KLASIFIKAČNÍ METODA k NEJBLIŽŠÍCH SOUSEDŮ A HLOUBKA DAT

Ondřej Vencálek
Přírodovědecká fakulta Univerzity Palackého v Olomouci

12.9.2012

Obsah

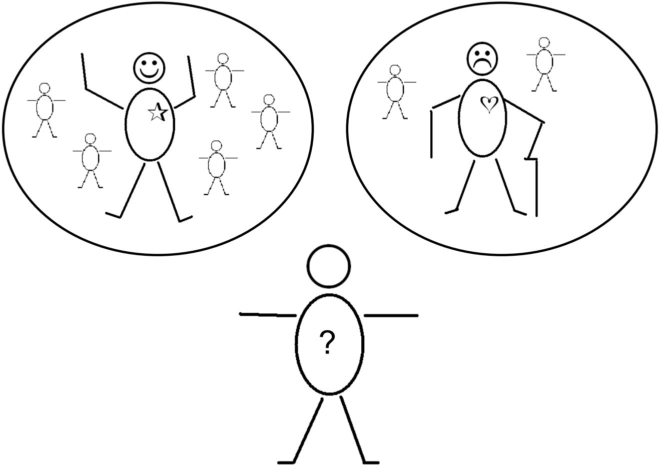
Metoda k nejbližších sousedů

- Úloha klasifikace
- kNN a jádrové odhady hustoty - dvě strany téže mince

Metoda k nejbližších sousedů a hloubka

- Přístup založený na „distribučním“okolí
- Symetrizační přístup
- DD přístup

Úloha klasifikace



$$\mathbf{X} = (X_1, \ldots, X_m) = (vek, BMI, systol.tlak, \ldots)$$

$\mathbf{X} \sim P_1$ (hustota f_1) $\qquad$ $\mathbf{X} \sim P_2$ (hustota f_2)

$$d : \mathbb{R}^m \to \{1, 2\}$$

Optimalita funkce d

- minimalizace pravděpodobnosti chybného zařazení

$$\sum_{i=1}^{K} \mathrm{P}(d(\mathbf{X}) \neq i | \mathbf{X} \sim P_i) \mathrm{P}(\mathbf{X} \sim P_i) \tag{1}$$

- minimalizace střední hodnoty ztráty

$$\sum_{i=1}^{K} \left[\sum_{j=1}^{K} \int_{\{\mathbf{y}: d(\mathbf{y})=j\}} z_{i,j} f_i(\mathbf{y}) \mathrm{d}\mathbf{y} \right] \mathrm{P}(\mathbf{X} \sim P_i) \tag{2}$$

  kde $z_{i,j}$ je ztráta, když objekt ze skupiny i přiřadíme do sk. j; pro $z_{i,j} = 1$ když $i \neq j$ a nula jinak se (2) redukuje na (1).

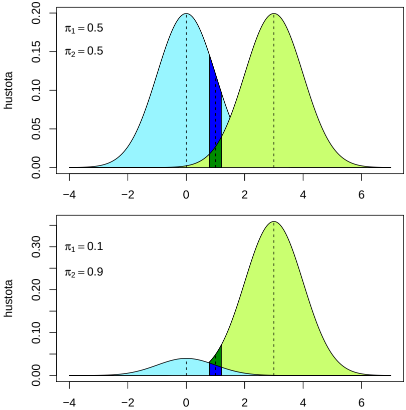
- minimalizace

$$\max_i \mathrm{P}(d(\mathbf{X}) \neq i | \mathbf{X} \sim P_i) \tag{3}$$

Hustoty f_i známé - Bayesovský klasifikátor

$$d(\mathbf{x}) = \arg\max_i \pi_i f_i(\mathbf{x}),$$

kde $\pi_i = \mathrm{P}(\mathbf{X} \sim P_i)$ ... apriorní pravděp.
dk. optimality viz *Antoch a Vorlíčková (1992)*.



