الگوسازی موضوعات بر پایهٔ روش بیز تغییراتی

وحید حیدری[۱]، سید محمود طاهری*[۲] و سیدمرتضی امینی[۳]
[۱] دانشکدهٔ علوم مهندسی، گروه الگوریتم‌ها و محاسبات، دانشگاه تهران
[۲] دانشکدهٔ علوم مهندسی، دانشگاه تهران
[۳] دانشکدهٔ ریاضی، آمار و علوم کامپیوتر، دانشکدگان علوم، دانشگاه تهران

چکیده

در این مقاله، بر پایهٔ روش بیز تغییراتی نشان می‌دهیم روش تخصیص پنهان دیریکله که یک الگوی احتمالاتی مولّد است و در پردازش زبان‌های طبیعی، متن‌کاوی، کاهش ابعاد، و زیست‌داده‌ورزی کاربرد دارد، نسبت به روش تحلیل معنایی پنهان احتمالاتی در الگوبندی داده‌ها عملکرد بهتری دارد. در این باره، نخست، یک الگوی بیزی را در الگوسازی موضوعات شرح می‌دهیم. آن‌گاه با روش بیز تغییراتی و الگوریتم امیدریاضی- بیشینه‌سازی (EM) شاخص‌های الگو را برآورد می‌کنیم. همچنین، تحلیل جدیدی بر پایهٔ الگوریتم تکرارشوندهٔ بیز تغییراتی (IVB) ارائه می‌دهیم و با استفاده از آن شاخص‌های الگو را برآورد می‌کنیم. سپس، الگوریتم EM- تغییراتی را، بر پایهٔ یک مجموعه‌دادهٔ نوشتاری از داده‌های واقعی در زمینهٔ تحلیل داده‌های خبری، پیاده‌سازی می‌کنیم. به‌علاوه، الگوبندی زبانی به‌دست‌آمده را بر اساس ملاک سرگشتگی بررسی می‌کنیم، و دقت خوشه‌بندی موضوعات و کاربرد کاهش ابعاد داده‌های حجیم را با کمک ماشین بردار پشتیبان می‌سنجیم. همچنین، در مقایسه‌ای دیگر، کاربرد الگوریتم پیشنهادی را در پالایش همکارانه بررسی می‌کنیم.

واژگان کلیدی: روش بیز تغییراتی، تخصیص پنهان دیریکله، الگوریتم امیدریاضی-بیشینه‌سازی، یادگیری ماشین، پردازش زبان‌های طبیعی

Topic Modeling Based on Variational Bayes Method

Vahid Heidari, Seyed Mahmoud Taheri* And Seyed Morteza Amini

Abstract

The Latent Dirichlet Allocation (LDA) model is a generative model with several applications in natural language processing, text mining, dimension reduction, and bioinformatics. It is a powerful technique in topic modeling in text mining, which is a data mining method to categorize documents by their topics.

Basic methods for topic modeling, including TF-IDF, unigram, and mixture of unigrams successfully deployed in modern search engines. Although these methods have some useful benefits, they do not provide much summarization and reduction. To overcome these shortcomings, the latent semantic analysis (LSA) has been proposed, which uses singular value decomposition (SVD) of word-document matrix to compress big collection of text corpora. User's search key words can be queried by making a pseudo-document vector. The next improvement step in topic modeling was probabilistic latent semantic analysis (PLSA), which has a close relation to LSA and matrix decomposition with SVD. By introducing the concept of exchangeability for the words in documents, the topic modeling has been proceeded beyond PLSA and leads to LDA model.

We consider a corpus $D = (W_1, \cdots, W_M)$ contains M documents, each document $W_d = (w_1, \cdots, w_N)$ has N words, and each word is an indicator from one of $\{1, 2, \cdots, V\}$ vocabularies. We defined a generative model for generation of each document as follows. For each document draw its topic θ from $\mathcal{D}ir(\alpha)$ and repeatedly for each $n = 1, \ldots, N$ draw topic of each word z_n from $\mathcal{M}ult(\theta)$ and draw each

* نویسندهٔ عهده‌دار مکاتبات
* Corresponding author

• تاریخ ارسال مقاله: ۱۴۰۰/۴/۲۰ • تاریخ پذیرش: ۱۴۰۱/۲/۲۱ • تاریخ انتشار: ۱۴۰۲/۷/۳۰ • نوع مطالعه: پژوهشی

word from the probability matrix of ϕ with probability of $P(w_n|z_n,\phi)$. We can repeat this procedure to generate whole documents of corpus. We want to find corpus related parameters α and ϕ as well as latent variables Z and θ for each document. Unfortunately, the posterior $P(W|\alpha,\phi)$ is intractable, and we have to choose an approximation scheme.

