Posudek oponenta bakalářské práce

Název: Bayesovské klasifikační a regresní stromy
Autor: Martin Dvořák

Shrnutí obsahu práce

Predložená bakalárska práca študenta Martina Dvořáka sa venuje klasifikačným a regresným stromom. Práca je formálne členená do štyroch kapitol: V prvej autor predstavuje problematiku klasifikačných stromov a definuje potrebné pojmy. V druhej kapitole sú následne diskutované modifikácie za účelom "klasifikácie" spojitej premennej—teda problematika tzv. regresných stromov. Tretia kapitola popisuje Bayesovské prístupy ako štandardné metódy bežne používané v praxi pri vytváraní klasifikačných, aj regresných stromov. Posledná, štvrtá kapitola stručne ilustruje použitie klasifikačných stromov pomocou programu R (Team Core, 2022) na reálnom príklade.

Vzhľadom k stanovenému cieľu práce (t.j., "*seznámit se s konstrukcí Bayesovských klasifikačních a regresních stromů*") by sa dala namietať určitá efektivita vypracovania, kde prvé dve kapitoly (takmer 40 strán textu) predstavujú vpodstate len akýsi úvod do danej problematiky, pričom samotné Bayesovské prístupy (základná a nosná téma samotnej práce) sú diskutované pomerne stručne—na deviatich stranách tretej kapitoly.

Téma práce považujem za pomerne náročné, hodne rozsiahlé a asi aj zbytočne komplikované pre študenta bakalárskeho štúdia (ktorý štandardne nie je ešte oboznámený ani s problematikou jednoduchej lineárnej regresie). Autor sa ale s danou problematikou popasoval dôstojne a dokázal ju určitým spôsobom aj zmysluplne spracovať a prezentovať. Práca je kompilačného charakteru (celkovo 60 strán textu predstavuje asi dvoj- až troj-násobok doporučeného a očakávaného rozsahu), formulovaný text inklinuje viacej k inžinierskej než matematickej terminológii. Z teoretického/matematického hľadiska ma práca viacero nedostatkov, čo pramení hlavne z komplexnosti celej problematiky. V práci sa často objavujú napr. nepresné, alebo nesprávne teoretické argumentácie a odvodenia, alebo nekonzistentné, či nejasné matematické formulácie (niekoľko konkrétnych prípadov je explicitne uvedených nižšie). Vzhľadom na celkový rozsah práce je taktiež trochu prekvapujúce, ako málo sú citované odborné zdroje, z ktorých autor v práci pravdepodobne vychádza (napr. v celej druhej kapitole o regresných stromoch sa objavuje pouze jediná referencia—James a kol., 2021). Na mnohých miestach by bolo určite vhodné doplniť relevantné zdroje (napr. pri zmienke o Giniho indexe diverzity, twoing kritériu, entropii, atď.).

Napriek všetkému vyššie uvedenému považujem predloženú prácu za zaujímavú a určite hodnú bakalárskej práce na MFF UK. Náročnosť spracovanej témy je výrazne nadpriemerná, samotné vypracovanie autorom skôr priemerné. Prácu jednoznačne doporučujem štátnicovej komisii uznať ako bakalársku prácu na MFF UK.

Hlavné pripomienky

- V Sekcii 1.3.1 autor definuje jednak teoretické pravdepodobnosti a následne príslušné empirické odhady. V použitom značení ale nijak nereflektuje rozdiel medzi teoretickými hodnotami a empirickými odhadmi. Celkovo sa v práci objavuje pomerne často značenie, z ktorého nie je zrejmé, či autor myslí teoretické vlasnosti, alebo odkazuje na empirické odhady spočítané z konkrétneho náhodného výberu;

- V úvode druhej kapitoly autor používa združené rozdelenie náhodného vektoru $(\boldsymbol{X}, Y)$ (str.31). Je logické preto predpokladať, že stredná hodnota v Definici 16 je myslená vzhľadom k tomuto združenému rozdeleniu. Čo ale predstavuje hodnota y v danom vzorci (Definice 16)?

- Stredná štvorcová chyba $R^*(\cdot)$ (viď Definice 16 a úvod Sekcie 2.1) vyžaduje v argumente funkciu (napr. $d : \mathcal{X} \to \mathbb{R}$). Ako je teda potrebné rozumieť výrazu

$$RE^*(d) = \frac{R^*(d)}{R^*(\mu)},$$

  kde $\mu = E[Y] \in \mathbb{R}$? Autor navyše tvrdí, že $R^*(\mu) = E[Y - \mu]^2$, pričom stredná hodnota v predchádzajúcom výraze je počítaná vzhľadom k marginálnemu rozdeleniu náhodnej veličiny Y (a nie združenému rozdeleniu $(\boldsymbol{X}, Y)$, ako to je formálne definované v Definici 16). Ak budeme pod výrazom $R^*(\mu)$ uvažovať strednú štvorcovú chybu pre funkciu $\mu : \mathcal{X} \to \mathbb{R}$ takú, že platí $\mu(\boldsymbol{x}) = E[Y]$ pre všetky $\boldsymbol{x} \in \mathcal{X}$ bude platiť rovnosť

$$E_{(\boldsymbol{X},Y)}[(Y - \mu(\boldsymbol{X}))^2] = E_Y[(Y - E[Y])^2],$$

  (kde stredná hodnota na ľavej strane je počítaná v zmysle Definice 16—teda vzhľadom k združenému rozdeleniu náhodného vektoru $(\boldsymbol{X}, Y)$—a pravá strana je počítaná v zmysle vzťahu uvedenom pred Definici 17—teda vzhľadom k marginálnemu rozdeleniu náhodnej veličiny Y)?

