Posudek oponenta diplomové práce

Autor/autorka práce:

Bc. Martin Váňa

Název práce:

Incremental News Clustering

Obsah práce:

Práce se zabývá pravděpodobnostními grafickými modely (z angl. probabilistic graphical models), zejména modelům založených na směsi pravděpodobnostních rozdělení (z angl. mixture models) a jejich bezparametrických verzí (z angl. non-parametric mixture models, nebo také infinite mixture models). Tyto modely jsou využity pro shlukování dokumentů. Dokumenty jsou reprezentovány pomocí metody Latentní sémantické analýzy (z angl. Latent Semantic Analysis - LSA), Latentní Dirichletovy alokace (z angl. Latent Dirichlet Allocation - LDA) a metody paragraph2vec. Diplomant demonstruje účinnost metod pomocí řady evaluačních mír na korpusu v anglickém jazyce.

Téma práce a navržené řešení přesahuje úroveň diplomové práce a spíše se blíží práci rigorózní. Diplomant musel proniknout do teoreticky poměrně náročné oblasti Bayesovské statistiky a její aplikaci pro zpracování přirozeného jazyka (z angl. natural language processing - NLP).

Poznámky a připomínky:

K práci mám řadu připomínek a poznámek, které nejsou ani tak způsobeny pochybením diplomanta, ale spíše náročností tématu diplomové práce.

Podstatnější věci:

- Práce se zaměřuje na Gaussovské směsi (z angl. Gaussian Mixture Models - GMM) a jejich bezparametrické verze schopné odhadnout počet shluků přímo z dat. Správné užití těchto metod předpokládá, že data následují Gaussovo rozdělení (v tomto případě vícerozměrné Gaussovo rozdělení, protože se jedná o vektory reálných čísel reprezentující význam dokumentu). U metod jako jsou word2vec či paragraph2vec je známo, že tento předpoklad přibližně splňují. U LDA tomu tak rozhodně není. Výstupem LDA jsou vektory pravděpodobností z Dirichletova rozdělení. V případě LSA bude silně záviset na vstupní matici (slovo krát dokument). Jelikož je shlukování prováděno na základě Gaussovských směsí, je velmi favorizován právě paragraph2vec, což potvrzují i experimenty v této práci. Pro shlukování ostatních modelů, by bylo vhodné zvolit jiná pravděpodobnostní rozdělení.
- Práci by prospělo provést testy normality dat (jak hodně data odpovídají Gaussovu rozdělení), případně testy i pro další pravděpodobnostní rozdělení, pro ověření předpokladů modelu.
- Použití LDA pro reprezentaci dokumentu a následné shlukování pomocí směsi pravděpodobnostních rozdělení je koncepčně špatně. LDA je samo o sobě hierarchická směs pravděpodobnostních rozdělení používající Dirichletovo rozdělení jako prior. LDA samo o sobě provádí shlukování dokumentů. Používání GMM pro LDA reprezentaci, je ve skutečnosti shlukování následované shlukováním. Mnohem čistší řešení, by bylo rozšířit LDA o jednu skrytou vrstvu, která by odpovídala finálnímu shluku. Uznávám ale, že to opět přesahuje diplomovou práci.

Méně podstatné věci:

- Nekonzistence ve vzorcích: str. 14 – matice nejsou tučně; písmena N a D napříč prací označují různé věci včetně matic, skalárů, prvků v posloupnosti; symbol "-" použitý jako dolní index není definován (např.: X_{-i}).

Západočeská univerzita v Plzni
Fakulta aplikovaných věd
katedra informatiky a výpočetní techniky

- Není řečeno, jaké implementace LSA, LDA a paragprah2vec byly použity.
- Bylo by užitečné otestovat různé nastavení LSA – použití tfidf nebo vzájemné informace (z angl. pointwise mutual information) pro výpočet čísel ve vstupní matici (namísto one-hot vektorů) následované vycentrování ve sloupcích (či řádcích) pro lepší numerickou stabilitu PCA.
- Str. 8,9,10 – definice pojmů prior, likelihood, posterior, atd. by bylo vhodnější přesunout do zvláštní kapitoly na začátek práce a demonstrovat je na obecném případě.
- Str. 15 – v generativním procesu LDA chybí $\phi_k \sim Dir(\beta)$ pro každé téma k.
- Str. 23 a 26 v algoritmu 1 a 2 – není nutná normalizace – již se jedná o pravděpodobnosti
- Str. 25 obrázek 5.4 – CRP slouží jako prior namísto Dirichletova rozdělení. V obrázku by mělo být CRP. Dirichlet to nemůže být, protože má fixní dimenzi.
- V experimentech mi chybí porovnání s obyčejnou (parametrickou Gaussovskou směsí). Samozřejmě, že znalost optimálního počtu shluků je nefér vůči bezparametrickým modelům, ale experimenty by udávaly horní hranici úspěšnosti pro model s CRP jako prior.
- Chybí mi úvod do teorie kolem Dirichletovo procesů, která by pomohla zmírnit skok z parametrických směsí na bezparametrické (respektive skok z Dirichletova rozdělení na CRP).
- Str. 21 – Dirichletovo rozdělení není jediná volba prioru pro váhy jednotlivých pravděpodobnostních rozdělení ve směsi.
- Str. 21 – značení $\theta_k \sim H(\beta) = NIW$ je zavádějící. NIW (z angl. normal-inverse Wishart distribution) jsou ve skutečnosti dvě rozdělení – vícerozměrné Gaussovo rozdělení (conjugate prior samo k sobě při konstantní kovarianční matici) pro střední hodnotu μ a IW (z angl. inverse Wishart distribution) sloužící jako conjugate prior pro kovarianční matici Σ.
- Definice a důsledky použití conjugate prior rozdělení jsou v práci opomenuty, přestože se jedná o velmi zásadní věc v Bayesovské statistice - usnadňuje odvození maximum a-posteriori odhadu parametrů. Posteriorní rozdělení bude stejné jako apriorní, pouze s jinými parametry.

Formální úroveň:

Formální úroveň práce je bezproblémová. Autor dodržuje zažité typografické konvence. Práce je vysázena v systému LaTeX. Autor používá vektorovou grafiku.

Práce s literaturou:

Diplomantova práce s literaturou je na úrovni. Diplomant cituje významné konference a časopisy v dané oblasti. Jediná výtka je drobná nekonzistentnost v referencích. Křestní jména autorů jsou občas uvedena celá, jindy ve zkratce. Za zkratkou je občas tečka, jindy ne.

Splnění zadání:

Práce zcela splňuje zadání, a to nad rámec kritérií běžných pro diplomové práce. Výše uvedené poznámky se vztahují pouze k drobným nedostatkům a neměly by výrazně ovlivňovat hodnocení kvality díla jako takového.

Dotazy k práci:

- Plánujete pokračovat na postgraduální studium? Teorie kolem pravděpodobnostních grafických modelů využitých zejména pro strojové učení bez učitele má velký výzkumný potenciál na poli NLP.

Navrhuji hodnocení známkou **výborně** a práci doporučuji k obhajobě.

V Plzni 31.5.2018

Ing. Tomáš Brychcín, Ph.D.