In this paper, we utilize LDA for collection of discrete text corpora. We describe procedures for inference and parameter estimation. Since computing posterior distribution of hidden variables given a document is intractable to compute in general, we use approximate inference algorithm called variational Bayes method. The basic idea of variational Bayes is to consider a family of adjustable lower bound on the posterior and find the tightest possible one. To estimate the optimal hyper-parameters in the model, we use the empirical Bayes method, as well as a specialized Expectation-Maximization (EM) algorithm called the variational-EM algorithm. Also, we propose a new perspective of this problem with Iterative-Variational-Bayes (IVB) method.

We report the results of document modeling, text classification, and collaborative filtering. The topic modeling of LDA and PLSA models are compared on a Persian news data set. It is observed that LDA has perplexity between $9.1e2$ and $1.67e3$, while the PLSA has perplexity between $9.16e4$ and $2.27e5$, which shows domination of LDA over PLSA.

We apply the LDA model in dimension reduction for a document classification problem, along with the support vector machines (SVM) classification method. Two competitor models are compared, first trained on a low-dimensional representation provided by LDA and the second trained on all documents of corpus, with accuracies 94.3% and 97.61%, respectively, this means we lose accuracy but it remains in reasonable range when LDA model is used for dimensionality reduction.

Finally, we use the LDA and PLSA methods along with the collaborative filtering for MovieLens 1m data set. We observed that the predictive-perplexity of LDA changes from $2.9e5$ to $1.05e5$ while it changes from $8.90e7$ to $2.37e8$ for PLSA, again showing the domination of the LDA method.

Keywords: **Variational Bayes method, Latent Dirichlet allocation, Expectation-Maximization algorithm, Machine learning, Natural language processing**

۱- مقدمه

بررسی حجم بالای داده‌های متنی دیجیتال که امروزه در شبکه‌های اجتماعی، اینترنت و بسیاری پایگاه‌های داده ذخیره شده‌است، نیازمند روش‌هایی خودکار برای یافتن الگوهای درون داده‌هاست تا بتوان از آنها برای پیش‌بینی داده‌های آینده استفاده کرد، یا انواع دیگر تصمیم‌گیری را تحت شرایط نایقینی انجام داد. یادگیری ماشین، مجموعه روش‌هایی را برای دستیابی به این اهداف برای پژوهش‌گر فراهم می‌کند [۲۴].

یکی از مسائل مطرح در متن‌کاوی، الگوسازی موضوعات[1] است. الگوسازی موضوعات شامل الگوبندی گردایه‌ای از داده‌های گسستهٔ متنی و خوشه‌بندی متون یا مستندات بر اساس موضوع آنها است [۱۸]. از آغاز پژوهش در بازیابی اطلاعات در زمینهٔ الگوسازی موضوعات، پیشرفت‌های چشم‌گیری حاصل شده‌است. روش‌های پایه‌ای مانند TF-IDF[2]، یک-گرم[3] و آمیختهٔ یک-گرم‌ها با موفقیت در موتورهای جستجوی پیشرفته اینترنتی به‌کارگرفته‌شده‌است. آنها اگرچه مزایای متنوعی دارند، میزان کاهش ابعاد کمی فراهم می‌کنند و تحلیل ساختاری زیادی ارائه نمی‌دهند [۷].

برای حل این مشکلات روش تحلیل معنایی پنهان[4] (LSA) برای بازیابی اطلاعات ارائه شده که از تجزیهٔ مقدارتکین[5] (SVD) ماتریس مستندات-واژگان استفاده می‌کند و به میزان زیادی گردایه‌های بزرگ را فشرده‌سازی می‌کند. در این روش می‌توان ماتریس اولیه را با ترکیب خطی عوامل تجزیه‌شده تقریب زد و با ایجاد بردار شبه-مستند، پرسوجوهای کاربر را اجرا کرد [۸].

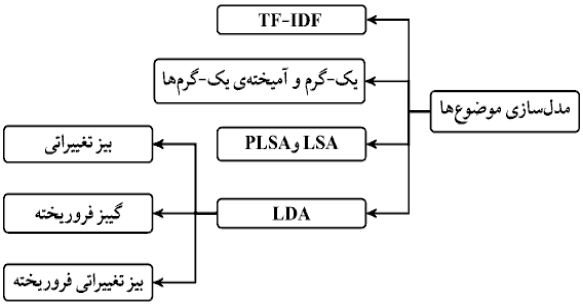
قدم بزرگ بعدی، ارائهٔ الگوی تحلیل معنایی پنهان احتمالاتی[6] (PLSA) است که نخستین‌بار در سه مقالهٔ [۱۳، ۱۴، ۱۵] مطرح شد و در آن‌ها نشان داده‌شد که PLSA ارتباط نزدیکی با LSA و کاهش ابعاد ماتریس با SVD دارد. با ارائهٔ مفهوم تبادل‌پذیری[7] الگوسازی موضوعات از PLSA فراتر رفت و به الگوی تخصیص پنهان دیریکله رسید [۷] که در این مقاله قصد داریم آن را معرفی و از آن استفاده کنیم.

