Projekt z Lineárneho programovania: Binárna klasifikácia dát – výber signifikantných atribútov a prehľad aplikácií

Jednou z typických úloh v strojovom učení (machine learning) je úloha klasifikácie dát. Dané sú dve množiny dát

$$\{x_1, \dots, x_N\}, \quad \{y_1, \dots, y_M\}.$$

V lineárnej klasifikácii ide o nájdenie nadroviny danej vektorom a a skalárom b, ktorá dané dáta separuje, t.j. platí

$$a^T x_i - b > 0,\ \forall i = 1, \dots, N, \quad a^T y_i - b < 0,\ \forall i = 1, \dots, M. \tag{1}$$

Keď že tieto nerovnosti sú homogénne v a, b (t.j. možno ich násobiť ľubovoľným kladným číslom a nič sa nezmení), možno ich ekvivalentne vyjadriť ako:

$$a^T x_i - b \geq 1,\ \forall i = 1, \dots, N, \quad a^T y_i - b \leq -1,\ \forall i = 1, \dots, M. \tag{2}$$

Takáto úloha sa dá riešiť ako úloha LP s premennými a, b a konštantnou (nulovou) účelovou funkciou.

a) Zdôvodnite, že nerovnosti (1) a (2) sú ekvivalentné. Naformulujte príslušnú úlohu prípustnosti LP, ktorou nájdeme separujúcu nadrovinu.[1] Otestujte na dátach zo súboru sp_ln_sp_data. Pre separovateľné dáta má táto úloha viac riešení – rôzne metódy teda vedú k rôznym riešeniam. Porovnajte riešenie pomocou simplexovej metódy a pomocou metód vnútorného bodu.

b) Dáta v súbore majú príliš veľa charakteristík. To sa dá vyriešiť nahradením konštantnej účelovej funkcie z a) účelovou funkciou $\|a\|_1$. Minimalizáciou vektora a v l_1 norme totiž získame riešenie s veľkým počtom nulových zložiek. Ak je $a_k = 0$, tak zrejme k-ty atribút nemá vplyv na klasifikáciu - neovplyvňuje nerovnosti v ohraničeniach. Naformulujte túto úlohu ako úlohu lineárneho programovania (po vhodnej transformácii) a riešte pre dáta z sp_ln_sp_data. Zistite, ktoré atribúty sú redundantné a ktoré signifikantné. (Za nulovú zložku považujeme zložku spĺňajúcu $|a_k| < \varepsilon$, kde ε je vhodná tolerancia, môžete zvoliť napr. 10^{-6}.)

c) Urobte prehľad možných aplikácií binárnej klasifikácie v reálnom živote.

d) Nepovinná nadstavba. Nájdite na internete vhodnú dátovú sadu, môžete využiť databázu https://archive.ics.uci.edu/ml/datasets.php alebo iné zdroje, a vyskúšajte a aplikovať postup z a) na tejto sade. V prípade reálnych dát sa môže stať, že dátové množiny nie sú separovateľné, no napriek tomu by sme ich chceli aspoň približne separovať. To sa dá urobiť zavedením doplnkových nezáporných premenných $u \geq 0_N, v \geq 0_M$ a "relaxovaním" pôvodných nerovností do tvaru

$$a^T x_i - b \geq 1 - u_i,\ \forall i = 1, \dots, N, \quad a^T y_i - b \leq -1 + v_i,\ \forall i = 1, \dots, M.$$

[1] Úloha prípustnosti je každá úloha, kde máme nájsť riešenie nejakého systému rovníc alebo nerovníc. Každú úlohu prípustnosti môžeme považovať za optimalizačnú úlohu s konštantnou (napr. nulovou) účelovou funkciou a riešiť pomocou optimalizačného softvéru.

Tento systém nerovníc bude mať vždy riešenie, aj pokiaľ dáta nie sú separovateľné - stačí zvoliť odchýlky u, v dosť veľké. Zároveň by sme chceli mať zložky u, v najmenšie možné, žo vedie na úlohu s účelovou funkciou

$$\sum_{i=1}^{N} u_i + \sum_{j=1}^{M} v_j.$$

Rozdeľte dátovú sadu na testovaciu a trénovaciu (napr. v pomere 1:3) a vyhodnoťte percentuálnu úspešnosť klasifikácie.
Aj v tomto prípade je možné redukovať počet atribútov pomocou l_1 normy. V prípade neseparovateľných dát to vedie na bi-kriteriálnu úlohu s účelovou funkciou

$$\sum_{i=1}^{N} u_i + \sum_{j=1}^{M} v_j + \mu \|a\|_1,$$

kde $\mu > 0$ je kladný parameter (relatívna váha). Pre rôzne hodnoty μ získame rôzne riešenia (sú to tzv. pareto-optimálne riešenia). Vyriešením veľkého množstva úloh pre rôzne hodnoty μ (napr. 100 hodnôt z intervalu $[10^{-2}, 10^{5}]$ v log. škále) získate trade-off medzi optimálnou hodnotou $\phi(\mu) := \sum_{i=1}^{N} u_i^*(\mu) + \sum_{j=1}^{M} v_j^*(\mu)$ a počtom signifikantných atribútov. Vykreslite trade-off $\phi(\mu)$ vs. kardinalita a.