KLASIFIKÁCIA ZAŠUMENÝCH DÁT

KATARÍNA CIMERMANOVÁ

katarina.cimermanova@gmail.com
ÚM SAV, Dúbravská cesta 9, 841 04 Bratislava

Klasifikácia viacrozmerných pozorovaní do dvoch tried je dôležitý problém. Existuje niekoľko klasifikačných metód na zatriedenie pozorovaných vektorov do jednej z dvoch tried, avšak v reálnom živote sú vektory pozorovaní zašumené. Riešením klasifikácie zašumených dát je robustná formulácia vychádzajúca z metódy oporných bodov. Formulácia je konvexný optimalizačný problém, ktorý je súčasťou problematiky kónického programovania druhého rádu. V robustnej formulácii sa predpokladá elipsoidálny model šumu. Nie je nutný predpoklad typu rozdelenia pozorovaných dát, predpokladá sa len konečnosť momentov druhého rádu.
Robustnú klasifikačnú metódu aplikujeme na analýzu vydychovaných plynov, kde sa budeme venovať klasifikácii dobrovoľníkov do skupiny fajčiarov a nefajčiarov.

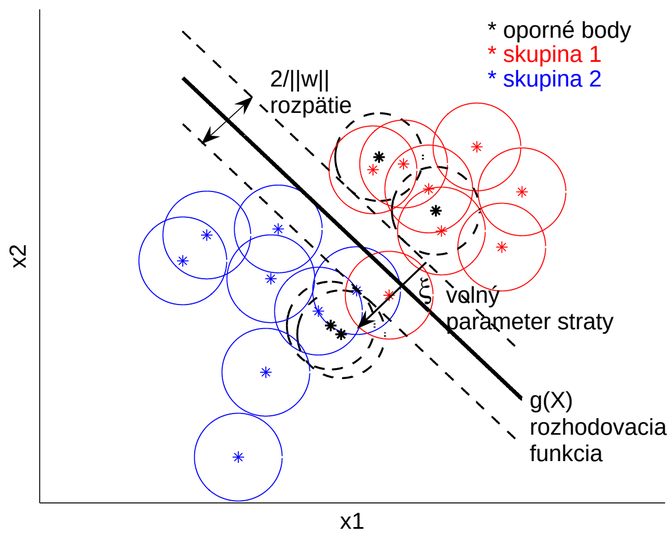
Predpokladajme, že naše namerané dáta $X \in R^N$ sú zašumené a skutočná hodnota je nejaký bod v špecifikovanom elipsoide, teda predpokladáme elipsoidálny model zašumenia. Nech

$$B(\overline{X}, \Sigma, \gamma) = \{X : (X - \overline{X})' \Sigma^{-1} (X - \overline{X}) \leq \gamma^2\}$$

je elipsoid, Σ je pozitívne semidefinitná matica a $\gamma \geq 0$. Riešením klasifikácie zašumených dát $X \in B(\overline{X}, \Sigma, \gamma)$ je nájdenie rozhodovacej funkcie

$$g(X) = \text{sign}\left(\langle w, X\rangle + b\right),$$

kde parametre w, b sú optimálne parametre pri hľadaní dvoch paralelných hyperrovín ku $g(X)$, ktorých rozpätie (*margin*) je $2/||w||$. Ide o maximalizáciu rozpätia tak, aby bol čo najmenší počet zle klasifikovaných pozorovaní, teda minimálna strata ďalej charakterizovaná voľnými (*slack*) parametrami straty $\xi \geq 0$. Body, pre ktoré platí $y(\langle w, B(\overline{X}, \Sigma, \gamma)\rangle + b) = 1$, kde $y = \{1, -1\}$ je kategorizácia pozorovania do jednej z dvoch tried, sa nazývajú oporné body (*support vectors*). Tieto body sú postačujúce pri popise rozhodovacej funkcie $g(X)$, predstavujú len malý zlomok všetkých dát. Riešením je optimalizačná úloha



kvadratického programovania

$$\min_{w,b,\xi} \quad \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} \xi_i$$
$$\text{s podm.} \quad y_i(\langle w, X_i\rangle + b) \geq 1 - \xi_i$$
$$\xi_i \geq 0,$$

