KLASIFIKACE NA ZÁKLADĚ HLOUBKY BODU

ONDŘEJ VENCÁLEK
ondrej.vencalek@upol.cz
Přírodovědecká fakulta Univerzity Palackého v Olomouci

ÚVOD

Hloubka dat je dnes již běžně využívaným neparametrickým přístupem v mnohorozměrné statistice. Tento příspěvek shrnuje možnosti využití hloubky dat v klasifikační úloze, jak byly navrženy v uplynulém desetiletí.

Klasifikační úlohou rozumíme situaci, kdy máme k dispozici náhodný výběr ze směsi J různých pravděpodobnostních rozdělení $P_1, \ldots, P_J$ na d-dimenzionálním reálném prostoru, tzv. tréninkovou množinu. Přitom u každého pozorování z tréninkové množiny je známo, ze které distribuce toto pozorování pochází. Úkolem je přiřadit nové pozorování $\mathbf{x}$ k některému z uvažovaných rozdělení. Pravidlo, které k $\mathbf{x}$ přiřadí vhodný index rozdělení $d(\mathbf{x}) \in \{1, \ldots, J\}$ se označuje termínem klasifikátor. Kritériem pro hodnocení klasifikátorů je velikost jejich *average misclassification rate*, definované vztahem

$$\Delta = \sum_{j=1}^{J} \pi_j P(d(X) \neq j | X \sim P_j)$$

Jsou-li známy apriorní pravděpodobnosti jednotlivých rozdělení $\pi_1, \ldots, \pi_J$ a jejich hustoty $f_j(\cdot)$, $j = 1, \ldots, J$, používá se obvykle tzv. Bayesovský optimální klasifikátor $d(\mathbf{x}) = \arg\max_{j=1,\ldots,J} \pi_j f_j(\mathbf{x})$, který minimalizuje *average misclassification rate*. V praxi však většinou hustoty (a často ani apriorní pravděpodobnosti) nejsou známy.

KLASIFIKACE POMOCÍ MAXIMÁLNÍ HLOUBKY

je zatím zřejmě nejrozšířenějším způsobem využití hloubky bodu v klasifikační úloze. Najdeme jej v článcích *Jörnsten (2004)*, *Ghosh a Chaudhuri (2005)*, *Mosler a Hoberg (2006)* nebo *Hartikainen a Oja(2006)*. Jde o plně neparametrický přístup založený na hloubce dat. Tento klasifikátor přiřadí nové pozorování $\mathbf{x}$ tomu rozdělení, vůči němuž má $\mathbf{x}$ největší hloubku (neboť větší hloubka znamená polohu blíže „centru“ rozdělení):

$$d(\mathbf{x}, TS) = \arg \max_{j=1\ldots,J} D_j(\mathbf{x}, TS),$$

kde $D_j(\mathbf{x}, TS)$ je odhad hloubky pozorování $\mathbf{x}$ vůči j-tému rozdělení P_j založený na bodech tréninkové množiny TS.

Toto klasifikační pravidlo nespecifikuje, která definice hloubky se má použít. Nejčastěji se uvažuje L_1-hloubka, zonoidová hloubka, případně poloprostorová hloubka. Použijeme-li Mahalanobisovu hloubku $D(\mathbf{x}) = [1 + (\mathbf{x} - \mu)^T \mathbf{\Sigma}^{-1} (\mathbf{x} - \mu)]^{-1}$, jedná se vlastně o minimalizaci Mahalanobisovy vzdálenosti a tudíž o klasickou (Fisherovu) diskriminační analýzu.

Ghosh a Chaudhuri ukázali, že klasifikátor založený na maximální hloubce bodu je při použití poloprostorové, simplexové, projekční nebo „majoritní“ hloubky asymptoticky ekvivalentní Bayesovskému optimálnímu klasifikátoru, jestliže jsou distribuce $P_1, \ldots, P_j$ eliptické, unimodální, liší se jen parametrem polohy (mají stejné varianční matice) a jejich apriorní pravděpodobnosti jsou si rovny.

