# Klasifikace podle minimální vzdálenosti – příklad

Bylo provedeno měření objemu hipokampu a objemu amygdaly u 3 pacientů s Alzheimerovou chorobou ($x\_{1},x\_{2},x\_{3}$) a 3 kontrolních subjektů ($x\_{4},x\_{5},x\_{6}$). Naměřené hodnoty byly (v řádcích) zaznamenány do matic $X\_{D}$ resp. $X\_{H}$ (označení D – diseased, H – healthy):

$$X\_{D}=\left[\begin{matrix}5&5\\6&2\\7&2\end{matrix}\right],$$

$$X\_{H}=\left[\begin{matrix}9&4\\7&3\\8&2\end{matrix}\right].$$

Určete, zda testovací subjekt $x\_{0}=\left[7 3\right]$ patří do skupiny pacientů či kontrolních subjektů pomocí klasifikace podle minimální vzdálenosti.

Řešení:

Postup:

1. zvolení metriky pro výpočet vzdáleností 2 bodů
2. zvolení metriky pro určení vzdálenosti mezi 2 množinami bodů

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

**1.1 Metoda *k* nejbližších sousedů + Euklidova metrika:**

Znázornění klasifikace podle nejbližšího souseda pomocí Euklidovy metriky je uvedeno na *Obrázku 1*.



*Obrázek 1*. Znázornění klasifikace podle nejbližšího souseda pomocí Euklidovy metriky. Modře je vyznačena množina bodů v obrazovém prostoru, které mají od testovacího subjektu stejnou vzdálenost.

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

 $d\_{E}\left(x\_{1},x\_{0}\right)=\sqrt{\left(x\_{11}-x\_{01}\right)^{2}+\left(x\_{12}-x\_{02}\right)^{2}}=\sqrt{\left(5-7\right)^{2}+\left(5-4\right)^{2}}=\sqrt{4+1}=2,24$

 $d\_{E}\left(x\_{2},x\_{0}\right)=\sqrt{\left(x\_{21}-x\_{01}\right)^{2}+\left(x\_{22}-x\_{02}\right)^{2}}=\sqrt{\left(6-7\right)^{2}+\left(2-4\right)^{2}}=\sqrt{1+4}=2,24$

 $d\_{E}\left(x\_{3},x\_{0}\right)=\sqrt{\left(x\_{31}-x\_{01}\right)^{2}+\left(x\_{32}-x\_{02}\right)^{2}}=\sqrt{\left(7-7\right)^{2}+\left(2-4\right)^{2}}=\sqrt{0+4}=2$

 $d\_{E}\left(x\_{4},x\_{0}\right)=\sqrt{\left(x\_{41}-x\_{01}\right)^{2}+\left(x\_{42}-x\_{02}\right)^{2}}=\sqrt{\left(9-7\right)^{2}+\left(4-4\right)^{2}}=\sqrt{4+0}=2$

 $d\_{E}\left(x\_{5},x\_{0}\right)=\sqrt{\left(x\_{51}-x\_{01}\right)^{2}+\left(x\_{52}-x\_{02}\right)^{2}}=\sqrt{\left(7-7\right)^{2}+\left(3-4\right)^{2}}=\sqrt{0+1}=1$

 $d\_{E}\left(x\_{6},x\_{0}\right)=\sqrt{\left(x\_{61}-x\_{01}\right)^{2}+\left(x\_{62}-x\_{02}\right)^{2}}=\sqrt{\left(8-7\right)^{2}+\left(2-4\right)^{2}}=\sqrt{1+4}=2,24$

Seřazení vzdáleností:

$$d\_{E}\left(x\_{5},x\_{0}\right)<d\_{E}\left(x\_{3},x\_{0}\right)\leq d\_{E}\left(x\_{4},x\_{0}\right)<d\_{E}\left(x\_{1},x\_{0}\right)\leq d\_{E}\left(x\_{2},x\_{0}\right)\leq d\_{E}\left(x\_{6},x\_{0}\right)$$

pro $k=1$: nejbližší soused bodu $x\_{0}$je bod$x\_{5}$, protože $d\_{E}\left(x\_{5},x\_{0}\right)$je nejmenší → testovací subjekt bude zařazen do třídy kontrolních subjektů;

lze rovněž zapsat jako: $d\_{NN}\left(D,x\_{0}\right)=min d\_{E}\left(x\_{i},x\_{0}\right)=2$, kde $i=1,2,3$, a $d\_{NN}\left(H,x\_{0}\right)=min d\_{E}\left(x\_{i},x\_{0}\right)=1$, kde $i=4,5,6$; protože $d\_{NN}\left(H,x\_{0}\right)<d\_{NN}\left(D,x\_{0}\right)$, testovacísubjekt bude zařazen do třídy kontrolních subjektů

pro $k=2$: nelze rozhodnout

pro $k=3$: subjekt zařazen do třídy kontrolních subjektů, protože mezi 3 nejbližšími sousedy je 1 pacient a 2 kontrolní subjekty

pro $k=4$: nelze rozhodnout

pro $k=5$: nelze rozhodnout

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 – v tomto konkrétním případě byl ale testovací subjekt vždy zařazen do třídy kontrolních subjektů)