در بین روش‌های یادگیری ماشین برای الگوسازی موضوعات، به روش‌های احتمالاتی به دلیل توانایی آنها در زمینهٔ متن‌کاوی، مانند بازیابی، خلاصه‌سازی، دسته‌بندی و خوشه‌بندی، در سال‌های اخیر توجه شده‌است [۲۱]. اما طراحی الگوریتم‌هایی که از روش‌های احتمالاتی با رویکرد

[1] Topic modeling
[2] Term frequency-inverse document frequency
[3] Unigram
[4] Latent semantic analysis
[5] Singular value decomposition
[6] Probability LSA
[7] Exchangeability

تمام‌بیزی یا رویکردهای مبتنی‌بر درست‌نمایی استفاده می‌کنند، اغلب به نتایجی می‌رسند که رام‌نشدنی[1] هستند. از این رو برای حل چنین مسائلی از یادگیری ماشین از روش‌های تقریبی استفاده می‌شود [۲۶].

یکی از روش‌های تقریبی که با آن می‌توان الگوریتم‌های رام‌شدنی (از نظر محاسباتی) طراحی کرد، بیز تغییراتی[2] است. از کاربردهای این روش می‌توان کاهش ابعاد، تحلیل عوامل در تصویربرداری پزشکی، پالایش برخط داده‌های دورافتاده و دنبال‌کردن فرایند غیرمانا را نام برد [۲۶]. همچنین، روش‌های تقریبی دیگر مانند زنجیر مارکوف مونت کارلو[3] و نمونه‌برداری گیبز می‌تواند استفاده شود؛ که نیاز به توزیع‌های تمام‌شرطی شاخص‌ها برای تقریب‌زدن توزیع پسینی دارد. دربارهٔ روش تخصیص پنهان دیریکله[4] (LDA) می‌توان ساده‌سازی‌هایی انجام داد تا نمونه‌برداری با سرعت بیشتری به هم‌گرایی برسد. روش گیبز فروریخته[5] باتوجه‌به این که توزیع‌های پیشینی مزدوج برای شاخص‌ها درنظرگرفته‌می‌شود، شاخص‌ها را با انتگرال‌گیری حذف می‌کند و نمونه‌بردار گیبز ساده‌تری برای متغیرهای پنهان می‌سازد. پیشنهاد گیبز فروریخته را می‌توان با بیز تغییراتی ترکیب کرد و به روش‌های بیز تغییراتی فروریخته دست یافت [۱۰]. در (شکل- ۱ برخی روش‌های رایج در الگوسازی موضوعات نمایش داده شده‌است.



(شکل- ۱) روش‌های رایج در الگوسازی موضوعات
(Figure-1): Common methods in topic modeling

در سال‌های اخیر پژوهش‌های بسیاری بر اساس رویکرد LDA انجام شده‌است. به‌ویژه در زمینهٔ یادگیری ماشین، داده‌کاوی و متن‌کاوی، الگوریتم‌های بسیاری مبتنی‌بر این روش ارائه و در زمینه‌های مختلفی به‌کارگرفته‌شده‌است. از کاربردهای LDA می‌توان کشف موضوعات در متون علمی، تحلیل رمزهای منبع در مهندسی نرم‌افزار، دسته‌بندی و برچسب‌گذاری تصاویر، بازیابی و پردازش اطلاعات صوتی و موسیقی، تحلیل‌های علوم سیاسی، تحلیل داده‌های جغرافیایی و مکانی، و پیش‌بینی وقوع جرم را نام برد[۱۵]. از کاربردهای الگوی مشابه LDA در زیست‌داده‌ورزی نیز می‌توان تعیین ساختار جمعیت از نمونه‌های ژنتیکی را ذکر کرد [۲۵].

با ظهور شبکه‌های اجتماعی، به‌ویژه سکوهای ریزبلاگ[6] مانند توییتر، پژوهش‌ها در زمینهٔ سنجش افکار عمومی شکل تازه‌ای گرفته‌است. برای مثال MF-LDA می‌تواند سوگیری در تمایلات جمعی را دنبال کند و موضوعات داغ موردتوجه عموم مردم را در بازه‌های زمانی مشخص پیدا کند [۹]. تحلیل نظرات [۲۰]، الگوبندی نویسنده- موضوع [۱۹] و توصیه‌گرهای هشتگ [۱۱] نیز از دیگر زمینه‌های مورد علاقه‌ی پژوهشگران در ریزبلاگ‌ها هستند.

