Posudek oponenta diplomové práce

Název: EM algorithm for truncated Gaussian mixtures
Autor: Bc. Adéla Nguyenová

Shrnutí obsahu práce

Predložená diplomová práca popisuje využitie EM algoritmu v úlohe maximalizovania vierohodnosti useknutej zmesi gaussovských rozdelení. V kapitole 1 je predstavený EM algoritmus pre gaussovské zmesi bez useknutia, v hlavnej kapitole 2 je popísaná jeho modifikácia v prípade useknutých rozdelení. Rôzne praktické problémy spojené s implementáciou EM algoritmu sú diskutované v kapitole 3, a nakoniec obšírna kapitola 4 predstavuje radu príkladov a simulačných štúdií.

Celkové hodnocení práce

Téma práce. Téma práce je zaujímavá a dôležitá. Model má mnohé aplikácie v analýze dát.

Vlastní příspěvek. Najdôležitejšou časťou práce je sekcia 2.3, kde autorka nachádza chybu v odbornom článku Lee a Scott (2012, ďalej [**LS**]), a opravuje ju. Ukazuje sa, že [**LS**] v M-kroku EM algoritmu získavajú explicitné, ale nesprávne riešenia vierohodnostných rovníc. Ako je ukázané v sekcii 2.3.1 predloženej práce, takéto riešenia nie je možné vyjadriť explicitne, a M-krok je teda nutné riešiť numericky. Táto vlastná modifikácia EM algoritmu je v kapitole 4 porovnaná s pôvodným algoritmom z článku [**LS**].

Matematická úroveň. Matematická úroveň práce je podpriemerná. Práca obsahuje množstvo — najmä menších — nezrovnalostí, nekonzistencií a preklepov v značení, ale aj chýb. Ako problematická sa mi zdá najmä sekcia 2.1 s odvodením momentov useknutého rozdelenia, kde by sa argumentácia a formálny zápis dali podstatne zlepšiť.

Práce se zdroji. Hlavným zdrojom práce je článok [**LS**]. Ďalšie zdroje sú citované, nie vždy však vhodne a správne. Napríklad, v popise adaptívneho EM algoritmu v sekcii 3.1.3 je uvedený popis metódy bez akéhokoľvek zdôvodnenia, alebo odkazu na literatúru. Alebo, na str. 28 je citovaný Hurvich a Tsai bez odkazu do zoznamu literatúry. Theorem 3 v prílohe nie je z práce odkazovaný. Zoznam použitej literatúry je neúplný a nekonzistentný. Odkazy do kníh by mali obsahovať čísla tvrdení.

Formální úprava. Práca je písaná anglicky, jej jazyková úroveň je zväčša dobrá. Text však obsahuje veľké množstvo preklepov a chýb v značení, ktoré významne sťažujú čítanie. Numerická časť práce v kapitole 4 a v prílohe má až 60 strán, kde veľká časť výsledkov nie je zaujímavá. Vhodnejšia by asi bola selekcia iba tých najdôležitejších príkladov, ktoré ukazujú zásadné rozdiely medzi použitými algoritmami, a ich dôsledná interpretácia.

Připomínky a otázky

Uvítam, ak sa autorka pri obhajobe vyjadrí k nasledujúcim otázkam:

1. Akým spôsobom súvisia odhady z [**LS**] so skutočnými vierohodnostnými rovnicami? Akým spôsobom [**LS**] zjednodušili tieto rovnice tak, aby získali riešenia zo sekcie 2.3.1?
2. V sekcii 2.2 je problém zjednodušený reparametrizáciou z π_k na η_k vo vzťahu (2.7). Parametre η_k však závisia aj na $\boldsymbol{\theta}$. Nie je teda nutné pri diferencovaní vierohodnosti v M-kroku na str. 24 derivovať aj výrazy s η_k podľa $\boldsymbol{\mu}_k$ a $\boldsymbol{\Sigma}_k$?