**1.2 Metoda průměrné vazby + Euklidova metrika:**

 $d\_{GA}\left(D,x\_{0}\right)=\frac{d\_{E}\left(x\_{1},x\_{0}\right)+d\_{E}\left(x\_{2},x\_{0}\right)+d\_{E}\left(x\_{3},x\_{0}\right)}{3}=\frac{2,24+2,24+2}{3}=2,16$

 $d\_{GA}\left(H,x\_{0}\right)=\frac{d\_{E}\left(x\_{4},x\_{0}\right)+d\_{E}\left(x\_{5},x\_{0}\right)+d\_{E}\left(x\_{6},x\_{0}\right)}{3}=\frac{2+1+2,24}{3}=1,75$

Protože $d\_{GA}\left(D,x\_{0}\right)>d\_{GA}\left(H,x\_{0}\right)$, testovacísubjekt bude zařazen do třídy kontrolních subjektů.

**1.3 Centroidová metoda + Euklidova metrika:**

$\overbar{x}\_{D}=\left[\begin{matrix}\frac{1}{n\_{D}}\sum\_{i=1}^{n\_{D}}x\_{i1}&\frac{1}{n\_{D}}\sum\_{i=1}^{n\_{D}}x\_{i2}\end{matrix}\right]=\left[\begin{matrix}\frac{1}{3}\left(5+6+7\right)&\frac{1}{3}\left(5+2+2\right)\end{matrix}\right]=\left[\begin{matrix}6&3\end{matrix}\right]$ – centroid pacientů

 $\overbar{x}\_{H}=\left[\begin{matrix}\frac{1}{n\_{H}}\sum\_{i=1}^{n\_{H}}x\_{i1}&\frac{1}{n\_{H}}\sum\_{i=1}^{n\_{H}}x\_{i2}\end{matrix}\right]=\left[\begin{matrix}\frac{1}{3}\left(9+7+8\right)&\frac{1}{3}\left(4+3+2\right)\end{matrix}\right]=\left[\begin{matrix}8&3\end{matrix}\right]$ – centroid kontrol

 $d\_{CE}\left(D,x\_{0}\right)=d\_{E}\left(\overbar{x}\_{D},x\_{0}\right)=\sqrt{\left(6-7\right)^{2}+\left(3-4\right)^{2}}=\sqrt{1+1}=1,41$

 $d\_{CE}\left(H,x\_{0}\right)=d\_{E}\left(\overbar{x}\_{H},x\_{0}\right)=\sqrt{\left(8-7\right)^{2}+\left(3-4\right)^{2}}=\sqrt{1+1}=1,41$

Protože $d\_{CE}\left(D,x\_{0}\right)=d\_{CE}\left(H,x\_{0}\right)$, nelze rozhodnout, do které třídy bude subjekt zařazen.

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



*Obrázek 2*. Ilustrace klasifikace testovacího subjektu pomocí centroidové metody, přičemž vzdálenosti testovacího subjektu od centroidů skupin jsou počítány pomocí Euklidovy metriky. Je patrné, že nelze rozhodnout, do jaké třídy máme testovací subjekt zařadit., protože jeho Euklidova vzdálenost od centroidu pacientů je stejná jako od centroidu kontrol.

**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: $\tilde{x}\_{D}=x\_{2}=\left[\begin{matrix}6&2\end{matrix}\right]$

medoid pro kontroly: $\tilde{x}\_{H}=x\_{5} nebo x\_{6 }(náhodně zvolíme například první v pořadí x\_{5}=\left[\begin{matrix}7&3\end{matrix}\right])$

 $d\_{CEM}\left(D,x\_{0}\right)=d\_{E}\left(\tilde{x}\_{D},x\_{0}\right)=d\_{E}\left(x\_{2},x\_{0}\right)=\sqrt{\left(6-7\right)^{2}+\left(2-4\right)^{2}}=\sqrt{1+4}=2,24$

 $d\_{CEM}\left(H,x\_{0}\right)=d\_{E}\left(\tilde{x}\_{H},x\_{0}\right)=d\_{E}\left(x\_{5},x\_{0}\right)=\sqrt{\left(7-7\right)^{2}+\left(3-4\right)^{2}}=\sqrt{0+1}=1$

Protože $d\_{CEM}\left(D,x\_{0}\right)>d\_{CEM}\left(H,x\_{0}\right)$, testovacísubjekt bude zařazen do třídy kontrolních subjektů.

