Posudek oponenta diplomové práce

Název: Robustní regrese a robustní neuronové sítě
Autor: Bc. Patrik Janáček

Shrnutí obsahu práce

Predložená diplomová práca bakalára Patrika Janáčka sa v zásade venuje rôznym robustným postupom bežne známym z literatúry, navrhnutých pre odhadovanie (parametrických aj neparametrických) regresných modelov. Práca je koncepčne rozdelená do troch základných častí: V prvej časti (Kapitoly 1 až 3) autor zhrňuje základné poznatky z oblasti klasickej (parametrickej) lineárnej regresie, popisuje základné teoretické vlastnosti získaných odhadov a uvádza niekoľko možnosti robustifikácie odhadovacích algoritmov. V druhej časti práce (Kapitola 4) autor definuje neuronové siete, vysvetľuje ich architektúru a motivuje ich využitie v oblasti nelinárnej (parametrickej) regresii, pričom popisuje aj konkrétny spôsob trénovania/učenia neuronových sieti a stručne vysvetľuje princíp tzv. "backpropagation" algoritmu. V tretej časti (Kapitola 5) sú neuronové siete aplikované prostredníctvom simulačnej štúdie, kde autor porovnáva kvalitu predikcie v neparametrickom nelineárnom regresnom modeli. Implementácia empirickej časti diplomovej práce je spracovaná prostredníctvom štandardne dostupných knižníc pre programovací jazyk Python.

Celkovo mi príde základná téma práce zaujímavá, dostatočne náročná a určite vhodná pre diplomovú prácu na MFF UK. Na druhej strane si ale myslím, že celkový tématický záber spracovaný a prezentovaný v práci je zbytočne rozsiahlý a vypracovanie žiaľ napriek značnému rozsahu (takmer 70 strán) príliš povrchné. Tomu zodpovedá aj pomerne časté používanie nezavedeného značenia, nedefinovaných, resp. nevysvetlených pojmov, alebo formulácii, ktoré nedávajú príliš dobrý zmysel. V práci trochu chýba aj základná koncepcia, resp. jasne formulovaný cieľ, ktorý by sa autor snažil postupne naplniť. Z abstraktu a úvodu sa čitateľ dozvie, že hlavným cieľom je "*predstaviť niekoľko robustných alternatív klasickej metódy najmenších štvorcov*" (i.e., lineárny parametrický model) a inšpirujúc sa týmito odhadmi neznámych parametrov "*predstaviť robustné neuronové siete a porovnať ich pomocou simulačnej štúdie.*" Z tohto pohľadu sú vyššie spomínané tri časti práce dosť samostatné, logicky či koncepčne nepreviazané a vôbec nie je jasné, čo má byť tým hlavným spoločným spájajúcim prvom (resp. čo je hlavná téma a cieľ práce). V prvej časti sa totiž autor explicitne venuje **parametrickému lineárnemu modelu** a konkrétne pojednáva o teoretických vlastnostiach parametrických odhadov (napr., konzistencia, asymptotická normalita, alebo bod zlyhania) V druhej časti, avšak s použitím inej terminológie, iného značenia, aj iného názvoslovia, autor predstavuje **nelineárny parametrický model**, diskutuje spôsob odhadovania neznámych parametrov, ale akékoľvek teoretické vlastnosti získaných odhadov sú už ignorované. Na záver je práca ukončená empirickou časťou, ktorá ale opäť s predchádzajúcimi dvoma časťami sotva súvisí, keďže autor tentokrát aplikuje **neparametrický a nelineárný model** a namiesto odhadovania neznámych parametrov dokonca rieši (bez akéhokoľvek vysvetlenia prečo) predikčné schopnosti uvažovaných neuronových sieti.

Z môjho pohľadu by si každá z troch spomínaných časti práce zaslúžila samostatné vypracovanie s výrazne väčším autorovým dôrazom pre jednotlivé teoretické, algoritmické, ale aj empirické detaily. Napriek tomu považujem prácu za pomerne kvalitnu a zajímavo napísanú. Práca je kompilačného charakteru z mnohých zdrojov, ktoré musel autor samostatne naštudovať a pochopiť. Ako hlavný prínos možno hodnotiť empirickú časť (Kapitola 5). Formálna úprava práce je na dobrej úrovni, použité zdroje sú citované na konci práce.

Prácu doporučujem štátnicovej komisii uznať ako diplomovú prácu na MFF UK.

