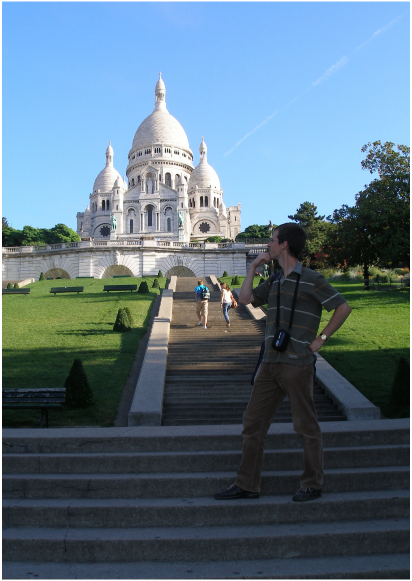
Klasifikace pomocí hloubky dat – nové nápady

Ondřej Vencálek
Univerzita Palackého v Olomouci

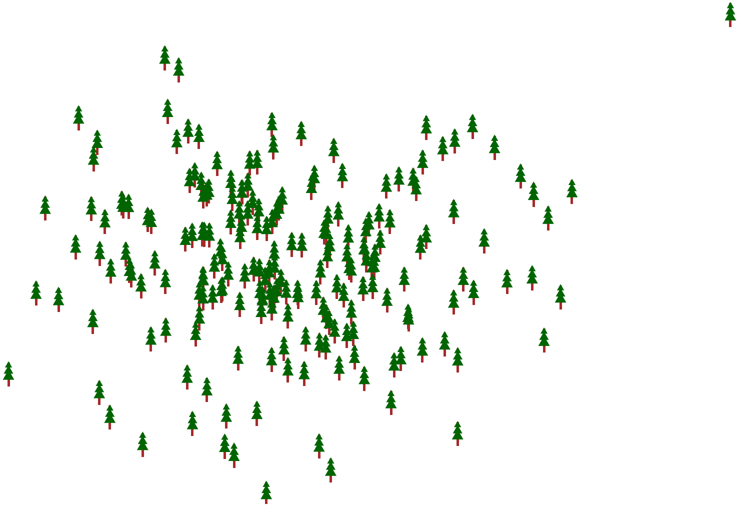
ROBUST – Rybník, 26. ledna 2018



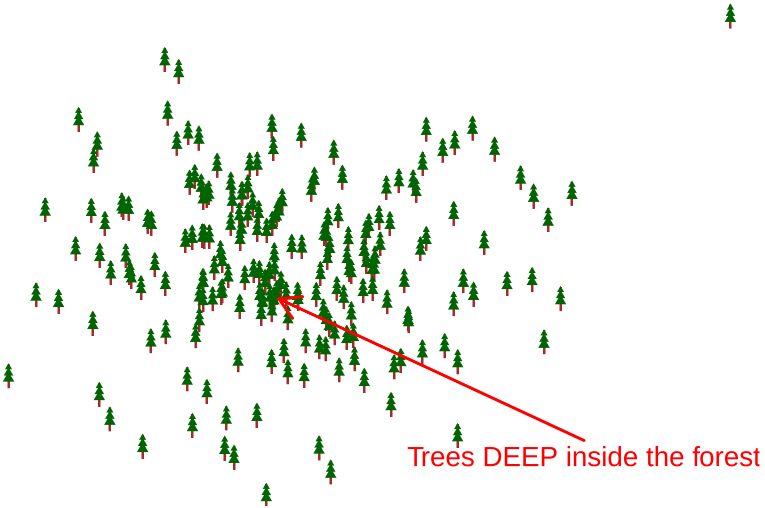
Z. Fabián: Vzpomínka na Compstat 2010 aneb večeře na Pařížské radnici, Informační Bulletin ČStS, 4/2017

Nápad pořádat konferenci v Paříži se dustojně řadí k nápadu pořádat mistrovství světa v Kataru nebo letní olympiádu v Letňanech. . . .

