Využití hloubky dat pro klasifikaci

– globální a lokální přístupy

Ondřej Vencálek
Přírodovědecká Fakulta Univerzity Palackého v Olomouci.

2014-01-24, Robust (Jetřichovice)

Poloprostorová hloubka

Motivace: charakteristika „odlehlosti“ („centrality“).
Poloprostorová hloubka bodu $x \in \mathbb{R}^1$ vzhledem
k pravděpodobnostní míře P na $\mathbb{R}^1$ je definována vztahem

$$D(x; P) = \min\left(P(X \leq x), P(X \geq x)\right), \text{kde } X \sim P.$$

Poloprostorová hloubka bodu $\mathbf{x} \in \mathbb{R}^m$ ($d \in \mathbb{N}$) vzhledem
k pravděpodobnostní míře P na $\mathbb{R}^m$ je definována vztahem

$$D(\mathbf{x}; P) = \inf_{\mathbf{u}:\|\mathbf{u}\|=1} D(\mathbf{u}^T\mathbf{x}; P).$$

Ekvivalentně:

$$D(\mathbf{x}; P) = \inf_{\mathbb{H}} \left\{P(\mathbb{H}) : \mathbb{H} \text{ uzavřený poloprostor v } \mathbb{R}^m : \mathbf{x} \in \mathbb{H}\right\}.$$

Vlastnosti poloprostorové hloubky

Označme symbolem $\mathcal{P}$ třídu všech pravděpodobnostních rozdělení na $\mathbb{R}^d$ a $P_{\mathbf{X}}$ rozdělení náhodného vektoru $\mathbf{X}$. Poloprostorová hloubka má následující vlastnosti:

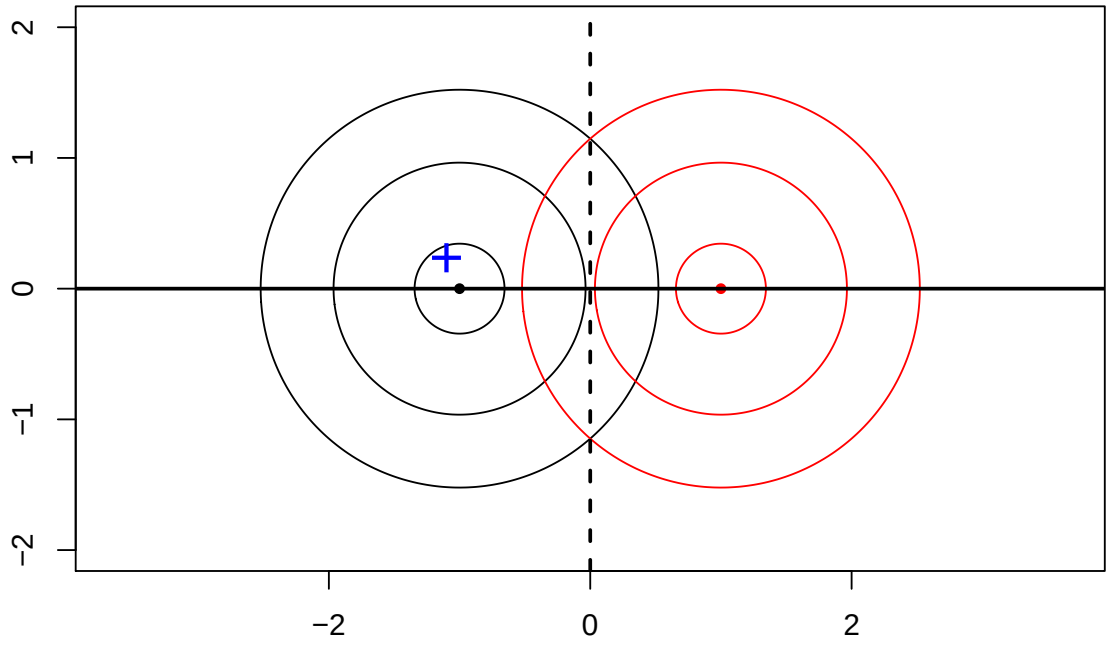
1. $D(\mathbf{Ax}+\mathbf{b}; P_{\mathbf{AX}+\mathbf{b}}) = D(\mathbf{x}; P_{\mathbf{X}})$ pro libovolný náhodný vektor $\mathbf{X}$ na $\mathbb{R}^d$, pozitivně definitní matici $\mathbf{A}$ $(d \times d)$ a d–rozměrný vektor $\mathbf{b}$;
2. $D(\mathbf{x}; P) \to 0$ pro $\|\mathbf{x}\| \to \infty$ pro všechna $P \in \mathcal{P}$;
3. $D(\boldsymbol{\theta}; P) = \sup_{\mathbf{x}\in\mathbb{R}^d} D(\mathbf{x}; P)$ pro všechna $P \in \mathcal{P}$ s centrem v $\boldsymbol{\theta}$;
4. $D(\mathbf{x}; P) \leq D(\boldsymbol{\theta} + \alpha(\mathbf{x}-\boldsymbol{\theta}), P)$ pro všechna $\alpha \in [0,1]$ a rozdělení $P \in \mathcal{P}$ s nejhlubším bodem $\boldsymbol{\theta}$.

Přehled hloubkových funkcí

Zhruba od přelomu 80. a 90. let začíná být hlobka chápána jako "libovolná" funkce poskytující uspořádání bodů podle jejich centrality. Postupně bylo vymyšleno mnoho různých funkcí hloubek. Začaly se vyvíjet inferenční metody založené na uspořádání pomocí hloubky.

- Half-space depth (poloprostorová hloubka)
- Simplicial depth (simplexová hloubka)
- L_1 depth (=spatial depth)
- Zonoid depth
- Mahalanobis depth, projection depth, covex hull peeling depth, likelihood depth, Oja depth, simplicial volume depth, L^p–depth, majority depth, různá zobecnění jako regression depth, tangent depth, proximity graph data measures