Hustoty f_i neznámé

- $d(\mathbf{x}) = \arg\max_i \pi_i f_i(\mathbf{x})$
- $d(\mathbf{x}) = \arg\max_i \widehat{\pi}_i \widehat{f_i(\mathbf{x})}$

Tréningová (trénovací) množina:

Jméno	(věk, BMI, tlak1, ...)	skupina
Jan Novák	(51, 30, 140,...)	♡
Karel Lucemburk	(32, 25, 135,...)	★
Jindřich Feo	(65, 23, 165,...)	♡
Petr Petr	(40, 31, 150,...)	♡
Jakub Hodný	(54, 22, 130,...)	★
⋮		

Skupina ♡

Jméno	(věk, BMI, tlak1, ...)
Jan Novák	(51, 30, 140,...)
Jindřich Feo	(65, 23, 165,...)
Petr Petr	(40, 31, 150,...)
⋮	

Skupina 1

1	$\boldsymbol{x}_{1,1}$
2	$\boldsymbol{x}_{1,2}$
⋮	
n_1	$\boldsymbol{x}_{1,n_1}$

Skupina ★

Jméno	(věk, BMI, tlak1, ...)
Karel Lucemburk	(32, 25, 135,...)
Jakub Hodný	(54, 22, 130,...)
⋮	

Skupina 2

1	$\boldsymbol{x}_{2,1}$
2	$\boldsymbol{x}_{2,2}$
⋮	
n_2	$\boldsymbol{x}_{2,n_2}$

Značení:

$TS_i = \{\mathbf{x}_{i,j}, j = 1, \ldots, n_i\}$ pro $i = 1, \ldots, K$

... i-tá část tréningové množiny

$n = \sum_i n_i$

... celkový počet prvků tréningové množiny

Odhad hustoty f_i

- parametrický přístup
  - LDA (Fisher 1936), QDA
- neparametrický přístup
  - jádrový odhad hustoty (Rosenblatt 1956, Parzen 1962)
  - metoda k nejbližších sousedů (kNN) (Fix a Hodges 1951)

Neparametrický přístup:
mějme bod $\mathbf{x} \in R^m$ a nějaké jeho okolí $L_i(\mathbf{x})$, pak odhad $f_i(\mathbf{x})$ můžeme založit na aproximaci

$$\mathrm{P}(\mathbf{X} \in L_i(\mathbf{x}) | \mathbf{X} \sim P_i) = \int_{L_i(\mathbf{x})} f_i(\mathbf{y}) d\mathbf{y} \cong f_i(\mathbf{x}) \lambda(L_i(\mathbf{x}))$$

$$\boxed{\widehat{f_i(\mathbf{x})} = \frac{k_i}{n_i \lambda(L_i(\mathbf{x}))},}$$

kde k_i ... počet bodů z TS_i, které náleží $L_i(\mathbf{x})$,
n_i ... počet všech bodů z TS_i.

Jádrové odhady hustoty

$\widehat{f_i(\mathbf{x})} = \frac{k_i}{n_i \lambda(L_i(\mathbf{x}))}$,

Nechť $L_1(\mathbf{x}) = L_2(\mathbf{x}) = \ldots = L_K(\mathbf{x}) =: L(\mathbf{x})$
kde $L(\mathbf{x})$ je takové okolí bodu, že $\lambda(L(\mathbf{x})) = V$ (konst.)

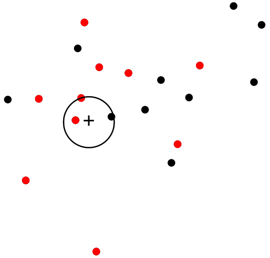
$$
\begin{aligned}
d(\mathbf{x}) = \arg\max_i \widehat{\pi}_i \widehat{f_i(\mathbf{x})} &= \arg\max_i \widehat{\pi}_i \frac{1}{n_i \lambda(L(\mathbf{x}))} \sum_{j=1}^{n_i} I_{[\mathbf{x}_{i,j} \in L(\mathbf{x})]} \\
&= \arg\max_i \widehat{\pi}_i \frac{1}{n_i \lambda(L(\mathbf{x}))} \sum_{j=1}^{n_i} Ker\left(\mathbf{x}, \mathbf{x}_{i,j}, L(\mathbf{x})\right)
\end{aligned}
$$

Metoda k nejbližších sousedů

$\widehat{f_i(\mathbf{x})} = \frac{k_i}{n_i \lambda(L_i(\mathbf{x}))}$,

Nechť $L_1(\mathbf{x}) = L_2(\mathbf{x}) = \ldots = L_K(\mathbf{x}) =: L(\mathbf{x})$
kde $L(\mathbf{x})$ je takové okolí bodu, že $\sum_i k_i = k$ (konst.)