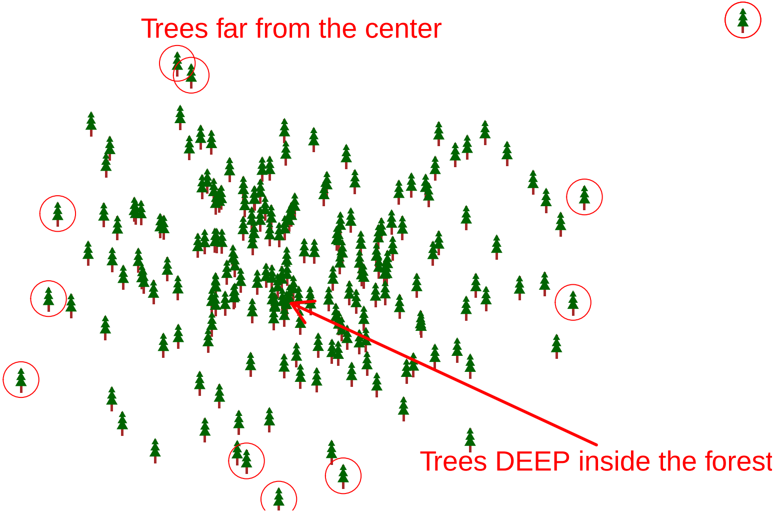
It is **deep** inside = it has high **depth**



It is **deep** inside = it has high **depth**



It is **deep** inside = it has high **depth**

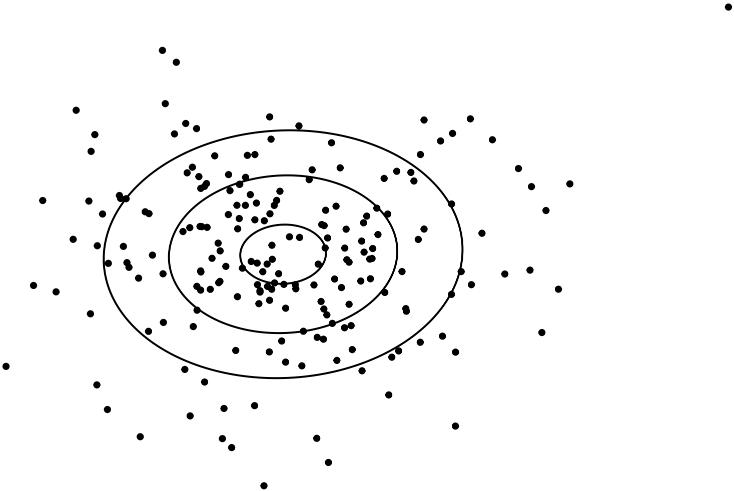


It is **deep** inside = it has high **depth**

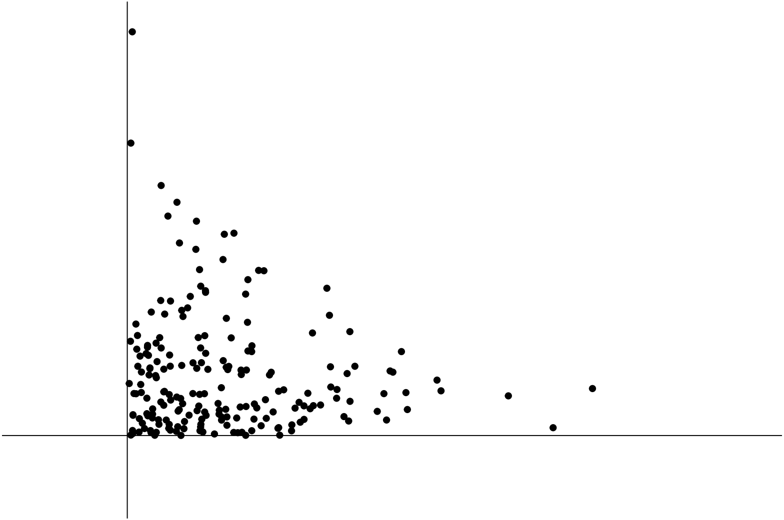
Outliers in $\mathbb{R}^1$

Outliers in $\mathbb{R}^2$ – any idea?

