Posudek vedoucího na diplomovou práci

Autor diplomové práce: Ondřej Lepša

Název diplomové práce: Využití bootstrapu a křížové validace v odhadu predikční chyby regresních modelů

Cílem diplomové práce je posoudit schopnost moderních výpočetních metod bootstrapu a křížové validace odhadovat predikční chybu regresních modelů a vybrat optimální model (z širší skupiny modelů) z hlediska predikčních schopností. Diplomová práce má rozsah 64 stran včetně příloh (avšak do délky nepočítám úvodní stránky obsahující titulní list, abstrakty apod.).

Autor diplomovou práci dělí na několik částí. Nejprve se zabývá popisem klasických přístupů (založených na upraveném indexu determinace či informačních kritériích) k volbě optimálního modelu (z širší skupiny modelů) z hlediska predikčních schopností. Následně popisuje přístupy založené na moderních simulačních metodách křížové validace a bootstrapu. Princip fungování těchto moderních metod demonstruje naprogramováním vlastních skriptů v softwaru R a jejich aplikací na jednoduchý datový soubor. Pro posouzení skutečných vlastností těchto moderních přístupů využívá autor Monte Carlo simulace inspirované článkem dostupným v literatuře. Veškeré potřebné skripty naprogramoval autor v softwaru R (skripty jsou přiloženy jako elektronická příloha k práci). Autor následně vyhodnotil výsledky a vlastnosti moderních přístupů a srovnal je s přístupy klasickými.

Na diplomové práci oceňuji především to, že se autor zabýval relativně náročnější problematikou, načetl pro tyto potřeby relativně rozsáhlé množství textu a literatury a musel naprogramovat veškeré potřebné skripty pro provedení Monte Carlo simulací. V těchto ohledech oceňuji množství práce, které autor odvedl. Měl bych přesto několik komentářů k několika bodům, ve kterých vidím nedostatky práce.

Textu by prospělo striktnější a korektnější rozlišování mezi skutečnou predikční chybou a jejím odhadem. Ač je toto rozlišení v kontextu práce extrémně důležité, není autorem vždy optimálně dodržované. Např. na str. 25 v prvním odstavci by bylo lepší psát „… způsobit, že hodnota *odhadu* predikční chyby …“ a nikoliv jen „… způsobit, že hodnota predikční chyby …“. V některých místech práce se zdá, že autor tyto pojmy používá poněkud volně, někdy mluví o predikční chybě jako o skutečné predikční chybě a někdy jí myslí pouze její odhad, a někdy mluví o odhadu ve chvíli, kdy se odkazuje na skutečnou predikční chybu. To místy vede až k nekorektnímu použití pojmů, viz např. str. 31 nahoře, kde autor výraz ve vzorci 32 nazývá *odhadem* predikční chyby. Je též otázkou, zda by nebylo vhodnější a pro srozumitelnost textu lepší rozlišovat mezi skutečnou predikční chybou a jejím odhadem tak, že skutečná predikční chyba by byla značena bez stříšky, zatímco její odhad se stříškou.

Některá tvrzení či vysvětlení v textu jsou nepřesná nebo je nejasné, co přesně v *daném kontextu* znamenají (viz např. „Takové modely nemusí být nutně na škodu, ale s rostoucím počtem proměnných je interpretace značně ztížená“ na str. 7, nebo „Jejich výhodou je sice, že se mnohem lépe přizpůsobují datům bez ohledu na to, jestli se jedná o data ze současného datového souboru nebo budoucího.“ na str. 32, ale analogicky i na jiných místech práce).

Některé zápisy rovnic jsou chybné, viz např. rovnice 15 na str. 11, kde je položeno rovnítko mezi výběrovou charakteristiku, konkrétně průměrnou čtvercovou chybu, a charakteristiku populační, očekávanou čtvercovou chybu. Chyby jsou např. také v zápise rovnic 10, 24 a 41. Místy není dodržováno konzistentní značení – jednou je ta samá proměnná značena kurzívou, jindy nikoliv (viz např. n vs. *n* na str. 17 v prvním odstavci, nebo H vs. *H* na str. 18, atd.). Nejasná jsou též některá značení (co je např. *M* v rovnici 47?).