از پژوهش‌های انجام‌شده در پردازش زبان فارسی می‌توان [۴] را ذکر کرد که از الگویی احتمالاتی برای تجزیه‌گر قواعد گرامری برای رفع ابهام و ترجمهٔ ماشینی افعال چندکلمه‌ای انگلیسی به فارسی استفاده می‌کند. همچنین، استخراج گرامر با خوشه‌بندی ناپارامتری بر مبنای الگوی بیزی و فرایند دیریکله [۵]، استخراج فعل بر پایهٔ الگوریتم امید ریاضی- بیشینه‌سازی [۲]، و نیز در زمینهٔ نظرکاوی حس‌نگار و شبکهٔ واژگان حسی فارسی فردوس‌نت برای استخراج بار احساسی نظرات کاربران [۱] را می‌توان نام برد. از موارد استفادهٔ LDA نیز می‌توان به [۶] اشاره کرد که دربارهٔ تحلیل معنایی واژگان و موضوع اشعار کهن فارسی انجام شده‌است، همچنین، در [۳] الگویی برای ابهام‌زدایی از واژگان فارسی که دارای چندین معنی هستند، ارائه شده است.

ادامهٔ ساختار مقالهٔ حاضر به شرح زیر است:

در بخش ۲ تخصیص پنهان دیریکله را معرفی می‌کنیم. بخش ۳ به برآورد شاخص‌های الگو بر اساس روش بیز تغییراتی و الگوریتم امیدریاضی- بیشینه‌سازی تغییراتی می‌پردازد. در بخش ۴ الگوریتم تکرارشوندهٔ بیز تغییراتی شرح داده‌می‌شود. در بخش ۵ از روش ارائه‌شده برای تحلیل داده‌های متنی واقعی در چند کاربرد مختلف استفاده می‌کنیم. در پایان، در بخش ۶ بحث و نتیجه‌گیری ارائه می‌شود. شرح تفضیلی روش‌ها در قسمت پیوست‌ها آورده شده است.

[1] Intractable
[2] Variational Bayes
[3] Markov chain Monte Carlo
[4] Latent Dirichlet Allocation
[5] Collapsed
[6] Microblog

۲- الگوی تخصیص پنهان دیریکله

به‌منظور برآورد شاخص‌های الگو با الگوریتم EM، به الگوی احتمال توأم مشاهدات و متغیرهای پنهان به شرط شاخص‌های الگو و نیز به احتمال حاشیه‌ای مشاهده‌ها به شرط شاخص‌های الگو نیاز داریم. در [۷] یک الگوی احتمالاتی مولد برای به‌دست‌آوردن این دو احتمال به نام الگوی تخصیص پنهان دیریکله (LDA) ارائه می‌شود. در ادامه راهکار ارائه‌شده در [۷] را شرح می‌دهیم.

الگوی LDA در کاربردهای مختلفی استفاده می‌شود، ولی ازآنجاکه LDA نخستین بار با هدف الگوبندی متن استفاده شده‌است، اصطلاحات مربوط به آن نیز مرتبط با متن تعریف می‌شوند. نخست، این اصطلاحات را تعریف می‌کنیم:

- به مجموعه‌ای گسسته از تمام واژه‌های مشاهده‌شده در پیکره «واژه‌نامه» گفته‌می‌شود.
- یک «واژه» داده‌ای گسسته و یکی از اقلام واژه‌نامه است که با یک نشانگر از مجموعهٔ گسستهٔ $\{1,2,\cdots,V\}$ تعریف می‌شود. هر واژه را با برداری یکانی نشان می‌دهیم. واژهٔ $w=v$ برداری است که مؤلفهٔ v-ام آن برابر ۱ است و سایر مؤلفه‌ها برابر صفر هستند. نشانگر $I[\![w=v]\!]$ زمانی ۱ می‌شود که مؤلفهٔ v-ام واژهٔ w برابر ۱ باشد و درغیراین‌صورت صفر می‌شود.
- یک «مستند» دنباله‌ای از N واژه است که با $W=(w_1,w_2,\cdots,w_N)$ نشان می‌دهیم و در آن w_n واژهٔ n-ام از دنباله است.
- هر «پیکره[1]» گردایه‌ای از M مستند است که با $D=\{W_1,W_2,\cdots,W_M\}$ نشان داده می‌شود.

