OPONENTSKÝ POSUDEK NA BAKALÁŘSKOU PRÁCI

Název: Aplikace EM algoritmu
Autor: Antonín Komora

Shrnutí:

Bakalářská práce Antonína Komory představuje stručný úvod do problematiky odhadování parametrů pomocí EM algoritmu doplněný konkrétními příklady a pojednáním o podmínkách konvergence algoritmu.

Práce je napsaná pečlivě, pěkně a srozumitelně, má dobrou grafickou úpravu a minimum překlepů a pravopisných chyb. Zadané téma bylo předloženou prací bezpochyby naplněno. Zpracované příklady prokazují, že uchazeč tématu porozuměl a studované metody umí aplikovat v praxi.

Práce Antonína Komory bohužel obsahuje i nemalé množství věcných chyb a nejasných či nepřesných formulací. Část z nich je zmíněna v konkrétních připomínkách níže; jako nejvážnější chápu připomínky 1 – 3.

Práci Antonína Komory považuji celkově za průměrnou a doporučuji ji uznat jako práci bakalářskou.

Konkrétní připomínky:

1. Data o tygrech na str. 6 nejsou v souladu s modelem (2.1). Zatímco model implikuje, že počet čínských tygrů by měl být roven zhruba třetině počtu tygrů bengálských, pozorovaná data uvádějí 1700 tygrů bengálských a jen 60 tygrů čínských. Z čeho vůbec model (2.1) vychází a proč tak výrazně odporuje datům?
2. Ve druhém příkladě v kapitole 2.4.1, se v kroku E podmiňuje úplnými daty – to ovšem nedává smysl, poněvadž pak by podmíněné střední hodnoty počítané dále byly triviálně rovny jednotlivým postačujícím statistikám a odhady v kroku M by nešly spočítat.
3. V bodě (iv) na konci str. 15 se mluví o tom, že „jev $[X_i \mid Y_i = y]$ má rozdělení $N(\cdot, \cdot)$“. Co to má být za jev a jak by jev mohl mít normální rozdělení?
4. Abstrakt: „EM algoritmus je velmi cenným nástrojem pro výpočty statistických problémů“. Nešlo by říct přesněji, jaké statistické problémy se EM algoritmem dají řešit?
5. Anglický abstrakt: Formulace vět nepřipomínají angličtinu, jedná se o slovosled přejatý z češtiny.
6. Str. 4: „EM algoritmus ... je iteračním algoritmem, který je jedním ze způsobů výpočtu statistických problémů“. To je stylisticky špatná a nicneříkající věta.
7. Str. 6: „budeme θ chápat jako parametr pro výpočet přesné podoby pravděpodobností, které odpovídají vstupním datům“. Má pravděpodobnost i nepřesnou podobu? Co se míní tím, že pravděpodobnosti mají odpovídat vstupním datům?
8. Str. 8: „Vynásobíme-li tuto rovnici jejími jmenovateli“. Co to je jmenovatel rovnice?
9. Str. 8: „Symbol $\mathbb{E}_a$ *značí* střední hodnotu, která je podmíněná nějakou pevnou hodnotou a“. Co se stane se střední hodnotou, když ji podmíníme konstantou?
10. Str. 8: „jsou ve své podstatě nahrazeny“. Jak, ve své podstatě?
11. Str. 8: „L_c je lineární funkcí našich jediných dvou neznámých x_2 a x_3“. L_c je pouze funkcí x_3, nikoli x_2.
12. Str.10: „v tomto případě až pátého“. Nevidím tam pátý stupeň.
13. Str. 11: „kde ν_i jsou čítací míry takové, že $\nu_2 \times \nu_3$ je definováno ve smyslu $x_2 + x_3 = y_2$ a zbývající ν_i jsou také definována ve smyslu (2.10).“ Jednak tato věta nevysvětluje ν_i v obecné situaci a jednak není jasné, jak definovat míru (nebo součin měr) pomocí součtu jakýchsi proměnných.
14. Str. 13: Z čeho plyne (2.13)?
15. Str. 13: Asi je třeba o funkcích $a(\boldsymbol{\theta})$ a $b(\boldsymbol{x})$ aspoň něco předpokládat.
16. Str. 13: „kde Ω je konvexní parametrický prostor všech $\boldsymbol{\theta}$ takových, že platí...“ Já bych řekl, že Ω je dáno předem, nikoli že je definováno nějakou nerovností. Navíc tady asi nemá být Lebesgueův integrál.
17. Str. 13: „Krok E: vynecháme členy, které neobsahují $\boldsymbol{\theta}$.“ Není řečeno, z čeho se ty členy mají vynechávat.
18. Str. 14: Hustota $\boldsymbol{W}_i$ vzhledem k Lebesgueově míře existuje pouze za určitých podmínek, nikoli vždycky.
19. Str. 14: Parametrický prostor Ω: hodnoty σ_{12} nejsou v celém $\mathbb{R}$, ale mají rozmezí plynoucí z Cauchy-Schwartzovy nerovnosti.
20. Str. 17: Krok M by měl být zřejmý, není třeba jej odvozovat.

21. Str. 19: Není snad třeba odvozovat ML odhad σ_{12}.
22. Str. 20: „Je tedy patrné, že vynecháme-li při výpočtu maximálně věrohodného odhadu neúplná pozorování, dopouštíme se tím velmi závažné chyby.“ Víte, o jaký druh chyby jde? Porušuje se zde nestrannost a/nebo konsistence, nebo je ta chyba v něčem jiném?
23. Str. 22: „Obecně to však předpokládat nelze, čímž bychom se při maximalizaci v kroku M mohli dostat do značných potíží.“ Význam věty vedlejší není jasný. Navíc na to, že MLE nelze vyjádřit explicitně, narazíme už u rozdělení exponenciálního typu (např. gama rozdělení). Je to snad nějaký nepřekonatelný problém?
24. Str. 25: „blížíme-li se středními hodnotami (μ_1, μ_2) k nějaké reálné hodnotě a zároveň rozptyly $(\sigma_{11}, \sigma_{22})$ směřují k 0, pak se hodnota věrohodnostní funkce blíží nekonečnu.“ Tomu nevěřím. To by pak maximálně věrohodný odhad rozptylu normálního rozdělení byl vždycky 0, ne?

doc. Mgr. Michal Kulich, PhD.
KPMS MFF UK
29. srpna 2011