Pro $\widehat{\pi}_i = \frac{n_i}{n}$ je $\arg\max_i \widehat{\pi}_i \widehat{f_i(\mathbf{x})} = \arg\max_i k_i$.



Kolika sousedů se ptát aneb jak zvolit k

(P1) $\lim_{n_i \to \infty} k = \infty$

(P1) $\lim_{n_i \to \infty} k / n_i = 0$

$$\forall \mathbf{x} \in \mathbb{R}^m \quad \widehat{f}_i(\mathbf{x}) = \frac{k_i}{n_i \lambda(L_i(\mathbf{x}))} \xrightarrow{P} f_i(\mathbf{x}) \text{ pro } n_i \to \infty \Leftrightarrow (P1)\&(P2)$$

Pozor: pro pevné $k \in \mathbb{N}$ a $n \in \mathbb{N}$ je $\int_{\mathbb{R}^m} \widehat{f_i(\mathbf{x})} d\mathbf{x} \neq 1$.
Příklad: 1-NN v $\mathbb{R}^1$:

$$\int_{x_{max}}^{\infty} \frac{1}{n(x - x_{max})} dx = \frac{1}{n} \int_0^{\infty} \frac{1}{x} dx = \infty.$$

O nejbližším sousedovi aneb když $k = 1$

$\mathrm{E}_{1NN} = \mathrm{P}(\text{chybné zařazení pomocí 1-NN})$
$\mathrm{E}_{Bayes} = \mathrm{P}(\text{chybné zařazení pomocí Bayes. klasifikátoru})$

$$\mathrm{E}_{1NN} \leq 2\mathrm{E}_{Bayes}$$

Přesněji: pro $K \geq 2$ skupin platí

$$\mathrm{E}_{1NN} \leq \mathrm{E}_{Bayes}\left(2 - \frac{K}{K-1}\mathrm{E}_{Bayes}\right)$$

dk. viz *Hand (1981)*.

Obsah

Metoda k nejbližších sousedů

- Úloha klasifikace
- kNN a jádrové odhady hustoty - dvě strany téže mince

Metoda k nejbližších sousedů a hloubka

- Přístup založený na „distribučním“okolí
- Symetrizační přístup
- DD přístup

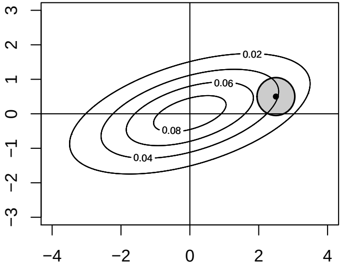
*k*NN s využitím hloubky, přístup „distribučního“ okolí

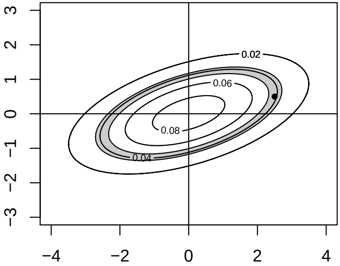
Připomeňme:

$$P(\mathbf{X} \in L(\mathbf{x})) \cong f(\mathbf{x}) \cdot \lambda_d (L(\mathbf{x})),$$

kde $L(\mathbf{x})$ je nějaké okolí bodu $\mathbf{x}$:

- $L(\mathbf{x}) = \left\{\mathbf{y} \in \mathbb{R}^d : \|\mathbf{x} - \mathbf{y}\| < \eta\right\}$
- $L(\mathbf{x}; P) = \left\{\mathbf{y} \in \mathbb{R}^d : |f(\mathbf{x}; P) - f(\mathbf{y}; P)| < \eta\right\}.$