3. Je možné z výsledkov simulačnej štúdie usúdiť, kedy bude algoritmus z [**LS**] zlyhávať?
4. Ako je možné, že napr. v tabuľke 4.11 vychádza Kullback-Leiblerove skóre záporné?
5. Ovplyvňuje odhalená chyba v [**LS**] aj odvodenia v cenzorovanom prípade v sekciách 4 a 5 článku [**LS**]?

Nasleduje niekoľko ďalších pripomienok k práci:

1. Finálne odvodenie druhej vierohodnostnej rovnice na str. 24 je príliš rýchle. Ako sme vo výraze získali druhý moment useknutého rozdelenia?
2. Celá sekcia 2.1 obsahuje veľké množstvo chýb a nezavedeného značenia. Napríklad, nikde v texte nie je vysvetlené, čo znamená $[\boldsymbol{s}, \boldsymbol{t}]$ alebo $\int_{\boldsymbol{s}}^{\boldsymbol{t}}$ pre $\boldsymbol{s}, \boldsymbol{t} \in \mathbb{R}^d$. Funkcia g definovaná na str. 18 nie je hustota pre $\boldsymbol{y} \in \mathbb{R}^d$ (rovnako celá rada chýbajúcich indikátorov pre $\boldsymbol{x}$ a $\boldsymbol{y}$ v celej kapitole 2). Ak sa píše o gradiente alebo Jakobiáne funkcie viacerých premenných, musí byť upresnené voči akej premennej derivujeme. Čo znamená $\mathbb{E}\,(Y_i, Y_j)$ na str. 20?
3. Celý zdĺhavý výpočet pre druhý moment useknutého rozdelenia v sekcii 2.1 je zbytočný, pretože na str. 20 je výpočet náhle ukončený odkazom do literatúry s finálnym výsledkom.
4. V teste navrhnutom v sekcii 3.1.1 nájdeme iba testovú štatistiku. Ako vyzerá kritický obor tohto testu?
5. Akým spôsobom sa v simuláciách odhaduje KL skóre? Z akého rozdelenia sa v rovnici (4.1) generujú body $\boldsymbol{x}^n$?
6. Str. 13: Čo znamená „linear rate of convergence“?
7. Str. 56: Čo znamená „confidence covariance ellipse“?
8. Nechýba vo formuli (2.1) znamienko mínus v poslednom vzorci pred gradientom?
9. Definition 2 na str. 35 je identická so zdrojom Taboga (2021). Problémom je, že Taboga používa odlišné značenie a terminológiu, ktoré v práci nepôsobia vhodne.
10. Celá sekcia 1.2.1 vrátane Theorem 1 s dôkazom a znenie Theorem 2 sú prevzaté zo skrípt Omelka (2021, Theorem 11 a 12). Je tu však nesprávne odpísaná nerovnosť pre Q vo formuli nasledujúcej (1.11).
11. Vzťah pre Jakobián po (1.12) na str. 13 nie je správne, porovnajte s Omelka (2021, vzorec (85)).
12. Výrazy ako rovnica (1.8) na str. 10 nie sú formálne správne. Na ľavej strane máme parameter $\boldsymbol{\theta}$, na pravej strane jeho odhad.
13. Hustotám ako f často chýbajú argumenty parametrov (napr. str. 10, vzorec (2.2), ale aj inde).
14. V celej práci nachádzame množstvo drobných problémov so značením vektorov pomocou tučného písma; $\boldsymbol{\theta_k}$ nie je to isté ako $\boldsymbol{\theta}_k$ alebo θ_k. Rovnice často nie sú ukončované bodkami. Iba tie rovnice, na ktoré sa ďalej v texte odkazujeme, majú byť číslované.

Závěr

V práci oceňujem najmä poukázanie na dôležitú chybu v odbornej literatúre, a jej opravenie. Celkove, text však trpí radou nedostatkov ktoré sťažujú čítanie a pochopenie. Napriek tomu si myslím, že prácu **je možné uznať** ako diplomovú prácu.

Stanislav Nagy
KPMS MFF UK
4. augusta 2022