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

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

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

 $d\_{H}\left(x\_{1},x\_{0}\right)=\left|x\_{11}-x\_{01}\right|+\left|x\_{12}-x\_{02}\right|=\left|5-7\right|+\left|5-4\right|=2+1=3$

 $d\_{H}\left(x\_{2},x\_{0}\right)=\left|x\_{21}-x\_{01}\right|+\left|x\_{22}-x\_{02}\right|=\left|6-7\right|+\left|2-4\right|=1+2=3$

 $d\_{H}\left(x\_{3},x\_{0}\right)=\left|x\_{31}-x\_{01}\right|+\left|x\_{32}-x\_{02}\right|=\left|7-7\right|+\left|2-4\right|=0+2=2$

 $d\_{H}\left(x\_{4},x\_{0}\right)=\left|x\_{41}-x\_{01}\right|+\left|x\_{42}-x\_{02}\right|=\left|9-7\right|+\left|4-4\right|=2+0=2$

 $d\_{H}\left(x\_{5},x\_{0}\right)=\left|x\_{51}-x\_{01}\right|+\left|x\_{52}-x\_{02}\right|=\left|7-7\right|+\left|3-4\right|=0+1=1$

 $d\_{H}\left(x\_{6},x\_{0}\right)=\left|x\_{61}-x\_{01}\right|+\left|x\_{62}-x\_{02}\right|=\left|8-7\right|+\left|2-4\right|=1+2=3$

Seřazení vzdáleností:

$$d\_{H}\left(x\_{5},x\_{0}\right)<d\_{H}\left(x\_{3},x\_{0}\right)\leq d\_{H}\left(x\_{4},x\_{0}\right)<d\_{H}\left(x\_{1},x\_{0}\right)\leq d\_{H}\left(x\_{2},x\_{0}\right)\leq d\_{E}\left(x\_{6},x\_{0}\right)$$

pro $k=1$: nejbližší soused bodu $x\_{0}$je bod$x\_{5}$, protože $d\_{H}\left(x\_{5},x\_{0}\right)$je nejmenší → testovací subjekt bude zařazen do třídy kontrolních subjektů

pro $k=2$: nelze rozhodnout

pro $k=3$: subjekt zařazen do třídy kontrolních subjektů, protože mezi 3 nejbližšími sousedy je 1 pacient a 2 kontrolní subjekty

pro $k=4$: nelze rozhodnout

pro $k=5$: nelze rozhodnout

pro $k=6$: nelze rozhodnout



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

**2.2 Metoda průměrné vazby + Hammingova (manhattanská) metrika:**

 $d\_{GA}\left(D,x\_{0}\right)=\frac{d\_{H}\left(x\_{1},x\_{0}\right)+d\_{H}\left(x\_{2},x\_{0}\right)+d\_{H}\left(x\_{3},x\_{0}\right)}{3}=\frac{3+3+2}{3}=2,67$

 $d\_{GA}\left(H,x\_{0}\right)=\frac{d\_{H}\left(x\_{4},x\_{0}\right)+d\_{H}\left(x\_{5},x\_{0}\right)+d\_{H}\left(x\_{6},x\_{0}\right)}{3}=\frac{2+1+3}{3}=2$

Protože $d\_{GA}\left(D,x\_{0}\right)>d\_{GA}\left(H,x\_{0}\right)$, testovacísubjekt bude zařazen do třídy kontrolních subjektů.

**2.3 Centroidová metoda + Hammingova (manhattanská) metrika:**

 $d\_{CE}\left(D,x\_{0}\right)=d\_{H}\left(\overbar{x}\_{D},x\_{0}\right)=\left|6-7\right|+\left|3-4\right|=1+1=2$

 $d\_{CE}\left(H,x\_{0}\right)=d\_{H}\left(\overbar{x}\_{H},x\_{0}\right)=\left|8-7\right|+\left|3-4\right|=1+1=2$

Protože $d\_{CE}\left(D,x\_{0}\right)=d\_{CE}\left(H,x\_{0}\right)$, nelze rozhodnout, do které třídy bude testovací subjekt zařazen.

