POSUDOK OPONENTA BAKALÁRSKEJ PRÁCE

Názov: Porovnání metod pro diverzifikaci 0-1 proměnných
Autor: Nikita Kan

ZHRNUTIE OBSAHU PRÁCE

Obsahom práce je popis viacerých prediktívnych štatistických metód regresného charakteru: hodnoty premennej typu odozva (závislá premenná) sú predikované z hodnôt viacerých kovariátov (nezávislých premenných, regresorov). Všetky premenné môžu byť tak numerickej, kvantitatívnej povahy, ako aj kategorického, kvalitatívneho charakteru; podľa typu odozvy autor rezervuje termín „predikcia“ na predikciu kvantitatívnych premenných, zakiaľ čo predikciu premenných kvalitatívneho charakteru nazýva „klasifikáciou“. Predikčné metódy popisované v práci sú logistická regresia, regresia/klasifikácia založené na rozhodovacích stromoch (CART), včítane špecializácie CHAID na dáta pozostávajúce výlučne z kategorizovaných premenných; v závere autor popisuje algoritmus Discrete AdaBoost, ktorý představuje „recyklovanú“ verziu predikcie pomocou jednoduchých klasifikačných stromov. Záverom práce sú uvedené metódy implementované v jazyku R a aplikované na dáta z praxe.

Všetko toto viac-menej zodpovedá zadaniu práce tak, ako ho formuloval jej školiteľ.

CELKOVÉ HODNOTENIE PRÁCE

V prvom rade, je tak trochu otázne, do akej miery bola uvedené zadanie adekvátne na bakalárskej úrovni, keď študenti ešte nie sú tak celkom vybavení dostatočnými znalosťami na spracovanie zadania tohto typu. Práca pôsobí skorej dojmom práce diplomovej, no ak by sme ju posudzovali z tohto hľadiska, tak by sme mohli ešte nájsť niektoré nedotiahnuté aspekty; no toto sa jej nedá vyčítať ako práci bakalárskej, a v konečnom dôsledku sa dá len konštatovať, že tak ako bolo zadanie formulované, tak bolo aj splnené.

Vlastným príspevkom autora je aplikácia popisovaných metód na dátový súbor z praxe; čo sa implementácie týka, tá je realizovaná viac-menej knižnicami existujúcimi v prostredí R, samotná autorova implementácia sa obmedzuje na niektoré špeciálne aspekty a úkony. Kód inak vyzerá vcelku kultivovane, a dá sa vcelku aj veriť, že je správny. Čo sa týka analyzovaných dát z praxe, tieto nastoľujú istý dáta-analytický problém; autor si ho uvedomuje, ale ďalej už s ním – už aj vzhľadom na existujúci rozsah práce – nič nepodniká. Z tohto hľadiska by sa mohla dátovo-analytická časť práce považovať za trochu nedotiahnutú – ale ako je už uvedené vyššie, toto by mohlo byť kritikou práce diplomovej, no táto je bakalárska.

Čo sa týka spracovania, matematickej úrovne, práce so zdrojmi a formálnej úpravy, tuná sa nedajú vytknúť nijaké nedostatky. Text je napísaný kultivovaným spôsobom, a to ako po štylistickej, tak aj matematickej stránke; v prezentácii matematických formulí a citácii zdrojov som nenašiel žiadne nedostatky. [Akurát v prvom odstavci na str. 6 mi udrela do očí istá kostrbatosť, veta „vychádzíme z práce Hosmera, Lemeshowa a Sturdivanta (Hosmer, Lemeshow a Sturdivant, 2013)“ by sa dala napísať priamočiarejšie „vychádzame z práce Hosmer, Lemeshow a Sturdivant (2013).“] Takže aj formálna úprava práce je výborná – čo zakladá moje veľmi pozitívne hodonotenie, založené na názore, že v práci bakalárskej treba klásť dôraz hlavne na profesionálny výzor textu.

ZÁVER

Záverom teda môžem skonštatovať, že prácu považujem za veľmi dobrú a odporúčam ju uznať ako bakalársku prácu. Celkové hodnotenie oznámim predsedovi komisie – a navrhujem, aby tiež záviselo aj od toho, ako bude uchádzač schopný zodpovedať otázky položené nižšie.

OTÁZKY

1. V práci na viacerých miestach uvádzate prístupy na nastavenie kontrolných parametrov: (i) rozdelenie dát na trénovaciu a testovaciu časť (angl. sample splitting; str. 19 dolu, druhý odstavec v kóde na str. 60);
   (ii) metóda „krížovej validácie“ (angl. cross-validation; str. 20, posledný odstavec Sekcie 3.2, str. 24, posledný odstavec Sekcie 3.4, str. 25, posledný odstavec Sekcie 3.5.).
   Popíšte, prosím, ako fungujú oba prístupy: s dôrazom na princípy, ale tiež aj na isté podstatné detaily.
2. Ako špeciálny prípad „krížovej validácie“: mohli by ste detailnejšie vysvetliť ako bude konkrétne aplikovaná na výber parametra M v boostingu? (Sekcia 3.5.)
3. Ak rozdeľujete (sample- splitting) dáta tak, ako je to implementované v druhom odstavci kódu na str. 60 – aký problém nastáva pre konkrétny dátový súbor opísaný v Sekcii 5.1.?

PRAHA, 5. 6. 2023

IVAN MIZERA
KPMS MFF UK
imizera@me.com