الگوی LDA فرایند مولد الگوریتم- را برای هر مستند W در پیکره D درنظرمی‌گیرد، که در آن z_n یک نشانگر از موضوعات $\{1,2,\cdots,K\}$ است و $I[\![z_n=k]\!]$ برابر ۱ می‌شود، اگر واژهٔ n-ام، موضوعی برابر k داشته‌باشد و درغیراین‌صورت صفر می‌شود. نخستین الگوریتم این فرایند به شرح زیر است:

الگوریتم- ۱: فرایند مولد LDA

۱. **N را به تصادف از توزیع $N\sim \mathcal{P}oisson(\zeta)$ تولید کنید.**
۲. **بردار $(\theta_1,\cdots,\theta_K)$ را به تصادف از توزیع $(\theta_1,\cdots,\theta_K)\sim \mathcal{D}ir(\alpha_1,\cdots,\alpha_K)$ تولید کنید.**
۳. **برای Nواژهٔ w_n گام‌های زیر را انجام دهید:**
۴. **یک موضوع از $z_n\sim \mathcal{M}ult(\theta)$ انتخاب کنید.**
۵. **یک واژه از ماتریس احتمال $P(w_n|z_n,\phi)$ انتخاب کنید.**

در این الگوریتم توزیع $\mathcal{D}ir(\alpha_1,\cdots,\alpha_K)$، توزیع دیریکله نام دارد. بردار تصادفی $\theta=(\theta_1,\cdots,\theta_K)$ با توزیع دیریکلهٔ K-بُعدی می‌تواند مقادیری روی $(K-1)$-سادک[2] بگیرد (بردار K مؤلفه‌ای θ روی $(K-1)$-سادک قرار دارد، اگر $\sum_{j=1}^{K}\theta_j=1,\theta_j\geq 0$)، و دارای چگالی احتمال زیر است:

$$P(\theta|\alpha)=\frac{\Gamma(\alpha_0)}{\prod_{j=1}^{K}\Gamma(\alpha_j)}\prod_{j=1}^{K}\theta_j^{\alpha_j-1}, \qquad (۱)$$

که در آن α یک بردار K مؤلفه‌ای با شرط $\alpha_j>0$، و $\Gamma(x)$ تابع گاما است، و $\alpha_0=\sum_{j=1}^{K}\alpha_j$. تابع چگالی (۱)، یک توزیع پیشینی مزدوج برای توزیع چندجمله‌ای است. بر اساس الگوی ارائه‌شدهٔ این الگوریتم، این احتمال را که واژهٔ n-ام از موضوع k-ام نمونه‌برداری شده‌باشد، با $z_n\sim \mathcal{M}ult(\theta)$ نشان دادیم که از رابطهٔ $P(z_n=k|\theta)=\theta_k$، و احتمال این‌که واژهٔ n-ام برابر v باشد، به شرطی که از موضوع k-ام انتخاب شده‌باشد، از رابطهٔ $P(w_n=v|z_n=k,\phi)=\phi_{kv}$ محاسبه می‌شوند.

با داشتن شاخص‌های α و ϕ می‌توان احتمال توأم آمیختهٔ موضوعات θ، روی مجموعه N موضوع z و مجموعه N واژهٔ مستند W را به‌صورت زیر نوشت:

$$P(\theta,Z,W|\alpha,\phi)=P(\theta|\alpha)\prod_{n=1}^{N}P(z_n|\theta)\,P(w_n|z_n,\phi).$$

با انتگرال‌گیری روی θ و جمع روی Z می‌توانیم توزیع حاشیه‌ای هر مستند W را این گونه به‌دست‌آوریم:

$$P(W|\alpha,\phi)=\int_{\Theta}\sum_{Z}P(\theta|\alpha)\times\left(\prod_{n=1}^{N}P(z_n|\theta)P(w_n|z_n,\phi)\right)d\theta. \qquad (۲)$$

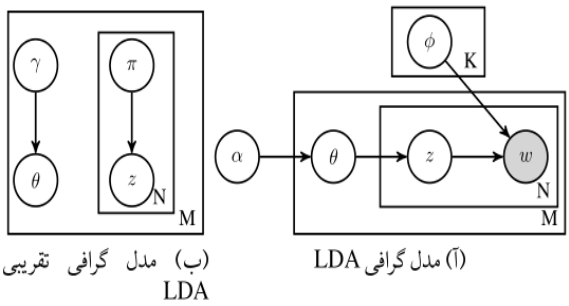
سرانجام، با ضرب احتمالات حاشیه‌ای تک‌تک مستندات می‌توانیم احتمال کل پیکرهٔ اسناد (D) را به‌دست‌آوریم:

$$P(D|\alpha,\phi)=\prod_{d=1}^{M}\int_{\Theta}\sum_{Z}P(\theta_d|\alpha)\times\left(\prod_{n=1}^{N_d}P(z_{dn}|\theta_d)P(w_{dn}|z_{dn},\phi)\right)d\theta. \qquad (۳)$$

[1] Corpus

[2] Simplex

الگوی گرافی LDA در شکل (۲) نشان داده شده‌است. این شکل مشخص می‌کند که LDA سه‌سطحی است. شاخص‌های α و ϕ در سطح پیکرهٔ اسناد هستند و فرض می‌شود در زمان ساختن پیکرهٔ اسناد فقط یک بار نمونه‌برداری می‌شوند. متغیرهای θ_d در سطح مستند هستند و به‌ازای هر مستند یک بار نمونه‌برداری می‌شوند. سرانجام متغیرهای z_{dn} و w_{dn} در سطح واژه هستند و به‌ازای هر واژهٔ درون یک مستند، نمونه‌برداری می‌شوند.



