Posudek oponenta bakalářské práce

Název: Klasifikace založená na směsových modelech
Autor: Lucie Janečková

Shrnutí obsahu práce

Bakalárska práca študentky Lucie Janečkové sa venuje problému klasifikácie dat na základe parametrického populačného modelu—konečnej zmesi normálnych (t.j., Gaussových) rozdelení. Jedná sa o revidovanú verziu práce, ktorá bola neúspešne obhájovaná v septembri 2022.

Hlavné explicitne uvedené námietky v neprospech obhajoby pôvodnej verzie boli "*pomerne veľké množstvo nedostatkov rôznej závažnosti—t.j., chybné matematické odvodenia, nezavedené alebo nevysvetlené značenie, nelogické a nezmyselné formulácie, ale aj značne nedokončená štvrtá—empirická kapitola s praktickou ilustráciou na reálnych datach*". Explicitne zmieneným nedostatkom práce bol aj fakt, že na rozdiel od stanoveného cieľa práce "*definovať zmesový model viacrozmerných normálnych rozdelení a popísať štandardné metódy odhadovania neznámych parametrov*" autorka viac-menej výhradne pracovala so zmesami jednorozmernými.

Predložená revidovaná verzia pôvodnej bakalárskej práce predstavuje v tomto smere výrazné zlepšenie. Pozitívne hodnotím hlavne snahu autorky preformulovať niektoré tvrdenia a modifikovať použité značenie a niektoré časti textu tak, aby korespondovali s modelom mnohorozmernej zmesy. Autorka taktiež celkom zdarne odstránila väčšinu nedostatkov explicitne vytknutých v oponentskom posudku vzhľadom k prvým dvom kapitolám. Použité značenie je viac-menej konzistentné, formulovaný text je súvislý a zmysluplný a celkovo sa jedná o velmi pekne spracovanú a napísanú časť práce. V prvých dvoch kapitolách sa po dodatočnej revízii objavuje len minimum drobnosti, nejasnosti, či preklepov.

Je preto na škodu veci, že tretiu kapitolu sa autorka rozhodla namiesto dôslednej revízie radšej kompletne prepracovať a na výsledku sa to žiaľ negatívne prejavilo. Tretia kapitola totiž pôsobí opäť dojmom, že by ešte potrebovala dodatočnú detailnú revíziu. Viacmenej väčšinu pripomienok z pôvodného oponentského posudku je možne opäť doslovne zopakovať: značenie, ktoré je nekonzistentné so zbytkom práce, používanie symbolov, ktoré označujú na rôznych miestach rôzne kvantity, nezmyselné matematické formulácie, nerozlišovanie medzi vektorovými, maticovými a jednorozmernými hodnotami, nerozlíšovanie medzi neznámými parametrami a ich odhadmi, alebo opätovné poprehadzované používanie symbolov $\boldsymbol{y}, \boldsymbol{y}_j, \boldsymbol{Y}, \boldsymbol{Y}_j$—z formulovaného textu totiž vôbec nie je jasné, či autorka správne chápe rozdiel medzi náhodnými veličinami (t.j., $\boldsymbol{Y}$, alebo $\boldsymbol{Y}_j$) a ich realizáciami—ako ich autorka explicitne v práci definuje pomocou symbolov $\boldsymbol{y}$ a $\boldsymbol{y}_j$ (viď niektoré konkrétne pripomienky explicitne uvedené nižšie). Z tretej kapitoly je síce evidentná snaha autorky podrobne doplniť a detailne vysvetliť EM algoritmus, čo hodnotím pozitívne, ale jednoznačne by bolo vhodné tejto snahe venovať výrazne viacej času aj potrebnej pozornosti.

Posledná, štvrtá kapitola je opäť výrazným zlepšením oproti predchádzajúcej verzii. Autorka zapracovala väčsšinu pripomienok z oponentského posudku—doplnila základné informácie o použitom datovom súbore, prezentovaná je aj drobná exploratívna analýza (hoci autorka používa iný datový súbor ako v pôvodnej verzii práce a chýba akékoľvek vysvetlenie prečo) a explicitne uvedená je aj úspešnosť použitého klasifikátora (i keď na rozdiel od pôvodnej verzie práce autorka tentokrát vôbec neposkytla zdrovový kód a výsledky klasifikácie vyzerajú byť bez dodatočného vysvetlenia trochu pochybné).

Celkovo považujem prácu za slabú a podpriemernú. Z revidovaných časti práce je na jednej strane zrejmé, že autorka je bez problémov schopná formulovať logický, zmysluplný a aj dobre a súvislo čítateľný matematický text. Na druhej strane novo-doplnené časti práce (Kapitola 3) pôsobia odfláknutým dojmom, akoby boli vypracované narýchlo a bez akéhokoľvek zamýšlania sa nad významom formulovaných viet.

