

Posudek bakalářské práce

Šárka Brodinová - Diskriminační analýza pro kompoziční data

Předložená práce spojuje dvě oblasti statistiky - diskriminační analýzu a analýzu kompozičních dat - které zřejmě nepatří do základního kurzu pravděpodobnosti a statistiky. Téma je to jistě zajímavé a lze kladně hodnotit, že se do takové práce studentka vůbec pustila. Jako oponent se však spíše než na chválu soustředím na problematická místa předložené práce.

Hlavním nedostatkem předložené práce je značné množství drobných nepřesností ve značení, které místy dosti snižují srozumitelnost textu pro čtenáře, který není s problematikou dopředu obeznámen. Např. str. 10: vzorec (6) ve jmenovateli má být k místo i , str. 12: v nečíslovaném vzorci mezi (12) a (13) vypadlo $\mathbf{x}_0$, ve vzorci (13) vypadlo $\mathbf{S}^{-1}$, str. 13: ve vzorci (14) špatně indexy 1,2, též ve vzorci ve větě 1.2, kde navíc chybí index 0 u $\mathbf{x}$, str. 15: zapletl se nový symbol g místo k , str. 20: chybí logaritmy ve vzorci pro alr-transformaci, str. 21: F má být matice $(D-1) \times D$, ale pak se píše $F = [I_{D-1}, -I_{D-1}]$, kde I_{D-1} je $(D-1)$ rozměrný vektor jedniček, navíc na str. 22 se symbol I_{D-1} používá ve významu $(D-1)$ rozměrné jednotkové matice.

Druhá připomínka se týká aplikační části, v níž postrádám detailnější komentář výsledků provedených analýz. Na co jsme vlastně přišli? Že podle toho, jakou část příjmu člověk vynaloží na oblečení, je možno soudit, jde-li o muže nebo o ženu? Nejsou snad jiné způsoby, jak tuto informaci zjistit? A je-li naším zjištěním, že se struktura výdajů u mužů a u žen významně liší, nebylo možná lépe použít nějakou „kompoziční“ verzi dvouvýběrového testu? Chybí mi diskuse dosažených výsledků, která by se zamýšlela nad otázkami typu „je zjištěná rozdílnost struktury výdajů faktická, nebo jde jen o to, že si muži a ženy uvědomují spíše různé typy vydání?“ např. může dojít k tomu, že muži si spíše uvědomují účty za bydlení, neboť většinou platí účty, ale průměrné vydání na osobu v domácnosti je stejné. Dala by se tato hypotéza nějak ověřit? Nabízí se i pochybnost, zda údaje získané (z části) dotazníkem na Facebooku můžeme považovat za „nezávislá“ pozorování.

K obhajobě mám ještě jednu otázku na autorku. Na straně 6 se píše o použitelnosti diskriminační analýzy v bankovníctví. Znáte nějaký příklad takového použití?

Práci doporučuji k obhajobě a navrhuji hodnotit stupněm C.

Ondřej Vencálek

Drobnější chyby či nepřesnosti:

- str. 4: Špatná formulace druhé věty.
- str. 6: „do jedné z předem známých skupin“ spíše než „do předem známé skupiny“?
- str. 7: Formulace „Nechť soubor objektů první třídy ...“ není srozumitelná.
- str. 7: Má být $R_2 = \Omega - R_1$.
- str. 9: Chybí závorky $f_1(\mathbf{x})$, $f_2(\mathbf{x})$.
- str. 9: Chybně index označující rozměr vektoru $N_p(\dots)$.
- str. 10: Místo velké K má být k .
- str. 10: V definici $\mathbf{S}_i$ má být „S“ vysázeno tučně (je používáno dále).
- str. 10: Co znamená „příslušný α -kvantil“? Když volím $\alpha = 0,05$ a je $p = 2, k = 2$, jaký kvantil použiji?
- str. 12: Symbol $\bar{\mathbf{x}}_1$ by měl být vysázen stejně ve vzorcích (11) a (13).
- str. 14: Ve vzorci pro $f_i(\mathbf{x})$ by mělo být místo Σ spíše Σ_i .
- str. 15: Některé symboly ve vzorci (17) mají být vysázeny tučně.
- str. 16: Jde spíše o odhad pravděpodobnosti chybné klasifikace.
- str. 19: Na posledním řádku chybí závorka.
- str. 21: „všechna data“
- str. 25: r-project.org
- str. 29: Spíše jsme schopni pohlaví „odhadnout“ než „určit“.
- str. 30: První dvě věty kapitoly 4.2 by měly být spojeny v jedinou.
- str. 31: Umístění první čárky v poslední větě je poněkud matoucí.