(ب) مدل گرافی تقریبی LDA (آ) مدل گرافی LDA

(شکل-۲): الگوی گرافی LDA به‌همراه مدل ساده‌شدهٔ آن [۱۰].

(Figure-2) Graphical model of LDA, and its approximated model

در [۷] برای الگوی LDA چندین ساده‌سازی و نکته به شرح زیر مطرح شده‌است:

- ابعاد توزیع دیریکله ازپیش‌معلوم، ثابت و مقدار آن برابر K است.
- احتمال شرطی واژگان در هر موضوع با یک ماتریس $K \times V$ الگوبندی می‌شود که آن را ϕ می‌نامیم. ماتریس شاخص‌های ϕ نخست ثابت فرض می‌شود و باید برآورد شود.
- توزیع پواسون یک فرض اساسی نیست و از هر توزیع احتمالی که طول مستند را به واقعیت نزدیک کند، می‌توان بهره گرفت. چون N به هیچ‌یک از متغیرهای مولد مستند (θ و z) وابسته نیست، یک متغیر کمکی[1] محسوب می‌شود و اغلب تصادفی‌بودن آن نادیده گرفته‌می‌شود.

۲-۱- الگوی تحلیل معنایی پنهان احتمالاتی (PLSA)

ازآنجاکه در ادامه الگوی تخصیص پنهان دیریکله (LDA) با الگوی تحلیل معنایی پنهان احتمالاتی (PLSA) مقایسه می‌شود، آن را کوتاه معرفی می‌کنیم.

الگوی PLSA یک الگوی نیمه-مولد است [۷] و فرایند تولید پیکره برای آن در [۱۵] به این صورت پیشنهاد شده‌است که نخست، با احتمال $P(d_i)$ یک مستند، سپس، برای آن مستند یک موضوع با احتمال $P(z_k|d_i)$، و پس از آن، یک واژه با احتمال $P(w_j|z_k)$ انتخاب می‌شود و این فرایند تا تولید کامل پیکره ادامه پیدا می‌کند. با توجه به فرایند تولید پیکره، احتمال توأم مشاهدهٔ (d_i, w_j) به صورت زیر نوشته می‌شود:

$$P(d_i, w_j) = P(d_i)P(w_j|d_i) = \sum_{k=1}^{K} P(w_j|z_k)\,P(z_k|d_i)$$

که در آن $P(w_j|z_k) = \phi_{z_k w_j}$، بردار ϕ_k توزیع احتمال واژگان در موضوع k، $P(z_k|d_i) = \theta_{d_i z_k}$ و بردار θ_{d_i} توزیع احتمال موضوعات برای مستند d_i است. یک الگوریتم EM برای برآورد شاخص‌ها و متغیرهای پنهان در [۱۵] برای PLSA پیشنهاد شده‌است.

۳- استنباط و برآورد شاخص‌های الگو

در رابطهٔ (۳) مشاهده می‌شود که عملگرهای جمع و انتگرال با لگاریتم طبیعی گرفتن از توزیع حاشیه‌ای، داخل لگاریتم قرار می‌گیرند، در نتیجهٔ بهینه‌سازی آن دشوار است و به یک جواب بسته نمی‌رسیم. نویسندگان [۷] بیان می‌کنند این تابع به‌خاطر جفت‌شدگی[2] بین θ و ϕ در مجموع‌یابی روی متغیرهای پنهان موضوعات رام‌نشدنی است و محاسبهٔ دقیق شاخص‌های آن ممکن نیست. اما، برای رسیدن به جواب‌های تا حدی قابل‌قبول، می‌توان روش‌های تقریبی مانند تقریب لاپلاس، تقریب تغییراتی و زنجیر مارکوف مونت کارلو را به‌کار برد. در ادامه، استنباط بیز تغییراتی را برای LDA شرح می‌دهیم.