- Celkovo je spôsob zavedeného a používaného značenia trochu zmätočný. Napr. v práci je hodne náročné rozlíšovať a správne sledovať, kedy autor myslí teoretickú hodnotu a kedy jej príslušnú empirickú verziu. Niektoré teoretické hodnoty sú zavedené v "*hviezdičkovanej* $(*)$" verzii—napr. (teoretická) stredná štvorcová chyba $R^*(d)$ v Definici 5. Ale teoretická regresná funkcia je značená symbolom d (namiesto očakávaného d^*). Značenie sa reverzne obráti, ked prejdeme k empirickým náprotivkom—odhad strednej štvorcovej chyby je značený symbolom $R(\cdot)$ (viď resubstitučný odhad v úvode str.32), zatiaľ čo empirický odhadnutá teoretická funkcia d je zase označená symbolom d^*. Podobných, prípadne hodne analogických (a z určitého hľadiska až nelogických) prípadov značenia je v práci pomerne veľa.

- Vo výraze (2.1) je niečo nesprávne. V prvom rade hodnota $Var[\overline{y}]$ je rozptyl výberového priemeru, teda $Var[\overline{y}] = \frac{1}{n} VarY$ a jedná sa o teoretickú, ale nenáhodnú hodnotu. Naproti tomu na pravej strane rovnosti je výberový rozptyl náhodnej veličiny Y a teda sa jedná o empirickú, ale hlavne náhodnú veličinu. Ako ma byť výraz (2.1) správne?

- V empirickej časti (t.j. štvrtá kapitola) by bolo vhodné začať s dostatočne podrobným popisom datového súboru a základného problému, kôli ktorému celý klasifikačný strom vytvárame. Následovať by mala aspoň stručná (základná) exploratívna analýza dat (napr. z textu vôbec nie je zrejmé, v akých hodnotách sa jednotlivé premenné pohybujú, aka je štruktúra jednotlivých pozorovaní, prípadne aké je (absolútne/relatívne) zastúpenie v jednotlivých kategoriách). Výsledky sú prezentované veľmi rozpačito, nekonzistentne a navyše chýba akákoľvek zmysluplná interpretácia prezentovaných výsledkov (napr. aj desatinné čisla autor zapisuje nekonzistentne, niekedy používa pre zápis desatinnú čiarku, inokedy desatinnú tečku).

  Mimochodom, zdrojový kod programu R by sa štandardne vôbec nemal v práci vyskytovať.

Drobné poznámky

- ❏ V práci autor často zamieňa symbol $\mathcal{X}$ so symbolom X (viď napr. Definice 1, alebo posledný odstavec na str.5, atď.). Vzhľadom k celkovej konzistencii používaného značenia by asi bolo vhodnejšie značiť príslušné podmnožiny $\mathcal{X}$ ako $\mathcal{X}_1, \mathcal{X}_2, \ldots$ namiesto autorom zavedeného značenia $X_1, X_2, \ldots$ (v Sekcii 1.2 na str.8);
- ❏ Naozaj predstavuje vektor $(\boldsymbol{X}, Y)$ "*nový výběr*", ako autor uvádza v Definici 2, alebo sa jedná o generický náhodný vektor z rozdelenia, z ktorého náhodný výber $\mathcal{L}$ pochádza?
- ❏ Čo presne znamená "*odhad odhadu* $R^* d^{(v)}$ v Definici 5?
- ❏ Je naozaj nutné, aby každý uvažovaný prediktor $x_m \in \{x_1, \ldots, x_M\}$ "*nabýval kategoriálních hodnot* v $\{b_1, \ldots, b_L\}$"? Resp. vyjadrené inými slovami, je nutné, aby každý prediktor bol rovnakého kategoriálneho typu (t.j., s tými istými kategóriami)? (str.9)
- ❏ Formulácia Definice 10 je hodne nematematická. Asi by bolo vhodné aspoň výraz "*maximum*" formalizovať presne s použitím vhodných matematických symbolov a zápisu;
- ❏ V Definici 10 autor spomína tzv. "*třídy* $\mathcal{F}$", chýba ale formálna definícia o aké triedy sa jedná, prípadne vhodný odkaz na literatúru;
- ❏ V poslednom odstavci na str.26 sa objavujú symboly ▲ a ▼. Nie je jasné, čo tieto symboly predstavujú;
- ❏ Čo predstavuje symbol $P(d\boldsymbol{x})$ na str.39? O akú pravdepodobnosť sa jedná?
- ❏ V práci sa objavujú preklepy, ktoré vyvolávajú dojem, že dokončovanie práce prebiehalo na poslednú možnú chvíľu a väčšina je veľmi rýchlo odhaliteľná, ak by si autor po sebe prácu aspoň raz poriadne prečítal;
- ❏ Štatistický program R by mal byť v práci taktiež korektne citovaný—napr. "*R (Core Team, 2022)*";
- ❏ V práci sa občas objavujú preklepy, rôzne chyby formátovania textu (napr. nesprávne úvodzovky, zdvojené slova, chýbajúce bodky na konci vety), nejasné formulácie (napr. odkazy na akýsi "*nový výběr*", pričom nie je zrejmé, či sa má jednať o náhodný výber, alebo nie), alebo neúplne, nejasné, alebo úplne chýbajúce popisky k obrázkom/tabuľkám.

Londýn, 18.08.2023

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