Některé pasáže by mohly být napsány srozumitelněji, viz např. pasáž na str. 41. Obdobně, v prvním odstavci sekce 3.7 se mluví o jakýchsi odhadech predikční chyby využívajících tréninkové chyby a je

psáno, že jsou nevychýlené. Význam tohoto slova (nevychýlené) mi však v příslušném kontextu není jasný. I některé další pasáže sekce 3.7 se jeví jako nekorektní či nejasně napsané (viz např. text nad rovnicí 32 nebo text pod Obrázkem 5).

V práci se vyskytují další drobnější nedostatky. Matici **X** by bylo možná vhodnější nazývat jako regresní maticí. Na str. 54 je odkazováno na rovnici 50, která je ale v příloze a s problematikou na str. 54 ve skutečnosti vůbec nesouvisí (autor se chtěl zřejmě odkázat na jinou rovnici, kterou zapomněl do textu ale napsat). Na str. 47 je v popisu kódu funkce *getTrueErr()* tvrzeno, že trueErr[i] odpovídá buď rovnici 44 nebo 46. Ani jedno z toho není zřejmě pravda (v rovnici 46 oproti kódu zřejmě chybí *N*, rovnice 44 taktéž zřejmě neodpovídá uvedenému kódu).

V práci se také vyskytují občasné překlepy a gramatické chyby („… pravděpodobností předpoklady …“ na str. 12, ale i jinde), slovosled či formulace věty odpovídající občas spíše anglickému jazyku (viz např. „Otázku představuje volba správného kritéria, …“ na str. 4, ale i na jiných místech), apod.

Některých dalších nejasností se týkají mé otázky níže.

Celkově lze shrnout, že vytyčené cíle diplomové práce se autorovi podařilo splnit, avšak v některých aspektech (uvedených výše) by v diplomové práci byl prostor pro vylepšení. Oceňuji však znovu to, že se autor zabýval obtížnějším tématem, které se mu podařilo jako celek – i přes výhrady uvedené výše – rozumně zpracovat a obdržet zajímavé a prakticky důležité výsledky, které odpovídají výsledkům dostupným v literatuře. Oceňuji také to, že autor si sám nastudoval relativně velké množství literatury a naučil se programovat v softwaru R (v okamžiku, kdy se psaním práce začínal, byl spíše začátečníkem, co se týče práce v R).

Prosím autora, aby v průběhu obhajoby diplomové práce zodpověděl následující dotazy (některé z nich se týkají dalších nejasností, na které jsem v práci narazil). Otázky řadím dle důležitosti:

- Nebylo by vhodnější pro posouzení kvality různých přístupů uvažovat v rovnici 42 na str. 38 raději průměrnou (průměr je počítán přes simulace) kvadratickou odchylku *odhadu aktuální* optimální dimenze od *skutečné aktuální* optimální dimenze a nikoliv od průměrné hodnoty této skutečné aktuální optimální dimenze, jak to činíte Vy v rovnici 42? Slovem „aktuální“ je myšleno „v rámci jednoho konkrétního vygenerovaného datového souboru“. Jak by to změnilo Vaše výsledky? Obdobná otázka se týká rovnice 45 na str. 40.
- Detailněji (a možná i jasněji než je to ve vlastní diplomové práci) vysvětlete důvod, proč klasická kritéria (upravený index determinace a informační kritéria) nejsou tak vhodná pro výběr optimálního modelu (ve srovnání s výpočetními metodami) a to i přesto, že v sobě obsahují „penalizační člen“.
- Na čem v obecnosti (tj. u obecných dat) závisí to, zda bude z hlediska přesnosti odhadu predikční chyby lepší křížová validace s $K = 5$ nebo $K = 10$? Detailně vysvětlete.
- V situaci $h = 1$ vysvětlete, zda skutečná aktuální optimální dimenze (tj. skutečná optimální dimenze pro konkrétní vygenerovaný soubor) může být nižší či případně vyšší než 3, a proč?
- V kódu funkce *Holdout*() na str. 19 je prováděn odhad modelu bez absolutního členu. Byl to záměr?

Diplomovou práci Ondřeje Lepši navrhuji hodnotit stupněm **velmi dobře**, v případě zdařilé obhajoby bych však souhlasil i se známkou výborně (neboť se jedná o poměrně náročnou práci).

26. ledna 2015,

Mgr. Milan Bašta, Ph.D.
Katedra statistiky a pravděpodobnosti, FIS, VŠE, Praha