Z celkového pohľadu ale autorka úspešne zapracovala všetky pripomienky explicitne uvedené v oponentskom posudku a v tomto smere odstránila všetky zásadné a podstatné námietky, ktoré zazneli v neprospech obhajoby bakalárskej práce v septembri 2022. Doporučujem preto štátnicovej komisii uznať predloženú (revidovanú) prácu ako bakalársku prácu na MFF UK.

Pripomienky

- Niekoľko drobnosti k prvým dvom kapitolám:
  - vo Vete 1 by asi bolo vhodné formálne definovať maticu $\mathbf{\Gamma}_k$ a tiež explicitne uviesť, že $\boldsymbol{\mu}_k$ je vektor stredných hodnôt príslušnej k-tej zmesy;
  - náhodný vektor $\boldsymbol{Y}_j$ asi nepredstavuje celkovú populáciu, ako je uvedené na str.3;
  - označenie podmienenej hustoty ako f_k (na str.5) je síce formálne ok, ale použitá formulácia nie je vhodná—symbol f_k bol totiž použitý už pri definícii mnohorozmernej zmesy v (1.1); Taktiež nie je jasné, prečo by mal argument hustoty $f(\boldsymbol{y}_j)$ závisieť na indexe j?
  - asi by bolo vhodnejšie napísať, že "*odhadneme pravdepodobnosti* $\tau_k(\boldsymbol{y}_j)$" namiesto uvedeného "*spočítame*";
  - Výraz $\sigma_1^2 \to 0$ použitý v Tvrdení 2 je zavádzajúci. V znení tvrdenia sa objavuje aj $n \in \mathbb{N}$ ako rozsah náhodného výberu a nie je preto jasné, v akom zmysle je uvedenú konvergenciu/limitu potrebné chápať;
  - parciálne derivácie uvedené na strane 9 by bolo vhodné vysvetliť trochu detailnejšie—symbol $\boldsymbol{\psi}$ označuje *sadu* parametrov a preto nie je zrejmé, ako presne dané výrazy chápať. Akú matematickú štruktúru má symbol $\mathbf{0}$ na pravej strane uvedených rovníc?

❑ Hlavné pripomienku ku Kapitole 3:

- Na mnohých miestach je použité značenie nekonzistentné (buď vzhľadom k danej kapitole, prípadne vzľadom ku kapitolám predchádzajúcim): napr., dva rôzne symboly ($\boldsymbol{z}_j$ a $\mathbf{z}_j$) pre tú istú realizáciu náhodného výberu $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n$; označenie jednorozmerného náhodného výberu $Y_1, \ldots, Y_n$ namiesto $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$; pravdepodobnosti $\tau_k(\boldsymbol{y}_j; \boldsymbol{\psi})$ používané v tretej kapitole vs. pravdepodobnosti $\tau_k(\boldsymbol{y}_j)$ definované v prvých dvoch kapitolách; podmienená a marginálna hustota f (str.11) vs. zmesová hustota f v (1.1); symbol 0 namiesto $\mathbf{0}$ (str.13); odhady zamené s príslušnými neznámymi parametrami (str.14); atď.;
- Prečo v argumente združenej hustoty náhodného vektoru $(\boldsymbol{Y}^\top, \boldsymbol{Z}^\top)^\top$ sú uvedené samotné náhodné vektory $\boldsymbol{Y}_j$ a $\boldsymbol{Z}_j$ (str.11)? Pre ktoré $j \in \{1, \ldots, n\}$ teda platí predpis hustoty? Nemal by byť zápis hustoty uvedený všeobecne pre všetky $\boldsymbol{x} \in \mathbb{R}^p$ a $\boldsymbol{z} \in \{0, 1\}^g$? Vzhľadom k akej miere je uvedená hustota definovaná?
- Vzhľadom k akému rozdeleniu je počítaná stredná hodnota $\mathsf{E}_{\boldsymbol{\psi}}$ na str.11?
  Čo konkrétne predstavuje podmienka $\boldsymbol{Y} = \boldsymbol{y}$ (resp. pre aké $\boldsymbol{y} \in \mathbb{R}^p$)?
- Ako rozumieť výrazu, že "*vypočteme hodnotu této funkce*" (str.12)? V ktorom konkrétnom bode hodnotu danej funkcie počítate? Následuje výraz, v ktorom na ľavej strane je uvedená podmienka $\boldsymbol{Y} = \boldsymbol{y}$ (teda pre obecnú hodnotu $\boldsymbol{y} \in \mathbb{R}^p$). Na pravej strane rovnosti sa ale objavujú podmienky $\boldsymbol{Y}_j = \boldsymbol{y}_j$, pre $j = 1, \ldots, n$ (teda pre jednu konkrétnu realizáciu náhodného výberu $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$). Ako tomuto zápisu správne rozumieť?
- Z formulácie na str.12 nie je jasné, ktorá kvantita je vlastne rozdelená na dve časti—rozdeľujete strednú hodnotu $\mathsf{E}_{\boldsymbol{\psi}}[\ell_c(\boldsymbol{\psi})|\boldsymbol{Y} = \boldsymbol{y}]$, alebo argument maxima vzhľadom k sade parametrov $\boldsymbol{\psi}$? Čo presne predstavuje $S(\boldsymbol{\omega}, \boldsymbol{\psi})$ a čo predstavuje $T(\boldsymbol{\xi}, \boldsymbol{\psi})$? A ak je sada parametrov rozdelená na dve disjunktné časti $\boldsymbol{\omega}$ a $\boldsymbol{\xi}$, prečo to následne značenie v $S(\boldsymbol{\omega}, \boldsymbol{\psi})$ a $T(\boldsymbol{\xi}, \boldsymbol{\psi})$ nereflektuje?
- Ako prakticky overiť podmienku na konci str.13? Vzhľadom k zavedenému značeniu sú kvantity $L(\boldsymbol{\psi}^{(m)})$ a $L(\boldsymbol{\psi}^{(m-1)})$ náhodné. Vierohodnosť $L(\cdot)$ formálne zavedená a definovaná na str.6 totíž využíva náhodné vektory $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$, nie ich konkrétne realizácie $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$. Opäť asi došlo k zámene použitých symbolov. Podobných problémov sa ale v Kapitole 3 objavuje omnoho viacej;
- Celkovo je prezentácia textu a jeho formátovanie v Kapitole 3 odfláknuté—text často presahuje cez okraj strany, použité sú nespárované zátvorky a vylepšiť by šlo aj zarovnanie matematických výrazov, formulácia textu a pod..

