FAKULTA INFORMAČNÍCH TECHNOLOGIÍ ČVUT V PRAZE

Posudek oponenta závěrečné práce

Student: Tung Anh Vu
Oponent práce: Mgr. Jan Starý, Ph.D.
Název práce: Metoda pohyblivých vážených nejmenších čtverců v Julia
Obor: Teoretická informatika

Datum vytvoření: 28. 5. 2018

Hodnotící kritérium:	*Způsob hodnocení - následující škálou 1 až 5:*
1. Náročnost a další komentář k zadání	*1=mimořádně náročné zadání,* *2=náročnější zadání,* ***3=průměrně náročné zadání,*** *4=lehčí, ale ještě dostatečně náročné zadání,* *5=nedostatečně náročné zadání*

Popis kritéria:
Podrobněji charakterizujte diplomovou (bakalářskou) práci a její případné návaznosti na předchozí nebo běžící projekty. Dále posuďte, čím je zadání této ZP náročné. (U obtížnější ZP lze dále tolerovat některé nedostatky, které by u ZP standardní obtížnosti tolerovány nebyly; a naopak u jednoduché ZP mohou být zjištěné nedostatky hodnoceny přísněji.)

Komentář:
Jedná se o implementaci zdokumentovaných algoritmů a struktur v jazyce Julia.

Hodnotící kritérium:	*Způsob hodnocení - následující škálou 1 až 4:*
2. Splnění zadání	***1=zadání splněno,*** *2=zadání splněno s menšími výhradami,* *3=zadání splněno s většími výhradami,* *4=zadání nesplněno*

Popis kritéria:
Posuďte, zda předložená ZP splňuje zadání. V komentáři uveďte body zadání, které nebyly zcela splněny, případně rozšíření ZP oproti původnímu zadání. Nebylo-li zadání zcela splněno, pokuste se posoudit závažnost, dopady a případně i příčiny jednotlivých nedostatků.

Komentář:
Práce splňuje všechny body zadání.

Hodnotící kritérium:	*Způsob hodnocení - následující škálou 1 až 4:*
3. Rozsah písemné zprávy	***1=splňuje požadavky,*** *2=splňuje požadavky s menšími výhradami,* *3=splňuje požadavky s většími výhradami,* *4=nesplňuje požadavky*

Popis kritéria:
Zhodnoťte přiměřenost rozsahu předložené ZP vzhledem k obsahu, tj. zda všechny části ZP jsou informačně bohaté a ZP neobsahuje zbytečné části.

Komentář:
Rozsah práce je odpovídající.

Hodnotící kritérium:	*Způsob hodnocení - bodové hodnocení 0 až 100 bodů (známka A až F):*
4. Věcná a logická úroveň práce	*85 (B)*

Popis kritéria:
Posuďte, zda předložená ZP je po věcné stránce v pořádku, případně vyskytují-li se v práci věcné chyby nebo nepřesnosti. Zhodnoťte dále logickou strukturu ZP, návaznosti jednotlivých kapitol a pochopitelnost textu pro čtenáře.

Komentář:
V Úvodu se při srovnání použitých metod trochu zmatečně zdůrazňuje, že metoda vážených čtverců nevyžaduje pravidelné rozložení vzorových dat - to nevyžaduje ani "obyčejná" metoda nejmenších čtverců.

V odhadu chyby na str. 5 je zřejmě překlep: odhadujeme hodnotu f(x_i) výsledné funkce f, nikoli f_i(x_i).
V závorce na začátku strany 6 chybí suma přes indexy j.
Na straně 7 se popíše speciální případ (ve vzdálenosti větší než epsilon považujeme chybu za nulovou), ale nezdůrazní se, že jiný než tento speciální případ zkoušet nebudeme.

Sekce 1.4 popisuje hledání sousedů v daném okolí: v 1.4.1 pomocí cell linked listu, v 1.4.2 pomocí K-d stromů.
Odhad složitosti na str. 9 bere v potaz dimenzi d definičního oboru R^d, kdežto u naivního hledání ne, těžko je pak srovnávat. Množina S(G) v algoritmu 1.3 má zřejmě být S(C).

Podstatná chyba je u popisu K-d stromů (1.4.2): uspořádání vektorů z R^k po složkách není lineární, takže "nebýt menší než B_l" neznamená "být větší než B_l" a podobně zprava. To vzbuzuje pochyby o implementaci, na kterou se pak autor odvolává v 3.6 ohledně numerických nepřesností.