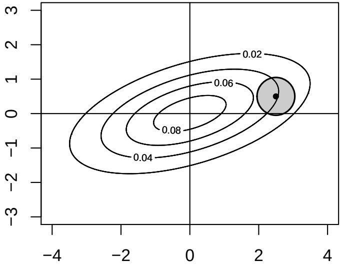
kNN s využitím hloubky, přístup „distribučního“ okolí

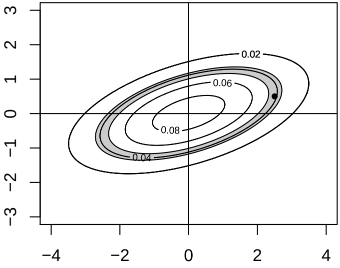
Připomeňme:

$$P(\mathbf{X} \in L(\mathbf{x})) \cong f(\mathbf{x}) \cdot \lambda_d \left(L(\mathbf{x}) \right),$$

kde $L(\mathbf{x})$ je nějaké okolí bodu $\mathbf{x}$:

- $L(\mathbf{x}) = \left\{ \mathbf{y} \in \mathbb{R}^d : \|\mathbf{x} - \mathbf{y}\| < \eta \right\}$
- $L(\mathbf{x}; P) = \left\{ \mathbf{y} \in \mathbb{R}^d : |D(\mathbf{x}; P) - D(\mathbf{y}; P)| < \eta \right\}.$





kNN s využitím hloubky, přístup „distribučního“ okolí

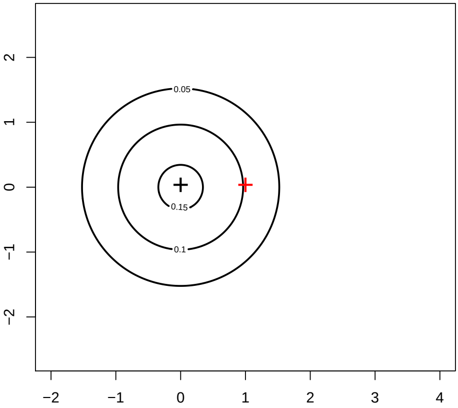
$$f_i(\mathbf{x}) = h_i(D(\mathbf{x}; P_i)) \qquad i = 1, \ldots, K$$

$$\widehat{\pi}_i \widehat{f}_i(\mathbf{x}) = \frac{n_i}{n} \frac{k_i}{n_i} \frac{1}{\widehat{\lambda}_d(L_i(\mathbf{x}))} = \frac{k_i}{n\widehat{\lambda}_d(L_i(\mathbf{x}))}$$

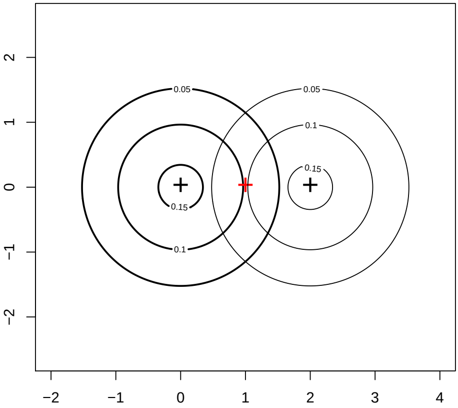
$$d(\mathbf{x}) = \arg \min_{i=1,\ldots,K} \widehat{\lambda}_d(L(\mathbf{x}, \widehat{P}_i)),$$

kde $L(\mathbf{x}, \widehat{P}_i)$ je „distribučním“ okolí bodu $\mathbf{x}$, které obsahuje právě k bodů z TS_i.