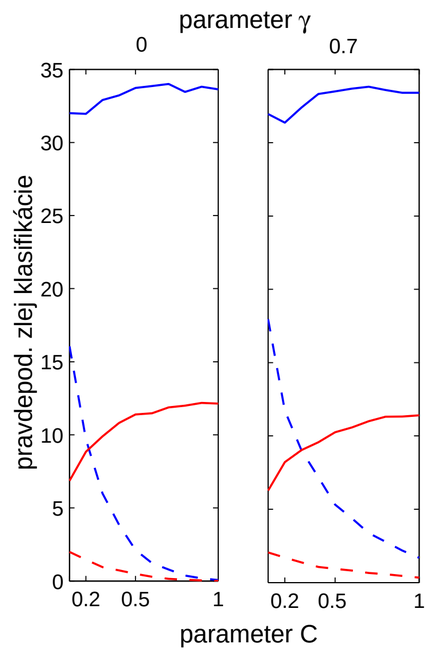
pre $\forall X \in B(\overline{X}, \Sigma, \gamma)$ a $i = 1, \ldots n$, kde parameter C je regularizačná konštanta, ktorá rieši kompromis medzi maximalizáciou rozpätia a stratou. Optimalizačná podmienka sa využitím Karush-Kuhn-Tuckerových podmienok dá prepísať na tvar $y_i \langle w, X_i\rangle = y_i \langle w, \overline{X}_i\rangle - \gamma_i ||\Sigma_i^{1/2} w||$. Potom nasledovná robustná formulácia je ekvivalentná s predchádzajúcou optimalizačnou úlohou

$$\min_{w,b,\xi} \quad \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} \xi_i$$
$$\text{s podm.} \quad y_i(\langle w, \overline{X}_i\rangle + b) \geq 1 - \xi_i + \gamma_i \|\Sigma_i^{\frac{1}{2}} w\|$$
$$\xi_i \geq 0,$$

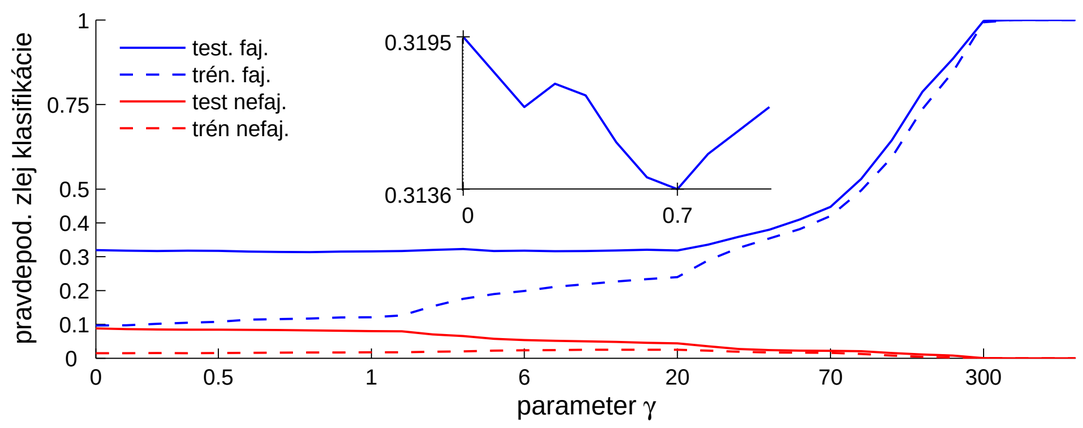
pre $i = 1, \ldots n$, kde robustnou ju robí nelineárny člen $||\Sigma_i^{\frac{1}{2}} w||$ nachádzajúci sa v obmedzujúcich podmienkach. Optimalizačná úloha sa rieši ako úloha kónického programovania druhého rádu (*second order cone programing*, SOCP)

$$\min_{w,b,\xi} \quad \sum_{i=1}^{n} \xi_i$$
$$\text{s podm.} \quad y_i(\langle w, \overline{X}_i\rangle + b) \geq 1 - \xi_i + \gamma_i \|\Sigma_i^{\frac{1}{2}} w\|$$
$$||w|| \leq W$$
$$\xi_i \geq 0,$$