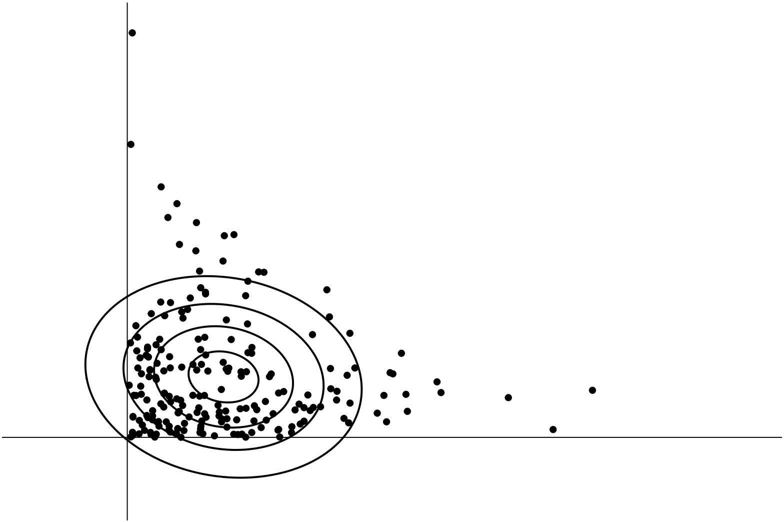
1st idea – Mahalanobist distance



1st idea – Mahalanobist distance – non-elliptical case



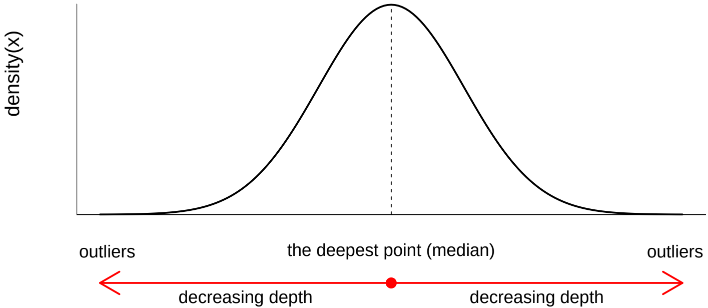
1st idea – Mahalanobist distance – non-elliptical case



Halfspace depth in $\mathbb{R}^1$

$$D(x) = \min\{P(X \leq x), P(X \geq x)\}$$

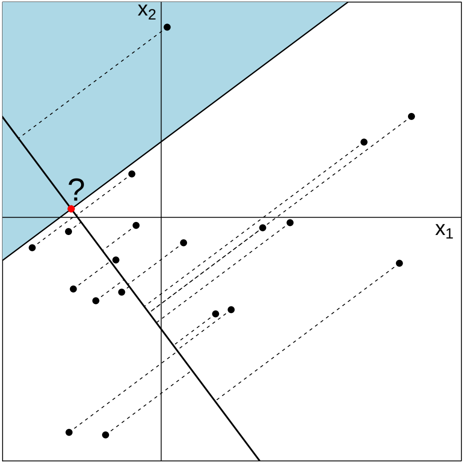
Example: univariate normal distribution $\mathrm{N}(\mu, \sigma^2)$



Halfspace depth in $\mathbb{R}^d$

projection pursuit approach

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



The **halfspace depth** of a point $\boldsymbol{x} \in \mathbb{R}^d$ with respect to a probability measure P on $\mathbb{R}^d$ is defined as the minimum probability mass carried by any closed halfspace containing $\boldsymbol{x}$, that is

$$D(\boldsymbol{x}; P) = \inf \{\mathsf{P}(H) : H \text{ a closed halfspace}, \boldsymbol{x} \in H\}$$

- The depth provides so called central-outward ordering of data
- *The deepest point* extends the notion of *median* into a higher dimensions.