Jak využít hloubku v úloze klasifikace



Klasifikace na základě maximální hloubky
(Maximal depth classifier):

$$d(\mathbf{x}) = \arg\max_{i=1,2} D(\mathbf{x}; \widehat{P}_i)$$

Klasifikace na základě maximální hloubky

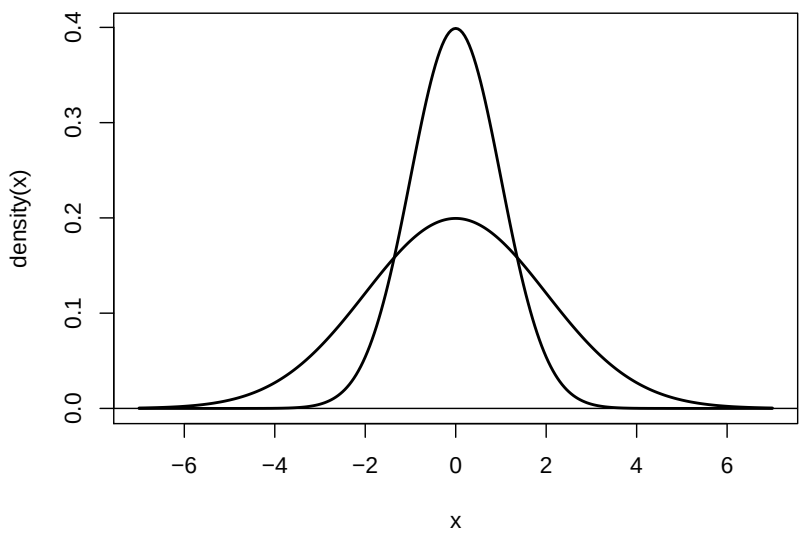
- L_1-hloubka – *Jörnsten (2004), Hartikainen and Oja (2006)*,
- poloprostorová hloubka – *Ghosh and Chaudhuri (2005)**.
- zonoidová hloubka (+Mahalan.) – *Mosler and Hoberg (2006)*,
- projekční hloubka – *Kosiorowski (2008), Hubert and Van der Veeken (2010), Dutta and Ghosh (2009/2012)**

Vlastnosti klasifikátoru založeného na maximální hloubce

Klasifikátor založený na maximální hloubce je Bayesovsky optimální za předpokladu, že rozdělení P_1, P_2

- jsou elipticky symetrická s hustotou klesající ze středu symetrie
- liší se pouze parametrem polohy,
- mají stejnou apriorní pravděp.: $\pi_1 = \pi_2 = 1/2$
- použitá hloubka je afinně invariantní.

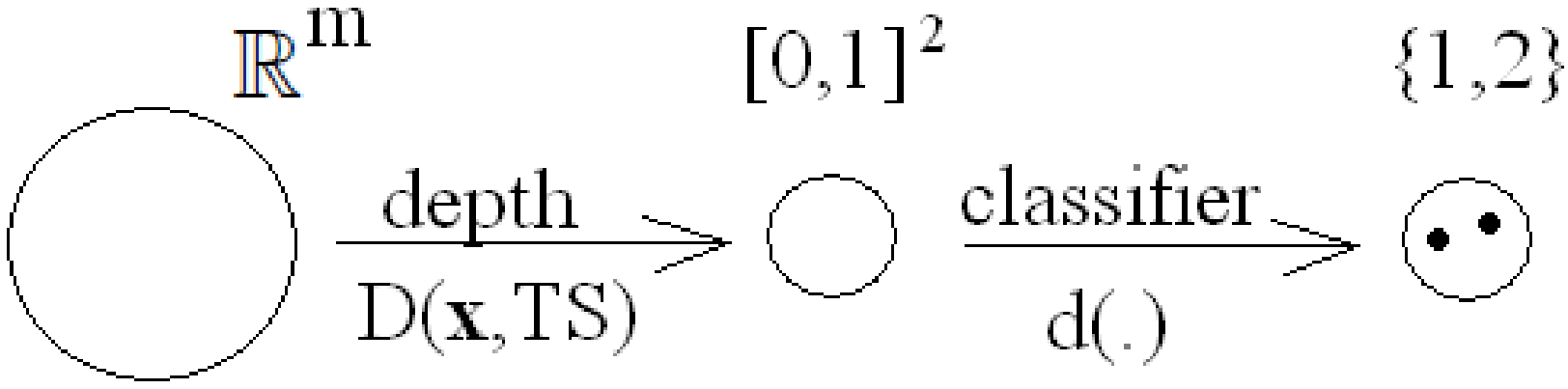
Problém různých disperzí:



$P_1 = N(0, \sigma_1^2)$,
$P_2 = N(0, \sigma_2^2)$,
$0 < \sigma_1 < \sigma_2$

Pro všechna $x \neq 0$ je
$D(x, P_1) < D(x, P_2)$

Schéma typické klasifikační procedury založené na hloubce



1. Jakou hloubku použít?
2. Jaký klasifikátor použít?
3. Jaké výhody může mít použití hloubky dat v klasifikaci oproti běžným metodám?

Klasifikátory založené na hloubce

- Globální hloubka + Globální klasifikátor:
  - Depth transvariation classifier (2008)
  - DD-plot classifier (2010/2012)
  - DDα- procedure (2012)
- Globální hloubka + Lokální klasifikátor
  - + jádrové odhady hustoty (2005)
  - + k nejbližších sousedů (2011, 2012, 2013)
- Lokální hloubka + Globální/Lokální klasifikátor

Globální hloubka + Globální klasifikátor I

Billor et al. (2008)
depth transvariation classifier
(maximal central area classifier)

$$d(\mathbf{x}) = \arg\max_{i=1,2} \mathrm{P}\left(D(\mathbf{X}_i; P_i) \leq D(\mathbf{x}; P_i)\right),$$

$$d(\mathbf{x}) = \arg\max_{i=1,2} \frac{1}{n_i} \sum_{j=1}^{n_i} I\left(D(\mathbf{X}_{i,j}, \widehat{P}_i) \leq D(\mathbf{x}; \widehat{P}_i)\right)$$

+ *Hubert and Van der Veeken (2010)*

Globální hloubka + Globální klasifikátor II

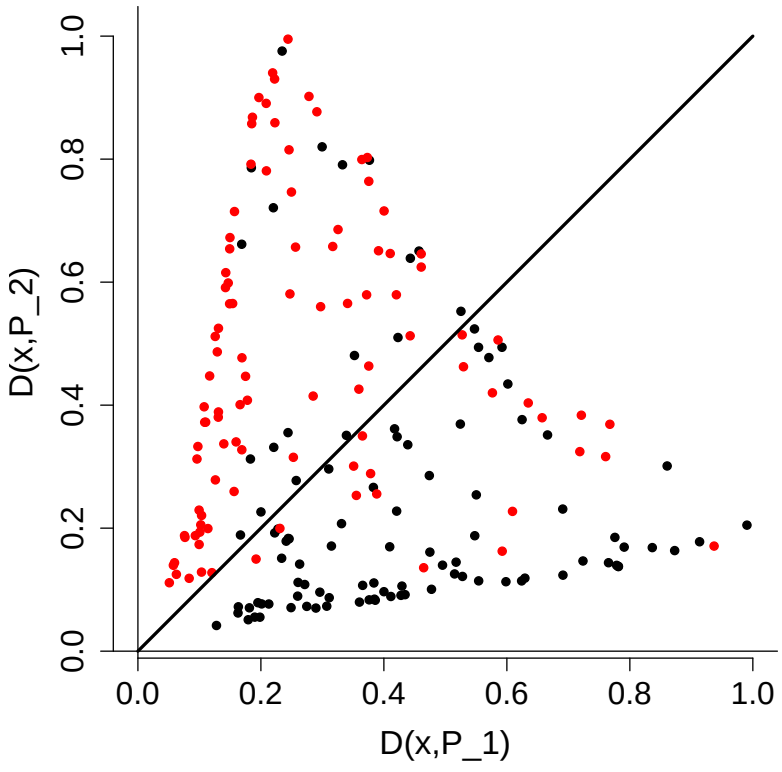
Li, Cuesta-Albertos, Liu (2010/2012)
DD-plot classifier

