

OPONENTSKÝ POSUDEK

Název diplomové práce: Additive Regression Models with Regression Splines
Jméno diplomanta: Matúš Maciak

Shrnutí:

Diplomová práce Matúše Maciaka se zabývá metodami odhadování spojité regresní funkce mnoha prediktorů pomocí splinů. Shrnuje metody pro odhadování jednorozměrných regresních funkcí (vyhlazovací spliny, regresní spliny pomocí B-splínové báze), pojednává o vícerozměrných aditivních modelech a rychlosti konvergence aditivních odhadů, zabývá se kritérii výběru modelu a krátce shrnuje některé adaptivní metody. Charakter práce je kompilační.

Práce je pečlivě zpracovaná a pěkně graficky upravená. Po formální stránce nejvíce trpí nedokonalou angličtinou; jazykových a gramatických chyb je v ní velmi mnoho. Metody, o nichž práce pojednává, jsou většinou pěkně vysvětleny. Faktických chyb ve značení a vzorcích je minimum. V práci je používán jeden praktický příklad, který je ale bohužel příliš málorozměrný, takže neilustruje některé důležité aspekty zkoumaných metod („curse of dimensionality“).

K práci Matúše Maciaka bych vyslovil dvě obecné připomínky.

1. Jednotlivé metody nejsou při výkladu dostatečně propojeny; pojednává se o nich odděleně a mnoho aspektů problematiky zůstává skryto. Tak např. vyhlazovací spliny a regresní spliny jsou úzce příbuzné metody, které se liší pouze formulací a přístupem k problému, ale jejich řešení spolu souvisí. Naopak B-splíny je pojem jiného druhu, nejedná se o zvláštní metodu *per se*, ale o jednu z mnoha možných bází pro generování regresních splinů, která je z určitých důvodů výpočetně a interpretačně výhodná. Pořád se ale jedná o regresní spliny. Není pravda, jak je řečeno na str. 13, že regresní spliny nemají vyhlazovací parametr. Jeho roli totiž hraje počet a rozmístění uzlů, což dává regresním splinům daleko větší flexibilitu, než jediný vyhlazovací parametr, který máme k dispozici u vyhlazovacích splinů. Tento rozdíl je do značné míry pomíjen. Problému hledání optimálního počtu a rozmístění uzlů není obecně věnována taková pozornost, jako hledání optimálního vyhlazovacího parametru.
2. Výklad metod zařazených ke konci práce (MARS, Projection Pursuit) je značně neúplný a místy nesrozumitelný. Např. u metody MARS je nejasná definice $B_v^{(n)}(\mathbf{x})$ (9.4) (Co je j ? Kam se ztratily ostatní složky \mathbf{x} ?). V rovnici (9.6) si zase čtenář není jistý, co je $B_v^{(n)}(x_j)$ a $B_k^{(n)}(x_j, x_k)$. Adaptivní aspekt MARSU se ztratil – není řečeno, jak rozhodnout o počtu a konkrétním tvaru zahrnutých bázevých funkcí. V Kapitole 10 (PPR) není promyšleno použití B-splínových bází. Jak zvolit V , jak vybrat uzly? B-splínové funkce jsou definovány na $\langle 0, 1 \rangle$: pak ovšem výrazem (10.1) nemůžeme modelovat záporné funkce a pokud je např. $\max f(\mathbf{x}) \geq 100$, pak musíme mít také $v \geq 100$.

Přes tyto obecné připomínky považuji diplomovou práci Matúše Maciaka za velmi dobrou a doporučuji ji přijmout k obhajobě.

Drobné připomínky:

- Str. 14: existuje i rekursivní vztah pro výpočet B-splinů
- Str. 21: příliš mnoho vykřičníků v textu; str. 22 nesrozumitelné slovo „indexknotsintervals“
- Str. 24: je možné použít interakce tvaru $f_i^*(x_i)f_j^*(x_j)$
- Kapitola 6: tyto výsledky odvodil Charles Stone?
- Kapitola 7: lze pro výběr uzlů použít výsledky kapitoly 8?
- Str. 38: není definováno σ^2 (aspoň ne zde)
- Str. 39: Jaký je vztah mezi krosvalidací a jackknifem? Bylo by možné aplikovat bootstrap?
- Obr. 10.1, 11.2: PPR dává špatné výsledky, ale lze ji jistě vyhladit více či méně, stejně jako všechny ostatní metody
- Str. 52–53: Nelze souhlasit s tvrzením, že „... PPR give(s) much more precise regression surface...“

Jazykové připomínky:

Jen výběr ze stran 1–4: *take* assumptions on; do not *depends*; let \mathbf{X}_i *denotes*; to know what kind of model *do* we want to fit; ...models are *taking in advance*; it is an important objective to *chose*; will be *mention*; generalization *into* more dimensions; The right positions of knots used to fit a model... (chybí konec věty).

Ve zbytku práce je frekvence jazykových chyb o něco nižší, ale stále značná. Na str. 53 mne zaujala vazba *wicked interpretability* (of PPR algorithm)¹.

Mgr. Michal Kulich, PhD.
KPMS MFF UK
9. května 2006



¹*wicked* = nemravný, hnusný, zkažený, zatracený