Posudek oponenta bakalářské práce

Název: Konformní predikce
Autor: Michaela Krynická

Shrnutí obsahu práce

Práce pojednává o tzv. *konformní predikci*. Jedná se o postup, pomocí kterého lze konstruovat predikční intervaly bez nutnosti silných předpokladů na model. S ohledem na tento aspekt předkládá autorka konformní predikci jako dobrou alternativu oproti predikci v kontextu lineární regrese, kde klasické rovnice pro predikční intervaly požadují normalitu dat.

Autorka nejprve definuje základní pojmy a připomíná dobře známé výsledky. Rovnice pro konstrukci predikčních intervalů konformní metodou jsou formulovány v názorném případu jednorozměrných nezávislých a stejně rozdělených dat ze spojitého rozdělení. V takovém případě můžeme elegantně využít pozorování, že pořadí každé veličiny v uspořádaném výběru má rovnoměrné rozdělení. Následně je metoda precizována v algoritmus, a to v základním případě a také v kontextu regresní úlohy, kdy kromě předešlých pozorování známe také následnou hodnotu volných proměných. Práce je ukončena vlastní simulací, která demonstruje, že dobré výsledky lze očekávat i za porušení zmíněného předpokladu normality dat.

Celkové hodnocení práce

Téma práce. Téma je přiměřeně náročné, je stavěné nad rámec základního učiva bakalářského studia a zadání bylo splněno.

Vlastní příspěvek. Vlastní příspěvek spočívá v rešerši, sjednocení značení a formulaci výsledků a je dostatečně specifikován.

Matematická úroveň. Práce obsahuje dva velmi jednoduché důkazy a dva formulované algoritmy, které ovšem neodpovídají svému popisu (viz Otázky). Definice jsou zpracovány rigorózně stejně jako doprovodný text.

Práce se zdroji. Zdroje, ze kterých studentka výchází, jsou uvedeny. Náznak plagiátorství není z textu patrný.

Formální úprava. Odpovídá úrovni bakalářské práce. Práce obsahuje některé jazykové nedostatky a nedokončené věty, avšak žádné z těchto nebrání porozumění textu.

Otázky

1. Někteří autoři (například Shafer & Vovk (2008)) uvažují namísto Vaší definice ze strany 14

$$r_i^x := A((X_1, \ldots, X_n), X_i), \quad i = 1, \ldots, n$$

raději definici

$$r_i^x := A((X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n), X_i), \quad i = 1, \ldots, n,$$

tedy porovnávají každé pozorování X_i s náhodným výběrem, ze kterého je dané X_i odebráno. Uvažujme nyní Váš příklad nonkonformity A ze strany 13, tedy Euklidovskou vzdálenost od výběrového průměru. Dávají různé definice r_i^x různé výsledky? Jak je to v případě obecného A?

2. V příkladu na straně 19 je uvedeno, že v případě modelu lineární regrese není potřeba pro všechna (tedy nespočetně mnoho) y ověřovat, zda náleží do predikčního intervalu. Důvod není jasně prezentován, prosím vysvětlete. Popište také užité značení ($\hat{Y}_i^y, \hat{Y}_{20}^y$ nejsou popsány).

3. V poznámce na straně 3 porovnáváte vlastnosti predikčního a konfidenčního intervalu, aniž by konfidenční interval byl definován. Jaká je jeho definice?

4. Oba prezentované algoritmy by měly dle popisu mít na výstupu predikční interval, avšak v takto prezentované formě tomu tak není. Prosím opravte algoritmy, aby nebyly vnitřně sporné.

5. Prosím popište přesně rozdíly mezi předpoklady na model v kapitole 2 oproti kapitole 1.4. V úvodu kapitoly uvádíte motivaci pro zobecnění metody, aby pokrývala i klasifikaci, nicméně hned v dalším odstavci se omezujete na spojitá rozdělení.

6. Na straně 6 citujete důkaz Tvrzení 1. Odkazovaná věta je ve verzi ze srpna 2023 formulována za jiných předpokladů. Ověřte, prosím, že Vaše tvrzení vyplývá z citovaného materiálu.

7. V Definici 6 uvažuje citovaný autor pojem *bag* místo množiny pro M. Tedy umožňuje, aby se prvky v M opakovaly. Diskutujte, zda za Vašich předpokladů jde o podstatný rozdíl.

Připomínky

1. Na straně 3 uvažujete náhodnou veličinu X, aniž by byla dále použita.

2. V první poznámce na straně 3 bychom měli opravit první větu "...s pravděpodobností *ale-spoň*...", abychom zachovali konzistenci s Definicí 1.

3. V druhé poznámce na straně 3 by $\eta_L(X)$ mělo být reálné číslo (η_L je zobrazení)

4. Formule v Definici 3 obsahuje nadbytečné odsazení.

5. Věta před Tvrzením 1 nedává dobrý smysl, neboť odkazovaná formule "prostě platí", není pravda, že formule by "platila asymptoticky".

6. Poslední formule na straně 6 neplyne pouze z Tvrzení 1 a (1.1), neboť tyto poskytují pouze jednu nerovnost. Jsou potřeba užít i rovnosti pod (1.1).

7. Str. 8, řádek 23 obsahuje nadbytečné závorky v indexech náhodného výběru

8. V prvním řádku v důkazu Věty 3 je nesprávně uvedeno, že R je funkce na reálné přímce (je opomenuto, že obsahuje náhodnou složku).

9. Řádek 22 na straně 13 je nedokončený.

10. Řádek 8 na straně 15 obsahuje překlep.

11. Na straně 17 nejdříven označíte sdružené rozdělení a následně prohlásíte, že toto rozdělení je zároveň rozdělením nového pozorování, což nedává smysl.

12. Poslední část textu na straně 2 nedává smysl a působí rozpracovaně.

Závěr

Práci považuji za průměrnou a doporučuji ji uznat jako bakalářskou práci.

Ondřej Týbl
Katedra pravděpodobnosti a matematické statistiky MFF UK
14. srpna 2023