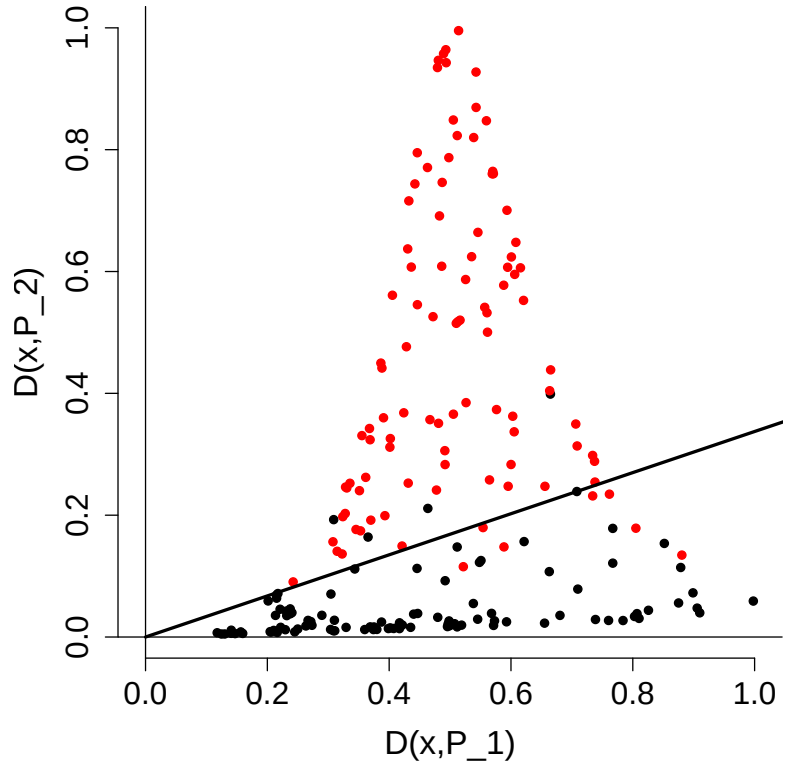
$$\begin{aligned} D(\mathbf{x}; \hat{P}_2) > \widehat{k} D(\mathbf{x}; \hat{P}_1) &\implies d(\mathbf{x}) = 2 \\ D(\mathbf{x}; \hat{P}_2) < \widehat{k} D(\mathbf{x}; \hat{P}_1) &\implies d(\mathbf{x}) = 1, \end{aligned} \tag{1}$$

kde $\widehat{k}$ je směrnice oddělující přímky odhadnutá minimalizací počtu chybných zařazení bodů tréningové množiny:
$\widehat{k} = \arg\min_k \widehat{\Delta}(k)$, kde

$$\widehat{\Delta}(k) = \widehat{\pi_1} \frac{1}{n_1} \sum_{i=1}^{n_1} I_{[D(\mathbf{x}_{1,i}; \hat{P}_2) > kD(\mathbf{x}_{1,i}; \hat{P}_1)]} + \widehat{\pi_2} \frac{1}{n_2} \sum_{j=1}^{n_2} I_{[D(\mathbf{x}_{2,j}; \hat{P}_2) < kD(\mathbf{x}_{2,j}; \hat{P}_1)]}.$$

Globální hloubka + Globální klasifikátor II





Vlevo: $P_1 = N\left((-1,0)^T, \mathbf{I}\right), P_2 = N\left((1,0)^T, \mathbf{I}\right)$

Vpravo: $P_1 = N\left(\mathbf{0}, 16\mathbf{I}\right), P_2 = N\left((2,2)^T, \mathbf{I}\right)$

Globální hloubka + Globální klasifikátor III

Lange, Mosler, Mozharovskyi (2012)
DDα-classifier

$$\left[D(\mathbf{x}, \hat{P}_1), D(\mathbf{x}, \hat{P}_2)\right]$$

$$\mathbf{z} := \left[D(\mathbf{x}, \hat{P}_1), D(\mathbf{x}, \hat{P}_2), D(\mathbf{x}, \hat{P}_1) \cdot D(\mathbf{x}, \hat{P}_2), D(\mathbf{x}, \hat{P}_1)^2, D(\mathbf{x}, \hat{P}_2)^2\right]$$

$$\text{feature space } \ldots \mathbf{Z} = \{\mathbf{z}_i, i = 1, \ldots, n_1 + n_2\}$$

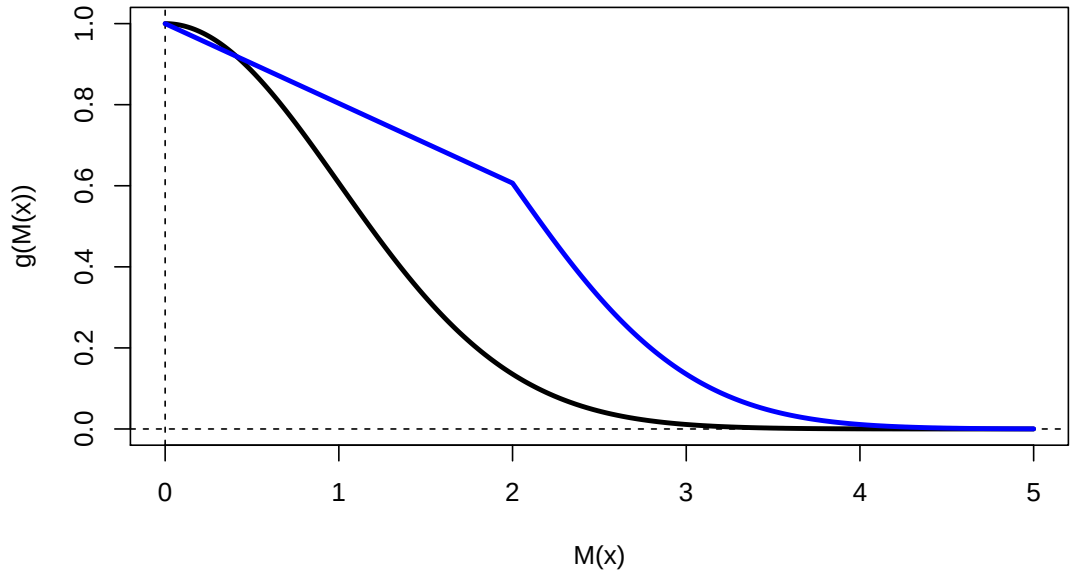
separating surface:
$$aD(\mathbf{x}, \hat{P}_1) + bD(\mathbf{x}, \hat{P}_2) + cD(\mathbf{x}, \hat{P}_1)D(\mathbf{x}, \hat{P}_2) + dD(\mathbf{x}, \hat{P}_1)^2 + eD(\mathbf{x}, \hat{P}_2)^2 = 0$$

