Oponentský posudek diplomové práce

Tereza Šolcová – Volba dopravního prostředku

Napsat celkové hodnocení předložené diplomové práce Bc. Terezy Šolcové pro mě není jednoduché. Optimista by o práci mohl soudit, že se autorka při jejím sepisování dopustila velkého množství nepřesností a chyb. Pesimista pak usoudí, že pro autorku zůstávají popsané metody záhadnou a podivnou magií, jíž nelze porozumět. Chci být optimistou, ale i kdyby se mi to podařilo, pořád vidím v práci to velké množství chyb. Níže uvedu několik příkladů chyb a nepřesností, které se v práci vyskytují.

Zásadní výtka se týká použití čtyř logistických regresí místo jedné multinomické. Hlavní problém tohoto přístupu spočívá v tom, že se autorka neudrží správné interpretace, jak se ukazuje již na straně 29, kde se píše „... v první logistické regresi budu sestavovat model, který se bude sažit zjistit, s jakou pravděpodobností zvolí respondent způsob dopravy chůzi.“ Ve skutečnosti jde o podmíněnou pravděpodobnost této volby. Tato chyba se opakuje v celé klíčové kapitole 4. Bylo-li cílem práce (jak se uvádí v Závěru na straně 47) „vytvoření modelu, který bude odhadovat pravděpodobnost, s jakou respondent zvolí určitý dopravní prostředek k uskutečnění své cesty“, pak musím konstatovat, že tohoto cíle nebylo dosaženo. Pokud jde o „zjišťování, které socioekonomické aspekty mají vliv na respondentovo rozhodnutí“, toto bylo v práci poměrně úspěšně provedeno.

Připomínky k teoretické části:

str. 11 Spojení „nezávislé a identické pokusy“ je v češtině poněkud neobvyklé. Jde zřejmě o originální překlad anglického „independent, identically distributed“.

str. 15 Tvrzení ve druhé větě o závislosti $\pi(x)$ na x platí jen při $\beta > 0$ (na to je upozorněno až níže na stránce).

str. 15 Popis obrázku 2: y-ová osa znázorňuje zřejmě pravděpodobnost, nikoliv odhad pravděpodobnosti.

str. 16 Věta „Jestliže jsou prvky výběru nezávislé, ... používá se místo funkce $L(\Theta)$ jejího logaritmu ...“ dává poněkud nešťastně dohromady dvě informace - o tvaru věrohodnosti při výběru ze spojitého rozdělení a o použití logaritmické transformace věrohodnosti.

str. 17 Větě „Validaci zavádíme, abychom byli schopni vypočítat chybu, která byla způsobena užitím statistické metody k odhadu výsledků na novém pozorování.“ nerozumím. Působí na mě jako špatný překlad, z nějž se vytráci obsah sdělení.

str. 18 Termín „Gross validace“ je snad jen vtipným překlepem. Nicméně cross-validace se do češtiny překládá jako křížová validace. Věta, že „... budeme užívat první uvedený případ validace, tzv. křížovou validaci.“ je tudíž poněkud matoucí (leave-one-out C.V. je též křížovou validací).

str. 18 $L_{reduced}$ je maximum věrohodnostní funkce, nikoliv věrohodnostní funkce samotná (L_{full} analogicky).

str. 18 Co znamená $T \sim \chi^2_{k_2-k_1}(0, 1-\alpha)$? Tento zápis je podivný. Co v něm dělá α? Nejsou vysvětleny symboly k_1, k_2. Navíc to, co se chtělo napsat, platí jen za platnosti nulové hypotézy, což je třeba uvést.

str. 19 Rozdělení W není předpoklad, je důsledkem asymptotického chování maximálně věrohodných odhadů.

str. 19 Místo „standardní chyba“ se v češtině používá termín „směrodatná odchylka“.

str. 19 Poslední věta na této straně chtěla asi upozornit na asymptotickou ekvivalenci druhé mocniny (!) statistiky W a logaritmu poměru věrohodnosti. Žel, její podoba je nesprávná.

Připomínky k praktické části:

str. 28 Pozorování typu „auto volí nejčastěji rodiny s dětmi“ jsou přinejmenším pochybná. To přece nemálo záleží na zastoupení těchto rodin v dotazníkovém šetření.

str. 28 V popisné části by bylo dobré ukázat histogram pro veličiny věk a vzdálenost. Jelikož tyto prediktory budou vstupovat do modelu („lineárně“), je dobré udělat si představu o jejich rozložení.

str. 29 Pochybuji, že by kritérium minimální MSE bylo smysluplné při hodnocení kvality modelu. To už jsme rovnou místo metody maximální věrohodnosti mohli použít metodu nejmenších čtverců.

str. 30 Zápis MSE je poněkud nezvyklý (y_1 vektor). Proč je navíc *celková* CV-chyba označena jako MSE (běžně *mean* squared error)? Zde nic neprůměrujeme.

str. 30 Overfittingu se dá bránit testováním hypotéz a používáním AIC (či BIC).

str. 32 Autorka konstatuje: „Vlastnictví řidičského průkazu významně ovlivňuje pravděpodobnost volby.“ Zde by se asi mělo říct, že ti, kteří nemají řidičský průkaz, by měli ve sto procentech případů z možností chůze a auto jako řidič volit tu první. (U všech dalších modelů to bude podobné).

str. 44 Nechápu, proč jsou testy poměrem věrohodnosti a kvalita proložení řešeny až dodatečně v kapitolách 4.5 a 4.6 a nejsou součástí kapitol 4.1 až 4.4.

str. 45 Závěrem kapitoly 4.5, v níž je testována platnost podmodelů navržených v předešlých kapitolách, autorka lapidárně konstatuje: „Bohužel jsem zjistila, že ani jeden test nevyšel statisticky významný.“ Unavený čtenář už si jen posteskne „Co dodat?“

str. 47 Nemyslím si, že by u modelu 4 nastala „chyba při agregaci dat“. Ten problém je opravdu zřejmě složitější a tak se povedlo vysvětlit jen menší část variability. To je přirozená věc.

I přes uvedené nedostatky práci doporučuji k obhajobě. Navrhuji hodnotit ji stupněm E.

V Olomouci 6. 1. 2015
Ondřej Vencálek