Slíbené vysvětlení aproxmačních funkcí ve třetím příkladě na str. 20 nepřijde, ale lze se domnívat, že jde o konstantní funkce, resp. funkce pracující s vyvolenou složkou R^2.

Na str. 20 dole se tvrdí, že levá strana uvažovné soustavy závisí pouze na předem daných vzorových datech, záleží ale také na bodě, a ve kterém pracujeme, a na chování chybové funkce na příslušném okolí.

Kapitola 3: Experimenty a testování má několik nedostatků. Především poměřujeme implementovanou metodu s metodu B-splines,
o které dosud nebyla vůbec řeč: čtenář neví, s čím se poměřujeme. Za druhé, měříme chybu vůči nějaké předem dané _analytické_funkci_ (sinus apod), zatímco dosud jsme jen prokládali konečnou množinu bodů.
I když takovou množinu tvoří samply předem dané funkce, jsou to zároveň hodnoty nekonečně mnoha jiných funkcí.
Například obrázek pak 3.3 ukazuje absolutní chybu vůči _jedné_ z těchto funkcí.
(Meli bychom zkoumat a měřit spíše konvergenci při postupně se zahušťujícím samplování,
při rostoucím stupni aproximujícího polynomu, a podobně.).
Dále, průměrnou chybu na daném rozsahu měříme pomocí integrálu. Sama numerická integrace je netriviální procedura zatížená netriviální chybou (nevíme, zda není větší než chyba, kterou nyní máme měřit).
Některé z váhových funkcí popsaných v 3.2 nejsou definované v nule,
zůstává tedy zásadní otázka, jak váží chybu přesně v zadaných bodech (tj. ve vzdálenosti nula).

Na straně 25 dole se bez dalších argumentů tvrdí, že přesnost se zhoršila kvůli nerovnoměrném rozdělení dat
(o pár stran dále to naopak znamená zpřesnění). Přitom popsaná metoda pracuje s rovnoměrně i nerovnoměrně rozdělenými daty stejně
(o B-splines není v tomto ohledu řečeno nic). Trochu pošetile se také tvrdí, že vychýlením vzorových dat jich bude v okolí některých bodů méně (jistě, ale jinde jich tedy zase bude více).

Aproximace derivací by zasloužila nějaký úvod: metoda dostává na vstupu jen data,
a neví nic o tom, zda pocházejí z nějaké funkce, z její derivace, či z nějakého diskrétního měření.
Ostatně sama zkoumaná funkce sin(2x)/2 + 10 je derivací (nějaké jiné funkce).
Smysl může mít aproximovat derivaci nějaké předem dané funkce, jejíž _hodnoty_ známe ve vybraných bodech; od čtenáře se očekává, že se to dovtípí sám.

U časových měření (3.3.1) se sice zvětšuje hustota dat, ale současně se zmenšuje parametr epsilon
(za kterým považujeme chybu za nulovou) "aby se vždy používalo stejné množství dat". To jde myslím proti smyslu takového měření: jak se výpočet prodlužuje, když zpracováváme _větší_ množství dat?

Je otázka, proč tabulka časových údajů končí u řádově desítek tisíc samplů a řádově desítek sekund.
Jakou chybou je zatíženo toto měření času? Proč nepokračujeme na miliony bodů a hodiny výpočtů?
(Tam by byla chyba v sekundách zanedbatelná, tady může znamenat podstatné vychýlení).

V sekci 3.4 se chyba měření chyby nastavuje na jednu desetinu,
což má u funkce s hodnotami v intervalu [0,1] dosti malou výpovědní hodnotu.
Výpočet by údajně jinak nedoběhl do deseti minut - proč ho nenecháme běžet den?
Parametr epsilon nastavujeme na 0.625 (proč?), což při rozložení dat po 0.5 znamená,
že v každém bodě uvažujeme ze všech samplů jen dva nejbližší.
("Křivky" B-splines zde mají být zřejmě "plochy".) Oproti předchozím argumentům
se nyní výpočet derivace _zpřesnil_ vychýlením vzorových dat.