Globální hloubka + Lokální klasifikátor I

Ghosh, Chaudhuri (2005) + jadrové odhady hustoty

Hustota elipticky symetrických rozdělení:

$$f(\mathbf{x}) = k \cdot g(M(\mathbf{x})), \quad \text{kde } M(\mathbf{x}) \text{ je Mahalanobisova vzdálenost}$$

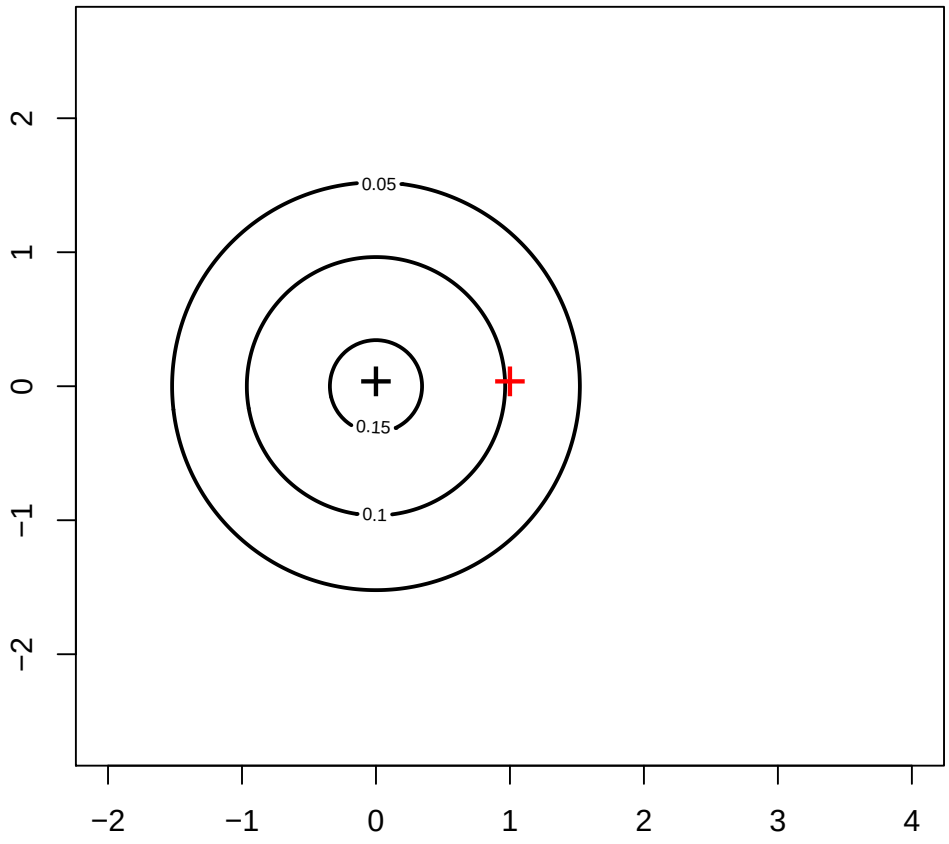


Bayesovský klasifikátor:

$$d(\mathbf{x}) = \arg\max_{i=1,2} \pi_i \theta_i \left(D(\mathbf{x}; P_i)\right)$$

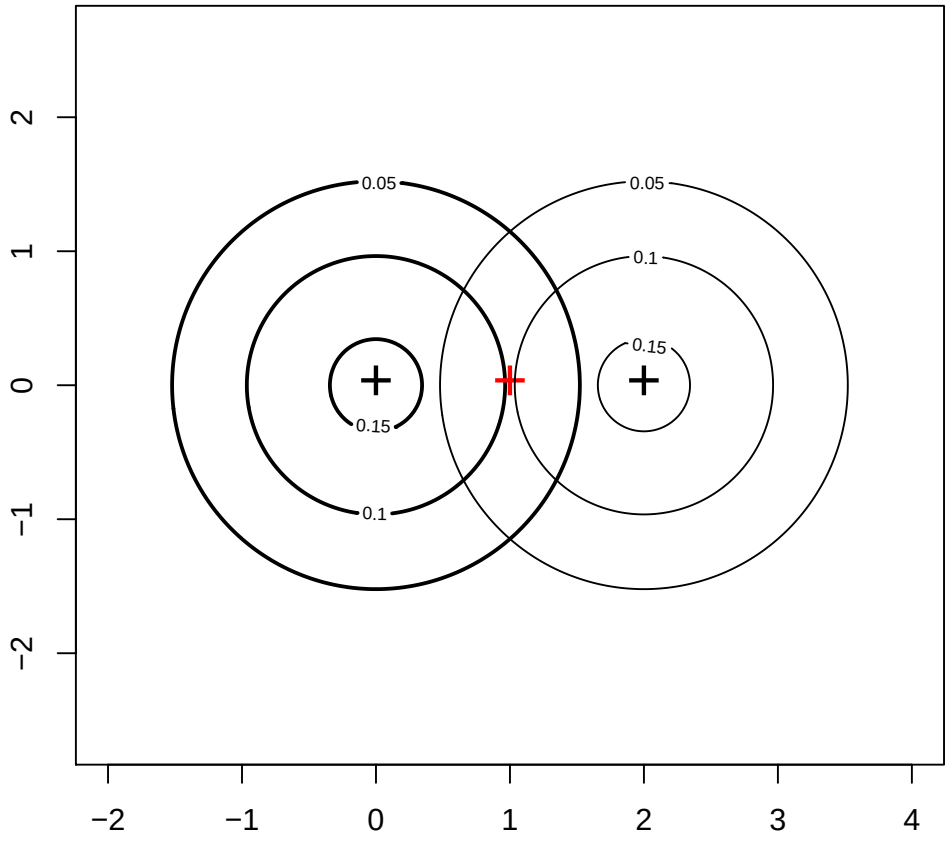
Globální hloubka + Lokální klasifikátor II