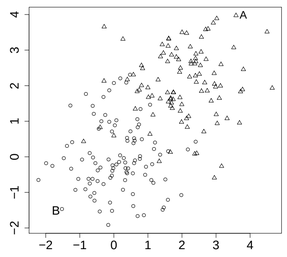
Problém nulové empirické hloubky

Při použití většiny známých hloubkových funkcí má poměrně velká část vícerozměrných pozorování nulovou empirickou hloubku vzhledem ke všem uvažovaným distribucím. Známý jev takzvané „řídkosti“ pozorování ve vícerozměrném prostoru ilustruje následující simulace: generujeme postupně z 2, 3, 4, 5 a 10-dimenzionálního standardního normálního rozdělení náhodné výběry o rozsahu 20, 50, 100 a 500 bodů. Pomocí quickhull algoritmu implementovaného v knihovně `R: geometry` zjistíme, kolik bodů z tohoto náhodného výběru se nachází na hranici konvexního obalu dat. Výsledky v podobě mediánu těchto počtů při 1000 opakováních simulace shrnuje Tabulka 1:

dimenze	rozsah výběru			
	20	50	100	500
2	7	8	9	11
3	12	17	22	32
4	16	28	39	70
5	19	37	57	124
10	20	50	98	

Tabulka 1: Medián počtu bodů na hranici konvexního obalu náhodného výběru z $N(0, I)$.

Naléhavost tohoto problému můžeme ilustrovat Obrázkem 1. Uvažujme dvě dvourozměrná normální rozdělení s jednotkovou varianční maticí a středními hodnotami $\mu_1 = (0, 0)^T$ a $\mu_2 = (2, 2)^T$. Z obou rozdělení máme náhodný výběr o rozsahu 100. Klasifikace bodů A a B (viz obrázek) je metodou maxiální hloubky při použití např. poloprostorové hloubky nemožná, neboť oba body mají empirickou hloubku vůči oběma rozdělením nulovou.



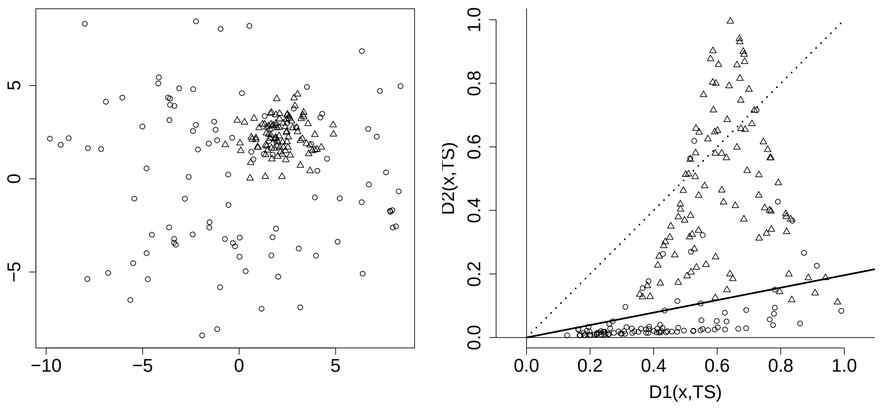
Obrázek 1: Příklad bodů (A, B), které nelze klasifikovat metodou maximální hloubky, neboť mají nulovou empirickou hloubku vůči oběma rozdělením.

Řešení problému nulové empirické hloubky:

- Použití hloubky, jejíž empirická verze je všude nenulová, např. Mahalanobisovy nebo L_1-hloubky (viz *Mosler a Hoberg*).
- Použití jiné klasifikační metody např. nejbližšího sousedního pozorování (viz *Ghosh a Chaudhuri*).
- Použití dalších pravidel jako např. minimalizace Euklidovské vzdálenosti od nejhlubšího bodu, případně extrapolace centrálních oblastí (je předmětem současného výzkumu).

Problém různých variančních matic

Tento problém snad nejlépe ukazuje příklad z článku *Li, Cuesta-Albertos, Liu* uvažující rozdělení $P_1 = N_2\left((0, 0)^T, 16I\right)$ a $P_2 = N_2\left((2, 2)^T, I\right)$.



Obrázek 2a (vlevo): náhodný výběr z rozdělení $P_1 = N_2\left((0, 0)^T, 16I\right)$ (kolečka) a $P_2 = N_2\left((2, 2)^T, I\right)$ (trojúhelníčky), 2b (vpravo): DD-plot, tedy graf hloubek jednotlivých bodů tréninkové množiny vůči první a druhé skupině bodů.