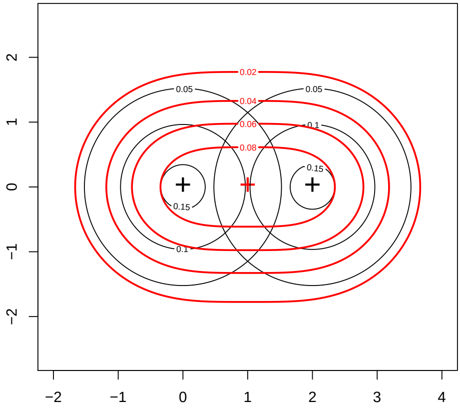
kNN s využitím hloubky, symetrizační přístup



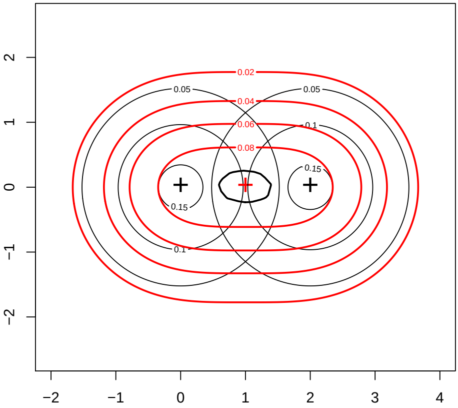
kNN s využitím hloubky, symetrizační přístup



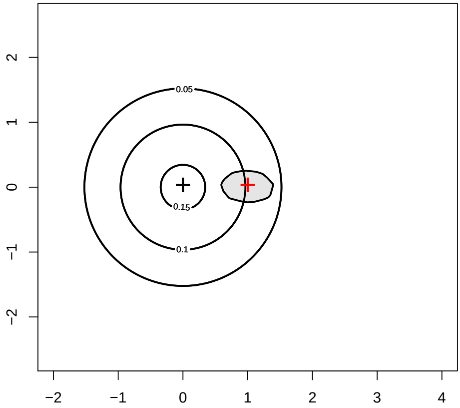
kNN s využitím hloubky, symetrizační přístup



*k*NN s využitím hloubky, symetrizační přístup



kNN s využitím hloubky, symetrizační přístup

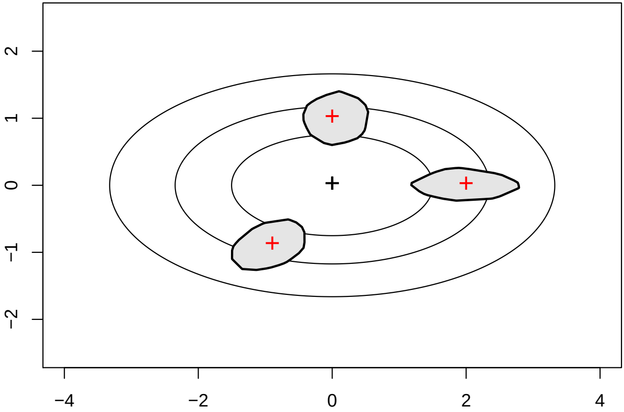


kNN s využitím hloubky, symetrizační přístup

Příklad:

Okolí bodů [0,1], [2,0] a $[\frac{2}{5}\sqrt{5}, \frac{2}{5}\sqrt{5}]$ vzhledem k rozdělení

$$N_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 4 & 0 \\ 0 & 1 \end{pmatrix}\right).$$



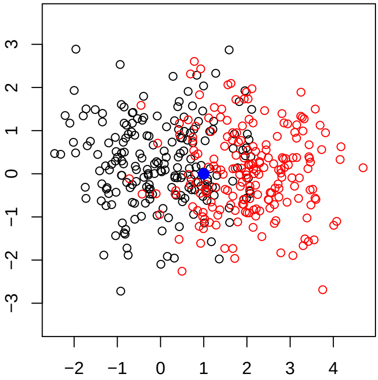
kNN s využitím hloubky, symetrizační přístup

Značení:
$\mathbf{X}_1, \ldots, \mathbf{X}_n$ ... všechny body tréningové množiny
$\mathbf{x}$ ... nové pozorování

