Hloubka dat

kontury, klasifikace a konzistence

Daniel Hlubinka

Univerzita Karlova v Praze
Matematicko-fyzikální fakulta
Katedra pravděpodobnosti a matematické statistiky

Robust
Němčičky 2012

Hloubka

Co je vlastně **hloubka dat**?

- Zcela obecně: *přiřazení pořadí* mnohorozměrné náhodné veličině. Klidně i nekonečněrozměrné (funkcionální proměnné).
- Buď $X : (\Omega, \mathcal{F}, \mathsf{P}) \to (\mathbb{E}, \mathcal{E}, P_X)$ náhodná veličina. Hloubka je funkce rozdělení náhodné veličiny a bodů ve výběrovém prostoru:

$$D : (\mathbb{E}, \mathbb{P}_{\mathbb{E}}) \to \mathbb{R}_+ \quad (\text{příp. } \to [0, 1]).$$

- Můžeme používat různá značení. Nejčastěji $D(x, Q) = D_Q(x)$, případně $D(x)$ bude-li jasné o jaké rozdělení se jedná.

Serflingův průvodce po hloubce

- Liu (1990) udává několik *žádoucích vlastností* hloubky.

H1 Hloubka má být afinně invariantní funkcí.

H2 Hloubka má být maximální v centru symetrie rozdělení.

H3 Hloubka má klesat směrem od nejhlubšího bodu.

H4 Hloubka má jít nule pro body jdoucí k nekonečnu (od nejhlubšího bodu).

- Serfling a Zuo (2000) pak zkoumají jednotlivé hloubky s ohledem na H1–H4 a jako *statistickou hloubku* definují nezápornou omezenou funkci splňující H1–H4.

Serflingův průvodce po hloubce

- Serfling a Zuo (2000) dále dělí hloubku na několik typů:

A $D(x, P) = E_P h(x; X_1, \ldots, X_r)$, kde h je libovolná nezáporná omezená měřitelná funkce měřící *blízkost* bodu x k bodům $x_1, \ldots, x_r$.

B $D(x, P) = \left(1 + E_P h(x; X_1, \ldots, X_r)\right)^{-1}$, kde h je libovolná nezáporná neomezená měřitelná funkce měřící *vzdálenost* bodu x od bodů $x_1, \ldots, x_r$.

C $D(x, P) = \left(1 + O(x, P)\right)^{-1}$, kde $O(x, P)$ je funkce udávající *odlehlost* bodu x vzhledem k rozdělení P.

D $D(x, P; \mathcal{H}) = \inf_{H \in \mathcal{H}} P[x \in H]$, kde $\mathcal{H}$ je vhodná třída měřitelných množin.

Hloubka a kvantil

- Označme úrovňové množiny a kontury hloubky

$$L(D, P, q) = \{x : D(x, P) \leq q\},$$
$$C(D, P, q) = \overline{L(D, P, q)} \setminus L^{\circ}(D, P, q)$$

  kde L° je vnitřek množiny.
- Kontura hloubky může být použita jako *mnohorozměrná analogie kvantilu*.
- Je tedy žádoucí, aby definice hloubky použitá na jednorozměrná data definovala kvantil (ve skutečnosti dva symetrické kvantily).
- Vnoření jednotlivých úrovňových množin hloubky je samozřejmostí.

Hloubka a hustota

- Hustota je **lokální** charakteristikou. Naopak hloubka zohledňuje **globální** postavení bodu vůči rozdělení.
- Obecně kontury hloubky a hustoty nejsou stejné.
- Výjimkou jsou *elipticky symetrická unimodální rozdělení*. Pro ty je hodnota hloubky **jednoznačně určena** hodnotou hustoty a naopak.
- Tato vlastnost ale **neplatí** pro jiné symetrie, ani pro l_p symetrická rozdělení pro $p \neq 2$.
- Proto není možné úplně přímočaré použití hloubky například pro klasifikaci. Musíme vymýšlet *rafinované* postupy.

Hloubka a nekonečně mnoho rozměrů

- Výhodou hloubky je, že může být definována i pro **nekonečně rozměrná data** (funkcionální data).
- Nelze ale postupovat úplně přimočaře zobecněním například poloprostorové hloubky.
- Existují jednoduché příklady, kdy poloprostorová hloubka dává *skoro všem bodům nulovou hloubku* (teoretická, nejenom empirická).
- Hloubka pro funkcionální data je *inspirována* hloubkou pro konečně rozměrná data, ale potřebujeme jiné definice.

- Připomeňme:

$$HD(x) = \inf_{H, x \in \partial H} P(H) = \inf_{|u|=1} P[(X - x)^T u > 0]$$

- Ze *SZVČ* víme, že $P_n[(X - x)^T u > 0] \to P[(X - x)^T u > 0]$ pro všechna u.
- Zde potřebujeme **stejnoměrnou konvergenci**.
- Díky stejnoměrnému SZVČ pak dostaneme bodovou konvergenci $HD_n(x) \to HD(x)$ s.j.

- Příliš velká *neshoda* mezi hloubkou a hustotou pro rozdělení s *nekonvexními konturami hustoty* vedla k zavedení zobecněné (vážené) poloprostorové hloubky.
- Vážené zobecnění poloprostorové hloubky:

$$WD(x) = \inf_{|u|=1} Ew(X - x, u)$$

- Opět potřebujeme *stejnoměrný SZVČ*, abychom mohli dokázat $WD_n(x) \to WD(x)$ skoro jistě pro všechna x.
- Dá se ale ukázat i více: $\sup_x |WD_n(x) - WD(x)| \to 0$ s.j.

- Připomeňme:

  $SD(x) = P[x \in S]$, kde S je náhodný simplex z rozdělení P

- Problém konzistence je *řešitelný* pomocí U-statistik.
- Podobně lze dokazovat i konzistenci některých *funkcionálních hloubek* (založených na **pásech funkcí** v roli simplexů). I tam lze využít teorii U-statistik.