Paindaveine, Van Bever (2012) - kNN + symetrizace
(mimo schéma)



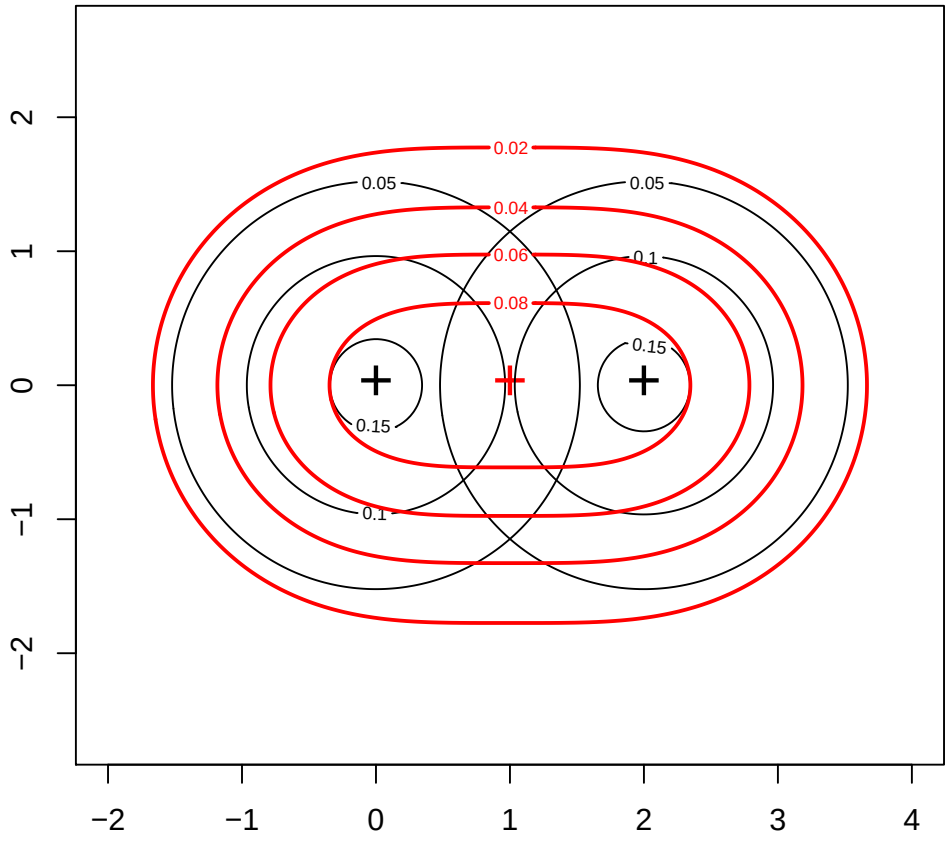
Globální hloubka + Lokální klasifikátor II

Paindaveine, Van Bever (2012) - kNN + symetrizace
(mimo schéma)



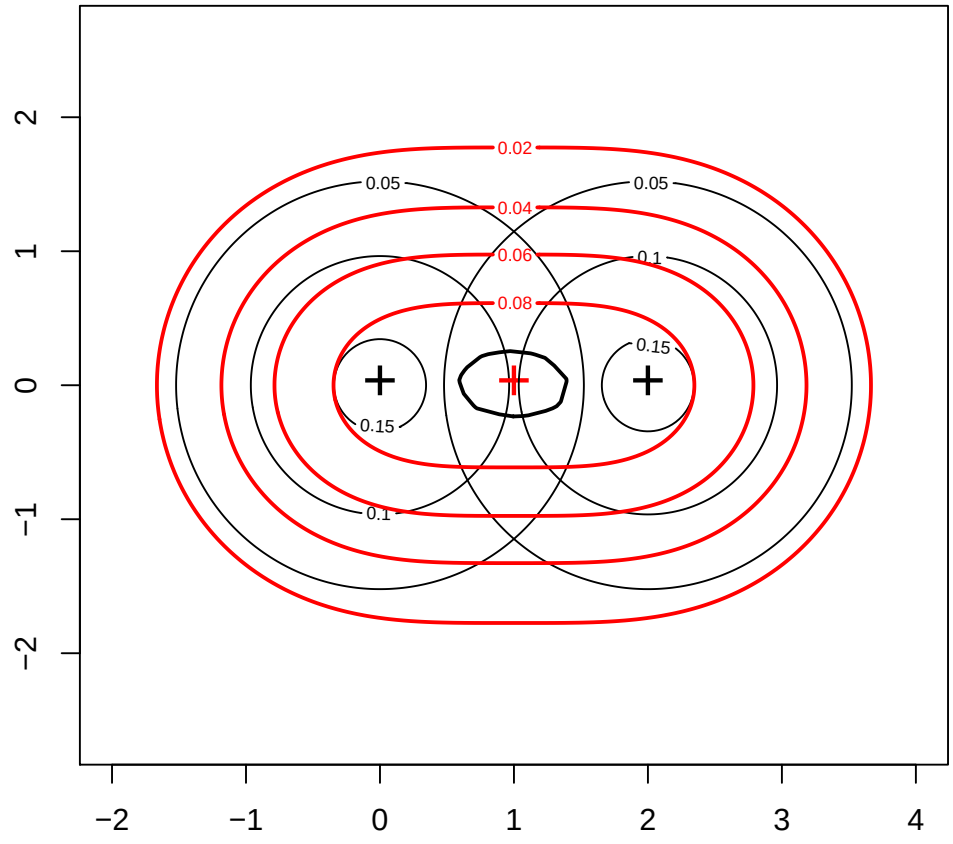
Globální hloubka + Lokální klasifikátor II

Paindaveine, Van Bever (2012) - kNN + symetrizace
(mimo schéma)