Pripomienky

- Práca má síce pomerne značný tématický záber, ale často je to na úkor zbytočnej, až prílišnej stručnosti. Mnoho veci v práci je nevysvetlených, nedefinovaných, niektoré formulácie (matematické i nematematické) sú nejasné, prípadne až nezmyselné/nesprávne. Občas nie je patrične vysvetlené, ako značenie spolu súvisí/nesúvisí. Nižšie prikladám niekoľko explicitných príkladov (pre zrejmú stručnosť pouze z Kapitoly 1, avšak zostávajúce kapitoly sú na tom dosť podobne):
  - formálne by malo byť vysvetlené, čo je $\boldsymbol{w}_K$ a ako súvisí s $\boldsymbol{w} \in \mathbb{W}$ (str.6);
  - není jasné, že funkce f je funkci z $\mathcal{X} \times \Theta$ do $\mathcal{Y}$ (str.7);
  - na str.8 je uvedené, že platí $\mathsf{E}R_{\mathcal{D}(\boldsymbol{\theta})} = \ldots = R(\boldsymbol{\theta})$, pričom ako argument autor uvádza, že "*pozorování jsou nezávislá, stejně rozdělená*". Asi úplnosť a korektnosť by asi bolo vhodné doplniť, že pozorovania sú rovnako rozdelené, ako generický vektor $(Y, \boldsymbol{X}^\top)^\top$;
  - bolo by vhodné doplniť, že $\delta \in \mathbb{N}$ (str.9);
  - není jasné/definvané, čo je to kapacita $d_{\mathcal{F}}$ (str.9);
  - nie je vhodné umiestňovať "footnote" k matematickému znaku (napr. $d_{\mathcal{F}}$ [1] na str.9, $\mathbb{N}^3$ na str.11, prípadne niekoľko ďalších výskytov;
  - Vapnik-Červoněnkisová dimenzia by mala byť minimálne stručné vysvetlená, keď ju autor používa (str.9);
  - není jasné, čo je to rizikový funkcionál (str.10);
  - skratka lasso by mala byť patrične vysvetlená, prípadne by mal byť doplnený odkaz (str.10);
  - asi by bolo vhodné použiť iný index než i vo výraze $i \in \{1, \ldots, k\}$. Index $i = 1, \ldots, n$ je už použitý pre označenie jednotlivých pozorovaní (str.11);
  - malo by byť špecifikované, z akej množiny je hyperparameter $\boldsymbol{\lambda}$ (str.11);
  - ktoré sú to *prípustné hodnoty* pre $\boldsymbol{\lambda}$? (str.11)
  - výraz $f_i(\boldsymbol{\lambda})$ na str.11 není definovaný a v návazností na predchádzajúcu poznámku ani nie je jasné, pre ake $i \in \mathbb{N}$ je myslený;
  - neni jasné, ako výraz $F(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} F_i(\boldsymbol{\theta})$ závisí na trénovacích datach (str.12);
  - z akej množiny sú hodnoty α_k? (str.12)
  - na str.12 je uvedený výraz $m = |f_i|$. Je zvláštne, že ľava strana nezávisí na k, zatiaľ čo pravá strana závisí;
  - neni jasné, čoho rozdelenie je označené ako $P(y, \boldsymbol{x})$ (str.13);
  - ako môžete "*minimalizovať očekávané riziko* $\mathsf{R}(\boldsymbol{\theta})$", keď je obecně toto riziko neznáme? (str.13);

Otázky k obhajobe

- Na str.16 autor uvádza, že daná "*věta platí i za porušení předpokladu shody rozptylů, pokud předpokládáme, že rozptyl je funkcí regresorů, var*$[Y|\boldsymbol{X}] = \sigma^2(\boldsymbol{X})$". Naozaj tento predpoklad pre platnosť danej vety postačuje?
- Ako presne sa myslí poznámka na str.23, ktorá je v práci formulovaná ako **Pozorování 8.**?
- Prečo je "*obecně ťažké říci, co je modelováno*"—viď záver na str.23. Ako (explicitne) vyzerá teoretická charakteristika, ktorú autor pomocou (3.1) resp. (3.2) odhaduje?
- Prečo boli v simulačnej štúdii v Kapitole 5 vzájomne porovnané jednotlivé algoritmy z hľadiska ich predikčných schopnosti, keď Kapitoly 1 až 4 pojednávali o teoretických vlastnostiach (a inferencii) odhadov neznámych parametrov v parametrických regresných modeloch? Prečo je uvažovaná funkcia v simuláciach analyticky definovaná ako neprametrická?

Praha, 21.01.2023

RNDr. Matúš Maciak, Ph.D.
maciak@karlin.mff.cuni.cz