۳-۱- برآورد شاخص‌های رویکرد LDA از طریق استنباط تغییراتی

پیشنهاد اصلی روش بیز تغییراتی برای تقریب‌زدن $\ln P(D|\alpha, \phi)$، استفاده از نابرابری جنسن[3] برای به‌دست‌آوردن کران پایین لگاریتم درست‌نمایی و سپس، بیشینه‌سازی آن است (نیاز به ذکر است نابرابری جنسن بیان می‌کند برای هر تابع مقعر f، $f(E[X]) \geq E[f(X)]$). در واقع خانواده‌هایی از کران‌های پایین برای مسئله درنظرگرفته‌می‌شوند که با شاخص‌های تغییراتی

[1] Ancillary variable

[2] Coupling

[3] Jensen's inequality

مشخص می‌شوند. با به‌کاربردن فرایند بهینه‌سازی، شاخص‌های تغییراتی طوری انتخاب می‌شوند که تا حد ممکن بهترین تقریب را برای توزیع پسینی ارائه کنند.

یک راه ساده برای پیداکردن خانواده‌هایی از کران‌های پایین، ایجاد تغییراتی در الگوی گرافی است، به‌طوری‌که یال‌ها و گره‌هایی از آن حذف کنیم. برای مثال الگوی گرافی شکل (۲-آ) را در نظر بگیرید. مشکل جفت‌شدگی بین θ و ϕ به خاطر وجود یال بین θ، z و w به‌وجودمی‌آید. با حذف این یال‌ها و گره‌های w و اضافه‌کردن شاخص‌های آزاد تغییراتی، گراف ساده‌شدهٔ شکل (۲-ب) حاصل می‌شود که خانواده‌ای از توزیع‌ها روی متغیرهای پنهان را نتیجه می‌دهد. در (د) نشان می‌دهیم با معرفی خانواده‌ای از توزیع‌های تغییراتی به صورت

$$Q(\theta, Z|D) = \prod_{d=1}^{M} Q(\theta_d|\gamma_d) \prod_{n=1}^{N} Q(z_{dn}|\pi_{dn}),$$

می‌توانیم درست‌نمایی را به شکل زیر تجزیه کنیم:

$$\ln P(D|\alpha, \phi) = KL[Q(\theta, Z|D) \parallel P(\theta, Z|D, \alpha, \phi)] + \varepsilon[Q(\theta, Z|D); \alpha, \beta],$$

که در آن $KL[Q(x)||P(x)]$ واگرایی کولبک-لایبلر[1] (KL) بین توزیع‌های $Q(x)$ و $P(x)$ است و به صورت $KL[Q(x)||P(x)] = \mathbb{E}_{Q(x)}\left[\ln \frac{Q(x)}{P(x)}\right]$ تعریف می‌شود و عبارت $\varepsilon[Q(\theta, Z|D); \alpha, \beta]$ یک کران پایین برای لگاریتم شواهد است که بیشینه‌سازی آن به‌مراتب راحت‌تر از بیشینه‌سازی مستقیم تابع درست‌نمایی است. همچنین، نشان می‌دهیم با بهینه‌سازی کران پایین لگاریتم شواهد و انتخاب مقادیر مناسب برای شاخص‌های تغییراتی، می‌توانیم از $Q(\theta, Z|D)$ به‌عنوان تقریب مناسبی برای $P(D, \theta, Z|\alpha, \phi)$ استفاده کنیم.

پس از اعمال تقریب بیز تغییراتی، مقادیر بهینه برای شاخص‌های تغییراتی به صورت:

$$\pi_{nk} \propto \phi_{kw_n} \exp\{\mathbb{E}_Q[\ln \theta_k|\gamma]\} \quad (۴)$$

$$\gamma_k = \alpha_k + \sum_{n=1}^{N} \pi_{nk}, \quad (۵)$$

حاصل می‌شوند که امید ریاضی شاخص چندجمله‌ای از رابطهٔ:

$$\mathbb{E}_Q[\ln \theta_k|\gamma] = \Psi(\gamma_k) - \Psi(\gamma_0), \quad (۶)$$

محاسبه می‌شود که در آن Ψ، تابع دایگاماست. مشاهده می‌شود همچنان به جوابی بسته نمی‌رسیم، زیرا محاسبهٔ هر یک از شاخص‌های تغییراتی به شاخص دیگر وابسته است، ولی به چارچوب یک الگوریتم تکرارشونده می‌رسیم که با ثابت درنظرگرفتن شاخص‌های الگو می‌توانیم متغیرهای پنهان را به صورت الگوریتم زیر برآورد کنیم:

الگوریتم (۲) استنباط تغییراتی LDA

۱. **برای همهٔ k و n مقدار اولیه را $\pi_{nk}^0 := 1/k$ قرار دهید.**
۲. **برای همهٔ k مقدار اولیه را $\gamma_k := \alpha_k + N/k$ قرار دهید.**
۳. **تا زمان رسیدن به هم‌گرایی انجام دهید:**
۴. **برای هر $n = 1$ تا N:**
۵. **برای هر $k = 1$ تا K:**
۶. $\pi_{nk}^{t+1} := \phi_{k\,w_n} \exp\{\Psi(\gamma_k^t)\}$
۷. **مقدار π_n^{t+1} را طوری نرمالیزه کنید که جمع مقادیرش ۱ شود.**
۸. $\gamma^{t+1} := \alpha + \sum_{n=1}^{N} \pi_n^{t+1}$

