \\ 5-9 \end{bmatrix}} = 2{,}08$$

Seřazení vzdáleností:

$$d_{MA}(\mathbf{x}_5, \mathbf{x}_0) < d_{MA}(\mathbf{x}_2, \mathbf{x}_0) \leq d_{MA}(\mathbf{x}_3, \mathbf{x}_0) < d_{MA}(\mathbf{x}_4, \mathbf{x}_0) < d_{MA}(\mathbf{x}_1, \mathbf{x}_0) < d_{MA}(\mathbf{x}_6, \mathbf{x}_0)$$

pro $k = 1$: nejbližší soused bodu $\mathbf{x}_0$ je bod $\mathbf{x}_5$, protože $d_{CA}(\mathbf{x}_5, \mathbf{x}_0)$ je nejmenší → testovací subjekt bude zařazen do třídy kontrolních subjektů

pro $k = 2$: nelze rozhodnout

pro $k = 3$: subjekt zařazen do třídy pacientů, protože mezi 3 nejbližšími sousedy jsou 2 pacienti a 1 kontrolní subjekt

pro $k = 4$: nelze rozhodnout
pro $k = 5$: subjekt zařazen do třídy pacientů, protože mezi 5 nejbližšími sousedy jsou 3 pacienti a 2 kontrolní subjekty
pro $k = 6$: nelze rozhodnout

5.2 Metoda průměrné vazby + Mahalanobisova metrika:

$$d_{GA}(\mathrm{D}, \mathbf{x}_0) = \frac{d_{MA}(\mathbf{x}_1,\mathbf{x}_0)+d_{MA}(\mathbf{x}_2,\mathbf{x}_0)+d_{MA}(\mathbf{x}_3,\mathbf{x}_0)}{3} = \frac{1{,}73+1+1}{3} = 1{,}24$$

$$d_{GA}(\mathrm{H}, \mathbf{x}_0) = \frac{d_{MA}(\mathbf{x}_4,\mathbf{x}_0)+d_{MA}(\mathbf{x}_5,\mathbf{x}_0)+d_{MA}(\mathbf{x}_6,\mathbf{x}_0)}{3} = \frac{1{,}53+0{,}58+2{,}08}{3} = 1{,}40$$

Protože $d_{GA}(\mathrm{D}, \mathbf{x}_0) < d_{GA}(\mathrm{H}, \mathbf{x}_0)$, testovací subjekt bude zařazen do třídy pacientů.

5.3 Centroidová metoda + Mahalanobisova metrika:

$$d_{CE}(\mathrm{D}, \mathbf{x}_0) = d_{MA}(\bar{\mathbf{x}}_D, \mathbf{x}_0) = \sqrt{[3-3{,}5 \quad 10-9] \cdot \begin{bmatrix} 4/3 & 1/3 \\ 1/3 & 1/3 \end{bmatrix} \cdot \begin{bmatrix} 3-3{,}5 \\ 10-9 \end{bmatrix}} = 0{,}58$$

$$d_{CE}(\mathrm{H}, \mathbf{x}_0) = d_{CA}(\bar{\mathbf{x}}_H, \mathbf{x}_0) = \sqrt{[4-3{,}5 \quad 7-9] \cdot \begin{bmatrix} 4/3 & 1/3 \\ 1/3 & 1/3 \end{bmatrix} \cdot \begin{bmatrix} 4-3{,}5 \\ 7-9 \end{bmatrix}} = 1$$

Protože $d_{CE}(\mathrm{D}, \mathbf{x}_0) < d_{CE}(\mathrm{H}, \mathbf{x}_0)$, testovací subjekt bude zařazen do třídy pacientů.

Centroidová metoda s využitím medoidu:

$$d_{CEM}(\mathrm{D}, \mathbf{x}_0) = d_{MA}(\tilde{\mathbf{x}}_D, \mathbf{x}_0) = d_{MA}(\mathbf{x}_2, \mathbf{x}_0) = 1$$

$$d_{CEM}(\mathrm{H}, \mathbf{x}_0) = d_{MA}(\tilde{\mathbf{x}}_H, \mathbf{x}_0) = d_{MA}(\mathbf{x}_4, \mathbf{x}_0) = 1{,}53$$

Protože $d_{CEM}(\mathrm{D}, \mathbf{x}_0) < d_{CEM}(\mathrm{H}, \mathbf{x}_0)$, testovací subjekt bude zařazen do třídy pacientů.

Výsledky uspořádáme do tabulky:

metrika	Euklidova	Hammingova	Čebyševova	Canberrská	Mahalanobisova
NN	H	H	H	H	H
3-NN	D	D	D	D	D
5-NN	D	-	D	H	D
GA	D	D	D	D	D
CE-centroid	D	D	D	D	D
CE-medoid	D	D	D	D	D

Je patrné, že výsledek klasifikace se může lišit při použití různých metrik vzdálenosti.