❑ Poznámky k empirickej časti:

- Prečo na rozdiel od predchádzajúcej verzie práce (datový súbor s celkovo 289 pozorovaniami) používa autorka v revidovanej verzii iný datový súbor so 156 pozorovaniami. Prečo?
- V empirickej časti práce (Kapitola 4), ktorá je oproti pôvodnej verzii výrazne dopracovaná (autorka doplnila popis datového súboru, stručnú exploratívnu analúzu, aj podrobnejšiu diskusiu výsledkov), stále chýbajú napr. relatívne počty zastúpenia jednotlivých kategórii. Podrobnejšie a detailnejšie by mohli byť uvedené aj popisky k obrázkom a tabuľkám (ideálne by každý obrázok a každá tabuľka mali so svojím popiskom tvoriť komplexnú a ucelenú informáciu a malo by byť zrejmé, o čo sa jedná);
- Trochu viacej pozornosti by si zaslúžila aj formálna upráva—napr. autorka používa dva rôzne symboly pre desatinnú čiarku, dva rôzne symboly pre znamienko mínus, prípadne v značení sa objavujú neznáme hodnoty parametrov namiesto ich príslušných empirických odhadov (str.18);
- Aký praktický význam má v exploratívnej analýze uvádzať odhadnutú kovariačnú maticu? Resp. aká presne je interpretácia odhadnutej kovariancie?
- Použitie pojmu *skutočnosť* napr. v Tabuľke 4.3 je trochu zavádzajúce. Autorka totiž stále pracuje len s náhodným výberom z celkovej populácie. Pod pojmom "skutočnosť" sa v štatistike často myslí neznáma vlastnosť vzhľadom k celkovej populácii;
- Keďže k práci nie je priložený zdrojový kód (na rozdiel od predchádzajúcej verzie, kde bol priložený zdrojový kód chybný), nie je možné skontrolovať prezentované výsledky. V každom prípade ale výsledky klasifikácie vyzerajú trochu pochybne—klasifikátor má zrejme tendenciu klasifikovať väčšinu pozorovaní do jedninej populácie—ako vzorky z Austrálie. Výsledných 58 správne zaradených subjektov (z celkových 156 pozorovaní) predstavuje len o niečo viac ako jednu tretinu—teda porovnateľne so zcela náhodným klasifikátorom.

  Z pohľadu na Obrázok 4.3 sa ale zdá, že populácia Austrálie by mohla byť najlepšie odlíšiteľná od zostávajúcich dvoch populácii a skôr by som očakával problém so správnym roztriedením populácii Berg a Burjati. Výsledky klasifikácie ale s týmto pozorovaním nekorespondujú. Má autorka nejaké zmysluplné vysvetlenie pre takéto získané a prezentované výsledky?

Tri Studničky, 29.08.2023

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