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



*Obrázek 4*. Ilustrace klasifikace testovacího subjektu pomocí centroidové metody, přičemž vzdálenosti testovacího subjektu od centroidů skupin jsou počítány pomocí Hammingovy (manhattanské) metriky. Je patrné, že nelze rozhodnout, do jaké třídy máme testovací subjekt zařadit., protože jeho Hammingova (manhattanská) vzdálenost od centroidu pacientů je stejná jako od centroidu kontrol.

**Centroidová metoda s využitím medoidu:**

 $d\_{CEM}\left(D,x\_{0}\right)=d\_{H}\left(\tilde{x}\_{D},x\_{0}\right)=d\_{H}\left(x\_{2},x\_{0}\right)=\left|6-7\right|+\left|2-4\right|=1+2=3$

 $d\_{CEM}\left(H,x\_{0}\right)=d\_{H}\left(\tilde{x}\_{H},x\_{0}\right)=d\_{H}\left(x\_{5},x\_{0}\right)=\left|7-7\right|+\left|3-4\right|=0+1=1$

Protože $d\_{CEM}\left(D,x\_{0}\right)>d\_{CEM}\left(H,x\_{0}\right)$, testovacísubjekt bude zařazen do třídy kontrolních subjektů.

**3.1 Metoda *k* nejbližších sousedů + Čebyševova metrika:**

Znázornění klasifikace podle nejbližšího souseda pomocí Čebyševovy metriky je uvedeno na *Obrázku 5*.

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

 $d\_{C}\left(x\_{1},x\_{0}\right)=max\left(\left|x\_{11}-x\_{01}\right|;\left|x\_{12}-x\_{02}\right|\right)=max\left(\left|5-7\right|;\left|5-4\right|\right)=max\left(2;1\right)=2$

 $d\_{C}\left(x\_{2},x\_{0}\right)=max\left(\left|x\_{21}-x\_{01}\right|;\left|x\_{22}-x\_{02}\right|\right)=max\left(\left|6-7\right|;\left|2-4\right|\right)=max\left(1;2\right)=2$

 $d\_{C}\left(x\_{3},x\_{0}\right)=max\left(\left|x\_{31}-x\_{01}\right|;\left|x\_{32}-x\_{02}\right|\right)=max\left(\left|7-7\right|;\left|2-4\right|\right)=max\left(0;2\right)=2$

 $d\_{C}\left(x\_{4},x\_{0}\right)=max\left(\left|x\_{41}-x\_{01}\right|;\left|x\_{42}-x\_{02}\right|\right)=max\left(\left|9-7\right|;\left|4-4\right|\right)=max\left(2;0\right)=2$

 $d\_{C}\left(x\_{5},x\_{0}\right)=max\left(\left|x\_{51}-x\_{01}\right|;\left|x\_{52}-x\_{02}\right|\right)=max\left(\left|7-7\right|;\left|3-4\right|\right)=max\left(0;1\right)=1$

 $d\_{C}\left(x\_{6},x\_{0}\right)=max\left(\left|x\_{61}-x\_{01}\right|;\left|x\_{62}-x\_{02}\right|\right)=max\left(\left|8-7\right|;\left|2-4\right|\right)=max\left(1;2\right)=2$

Seřazení vzdáleností:

$$d\_{C}\left(x\_{5},x\_{0}\right)<d\_{C}\left(x\_{1},x\_{0}\right)\leq d\_{C}\left(x\_{2},x\_{0}\right)\leq d\_{C}\left(x\_{3},x\_{0}\right)\leq d\_{C}\left(x\_{4},x\_{0}\right)\leq d\_{C}\left(x\_{6},x\_{0}\right)$$

pro $k=1$: nejbližší soused bodu $x\_{0}$je bod$x\_{5}$, protože $d\_{C}\left(x\_{5},x\_{0}\right)$je nejmenší → testovací subjekt bude zařazen do třídy kontrolních subjektů

pro $k=2$: nelze rozhodnout

pro $k=3$: nelze rozhodnout

pro $k=4$: nelze rozhodnout

pro $k=5$: nelze rozhodnout

pro $k=6$: nelze rozhodnout



*Obrázek 5*. Znázornění klasifikace podle nejbližšího souseda pomocí Čebyševovy metriky. Modře je vyznačena množina bodů v obrazovém prostoru, které mají od testovacího subjektu stejnou vzdálenost.