V sekci 3.6 (numerické nepřesnosti) se opatrně tvrdí, že různé výsledky při hledání sousedů K-d stromem

a cell linked listem mohou být způsobeny tím, že "se použilo nesprávné řešení problému hledání sousedů".
To ale není numerická chyba, ale spíše implementační chyba v jedné či druhé metodě (nebo v obou).
Fakt, že obě metody označují za sousedy v daném rozsahu různé body je pak vysvětlen "nesprávnou prací s floaty".
To je pro čísla v řádu "0.4" jednak těžko pravda, jednak není jasné, jak jinak bychom měli "pracovat s floaty"
než sčítat je, násobit a porovnávat. Jako magie působí formulace, že "implementace nepodporuje práci v jednorozměrném prostoru"
(co jiného mají floaty reprezentovat, než reálná čísla, tj. prvky R?) a "nemá uzpůsobené provnávní floatů v tomto prostoru"
(jak si má aplikace "uzpůsobit" porovnávání floatů a proč?)

Hodnotící kritérium:	*Způsob hodnocení - bodové hodnocení 0 až 100 bodů (známka A až F):*
5. Formální úroveň práce	*80 (B)*

Popis kritéria:
Posuďte správnost používání formálních zápisů obsažených v práci. Posuďte typografickou a jazykovou stránku ZP, viz Směrnice děkana č. 26/2017, článek 3.

Komentář:
Překlepů není více, než je obvyklé.
Stranu 2 tvoří jen přeteklé řádky, které by se při reformulaci vešly do Úvodu na předhozí straně.
"Třídní proměnné" jazyka Julia by měly zřejmě být "třídové".
Čestina občas pokulhává ("budeme se zabývat 2 řešeními", "na sobě kolmé přímky").

Hodnotící kritérium:	*Způsob hodnocení - bodové hodnocení 0 až 100 bodů (známka A až F):*
6. Práce se zdroji	*80 (B)*

Popis kritéria:
Vyjádřete se k aktivitě studenta při získávání a využívání studijních materiálů k řešení ZP. Charakterizujte výběr studijních pramenů. Posuďte, zda student využil všechny relevantní zdroje nebo zda se pokoušel řešit již vyřešené problémy. Ověřte, zda jsou všechny převzaté prvky řádně odlišeny od vlastních výsledků a úvah, zda nedošlo k porušení citační etiky a zda jsou bibliografické citace úplné a v souladu s citačními zvyklostmi a normami.

Komentář:
Literaturu tvoří odkazy na použté algoritmy a Julia knihovny.
U online zdrojů se z nějakého důvodu zkracují názvy měsíců ("Břez. 2018").

Hodnotící kritérium:	*Způsob hodnocení - bodové hodnocení 0 až 100 bodů (známka A až F):*
7. Hodnocení výsledků, publikační výstupy a ocenění	*89 (B)*

Popis kritéria:
Vyjádřete se k úrovni dosažených hlavních výsledků ZP, např. k úrovni teoretických výsledků, nebo k úrovni a funkčnosti technického nebo programového vytvořeného řešení, apod. Případně také zhodnoťte, zda software nebo zdrojové texty, které nevytvořil sám student, byly v ZP použity v souladu s licenčními podmínkami a autorským právem. Popište případnou publikační činnost a získaná ocenění související s řešením této ZP.

Komentář:
Výsledkem je funkční balíček pro Julia, implementující metodu pohyblivých vážených čtverců.

Hodnotící kritérium:	*Způsob hodnocení - nehodnotí se*
8. Komentář o využitelnosti výsledků	

Popis kritéria:
Uveďte, zda hlavní výsledky ZP rozšiřují již publikované známé výsledky a/nebo přinášející zcela nové poznatky. Uveďte možnosti využití výsledků ZP v praxi.

Komentář:
Vytvořený balíček je použitelný v Julia prostředí,
i když v implementaci hledání susedů očekávm chyby.

Hodnotící kritérium:	*Způsob hodnocení - nehodnotí se*
9. Otázky k obhajobě	

Popis kritéria:
Uveďte případné dotazy, které by měl student zodpovědět při obhajobě ZP před komisí (body oddělte odrážkami).

Otázky:
Nenalezl jste mezi odevzdáním a obhajobou implementační chybu v hledání sousedů
cell-linked-listem nebo K-d stromem?

Hodnotící kritérium:	*Způsob hodnocení - bodové hodnocení 0 až 100 bodů (známka A až F):*
10. Celkové hodnocení	*89 (B)*

Popis kritéria:
Shrňte stránky ZP studenta, které nejvíce ovlivnily Vaše celkové hodnocení. Celkové hodnocení **nemusí** být aritmetickým průměrem či jinou hodnotou vypočtenou z hodnocení v předchozích jednotlivých kritériích 1 až 9.

Text hodnocení:
I přes uvedené nedostatky je práce nadprůměrná.

Podpis oponenta práce: