VÁŽENÁ HLOUBKA DAT A JEJÍ VLASTNOSTI

ONDŘEJ VENCÁLEK

vencalek@karlin.mff.cuni.cz
Matematicko-fyzikální fakulta Univerzity Karlovy v Praze

Tento příspěvek z oblasti neparametrických metod analýzy mnohorozměrných dat představuje možnost zobecnění poloprostorové hloubky bodu, tzv. váženou poloprostorovou hloubku, a zabývá se jejími vlastnostmi v závislosti na volbě váhy v její definici. Speciálně nás zajímá nulovost hloubky bodů mimo nosič rozdělení.

POLOPROSTOROVÁ HLOUBKA A JEJÍ ZOBECNĚNÍ

Jeden z nejdůležitějších neparametrických přístupů práce s mnohorozměrnými daty je založen na tzv. hloubce dat. Jde vlastně o způsob, jak uspořádat prvky vícerozměrného prostoru. Zřejmě nejrozšířenější definici hloubky publikoval v roce 1975 J. Tukey:

Definice: Uvažujme $p-$rozměrný prostor $\mathbb{R}^p$ a bod $\boldsymbol{x} \in \mathbb{R}^p$. Nechť P je pravděpodobnostní míra na $\mathbb{R}^p$. Poloprostorová hloubka bodu $\boldsymbol{x}$ vzhledem k P je definována vztahem

$$D_P(\boldsymbol{x}) = \inf_{\boldsymbol{u}:\|\boldsymbol{u}\|=1} P\left\{\boldsymbol{y} : \boldsymbol{u}^T(\boldsymbol{y}-\boldsymbol{x}) \geq 0\right\}$$

Tedy poloprostorová hloubka bodu $\boldsymbol{x}$ je definována jako minimální pravděpodobnost uzavřeného poloprostoru obsahujícího tento bod. Navrhované zobecnění této definice, tedy vážená poloprostorová hloubka, místo pravděpodobnosti poloprostoru uvažuje váhovou funkci integrovanou na tomto poloprostoru podle příslušné pravděpodobnostní míry.

Definice 2: Uvažujme $p-$rozměrný prostor $\mathbb{R}^p$ a bod $\boldsymbol{x} = (x_1, \ldots, x_p) \in \mathbb{R}^p$.
Nechť P je pravděpodobnostní míra na $\mathbb{R}^p$.
Buď $w_+ : \mathbb{R}^p \to [0, \infty)$ ohraničená, měřitelná váhová funkce, taková, že $w_+(\boldsymbol{x}) = 0$ jestliže $x_p < 0$.
Nechť dále $w_-(\boldsymbol{x}) = w_-(x_1, \ldots, x_{p-1}, x_p) = w_+(x_1, \ldots, x_{p-1}, -x_p)$.
Váženou poloprostorovou hloubku bodu $\boldsymbol{x}$ vzhledem k P definujeme jako

$$D_{\boldsymbol{X}}(\boldsymbol{x}) = D_P(\boldsymbol{x}) = \inf_{\boldsymbol{A} \in O_p} \frac{\mathsf{E}_P w_+\left(\boldsymbol{A}\left(\boldsymbol{X}-\boldsymbol{x}\right)\right)}{\mathsf{E}_P w_-\left(\boldsymbol{A}\left(\boldsymbol{X}-\boldsymbol{x}\right)\right)}, \tag{1}$$

kde O_p označuje prostor všech ortogonálních $p \times p$ matic.
Definujme ještě hodnotu podílu dvou nul $0/0 = 1$.

Příklady volby váhové funkce pro dvourozměrný prostor:

- $w_+(\boldsymbol{x}) = 1$ pro $x_2 \geq 0, x_1 \in \mathbb{R}$,
- $w_+(\boldsymbol{x}) = 1$ pro $x_2 \geq 0, |x_1| < k$,
- $w_+(\boldsymbol{x}) = 1$ pro $x_2 \geq 0, |x_1| < kx_2$,
- $w_+(\boldsymbol{x}) = 1$ pro $x_2 \geq 0, |x_1| < k/x_2$,
- $w_+(\boldsymbol{x}) = \Phi(x_1)$ pro $x_2 \geq 0, x_1 \in \mathbb{R}$,

kde k je nezáporná konstanta, $\Phi(\cdot)$ distribuční funkce standardního normálního rozdělení $N(0,1)$. Vždy $w_+(\boldsymbol{x}) = 0$ pro všechna ostatní $\boldsymbol{x}$. Viz schéma vyznačující oblasti, kde je váhová funkce nenulová.

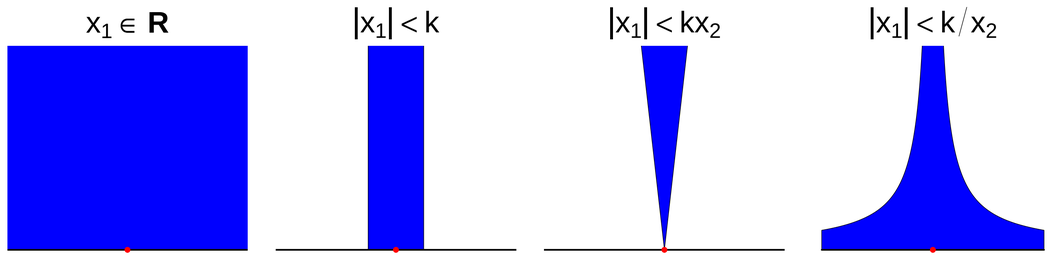


Schéma: příklady oblastí, kde je váhová funkce w_+ nenulová.

NULOVOST HLOUBKY V BODECH MIMO NOSIČ ROZDĚLENÍ

Jednou z přirozených vlastností, kterou od hloubkové funkce očekáváme, je její nulovost v bodech ležících mimo konvexní uzávěr nosiče rozdělení P. Připomeňme, že nosič rozdělení P (značíme jej $\mathsf{sp}(P)$) je nejmenší uzavřená množina s pravděpodobností 1, tedy

$$\mathsf{sp}(P) = \bigcap \{F \in \mathcal{F} : P(F) = 1\},$$

kde $\mathcal{F}$ je množina všech uzavřených podmnožin $\mathbb{R}^p$. Symbolem $\mathsf{csp}(P)$ pak označujeme uzavřený konvexní obal nosiče rozdělení $\mathsf{sp}(P)$. Pro názornost budeme uvažovat dvourozměrný prostor $\mathbb{R}^2$, všechny následující věty však mají jednoduché zobecnění pro vícerozměrný případ $\mathbb{R}^p$, kde $p \in \mathbb{N}$.

Uvažujme nejprve případ, kdy nosič rozdělení je konvexní. Podmínku na váhovou funkci, která je postačující pro to, aby body mimo nosič tohoto rozdělení měly nulovou hloubku, formuluje následující věta:

Věta 1: Nechť $\mathrm{sp}(P)$ je konvexní. Označme $W = \{\boldsymbol{y} : w_+(\boldsymbol{y}) > 0\}$.
Předpokládejme, že pro váhovou funkci platí:

$$\forall \boldsymbol{x} : x_1 = 0, x_2 > 0 \; \exists \, U \text{ okolí bodu } \boldsymbol{x} : U \subset W, \tag{2}$$

potom pro všechny $\boldsymbol{x} \notin \mathsf{sp}(P)$ platí $D_P(\boldsymbol{x}) = 0$.

Důkaz: V důkazu se využije invariance vážené poloprostorové hloubky vzhledem k posunutí a otočení. Uvažujme bod $\boldsymbol{x}_0 \notin \mathsf{sp}(P)$. Chceme ukázat, že jeho vážená hloubka bude při splnění podmínky (2) rovna nule.
Označme $\boldsymbol{x}_m = \arg\min\{|\boldsymbol{x} - \boldsymbol{x}_0| : \boldsymbol{x} \in \mathsf{sp}(P)\}$, tedy bod nosiče rozdělení P nejbližší k bodu $\boldsymbol{x}_0$. Existence a jednoznačnost tohoto bodu vyplývá z konvexity $\mathrm{sp}(P)$. Díky invarianci hloubky vzhledem k posunutí můžeme búno předpokládat, že $\boldsymbol{x}_0 = (0,0)$, a díky invarianci hloubky vzhledem k otočení můžeme búno předpokládat, že $\boldsymbol{x}_m = (0, v)$, kde $v = |\boldsymbol{x} - \boldsymbol{x}_0|$.
Nyní si stačí uvědomit, že pro takovéto otočení musí $\mathrm{sp}(P) \subset H_v \subset H_0$, kde $H_v = \{\boldsymbol{x} : x_2 \geq v\}$, $H_0 = \{\boldsymbol{x} : x_2 \geq 0\}$. Z předpokladu (2) vyplývá, že existuje $U_{\boldsymbol{x}_m}$ okolí bodu $\boldsymbol{x}_m$ takové, že $U_{\boldsymbol{x}_m} \subset W$ a tedy $U_{\boldsymbol{x}_m} \cap W \cap \mathsf{sp}(P) \neq \varnothing$, odtud $\mathsf{E}_P w_+\left(\boldsymbol{A}\left(\boldsymbol{X} - \boldsymbol{x}_0\right)\right) > 0$.
Z vlastnosti $\mathrm{sp}(P) \subset H_v \subset H_0$ naopak vyplývá, že $\mathsf{E}_P w_-\left(\boldsymbol{A}\left(\boldsymbol{X} - \boldsymbol{x}_0\right)\right) = 0$ a tedy $D_P(\boldsymbol{x}_0) = 0$.
□

Pro nekonvexní nosič rozdělení P je třeba buď předpokládat souvislost $\mathrm{sp}(P)$ nebo zesílit požadavek na tvar váhové funkce:

Věta 2: Nechť existuje $n \in \mathbb{N}$ takové, že $\mathrm{sp}(P) = \bigcup_{i=1}^{n} K_i$,
kde K_i (i=1,...,n) je souvislá podmnožina $\mathbb{R}^2$, přičemž $\mathrm{sp}(P)$ nemá žádný izolovaný bod.
Buď dále $m_{ij} = \min\left\{|\boldsymbol{x} - \boldsymbol{y}| : \boldsymbol{x} \in K_i, \boldsymbol{y} \in K_j\right\}$, $i, j = 1, \ldots, n$.
Uvažujme $m = \max_{1 \leq i,j \leq n} m_{ij}$.
Nechť pro váhovou funkci w_+ platí

$$\forall \boldsymbol{x} : |x_1| \leq m/2 \; \exists \, U_{\boldsymbol{x}} \text{ okolí } \boldsymbol{x} \text{ takové, že } U_{\boldsymbol{x}} \subset W. \tag{3}$$

Potom $\boldsymbol{x} \notin \mathsf{csp}(P) \Rightarrow D_P(\boldsymbol{x}) = 0$.

Uvědomme si, že pro speciální případ $n = 1$, tedy situaci, kdy nosič je souvislý nekonvexní, věta říká, že podmínka (2) zaručuje, že všechny body mimo konvexní obal nosiče rozdělení mají nulovou váženou poloprostorovou hloubku.

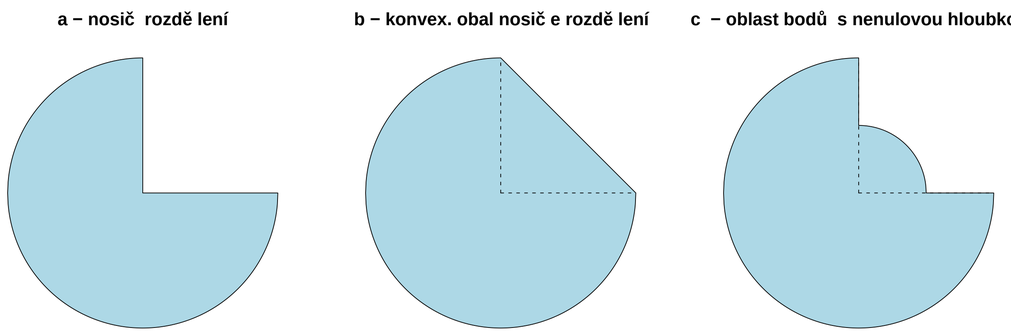
HLOUBKA BODŮ, KTERÉ JSOU V KONVEX. OBALU NOSIČE ROZDĚLENÍ, ALE MIMO NOSIČ SAMOTNÝ

Snadno se ukáže, že poloprostorová hloubka všech bodů v konvexním obalu nosiče rozdělení je nenulová (tedy i v bodech, které se nacházejí v konvexním obalu nosiče rozdělení, ale mimo nosič samotný). Vážená poloprostorová hloubka tuto diskutabilní vlastnost obecně nemá.

Uvažujme například rovnoměrné rozdělení na kruhové výseči vzniklé z kružnice se středem S a poloměrem r (viz obrázek 1a). Všechny body v konvexním obalu nosiče rozdělení (viz obrázek 1b) mají nenulovou poloprostorovou hloubku. Uvažujme nyní váženou poloprostorovou hloubku s váhovou funkcí ve tvaru

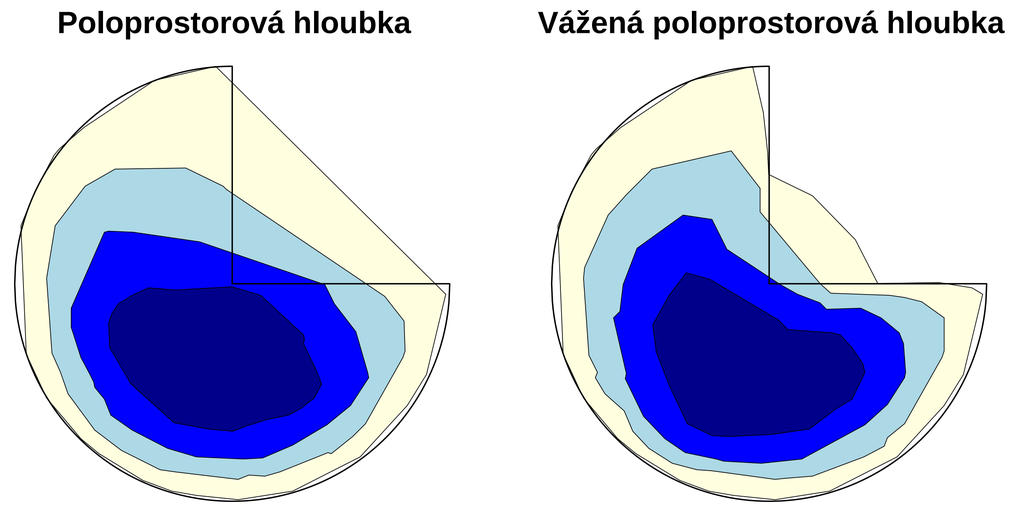
$$\begin{aligned} w_+(\boldsymbol{x}) &= 1 \quad \text{pro } |x_1| < h, \; x_2 \geq 0 \\ &= 0 \quad \text{jinak,} \end{aligned} \tag{4}$$

kde h je nějaká kladná konstanta menší než r. Pak oblast bodů s nenulovou váženou poloprostorovou hloubkou je sjednocením nosiče rozdělení a kružnice se středem S a poloměrem h (viz obrázek 1c).



Obrázek 1: nosič rovnoměrného rozdělení ve tvaru kruhové výseče (a); konvexní obal tohoto nosiče je oblastí bodů s nenulovou poloprostorovou hloubkou (b); oblast bodů s nenulovou váženou poloprostorovou hloubkou při použití váhové funkce (4), kde $h = r/2$ (c).

Výše uvedené skutečnosti ilustrujeme pomocí simulace. Bylo generováno 1000 bodů z rovnoměrného rozdělení na kruhové výseči (poloměr kruhu r). Srovnáme-li tvar oblastí 25, 50, 75 a 100% nejhlubších bodů pro poloprostorovou hloubku a váženou poloprostorovou hloubku s váhovou funkcí (4), viz obrázek 2, odhalíme velké rozdílností. Jednou z nejdůležitějších je nulovost hloubky velkého množství bodů v oblasti, která se nachází v konvexním obalu nosiče rozdělení, ale mimo nosič samotný, pro váženou poloprostorovou hloubku.



Obrázek 2: Empiricky zjištěné oblasti 25, 50, 75 a 100% nejhlubších bodů pro poloprostorovou (vlevo) a váženou poloprostorovou (vpravo) hloubku. Nosič rozdělení, ze kterého bylo generováno, je znázorněn silnější čarou.

Poděkování: Rád bych poděkoval docentu Danielu Hlubinkovi a Lukáši Kotíkovi, se kterými na práci o hloubce dat dlouhodobě spolupracuji. Tato práce byla podporována grantem GAČR 201/08/0486.

Literatura:

[1] Liu R.Y., Parelius J.M. a Singh K. (1999) *Multivariate analysis by data depth: descriptive statistics, graphics and inference.* The Annals of Statistics, Vol. 27, No. 3, 783–858.
[2] Zuo Y. a Serfling R. (2000) *General notion of statistical depth function.* Annals of Statistics, Vol. 28, 461–482.
[3] Tukey J. (1975) *Mathematics and picturing data.* In Proceedings of the 1975 International Congress of Mathematics, Vol. 2, 523–531.