**3.2 Metoda průměrné vazby + Čebyševova metrika:**

 $d\_{GA}\left(D,x\_{0}\right)=\frac{d\_{C}\left(x\_{1},x\_{0}\right)+d\_{C}\left(x\_{2},x\_{0}\right)+d\_{C}\left(x\_{3},x\_{0}\right)}{3}=\frac{2+2+2}{3}=2$

 $d\_{GA}\left(H,x\_{0}\right)=\frac{d\_{C}\left(x\_{4},x\_{0}\right)+d\_{C}\left(x\_{5},x\_{0}\right)+d\_{C}\left(x\_{6},x\_{0}\right)}{3}=\frac{2+1+2}{3}=1,67$

Protože $d\_{GA}\left(D,x\_{0}\right)>d\_{GA}\left(H,x\_{0}\right)$, testovacísubjekt bude zařazen do třídy testovacích subjektů.

**3.3 Centroidová metoda + Čebyševova metrika:**

 $d\_{CE}\left(D,x\_{0}\right)=d\_{C}\left(\overbar{x}\_{D},x\_{0}\right)=max\left(\left|6-7\right|;\left|3-4\right|\right)=max\left(1;1\right)=1$

 $d\_{CE}\left(H,x\_{0}\right)=d\_{C}\left(\overbar{x}\_{H},x\_{0}\right)=max⁡\left(\left|8-7\right|;\left|3-4\right|\right)=max\left(1;1\right)=1$

Protože $d\_{CE}\left(D,x\_{0}\right)=d\_{CE}\left(H,x\_{0}\right)$, nelze rozhodnout, do které třídy bude testovací subjekt zařazen.

**Centroidová metoda s využitím medoidu:**

 $d\_{CEM}\left(D,x\_{0}\right)=d\_{C}\left(\tilde{x}\_{D},x\_{0}\right)=d\_{C}\left(x\_{2},x\_{0}\right)=max\left(\left|6-7\right|;\left|2-4\right|\right)=max\left(1;2\right)=2$

 $d\_{CEM}\left(H,x\_{0}\right)=d\_{C}\left(\tilde{x}\_{H},x\_{0}\right)=d\_{C}\left(x\_{5},x\_{0}\right)=max\left(\left|7-7\right|;\left|3-4\right|\right)=max\left(0;1\right)=1$

Protože $d\_{CEM}\left(D,x\_{0}\right)>d\_{CEM}\left(H,x\_{0}\right)$, testovacísubjekt bude zařazen do třídy kontrolních subjektů.

**4.1 Metoda *k* nejbližších sousedů + Canberrská metrika:**

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

 $d\_{CA}\left(x\_{1},x\_{0}\right)=\frac{\left|x\_{11}-x\_{01}\right|}{\left|x\_{11}\right|+\left|x\_{01}\right|}+\frac{\left|x\_{12}-x\_{02}\right|}{\left|x\_{12}\right|+\left|x\_{02}\right|}=\frac{\left|5-7\right|}{\left|5\right|+\left|7\right|}+\frac{\left|5-4\right|}{\left|5\right|+\left|4\right|}=\frac{2}{12}+\frac{1}{9}=0,28$

 $d\_{CA}\left(x\_{2},x\_{0}\right)=\frac{\left|x\_{21}-x\_{01}\right|}{\left|x\_{21}\right|+\left|x\_{01}\right|}+\frac{\left|x\_{22}-x\_{02}\right|}{\left|x\_{22}\right|+\left|x\_{02}\right|}=\frac{\left|6-7\right|}{\left|6\right|+\left|7\right|}+\frac{\left|2-4\right|}{\left|2\right|+\left|4\right|}=\frac{1}{13}+\frac{2}{6}=0,41$

 $d\_{CA}\left(x\_{3},x\_{0}\right)=\frac{\left|x\_{31}-x\_{01}\right|}{\left|x\_{31}\right|+\left|x\_{01}\right|}+\frac{\left|x\_{32}-x\_{02}\right|}{\left|x\_{32}\right|+\left|x\_{02}\right|}=\frac{\left|7-7\right|}{\left|7\right|+\left|7\right|}+\frac{\left|2-4\right|}{\left|2\right|+\left|4\right|}=\frac{0}{14}+\frac{2}{6}=0,33$

 $d\_{CA}\left(x\_{4},x\_{0}\right)=\frac{\left|x\_{41}-x\_{01}\right|}{\left|x\_{41}\right|+\left|x\_{01}\right|}+\frac{\left|x\_{42}-x\_{02}\right|}{\left|x\_{42}\right|+\left|x\_{02}\right|}=\frac{\left|9-7\right|}{\left|9\right|+\left|7\right|}+\frac{\left|4-4\right|}{\left|4\right|+\left|4\right|}=\frac{2}{16}+\frac{0}{8}=0,13$