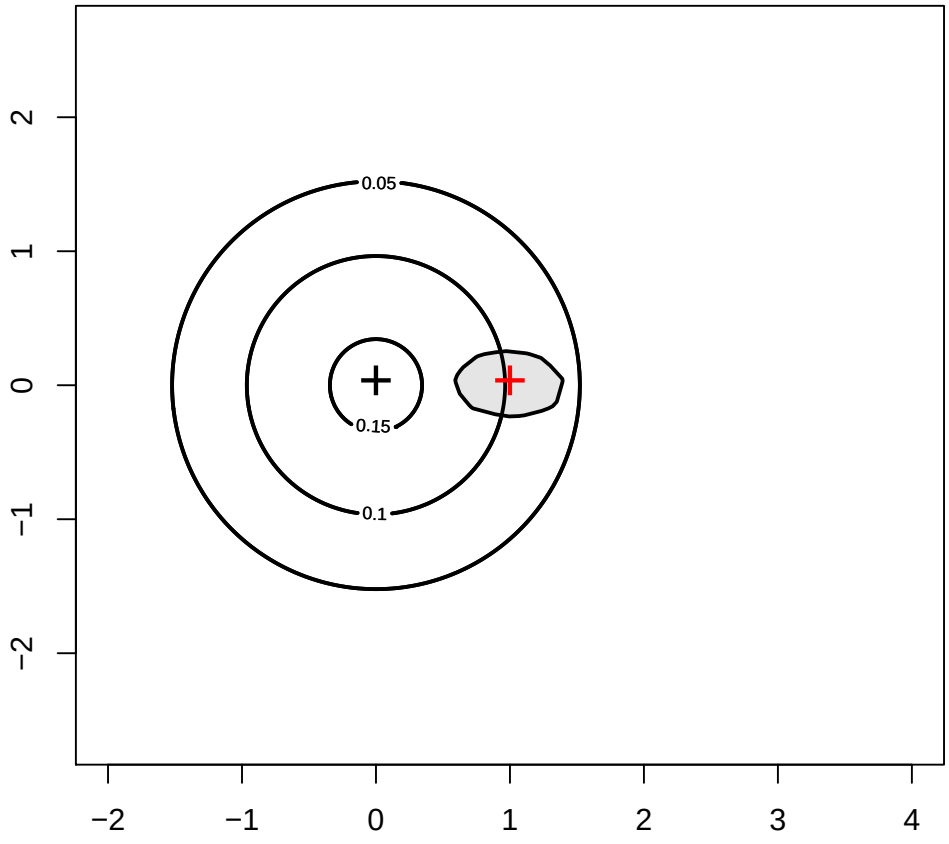
Globální hloubka + Lokální klasifikátor II

Paindaveine, Van Bever (2012) - kNN + symetrizace
(mimo schéma)



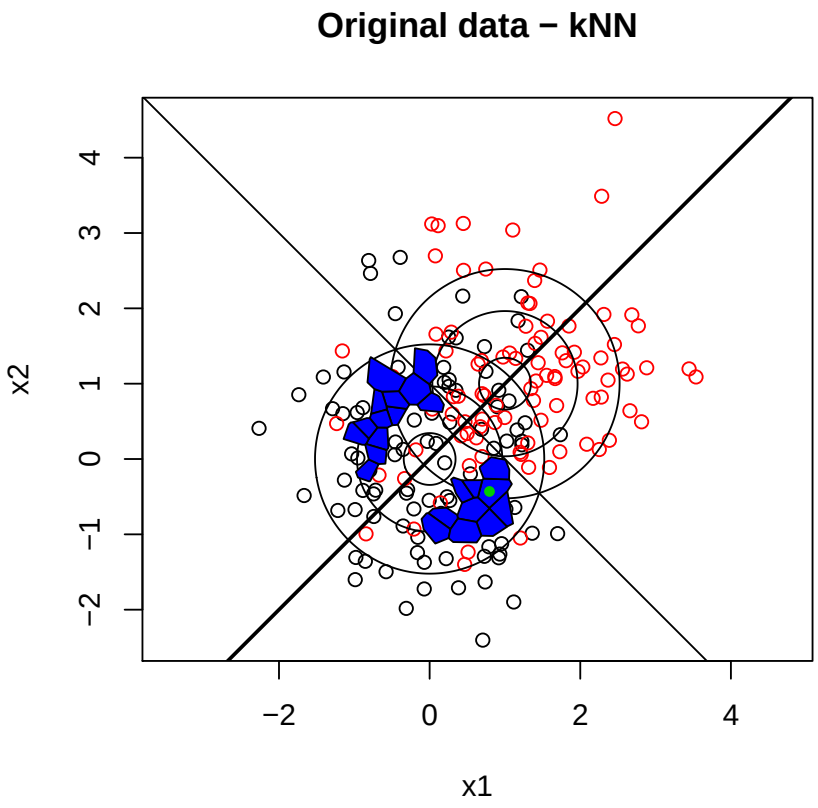
Globální hloubka + Lokální klasifikátor II

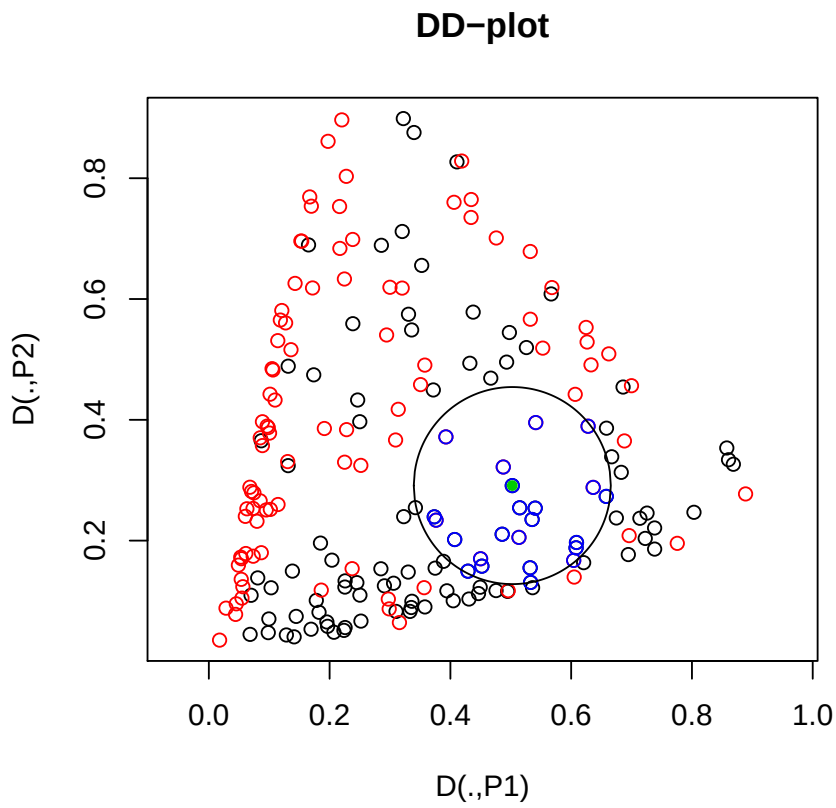
Paindaveine, Van Bever (2012) - kNN + symetrizace
(mimo schéma)



Globální hloubka + Lokální klasifikátor III

k nejbližších sousedů podle hloubky
k-Depth-Nearest Neighbour Method (k-Depth-NN):





Simulace I - dvě normální rozdělení lišící se (jen) disperzí

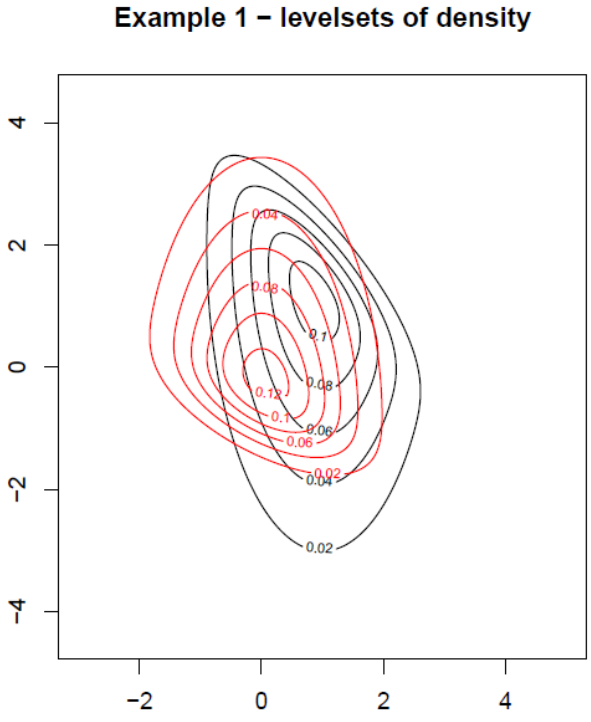
$$P_1 = N_d(\mathbf{0}, \mathbf{I})$$
$$P_2 = N_d(\mathbf{0}, v\mathbf{I})$$

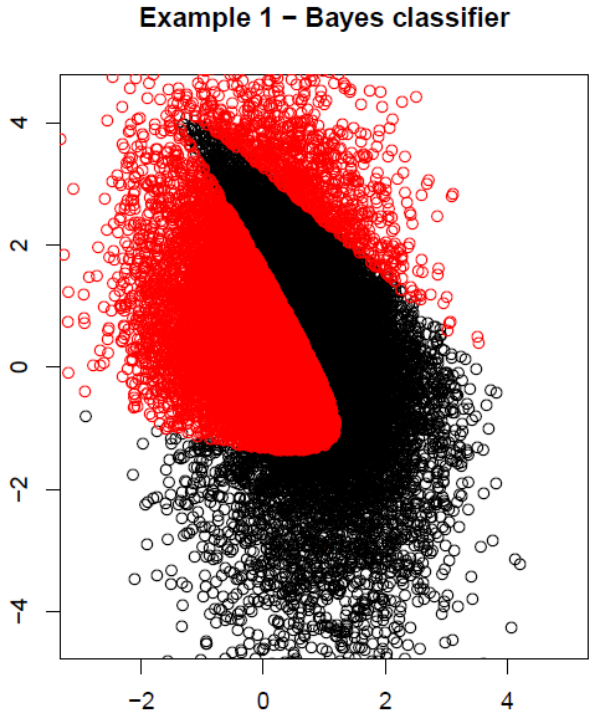
d	v	Bayes	kNN		$k-Depth-NN$	
		ER_B	ER_C	ER_C - ER_B	ER_M	ER_M - ER_B
2	2.25	0.356	0.402	0.046	0.362	0.006
	4	0.265	0.298	0.033	0.274	0.010
5	2.25	0.268	0.390	0.122	0.281	0.014
	4	0.148	0.268	0.120	0.156	0.008
10	2.25	0.186	0.408	0.222	0.192	0.006
	4	0.065	0.313	0.248	0.074	0.009

Chybovost: Average misclassification rates (ER)
Tři klasifikátory: Bayes (index B), kNN (index C) a k-Depth-NN (index M)
Symboly: d . . . dimenze, v . . . rozptyl.