Na Obrázku 2a jsou body náhodného výběru z P_1 znázorněny kolečkem a body z výběru z rozdělení P_2 trojúhelníčkem. Jejich hloubky vzhledem k oběma skupinám bodů jsou zaznamenány na Obrázku 2b. Klasifikátor založený na maximální hloubce bodu přiřazuje body pod osou prvního kvadrantu do první skupiny ($D_1(\mathbf{x}, TS) > D_2(\mathbf{x}, TS)$, zatímco body nad touto osou přiřazuje do druhé skupiny. Je vidět, že dělící přímka (osa kvadrantu) zdaleka není zvolena optimálně. Lepší řešení je znázorněno plnou čarou. Je tedy vidět, že metoda maximální hloubky není pro případ různých variancí vhodná.

Problém různých apriorních pravděpodobností

Ukazuje se, že metoda maximální hloubky je nevhodná, pokud není splněn předpoklad stejných apriorních pravděpodobností všech rozdělení.

KLASIFIKACE POMOCÍ DD-PLOTU

V zatím nepublikovaném článku *Li, Cuesta-Albertos a Liu* navrhli použít pro klasicikační úlohu tzv. DD-plot. Uvažujme pro jednoduchost jen dvě pravděpodobnostní rozdělení P_1 a P_2 s distribučními funkcemi F a G. Pak DD-plot (graf hloubek vůči těmto dvěma rozdělením) je definován vztahem:

$$DD(F, G) = \{(D_F(\mathbf{x}), D_G(\mathbf{x})), \mathbf{x} \in Z\},$$

kde $Z = \{X_1, \ldots, X_m\} \cup \{Y_1, \ldots, Y_n\}$ je sjednocení náhodných výběrů z F a G, $D_F(\mathbf{x})$ je (nějaká) hloubka bodu $\mathbf{x}$ vůči F a $D_G(\mathbf{x})$ je (nějaká) hloubka bodu $\mathbf{x}$ vůči G. Pokud F a G nejsou známy, použijeme jejich empirické verze F_m a G_n.

Snadno se nahlédne, že pro eliptické, unimodální rozdělení je Bayesovský optimální klasifikátor ekvivalentní pravidlu:

$$D_G(\mathbf{x}, TS) > r\left(D_F(\mathbf{x}, TS)\right) \Longrightarrow \text{přiřaďme } \mathbf{x} \text{ ke } G$$
$$D_G(\mathbf{x}, TS) < r\left(D_F(\mathbf{x}, TS)\right) \Longrightarrow \text{přiřaďme } \mathbf{x} \text{ k } F$$

Autoři předpokládají (z důvodu jednoduchosti), že $r(\cdot)$ je lineární. Navíc musí platit $r(0) = 0$, a tak docházejí k pravidlu:

$$D_2(\mathbf{x}, TS) > \widehat{k} D_1(\mathbf{x}, TS) \Longrightarrow d(\mathbf{x}, TS) = 2$$
$$D_2(\mathbf{x}, TS) < \widehat{k} D_1(\mathbf{x}, TS) \Longrightarrow d(\mathbf{x}, TS) = 1$$

kde $D_j(\mathbf{x}, TS)$ je odhad hloubky pozorování $\mathbf{x}$ vůči j-tému rozdělení P_j (j=1,2) založený na bodech tréninkové množiny TS a $\widehat{k}$ je odhad směrnice přímky minimalizující empirickou misclassification rate:
$\widehat{\Delta}(k) = \frac{\pi_1}{m} \sum_{i=1}^{m} I_{[D_2(X_i, TS) > kD(X_i, TS)]} + \frac{\pi_2}{n} \sum_{j=1}^{n} I_{[D_2(Y_j, TS) < kD(Y_j, TS)]}$

Li, Cuesta-Albertos a Liu ukázali, že klasifikátor založený na DD-plotu je při použití poloprostorové, simplexové, projekční nebo Mahalanobisovy hloubky asymptoticky ekvivalentní Bayesovskému optimálnímu klasifikátoru, jestliže jsou distribuce P_1, P_2 eliptické, unimodální, liší se jen parametrem polohy (mají stejné varianční matice) a jejich apriorní pravděpodobnosti jsou si rovny.

