

## Domáca úloha č. 2

2-INF-150: Strojové učenie, Zima 2010

Termín: 8.11.2010, M163 (pod dvere)

**1. Teória strojového učenia.** Uvažujme problém regresie nad množinou hypotéz

$$H = \{h_b : x \rightarrow 2x + b\}.$$

- Popíšte algoritmus, ktorý spočíta pre danú tréningovú množinu  $(x_1, y_1), \dots, (x_t, y_t)$  hypotézu, ktorá minimalizuje chybu charakterizovanú chybovou funkciou  $J(b) = \frac{1}{t} \sum_{i=1}^t (h_b(x_i) - y_i)^2$ .
- Uvažujme pravdepodobnostnú distribúciu  $P_{x,y}$  definovanú nasledujúcim spôsobom:
  - rozdelenie  $x$ -ov je rovnomerné na intervale  $[0, 100]$
  - pre dané  $x$  je  $\Pr(y = 2x + 3 | x) = 0.3$  a  $\Pr(y = 2x - 2 | x) = 0.7$  (iné hodnoty  $y$  sa v kombinácii s  $x$  nevyskytujú).

Áká je optimálna testovacia chyba pre množinu hypotéz  $H$  ak predpokladáme, že dáta sú nezávislými vzorkami z tejto distribúcie?

- Pre pravdepodobnostnú distribúciu dát  $P_{x,y}$  z časti b) a  $t = 1$  spočítajte očakávanú tréningovú a očakávanú testovaciu chybu.
- Pre pravdepodobnostnú distribúciu dát  $P_{x,y}$  z časti b) a všeobecné  $t$  spočítajte očakávanú tréningovú a testovaciu chybu. V akom vzťahu sú tieto chyby ku optimálnej testovacej chybe keď  $t \rightarrow \infty$ ? Vedeli by ste na základe vašich výsledkov zvoliť vhodnú veľkosť tréningovej množiny?

**2. Cukrovka u arizonských indiánov.** V tejto úlohe sa budeme zaoberať štúdiom výskytu cukrovky u arizonských indiánov. Keďže sa v medicíne vôbec nevyznáte, rozhodli ste sa ísť na to vedecky a aplikovať na tento problém poznatky zo strojového učenia.

V UCI repository na adrese <http://archive.ics.uci.edu/ml/datasets/Pima+Indians+Diabetes> nájdete data set "Pima Indians Diabetes". Tento data set obsahuje pre každú indiánku zaradenú do štúdie 8 kvantitatívnych atribútov a diagnostiku, či indiánka trpí cukrovkou alebo nie.

V každej časti zdokumentujte svoj postup a ukážte kusy kódu, ktoré ste použili na riešenie jednotlivých podúloh.

- Dáta náhodným spôsobom rozdeľte na 5 približne rovnako veľkých skupín. Prvé tri skupiny použijete v častiach b-d ako tréningovú množinu, posledné dve skupiny použijete na testovanie. Koľko diabetikov sa nachádza v každej skupine?
- Natrénujte a otestujte SVM pre rozlišovanie diabetikov od nediabetikov a vhodným spôsobom vyhodnoťte jeho úspešnosť. Porovnajme použitie lineárneho kernelu (obyčajný skalárny súčin) a gaussovského kernelu.
- Lineárne preškálujte dáta tak, aby každý atribút mal strednú hodnotu 0 a štandardnú odchýlku 1 (tzv. normalizácia). Potom zopakujte experiment z časti b). Viete zdôvodniť, prečo metóda funguje na normalizovaných dátach lepšie, ako na pôvodných?
- Knižnica, ktorú použijete na SVM, pravdepodobne používa rôzne vylepšenia oproti základnej metóde SVM z prednášky. Zistite aké a stručne popíšte parametre, ktoré je možné nastavovať a akým spôsobom sa prejavujú pri aplikácii algoritmu. (Môžete si vypomôcť odbornou literatúrou a jednoduchými experimentami s dátami.)

- e) Dáta obsahujú podstatne menej diabetikov ako nediabetikov. Môže tento fakt ovplyvniť aplikáciu metódy SVM? Akým spôsobom? (Hint: rozmýšľajte nad tým, čo by sa stalo, keby sme mali ešte oveľa viac nediabetikov, môžete tiež použiť experimenty s dátami.)

Môžete použiť ľubovoľný programovací jazyk (alebo ich kombináciu) a ľubovoľnú knižnicu, ktorá implementuje SVM. Odporúčame kombináciu Octave a LibSVM.

**3. (Bonus) Duálny duálny program.** Odvodte duálny Lagrangeov program od duálneho programu pre SVM prezentovaného na prednáške.

**Poznámka:** Bonusový príklad bude hodnotený oveľa prísnejšie ako zvyšok domácej úlohy a keďže je bonusový, nie je nutné ho riešiť na získanie plného počtu bodov. Čiastočné riešenia nezískajú žiadne body. Ak ste na pochybách, radšej investujte čas do príkladov 1 a 2.