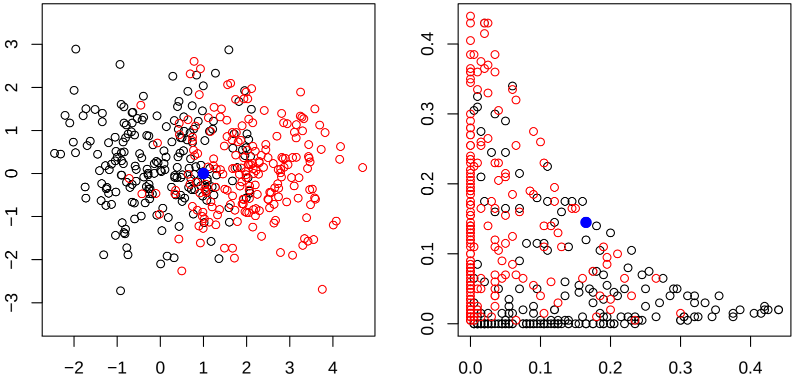
Postup:

1. „Reflexe": $\mathbf{X}_{n+i} := 2\mathbf{x} - \mathbf{X}_i$ pro $i = 1, \ldots, n$, body $\mathbf{X}_1, \ldots, \mathbf{X}_{2n}$ určují rozdělení $P_{\mathbf{x}}^{(n)}$.
2. Seřaďme body $\mathbf{X}_1, \ldots, \mathbf{X}_n$ tak, aby platilo $D(\mathbf{X}_{(1)}, P_{\mathbf{x}}^{(n)}) \geq D(\mathbf{X}_{(2)}, P_{\mathbf{x}}^{(n)}) \geq \ldots \geq D(\mathbf{X}_{(n)}, P_{\mathbf{x}}^{(n)})$.
3. Pro libovolné $k \in \{1, \ldots, n\}$ představují body $\mathbf{X}_{(i)}, i = 1, \ldots, k$, k nejbližších sousedů bodu $\mathbf{x}$.

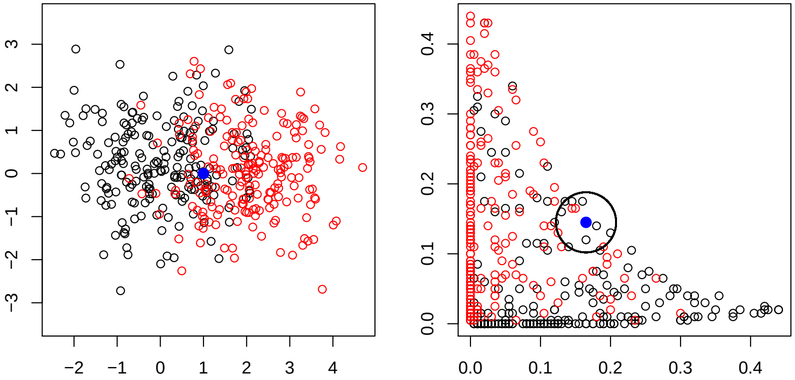
*k*NN s využitím hloubky, DD přístup



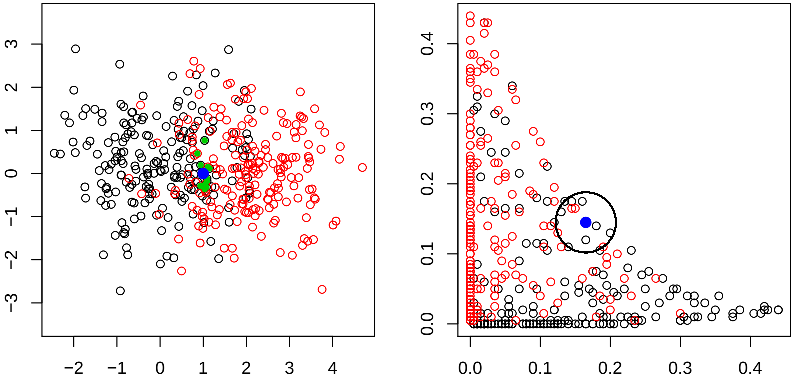
*k*NN s využitím hloubky, DD přístup



*k*NN s využitím hloubky, DD přístup



*k*NN s využitím hloubky, DD přístup



kNN s využitím hloubky, DD přístup