 $d\_{CA}\left(x\_{5},x\_{0}\right)=\frac{\left|x\_{51}-x\_{01}\right|}{\left|x\_{51}\right|+\left|x\_{01}\right|}+\frac{\left|x\_{52}-x\_{02}\right|}{\left|x\_{52}\right|+\left|x\_{02}\right|}=\frac{\left|7-7\right|}{\left|7\right|+\left|7\right|}+\frac{\left|3-4\right|}{\left|3\right|+\left|4\right|}=\frac{0}{14}+\frac{1}{7}=0,14$

 $d\_{CA}\left(x\_{6},x\_{0}\right)=\frac{\left|x\_{61}-x\_{01}\right|}{\left|x\_{61}\right|+\left|x\_{01}\right|}+\frac{\left|x\_{62}-x\_{02}\right|}{\left|x\_{62}\right|+\left|x\_{02}\right|}=\frac{\left|8-7\right|}{\left|8\right|+\left|7\right|}+\frac{\left|2-4\right|}{\left|2\right|+\left|4\right|}=\frac{1}{15}+\frac{2}{6}=0,40$

Seřazení vzdáleností:

$$d\_{CA}\left(x\_{4},x\_{0}\right)<d\_{CA}\left(x\_{5},x\_{0}\right)<d\_{CA}\left(x\_{1},x\_{0}\right)<d\_{CA}\left(x\_{3},x\_{0}\right)<d\_{CA}\left(x\_{6},x\_{0}\right)<d\_{CA}\left(x\_{2},x\_{0}\right)$$

pro $k=1$: nejbližší soused bodu $x\_{0}$je bod$x\_{4}$, protože $d\_{CA}\left(x\_{4},x\_{0}\right)$je nejmenší → testovací subjekt bude zařazen do třídy kontrolních subjektů

pro $k=2$: subjekt zařazen do třídy kontrolních subjektů, protože mezi 2 nejbližšími sousedy jsou 2 kontrolní subjekty a žádný pacient.

pro $k=3$: subjekt zařazen do třídy kontrolních subjektů, protože mezi 3 nejbližšími sousedy je 1 pacient a 2 kontrolní subjekty

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:**

 $d\_{GA}\left(D,x\_{0}\right)=\frac{d\_{CA}\left(x\_{1},x\_{0}\right)+d\_{CA}\left(x\_{2},x\_{0}\right)+d\_{CA}\left(x\_{3},x\_{0}\right)}{3}=\frac{0,28+0,41+0,33}{3}=0,34$

 $d\_{GA}\left(H,x\_{0}\right)=\frac{d\_{CA}\left(x\_{4},x\_{0}\right)+d\_{CA}\left(x\_{5},x\_{0}\right)+d\_{CA}\left(x\_{6},x\_{0}\right)}{3}=\frac{0,13+0,14+0,40}{3}=0,22$

Protože $d\_{GA}\left(D,x\_{0}\right)>d\_{GA}\left(H,x\_{0}\right)$, testovacísubjekt bude zařazen do třídy kontrolních subjektů.

**4.3 Centroidová metoda + Canberrská metrika:**

 $d\_{CE}\left(D,x\_{0}\right)=d\_{CA}\left(\overbar{x}\_{D},x\_{0}\right)=\frac{\left|6-7\right|}{\left|6\right|+\left|7\right|}+\frac{\left|3-4\right|}{\left|3\right|+\left|4\right|}=\frac{1}{13}+\frac{1}{7}=0,22$

 $d\_{CE}\left(H,x\_{0}\right)=d\_{CA}\left(\overbar{x}\_{H},x\_{0}\right)=\frac{\left|8-7\right|}{\left|8\right|+\left|7\right|}+\frac{\left|3-4\right|}{\left|3\right|+\left|4\right|}=\frac{1}{15}+\frac{1}{7}=0,21$

Protože $d\_{CE}\left(D,x\_{0}\right)>d\_{CE}\left(H,x\_{0}\right)$, testovacísubjekt bude zařazen do třídy kontrolních subjektů.