Problem of classification

The Bayes classifier

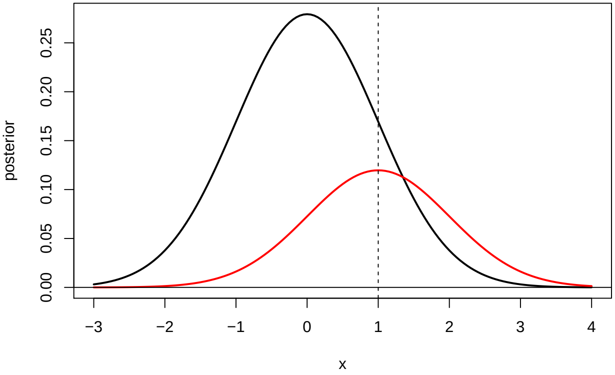
$$class(\boldsymbol{x}) = \arg\max_i f_i(\boldsymbol{x})\pi_i.$$

produces smallest possible number of misclassified points (total over all groups)

When Bayes misclassifies the centre of the distribution...

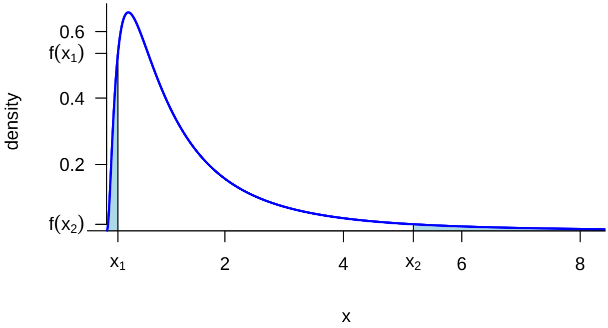
$$P_1 = N(0, 1), P_2 = N(1, 1)$$

$$\pi_1 = 0.7, \pi_2 = 0.3$$



Hey, Bayes, be fair to x_2!

lognormal distribution from $N(0,1)$



$$\mathrm{quantile}_{0.05} = x_1, \qquad f(x_1) = 0.53$$
$$\mathrm{quantile}_{0.95} = x_2, \qquad f(x_2) = 0.02$$

Basics in decision theory

- Distributions P_i defined on $\mathbb{R}^d$, with densities f_i and prior probabilities π_i, $i = 1, \ldots, K$.
- A classifier divides the space $\mathbb{R}^d$ into K disjoint parts $A_i, i = 1, \ldots, K$, $\bigcup_{i=1}^{K} A_i = \mathbb{R}^d$ such that any $\boldsymbol{x} \in \mathbb{R}^d$ is assigned to P_i iff $\boldsymbol{x} \in A_i$.
- Cost function

$$c_{ij}(\boldsymbol{x}) = \begin{cases} c_i(\boldsymbol{x}) & \text{if } j \neq i, \\ 0 & \text{if } j = i. \end{cases}$$

- Total cost

$$\min \sum_{i=1}^{K} \sum_{j \neq i} \int_{A_j} c_i(\boldsymbol{x}) f_i(\boldsymbol{x}) \pi_i d\boldsymbol{x}.$$

- Optimal classifier

$$class(\boldsymbol{x}) = \arg\max_i c_i(\boldsymbol{x}) f_i(\boldsymbol{x}) \pi_i.$$

Bayes classifier and its 2 new competitors

$$class(\boldsymbol{x}) = \arg\max_i c_i(\boldsymbol{x}) f_i(\boldsymbol{x}) \pi_i.$$

1. **Bayes classifier**: $c_i(\boldsymbol{x}) = 1$

$$class_B(\boldsymbol{x}) = \arg\max_i f_i(\boldsymbol{x}) \pi_i$$

2. **Depth-weighted classifier**: $c_i(\boldsymbol{x}) = D(\boldsymbol{x}; P_i)$

$$class_D(\boldsymbol{x}) = \arg\max_i D(\boldsymbol{x}; P_i) f_i(\boldsymbol{x}) \pi_i$$

