Posudek oponenta bakalářské práce

Název: Neighborhood components analysis and machine learning
Autor: Jan Hanousek

Shrnutí obsahu práce

Bakalárska práca študenta Jana Hanouska sa venuje problému klasifikácie pomocou metódy najbližších susedov. V úvode práce autor stručne popisuje primárny algoritmus v tejto oblasti, tzv. metodu k-teho najbližšieho suseda (KNN) a následne v ďalších dvoch kapitolách predstavuje niekoľko zobecnení, ktoré základný algoritmický postup v rôznych ohľadoch vylepšujú.

V štvrtej kapitole autor predstavuje vlastný klasifikačný prístup – dvojkrokový algoritmus založený na kombinácii lineárnej diskriminačnej analýzy a metódy k-teho najbližšieho suseda (resp. jej modifikáciach). V závere práce sú diskutované klasifikačné postupy porovnané vzhľadom k ich výpočetnej náročnosti pomocou autorom navrhnutej simulačnej štúdie.

Bakalárska práca je celkovo spracovaná ako súvislý a tématický aj logický korektne formulovaný text. Z matematického hľadiska prácu hodnotím ako jednoduchú a nenáročnu. Celkový dojem z práce ale kazí množstvo preklepov (hlavne v závere práce, kde je vidieť, že autor nestíhal), slabú formálnu úpravu a miestami aj angličtinu a často nekonzistentné, prípadne chybné, alebo nezavedené matematické značenie.

Otázky & pripomienky

Celkovo považujem prácu za dostatočnú a doporučujem ju uznať ako bakalársku prácu.

Otázky & pripomienky k obhajobe

- V definícii (1.4) nie je jasné, čo je "training set" a čo je "validation set". Asi by bolo vhodnejšie nenazývať celkovú množinu S ako "training set".
- Ako je definovaná "K-fold cross validation"? Z definície (1.5) nevyplýva, že množiny S_i pre $i = 1, \ldots, k$ by mali mať rovnakú mohutnosť.
- V definici (1.7) je K_x formálne množina indexov, nie množina susedov (t.j. podmnožina S).
- V druhej kapitole nie je jasné, kedy má autor na mysli samotný bod $x \in S$ a kedy referenčný bod $Ref(x) \in S$. Správne by sa malo byť uvedené buď: "the reference point of x", nebo "the reference point $Ref(x)$". Množina S^{-i} by mala byť taktiež formálne definovaná.
- Je rozdiel medzi vektorom w a $\boldsymbol{w}$ (napr. str. 8, 10)?
  Je rozdiel medzi maticou L a $\boldsymbol{L}$ (napr. str. 10, 11)?
- Ako sú definované množiny D_i, S_i, $NS_k(x_i)$ a $ND_k(x_i)$? Formálne sa nejedná o množiny potenciálných referenčných bodov, ale množiny indexov potenciálných referenčných bodov. Je samotný index i prvkom množiny D_i (nebo S_i)?

- ❏ Je množina S_i vo výraze (3.5) totožná s množinou S_i v definícii (1.4)?
- ❏ Aká lineárna transformácia je uvažovaná vo výraze (3.6)? Ako je definovaný výraz $d_{ij}(\boldsymbol{L})$?
- ❏ Výraz $\|\boldsymbol{L}\|_F^2$ na str. 11 by mal byť formálne zavedený.
- ❏ Ako je definovaná funkcia $k(x, y)$ na str.12?
- ❏ Ako sú definované pásy v obrázkoch 4.1 a 4.2?
- ❏ Čo znamená funkcia v poznámke na str. 17?
- ❏ Ako je definovaná "data set size" na obr. 5.1?
  Prečo hodnota n nadobúda neceločíselné hodnoty?
- ❏ Ako sú definované hypotézy a štatistický test, ktorý je uvedený na str. 20 a 21?

Praha, 15.06.2018

RNDr. Matúš Maciak, Ph.D.
maciak@karlin.mff.cuni.cz