**Centroidová metoda s využitím medoidu:**

 $d\_{CEM}\left(D,x\_{0}\right)=d\_{CA}\left(\tilde{x}\_{D},x\_{0}\right)=d\_{CA}\left(x\_{2},x\_{0}\right)=\frac{\left|6-7\right|}{\left|6\right|+\left|7\right|}+\frac{\left|2-4\right|}{\left|2\right|+\left|4\right|}=\frac{1}{13}+\frac{2}{6}=0,41$

 $d\_{CEM}\left(H,x\_{0}\right)=d\_{CA}\left(\tilde{x}\_{H},x\_{0}\right)=d\_{CA}\left(x\_{5},x\_{0}\right)=\frac{\left|7-7\right|}{\left|7\right|+\left|7\right|}+\frac{\left|3-4\right|}{\left|3\right|+\left|4\right|}=\frac{0}{14}+\frac{1}{7}=0,14$

Protože $d\_{CEM}\left(D,x\_{0}\right)>d\_{CEM}\left(H,x\_{0}\right)$, testovacísubjekt bude zařazen do třídy kontrolních subjektů.

**5.1 Metoda *k* nejbližších sousedů + Mahalanobisova metrika:**

Nejprve je potřeba vypočítat výběrové kovarianční matice pro třídu pacientů a kontrol, tzn. $S\_{D}=\left[\begin{matrix}1&-1,5\\-1,5&3\end{matrix}\right]$ a $S\_{H}=\left[\begin{matrix}1&0,5\\0,5&1\end{matrix}\right]$ (výpočet výběrových kovariančních matic lze nalézt ve Cvičení 1) a jejich inverzi její inverzi $S\_{D}^{-1}=\left[\begin{matrix}4&2\\2&{4}/{3}\end{matrix}\right]$ a $S\_{H}^{-1}=\left[\begin{matrix}{4}/{3}&{-2}/{3}\\{-2}/{3}&{4}/{3}\end{matrix}\right]$ .

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

 $d\_{MA}\left(x\_{1},x\_{0}\right)=\sqrt{\left(x\_{1}-x\_{0}\right)^{T}∙S\_{D}^{-1}∙\left(x\_{1}-x\_{0}\right)}=\sqrt{\left[\begin{matrix}5-7&5-4\end{matrix}\right]∙\left[\begin{matrix}4&2\\2&{4}/{3}\end{matrix}\right]∙\left[\begin{matrix}5-7\\5-4\end{matrix}\right]}=\sqrt{\frac{28}{3}}=3,06$

 $d\_{MA}\left(x\_{2},x\_{0}\right)=\sqrt{\left(x\_{2}-x\_{0}\right)^{T}∙S\_{D}^{-1}∙\left(x\_{2}-x\_{0}\right)}=\sqrt{\left[\begin{matrix}6-7&2-4\end{matrix}\right]∙\left[\begin{matrix}4&2\\2&{4}/{3}\end{matrix}\right]∙\left[\begin{matrix}6-7\\2-4\end{matrix}\right]}=\sqrt{\frac{52}{3}}=4,16$

 $d\_{MA}\left(x\_{3},x\_{0}\right)=\sqrt{\left(x\_{3}-x\_{0}\right)^{T}∙S\_{D}^{-1}∙\left(x\_{3}-x\_{0}\right)}=\sqrt{\left[\begin{matrix}7-7&2-4\end{matrix}\right]∙\left[\begin{matrix}4&2\\2&{4}/{3}\end{matrix}\right]∙\left[\begin{matrix}7-7\\2-4\end{matrix}\right]}=\sqrt{\frac{16}{3}}=2,31$

 $d\_{MA}\left(x\_{4},x\_{0}\right)=\sqrt{\left(x\_{4}-x\_{0}\right)^{T}∙S\_{H}^{-1}∙\left(x\_{4}-x\_{0}\right)}=\sqrt{\left[\begin{matrix}9-7&4-4\end{matrix}\right]∙\left[\begin{matrix}{4}/{3}&{-2}/{3}\\{-2}/{3}&{4}/{3}\end{matrix}\right]∙\left[\begin{matrix}9-7\\4-4\end{matrix}\right]}=\sqrt{\frac{16}{3}}=2,31$

 $d\_{MA}\left(x\_{5},x\_{0}\right)=\sqrt{\left(x\_{5}-x\_{0}\right)^{T}∙S\_{H}^{-1}∙\left(x\_{5}-x\_{0}\right)}=\sqrt{\left[\begin{matrix}7-7&3-4\end{matrix}\right]∙\left[\begin{matrix}{4}/{3}&{-2}/{3}\\{-2}/{3}&{4}/{3}\end{matrix}\right]∙\left[\begin{matrix}7-7\\3-4\end{matrix}\right]}=\sqrt{\frac{4}{3}}=1,15$