3. **Rank-weighted classifier**: $c_i(\boldsymbol{x}) = F_i(D(\boldsymbol{x}; P_i))$,
   where F_i is CDF of $D(\boldsymbol{X}; P_i), \boldsymbol{X} \sim P_i$

$$class_R(\boldsymbol{x}) = \arg\max_i F_i(D(\boldsymbol{x}; P_i)) f_i(\boldsymbol{x}) \pi_i$$

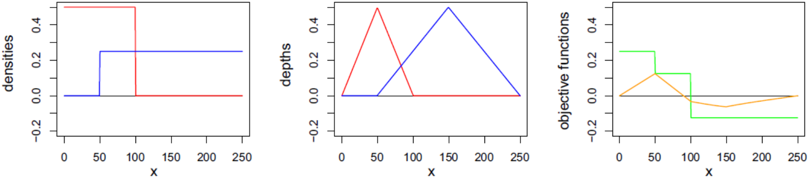
Toy example

- $P_1 = \mathit{Unif}[0, 100], \pi_1 = 0.5,$
  $P_2 = \mathit{Unif}[50, 250], \pi_2 = 0.5.$
- Classification on the overlapping part of supports:
  Bayes classifier:

  $$\mathit{class}_B(x) = 1 \quad \text{for } x \in [50, 100]$$

  Depth-weighted classifier:

  $$\mathit{class}_D(x) = \begin{cases} 1 & \text{for } x \in [50, 90), \\ 2 & \text{for } x \in (90, 100]. \end{cases}$$



densities *depths* *classifiers*

Example 1: Uniform distributions, red: $P_1 = Unif[0, 100]$; blue: $P_2 = Unif[50, 250]$, $\pi_1 = \pi_2 = 0.5$.

The Bayes classifier: $x = 100$, $P(class(x) \neq i | x \in P_i) = 0.125$,

$P(class(x) \neq 1 | x \in P_1) = 0$, $P(class(x) \neq 2 | x \in P_2) = 0.25$.

The depth-weighted classifier: $x = 90$, $P(class(x) \neq i | x \in P_i) = 0.15$,

$P(class(x) \neq 1 | x \in P_1) = 0.1$, $P(class(x) \neq 2 | x \in P_2) = 0.2$.

The first question (or two)

- To what extent do the new (depth-weighted) classifiers differ from the Bayes classifier?
- How large might the corresponding difference in the average misclassification rate be?

Difference in the case of elliptical symmetry

Let us assume:

- (P1): $f_i(\mathbf{x}) = k_i g(M_i(\mathbf{x}))$, where g is a strictly decreasing function, $k_i > 0$ are constants, and $M_i(\mathbf{x}) = \left((\mathbf{x} - \alpha_i)'\mathbf{B}_i^{-1}(\mathbf{x} - \alpha_i)\right)^{\frac{1}{2}}$ with $\mathbf{B}_i$ positive definite, $\alpha_i \in \mathbb{R}^d$, denotes the generalized distance of the point $\mathbf{x}$ from the center of the distribution P_i.
- (P2) $D(\mathbf{x})$ is an affine invariant depth. Then, given (P1), $D_i(\mathbf{x}) = D(\mathbf{x}; P_i)$ is a fixed decreasing function of generalized distance, that can be expressed as $D_i(\mathbf{x}) = h(M_i(\mathbf{x}))$, where h is a strictly decreasing function.

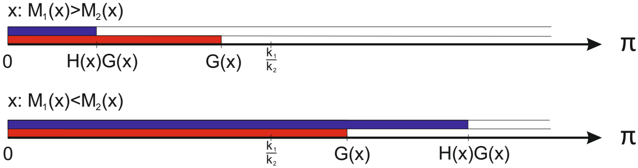
Denote

- $G(\mathbf{x}) = \frac{k_1 g(M_1(\mathbf{x}))}{k_2 g(M_2(\mathbf{x}))}$ as likelihood ratio,
- $H(\mathbf{x}) = \frac{h(M_1(\mathbf{x}))}{h(M_2(\mathbf{x}))}$ as depth ratio,
- $\pi = \frac{\pi_2}{\pi_1}$ as inverse prior ratio

Then

- The Bayes classifier assigns $\mathbf{x}$ to P_1 if $G(\mathbf{x}) > \pi$,
- The depth-weighted classifier assigns $\mathbf{x}$ to P_1 if $H(\mathbf{x})G(\mathbf{x}) > \pi$.

Difference in the case of elliptical symmetry



The classifiers differ when π is between $G(\mathbf{x})$ and $H(\mathbf{x})G(\mathbf{x})$.
For the fixed π, the region where the classifiers differ is

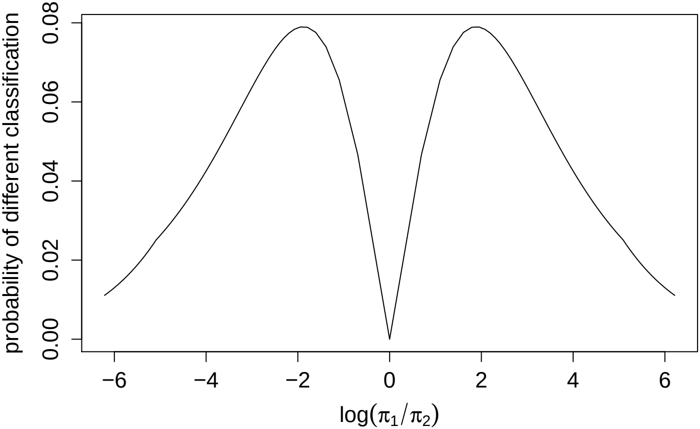
$$RD(\pi) = \left\{\mathbf{x} \in \mathbb{R}^d : H(\mathbf{x})G(\mathbf{x}) < \pi < G(\mathbf{x}) \text{ or } G(\mathbf{x}) < \pi < H(\mathbf{x})G(\mathbf{x})\right\}.$$

Let (P1) and (P2) hold for P_1 and P_2. Then

$$P(class_B(\boldsymbol{X}) \neq class_D(\boldsymbol{X})) = 0 \Leftrightarrow \pi_1 k_1 = \pi_2 k_2.$$

Difference in the case of elliptical symmetry – example

$P_1 = N(-1, 1), P_2 = N(1, 1)$

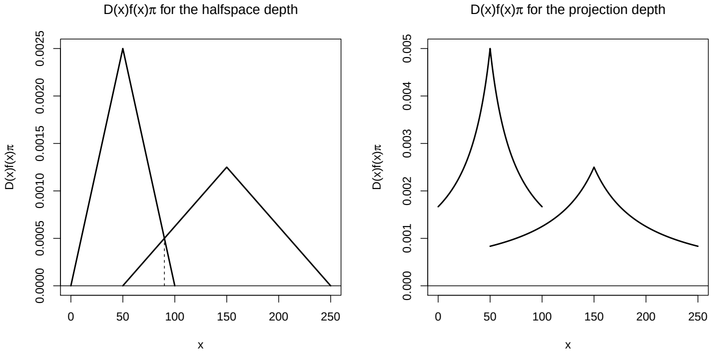


Next question

- To what extent does the performance of the new classifiers depend on the choice of depth function?
- Which depth function results in the smallest (biggest) differences from the Bayes classifier?

Difference between the depth-based classifiers using halfspace and projection depths

$$P_1 = Unif[0,100], \pi_1 = 0.5,$$
$$P_2 = Unif[50,250], \pi_2 = 0.5.$$



Rank-weighted classifier

- **Depth-weighted classifier**: $c_i(\boldsymbol{x}) = D(\boldsymbol{x}; P_i)$

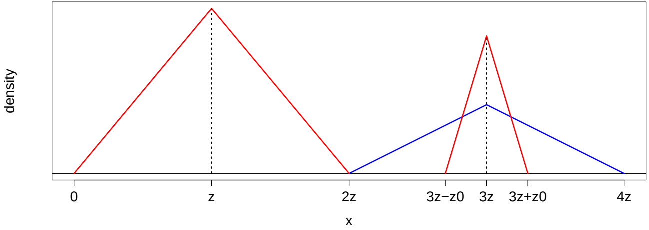
$$class_D(\boldsymbol{x}) = \arg\max_i D(\boldsymbol{x}; P_i) f_i(\boldsymbol{x}) \pi_i$$

- **Rank-weighted classifier**: $c_i(\boldsymbol{x}) = F_i(D(\boldsymbol{x}; P_i))$,
  where F_i is CDF of $D(\boldsymbol{X}; P_i), \boldsymbol{X} \sim P_i$

$$class_R(\boldsymbol{x}) = \arg\max_i F_i(D(\boldsymbol{x}; P_i)) f_i(\boldsymbol{x}) \pi_i$$

Let (P1) and (P2) hold for P_1 and P_2 and any two depth functions $D(\cdot)$ and $D^*(\cdot)$. Then $class_R(\mathbf{x})$ is the same for both $D(\cdot)$ and $D^*(\cdot)$ with probability one.

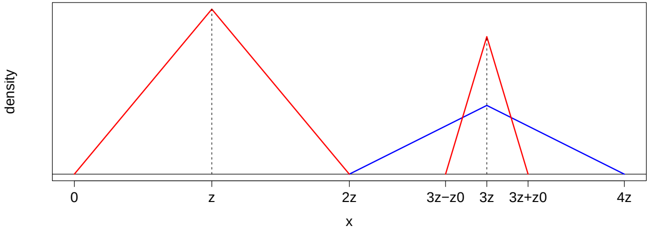
Robustness of the newly proposed classifiers



- $(1-\alpha)\pi_1$ for P_1,
- $\alpha\pi_1$ for the contamination of P_1
- π_2 for the non-contaminated P_2.

Assume $z_0 = z$, $\pi_2 < \alpha\pi_1$.

Robustness of the newly proposed classifiers



- $(1-\alpha)\pi_1$ for P_1,
- $\alpha\pi_1$ for the contamination of P_1
- π_2 for the non-contaminated P_2.

Assume $z_0 = z$, $\pi_2 < \alpha\pi_1$.

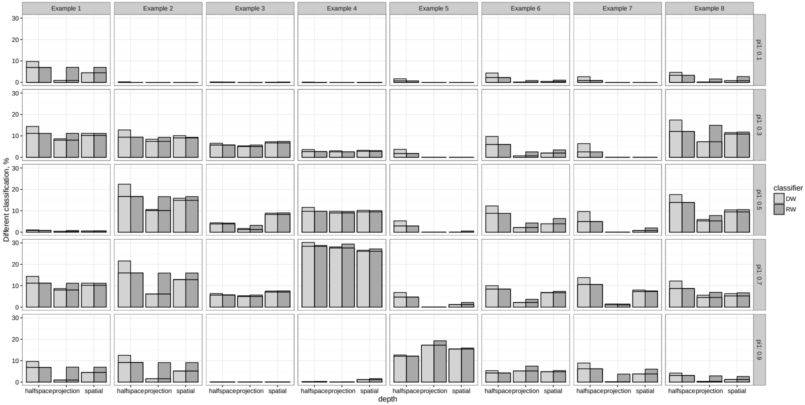
- $f_2(x)\pi_2 < f_1(x)\pi_1$ for all $x \in (2z, 4z)$.
  Misclass. rate for group 2 is 1.
- $D_2(x)f_2(x)\pi_2 > D_1(x)f_1(x)\pi_1$ for $x > 2z + \sqrt{2\pi_2}z$
  Misclass. rate for group 2 is π_2

Simulation study – settings

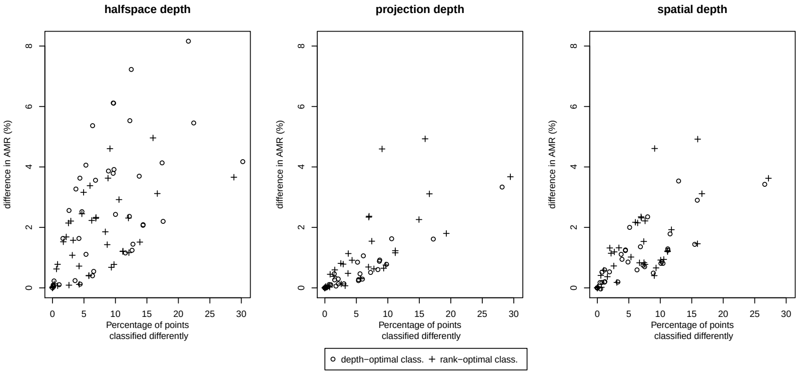
	Group 1		Group 2	
Ex.	Distribution	Parameters	Distribution	Parameters
1	Normal	$\mathbf{0}$, $\mathbf{\Sigma}_0$	Normal	$\mathbf{1}$, $\mathbf{\Sigma}_0$
2	Normal	$\mathbf{0}$, $\mathbf{\Sigma}_0$	Normal	$\mathbf{1}$, $4\mathbf{\Sigma}_0$
3	Cauchy	$\mathbf{0}$, $\mathbf{\Sigma}_0$	Cauchy	$\mathbf{1}$, $\mathbf{\Sigma}_0$
4	Cauchy	$\mathbf{0}$, $\mathbf{\Sigma}_0$	Cauchy	$\mathbf{1}$, $4\mathbf{\Sigma}_0$
5	Bivar. exponen.	1, 1	Shifted bivar. expon. $(+\mathbf{1})$	1, 1
6	Bivar. exponen.	1, 1/2	Shifted bivar. expon. $(+\mathbf{1})$	1/2, 1
7	Normal	$\mathbf{0}$, $\boldsymbol{I}$	Bivar. exponential	1, 1
8	Skewed normal	$\begin{pmatrix} 1 \\ 2 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 7 \end{pmatrix}, \begin{pmatrix} -2 \\ -5 \end{pmatrix}$	Skewed normal	$\begin{pmatrix} 0 \\ -1 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 5 \end{pmatrix}, \begin{pmatrix} 1 \\ 5 \end{pmatrix}$

Where $\mathbf{\Sigma}_0 = \begin{pmatrix} 1 & 1 \\ 1 & 4 \end{pmatrix}$.

Percentage of points classified differently than by the Bayes classifier



Increase of AMR versus percentage of points classified differently than by the Bayes classifier



Conclusion

- We suggest to weight misclassification cost according to the centrality of a misclassified point measured either by its depth w.r.t. the distribution from which it comes (depth-weighted classifier) or by its rank w.r.t. the distribution from which it comes (rank-weighted classifier).
- In particular cases, the new classifiers does not differ from the Bayes classifier. Simulation study showed that increase in AMR is much smaller than percentage of points classified differently.
- Both classifiers depend on the depth function which they are using. In particular cases rank-weighted classifier does not depend on the used depth function.
- Presence of the depth term in the depth-based classifiers may substantially increase robustness of the procedure.

ondrej.vencalek@upol.cz

Ondrej Vencalek
Department of Mathematical Analysis and Applications of Mathematics,
Faculty of Science, Palacky University in Olomouc,
17. listopadu 12, 771 46 Olomouc, Czech Republic

Happy end


dítěte:
Datum narození: 2.9.2010
Hodina narození: 10.43
Váha: 2360 g
Míra:
Srdečně blahopřejeme
k narození Vašeho dítěte