Simulace II - sešikmená normální rozdělení

Skewed-normal distributions: *Azzalini and Dalla Valle (1996)*

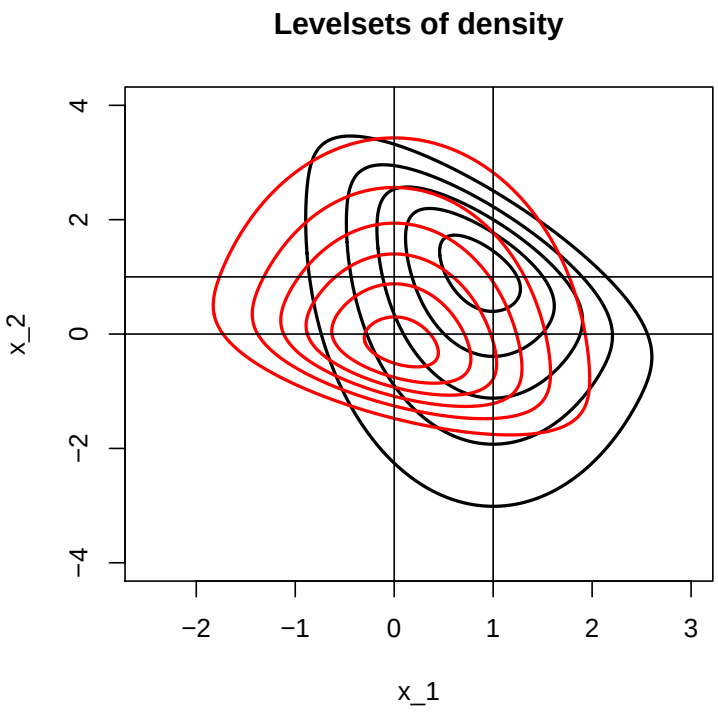


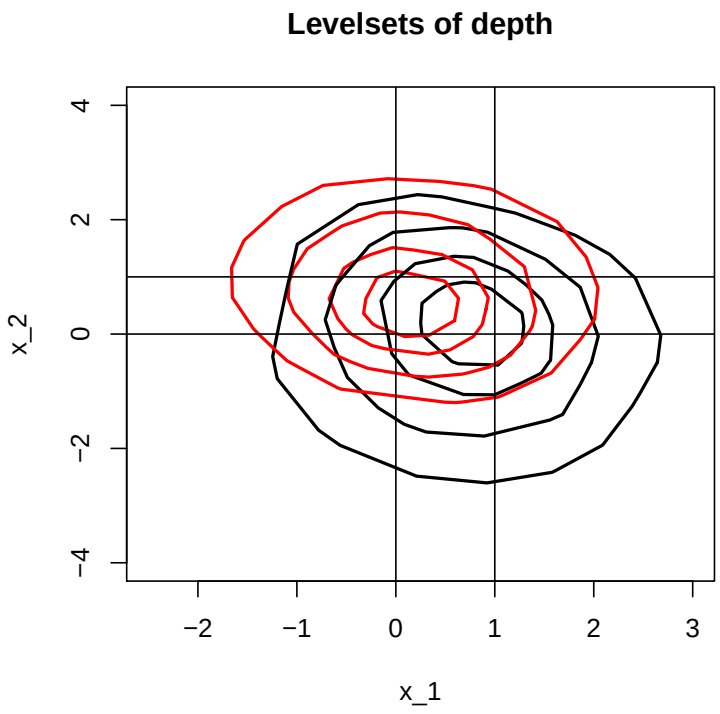


Average Misclassification Rate

$$\begin{aligned} \text{Bayes} &: 32,3\% \\ k\text{NN} &: 36,4\% \\ k\text{-Depth-NN} &: 42,8\% \end{aligned}$$

Kde je problém?





Lokální hloubka + Globální/Lokální klasifikátor

Dutta, Ghosh a Chaudhuri dokázali:

Mějme rozdělení na $\mathbb{R}^m$ s hustotou f ve tvaru $f(\mathbf{x}) = \phi(\|\mathbf{x}\|_p)$, kde ϕ je nějaká klesající funkce.
Levelsety *poloprostorové hloubky* odpovídají levelsetům hustoty právě tehdy, když $p = 2$.

Lokální hloubka + Globální/Lokální klasifikátor

- vážená poloprostorová hloubka + maximal depth classifier
  Hlubinka, Vencalek (2013)
- lokální prostorová hloubka + generalized additive models
  Dutta, Ghosh and Chaudhuri (2012)

Vzpomínka na minulý zimní Robust v Králíkách aneb kdeže loňské sněhy jsou ...

http://artax.karlin.mff.cuni.cz/∼venco2am/hloubka.html
http://artax.karlin.mff.cuni.cz/∼venco2am/DataDepth.html

ondrej.vencalek@upol.cz

Ondřej Vencálek
Katedra matematické analýzy a aplikací matematiky,
Přírodovědecká Fakulta Univerzity Palackého v Olomouci,
17. listopadu 12, 771 46 Olomouc, Czech Republic

Literatura: Maximal depth classifier

- Jörnsten, R. Clustering and classification based on the L_1 data depth. *Journal of Multivariate Analysis*, 2004, vol. 90, pp. 67 – 89.
- Hartikainen, A., Oja, H. On some parametric, nonparametric and semiparametric discrimination rules. In *Data depth: robust multivariate analysis, computational geometry and applications*. R. Y. Liu, R. Serfling, and D. L. Souvaine, eds. 1st edition. New York: AMS, 2006. pp. 61 – 70.
- Ghosh, A. K., Chaudhuri, P. On maximum depth and related classifiers. *Scandinavian Journal of Statistics*, 2005, vol. 32, pp. 327 – 350.
- Mosler, K., Hoberg, R. Data analysis and classification with the zonoid depth. In *Data depth: robust multivariate analysis, computational geometry and applications*. R. Y. Liu, R. Serfling, and D. L. Souvaine, eds. 1st edition. New York: American Mathematical Society, 2006. pp. 49 – 59.
- Kosiorowski, D. Robust classification and clustering based on the projection depth function. In *COMPSTAT 2008: proceedings in computational statistics: 18th symposium held in Porto, Portugal* P. Brito, editor. [CD-ROM]. Heidelberg: Physica, 2008. pp. 209 – 216.
- Hubert, M. and Van der Veeken, S. Robust classification for skewed data. *Advances in Data Analysis and Classification*, 2010, vol 4.4, pp. 239 – 254.
- Dutta, S. and Ghosh, A. K. On robust classification using projection depth. *Annals of the Institute of Statistical Mathematics*, 2012, 64, pp. 657 – 676.

Literatura: Globální hloubka + Globální klasifikátor

- Billor, N., et al. Classification based on depth transvariations. *Journal of Classification*, 2008, vol. 25, pp. 249 – 260.
- Li, J., Cuesta-Albertos, J. A., Liu, R. DD-classifier: nonparametric classification procedure based on DD-plot. *JASA*, 2012, Vol. 107, No. 498, pp. 737 – 753.
- Lange, T., Mosler, K. and Mozharovskyi, P. Fast nonparametric classification based on data depth. *Statistical Papers*, 2012, pp. 1 – 21.

Literatura: Globální hloubka + Lokální klasifikátor

- Vencalek, O. Weighted data depth and depth based classification. *PhD thesis*, [online] *http://artax.karlin.mff.cuni.cz/∼venco2am/DataDepth.html* , 2011.
- Paindaveine, D., Van Bever, G. Nonparametrically consistent depth-based classifiers. arXiv preprint arXiv:1204.2996, 2012.
- Vencalek, O. New depth-based modification of the *k*-nearest neighbour method. to appear in *Informacni bulletin Ceske statisticke spolecnosti*, 2013.
- Vencalek, O. k-Depth-nearest Neighbour Method and its Performance on Skew-normal Distributions. *Acta Univ. Palacki. Olomuc., Fac. rer. nat., Mathematica*, 2013, vol. 52, no. 2, pp. 121 – 129.

Literatura: Lokální hloubka + Globální/Lokální klasifikátor

- Hlubinka, D., Kotik, L., Vencalek, O. Weighted data depth. *Kybernetika*, 2010, vol. 46, no. 1, pp. 125 – 148.
- Hlubinka, D., Vencalek, O. Depth-Based Classification for Distributions with Nonconvex Support. *Journal of Probability and Statistics*, 2013.
- Dutta, S., Chaudhuri, P. and Ghosh, A. K. Some intriguing properties of Tukey's half-space depth. *Bernoulli*, 2011, vol. 17, pp. 1420 – 1434
- Dutta, S., Chaudhuri, P. and Ghosh, A. K. Classification using Localized Spatial Depth with Multiple Localization. Communicated for publication, 2012.
- Agostinelli, C., Romanazzi, M. Local depth. *Journal of Statistical Planning and Inference*, 2011. vol. 141, pp. 817 – 830.
- Paindaveine, D., Van Bever, G. From Depth to Local Depth: A Focus on Centrality. ECARES Working Papers, 2012.