 $d\_{MA}\left(x\_{6},x\_{0}\right)=\sqrt{\left(x\_{6}-x\_{0}\right)^{T}∙S\_{H}^{-1}∙\left(x\_{6}-x\_{0}\right)}=\sqrt{\left[\begin{matrix}8-7&2-4\end{matrix}\right]∙\left[\begin{matrix}{4}/{3}&{-2}/{3}\\{-2}/{3}&{4}/{3}\end{matrix}\right]∙\left[\begin{matrix}8-7\\2-4\end{matrix}\right]}=\sqrt{\frac{28}{3}}=3,06$

Seřazení vzdáleností:

$$d\_{MA}\left(x\_{5},x\_{0}\right)<d\_{MA}\left(x\_{3},x\_{0}\right)\leq d\_{MA}\left(x\_{4},x\_{0}\right)<d\_{MA}\left(x\_{1},x\_{0}\right)\leq d\_{MA}\left(x\_{6},x\_{0}\right)<d\_{MA}\left(x\_{2},x\_{0}\right)$$

pro $k=1$: nejbližší soused bodu $x\_{0}$je bod$x\_{5}$, protože $d\_{CA}\left(x\_{5},x\_{0}\right)$je nejmenší → testovací subjekt bude zařazen do třídy kontrolních subjektů

pro $k=2$: nelze rozhodnout

pro $k=3$: subjekt zařazen do třídy kontrolních subjektů, protože mezi 3 nejbližšími sousedy je 1 pacient a 2 kontrolní subjekty

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

**5.2 Metoda průměrné vazby + Mahalanobisova metrika:**

 $d\_{GA}\left(D,x\_{0}\right)=\frac{d\_{MA}\left(x\_{1},x\_{0}\right)+d\_{MA}\left(x\_{2},x\_{0}\right)+d\_{MA}\left(x\_{3},x\_{0}\right)}{3}=\frac{3,06+4,16+2,31}{3}=3,18$

 $d\_{GA}\left(H,x\_{0}\right)=\frac{d\_{MA}\left(x\_{4},x\_{0}\right)+d\_{MA}\left(x\_{5},x\_{0}\right)+d\_{MA}\left(x\_{6},x\_{0}\right)}{3}=\frac{2,31+1,15+3,06}{3}=2.17$

Protože $d\_{GA}\left(D,x\_{0}\right)>d\_{GA}\left(H,x\_{0}\right)$, testovacísubjekt bude zařazen do třídy kontrolních subjektů.

**5.3 Centroidová metoda + Mahalanobisova metrika:**

 $d\_{CE}\left(D,x\_{0}\right)=d\_{MA}\left(\overbar{x}\_{D},x\_{0}\right)=\sqrt{\left[\begin{matrix}6-7&3-4\end{matrix}\right]∙\left[\begin{matrix}4&2\\2&{4}/{3}\end{matrix}\right]∙\left[\begin{matrix}6-7\\3-4\end{matrix}\right]}=3,06$

 $d\_{CE}\left(H,x\_{0}\right)=d\_{CA}\left(\overbar{x}\_{H},x\_{0}\right)=\sqrt{\left[\begin{matrix}8-7&3-4\end{matrix}\right]∙\left[\begin{matrix}{4}/{3}&{-2}/{3}\\{-2}/{3}&{4}/{3}\end{matrix}\right]∙\left[\begin{matrix}8-7\\3-4\end{matrix}\right]}=2$

Protože $d\_{CE}\left(D,x\_{0}\right)>d\_{CE}\left(H,x\_{0}\right)$, testovacísubjekt bude zařazen do třídy kontrolních subjektů.

**Centroidová metoda s využitím medoidu:**

 $d\_{CEM}\left(D,x\_{0}\right)=d\_{MA}\left(\tilde{x}\_{D},x\_{0}\right)=d\_{MA}\left(x\_{2},x\_{0}\right)=4,16$

 $d\_{CEM}\left(H,x\_{0}\right)=d\_{MA}\left(\tilde{x}\_{H},x\_{0}\right)=d\_{MA}\left(x\_{5},x\_{0}\right)=1,15$

Protože $d\_{CEM}\left(D,x\_{0}\right)>d\_{CEM}\left(H,x\_{0}\right)$, testovacísubjekt bude zařazen do třídy kontrolních subjektů.

**Výsledky uspořádáme do tabulky:**

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

Je patrné, že výsledek klasifikace se může lišit při použití různých metrik vzdálenosti. Testovací subjekt sice nebyl v žádném ze sledovaných případů zařazen mezi pacienty, ale v mnoha případech nešlo o výsledné klasifikaci rozhodnout, což se v praxi nejčastěji řeší tak, že je testovaný subjekt zařazen do náhodně zvolené třídy.