pre $i = 1, \ldots n$, kde člen $||w||$ je presunutý do podmienky a ohraničený zhora konštantou W, ekvivalentnou s konštantou C.
Robustnú klasifikačnú metódu aplikujeme v analýze vydychovaných plynov (*Breath analysis*) na klasifikáciu dobrovoľníkov do skupiny fajčiarov a nefajčiarov (54 fajčiarov a 178 nefajčiarov). Pre každý subjekt ($i = 1, \ldots, n$) sú namerané koncentrácie prchavých organických zložiek s molekulovou hmotnosťou od 21 do 230 ($N = 208$) v jednotke počet molekúl na miliardu (*particles per billion*, ppb) pomocou on-line metódy hmotnostnej spektrometrie s protónovou prenosovou reakciou (*proton transfer reaction mass spectrometry*, PTR-MS).
Ďalej sledujeme závislosť empirickej pravdepodobnosti zlej klasifikácie subjektov skupiny fajčiarov a nefajčiarov pri trénovaní a testovaní klasifikačnej metódy pre kombinácie parametrov C = 0.1:0.1:1 a γ = 0:0.1:1. Výsledky sú priemery pravdepodobností chýb zo 100-krát náhodne rozdelenej databázy jednotlivých skupín na trénovaciu a testovaciu množinu v pomere 3:2. Predpokladáme $\Sigma_i = \sigma^2 I$, kde σ^2 predstavuje minimálny rozdiel medzi najvyššou a najnižšou nameranou koncentráciou zložiek v trénovacej množine. Z grafu vidieť, že klasifikačná metóda prednostne zatrieďuje subjekty z početnejšej skupiny. Pre rastúci parameter C chybovosť trénovania klesá a zároveň rastie pri testovaní. Pri vyšších hodnotách C dochádza k pretrénovaniu (*overfitting*) klasifikátora. Pre naše dáta sme za optimálny zvolili parameter $C = 0.2$, nakoľko pravdepodobnosť zle zatriedených fajčiarov z trénovacej množiny pre všetky parametre γ je najnižšia. Z grafu ďalej vidieť, že pravdepodobnosť zlého zatriedenia subjektov z testovacej množiny pri zvyšujúcom sa šume (parameter γ) pomaly klesá. Tým sme dokázali, že pravdepodobnosť zlého zatriedenia pri testovaní sa znižuje pri predpoklade, že aplikované dáta sú zašumené.



Pri veľkej hodnote parametra γ môže dochádzať k situácii, že dáta nie sú lineárne oddeliteľné. Preto sme pre parameter $C = 0.2$ zostrojili ďalšiu simulačnú metódu, kde pravdepodobnosť zlej klasifikácie bola odhadnutá na základe 100-krát náhodne rozdelenej skupiny fajčiarov a nefajčiarov na trénovaciu a testovaciu v pomere 3:2 pre



parameter γ = 0:500. Najlepšie výsledky dosahovala metóda pre parameter $\gamma = 0.7$ ($P_{testF} = 0.31$, $P_{testN} = 0.08$). Pre $\gamma \geq 20$ klasifikačná metóda nebola schopná oddeliť jednotlivé skupiny, pre $\gamma \geq 300$ zatrieďuje všetky subjekty do početnejšej skupiny.

Plány:

Coverova veta hovorí, že pravdepodobnosť, že triedy sú lineárne oddeliteľné rastie, keď sú premenné nelineárne mapované do priestoru s vyššou dimenziou. Implementovaním nelineárneho mapovania premenných zašumených dát do priestoru s vyššou dimenziou, by klasifikačná metóda mala dosiahnuť lepšie výsledky zatriedenia testovaných subjektov.

Poďakovanie:

Práca bola podporovaná EU projektom BAMOD: LSHC-CT-2005-019031 STREP, Agentúrou na podporu výskumu a vývoja (APVV), grant RPEU-0008-06, Vedeckou grantovou agentúrou Ministerstva školstva SR a Slovenskej akadémie vied (VEGA), grant 1/3016/06 a 2/7087/27.

Referencie:

[1] Bhattachryya, Ch. (2004) *Robust classification of noisy data using second order cone programming approach*. In Proceedings of Intelligent Sensing and Information Processing, 433 – 438.

[2] Shivaswamy, P.K., Bhattacharyya, Ch., Smola, A.J. (2006) *Second Order Cone Programming Approaches for Handling Missing and Uncertain Data*. Journal of Machine Learning Research, Vol. 7, 1283 – 1314.

[3] Sturm, J.F. (1999) *Using SeDuMi 1.02, a Matlab toolbox for optimization over symmetric cones*. Optimization Methods and Software, Vol. 11–12, 625 – 653.