KLASIFIKACE POMOCÍ DISTRIBUCÍ HLOUBEK

Billor a kol. navrhli tzv. „depth transvariation classifier“, který maximalizuje $F_{D_j}(D_j(\mathbf{x}, TS))$, kde $D_j(\mathbf{x}, TS)$ je hloubka pozorování $\mathbf{x}$ vůči j-té skupině bodů tréninkové množiny a $F_{D_j}(\cdot)$ je empirická distribuční funkce hloubek bodů j-té skupiny tréninkové množiny vůči celé této skupině. Snaží se tedy najít takovou skupinu, ve které podíl bodů z tréninkové skupiny s nižší hloubkou (vůči téže skupině) než bod $\mathbf{x}$ je co největší. Rozhodovací pravidlo tak má následující podobu:

$$d(\mathbf{x}, TS) = \arg \max_{j=1\ldots,J} \frac{1}{n_j} \sum_{i=1}^{n_j} I_{[D_j(\mathbf{X}_{ji}, TS) \leq D_j(\mathbf{x}, TS)]},$$

kde $D_j(\mathbf{X}_{ji}, TS)$ je hloubka i-tého bodu j-té skupiny tréninkové množiny vůči celé j-té skupině, $D_j(\mathbf{x}, TS)$ je hloubka pozorování $\mathbf{x}$ vůči j-té skupině bodů téninkové množiny a n_j je počet bodů j-té skupiny tréninkové množiny.

Vencálek ukázal, že klasifikátor založený na distribucích hloubek je při použití poloprostorové, simplexové, projekční nebo Mahalanobisovy hloubky asymptoticky ekvivalentní Bayesovskému optimálnímu klasifikátoru, jestliže jsou distribuce $P_1, \ldots, P_J$ eliptické, unimodální, liší se jen parametrem polohy (mají stejné varianční matice) a jejich apriorní pravděpodobnosti jsou si rovny.

SOFISTIKOVANĚJŠÍ KLASIFIKÁTORY

Všechny tři výše uvedené metody mají společné to, že jejich optimalita je dokázána jen pro velmi úzkou třídu klasifikačních úloh. Zejména předpoklady stejného rozdělení, stejných variančních matic a stejných apriorních pravděpodobností jsou hodně restriktivní a v praxi často neudržitelné.

Metody navržené *Ghoshem a Chaudhurim* a *Duttou a Ghoshem* jsou ekvivalentní Bayesovskému optimálnímu klasifikátoru za daleko mírnějších podmínek. Vycházejí ze skutečnosti, že pro elipticky symetrická rozdělení je Bayesovský klasifikátor ekvivaletní klasifikátoru $d(\mathbf{x}) = \arg\max_{j=1,\ldots,J} \pi_j \theta_j(D_j(\mathbf{x}))$, kde $\theta_j(\cdot)$ je nějaká funkce závislá na použité hloubce. Ukazují, že $\pi_j \theta_j(D_j(\mathbf{x}))$ má tvar:

$\lambda_j \rho_j(D_j(\mathbf{x})) \frac{D_j(\mathbf{x})^{d-3}}{(1 - D_j(\mathbf{x}))^{d-1}}$	pro projekční hloubku
$\lambda_j^* \rho_j^*(M_j(\mathbf{x})) \frac{1}{M_j(\mathbf{x})^{d-1}}$	pro Mahalanobisovu vzdálenost

kde λ_j, λ_j^* jsou konstanty a $\rho_j(\cdot), \rho_j^*(\cdot)$ jsou hustoty příslušných náhodných veličin $D_j(X)$ a $M_j(X)$. Tyto hustoty se pak odhadují jádrovým odhadem (jde o jednodimenzionální úlohu). Poznamenejme, že Mahalanobisova vzdálenost je známou funkcí Mahalanobisovy nebo poloprostorové hloubky.

Literatura:

[1] Ghosh A.K., Chaudhuri P. (2005) *On Maximum Depth and Related Classifiers.* Scandinavian Journal of Statistics, Vol. 32, 327–350.

[2] Li J., Cuesta-Albertos J.A., Liu R. *Nonparametric Classification Procedures Based on DD-plot.* rukopis

[3] Billor N. a kol. (2008) *Classification Based on Depth Transvariations.* Journal of Classification Vol. 25, 249–260.

[4] Dutta S., Ghosh A.K. (2009) *On Robust Classification Using Projection Depth.* technical report

[5] Jörnsten R. (2004) *Clustering and Classification Based on the L_1 Data Depth.* Journal of Multivariate Analysis Vol. 90, 67–89.

[6] Mosler K., Hoberg R. (2006) *Data Aanalysis and Classification with the Zonoid Depth.* in Data Depth: Robust Multivariate Analysis, Computational Geometry and Applications (Liu et al. Eds.), 49–59.

[7] Hartikainen A., Oja H. (2006) *On Some Parametric, Nonparametric and Semiparametric Discrimination Rules.* in Data Depth: Robust Multivariate Analysis, Computational Geometry and Applications (Liu et al. Eds.), 61–70.