Posudek oponenta bakalářské práce

Název: Statistické učení a určování rizikovosti klientů
Autor: Martin Šíma

Shrnutí obsahu práce

Práce se zabývá nejprve obecným popisem problematiky predikce založené na modelech statistického učení. Dále uvádí dle studijních materiálů využívaných k výuce statistiky základní přehled modelu lineární regrese. Hlavním obsahem práce je pak popis modelu logistické regrese obsahující vedle formulace modelu a odvození rovnic pro maximálně věrohodné odhady parametrů také přehled nástrojů využívaných pro testování submodelů, postupy pro výběr množiny regresorů a hodnocení kvality modelu.

Praktická část práce obsahuje detailní popis konkrétního skóringového modelu postaveného na reálných datech o úvěrových klientech české banky. Součástí je i podrobná analýza a interpretace získaných výsledků.

Celkové hodnocení práce

Téma práce. Téma práce považuji za přiměřené, v některých pasážích pro bakalářského studenta spíše náročnější. Zadání bylo zřejmě splněno.

Vlastní příspěvek. Vlastním příspěvkem autora je rozsáhlá praktická část práce obsahující prakticky využitelné výsledky.

Matematická úroveň. Matematická úroveň textu je kolísavá. Jisté nedostatky shledávám zejména v první obecné kapitole, kde některé nepřesné formulace týkající se podmíněných a nepodmíněných rozdělení budí pochybnosti, zda autor daná tvrzení správně chápe.

Práce se zdroji. Použité zdroje jsou řádně uvedeny v seznamu literatury a většinou i v textu používajícím příslušné pasáže. Pouze u podkapitoly 1.1 postrádám bližší specifikaci ohledně zdroje pro text, případně vymezení částí formuovaných samostatně autorem.

Formální úprava. Formální úprava práce je dobrá, vylepšit by se dala po jazykové stránce. Místy obsahuje drobné odchylky od souvislého textu psaného v gramaticky správně postavených větách.

Připomínky a otázky

1. První odstavec na str. 3 a dále: náhodný vektor $\mathbf{X}$ je místy nazýván náhodným vektorem, jindy zase náhodnou veličinou.
2. Opravdu je podmíněný rozptyl veličiny Y vzhledem k náhodnému vektoru $\mathbf{X}$ roven druhé mocnině náhodné veličiny ε, jak se uvádí na druhém řádku vztahu (1.2)?
3. K prvnímu odstavci části 1.1.1.: Můžete podrobněji vysvětlit, v jakém smyslu je $\hat{f}(\mathbf{x})$ náhodná veličina?
4. Vzhledem k jakému rozdělení se uvažuje střední hodnota v (1.5)? Poslední věta odstavce pod vztahem (1.5) se týká spíše výpočtu MSE jako průměru přes určitou množinu pozorovaných dat.

5. V důkazu Věty 2 se zbytečně rozepisuje vztah pro prostřední sčítanec na pravé straně dokazované rovnosti, který je zřejmý vzhledem k tomu, že $f(\mathbf{x})$ je konstanta.

6. U tabulky 1.2 a dalších tabulek uvedených v teoretické části práce by bylo vhodné uvést, že obsahují hodnoty vypočtené na základě datového souboru popsaného až v třetí kapitole.

7. Definice 8, str. 11: Pro ekvivalenci (1.10) a (1.11) by bylo třeba předpokladu nulové střední hodnoty chybového členu. Na následující straně se naopak vlastnosti chybové veličiny odvozují z předpokladu, že (1.10) a (1.11) platí současně.

8. Pokud se nadále chápe $\mathbf{X_i}$ jako náhodný vektor, měly by být pravděpodobnosti a momenty chybové veličiny v (1.14) brány jako podmíněné vzhledem k $\mathbf{X_i}$.

Závěr

Práci doporučuji uznat jako bakalářskou.

RNDr. Lucie Mazurová, Ph.D.
KPMS MFF UK
26.8.2021