گفتنی است در [۷] نکاتی دربارهٔ الگوریتم به شرح زیر مطرح شده‌است:

- روابط (۴) و (۵) تعبیرهای شهودی جالبی دارند. به‌روزرسانی توزیع دیریکله، یعنی توزیع پسینی دیریکله به شرط امید ریاضی مشاهدات تحت توزیع تغییراتی. به‌روزرسانی شاخص چندجمله‌ای شبیه قضیهٔ بیز به صورت $P(z_n|w_n) \propto P(w_n|z_n)P(z_n)$ است که در آن $P(z_n)$ با نمای امید ریاضی لگاریتم آن، تحت توزیع تغییراتی تقریب زده شده‌است.
- توزیع تغییراتی در واقع یک توزیع شرطی و تابعی از W است، زیرا مسئلهٔ بهینه‌سازی در رابطهٔ (۳) با ثابت درنظرگرفتن W انجام می‌شود تا بهینه‌سازی شاخص‌های (γ^*, π^*) را نتیجه دهد. پس توزیع تغییراتی حاصل را به‌صورت $Q(\theta, Z|\gamma^*(W), \pi^*(W))$ می‌نویسیم تا وابستگی به W را با صراحت نشان دهیم. در نتیجه، می‌توان گفت توزیع تغییراتی، تقریبی از توزیع پسینی $P(\theta, Z|W, \alpha, \phi)$ است.
- بهینه‌سازی شاخص‌های $(\gamma^*(W), \pi^*(W))$ ویژهٔ یک مستند است. به‌طور خاص، شاخص توزیع دیریکلهٔ $\gamma^*(W)$ را می‌توان به‌صورت نمایش یک مستند در سادک موضوع نشان داد که در واقع نشان‌دهندهٔ توزیع احتمال موضوعات برای یک مستند است و از آن به‌عنوان بردار کاهش ابعاد یافتهٔ یک مستند استفاده می‌شود.

[1] Kullback-Leibler divergence

- فرایند استنباط تغییراتی را می‌توان در الگوریتم خلاصه کرد. از شبه‌رمزها روشن است که در هر چرخهٔ تکرار برای رسیدن به نتیجهٔ مطلوب LDA، به تعداد $O((N+1)K)$ عملیات نیاز است. در عمل، تعداد چرخه‌های تکرار لازم برای یک مستند از مرتبهٔ تعداد واژگان درون مستند است. در نتیجه، برای استنباط نیاز به N^2K عملیات است.

۳-۲- برآورد شاخص‌های الگو از طریق استنباط بیز تجربی

حال یک روش بیز تجربی برای برآورد شاخص‌های الگو LDA بیان می‌کنیم. هدف از ارائهٔ استنباط بیز تجربی، بیشینه‌سازی تابع لگاریتم درست‌نمایی، یعنی $\ell(\alpha,\phi)=\sum_{d=1}^{M}\ln P(D|\alpha,\phi)$، است. مقادیر به‌دست‌آمده برای π و γ که از اجرای الگوریتم به‌دست‌می‌آیند، باعث بیشینه‌شدن کران پایین لگاریتم شواهد و پیرو آن بیشینه‌سازی تابع درست‌نمایی می‌شود. در گام بعدی با ثابت درنظرگرفتن آنها، کران پایین را نسبت به شاخص‌های α و ϕ بیشینه می‌کنیم. در نهایت، به الگوریتم تکرارشونده‌ای به صورت زیر می‌رسیم:

الگوریتم (۳) EM- تغییراتی

۱. **(گام-E) برای هر مستند مقادیر بهینهٔ شاخص‌های تغییراتی $\{\gamma_d^*, \pi_d^* : d \in D\}$ را یافته و بر اساس آن تقریبی از $\ell(\alpha,\phi)$ به‌دست‌آورید.**

۲. **(گام-M) تقریب به‌دست‌آمده در گام-E را نسبت به شاخص‌های الگو α و ϕ بیشینه کنید.**

دو گام الگوریتم تا زمانی که برای کران پایین روی لگاریتم درست‌نمایی به هم‌گرایی برسیم، به‌ترتیب تکرار می‌شوند.

در **پیوست ب** نشان می‌دهیم که برای به‌روزرسانی گام-M به شرط شاخص چندجمله‌ای ϕ به رابطهٔ زیر می‌رسیم:

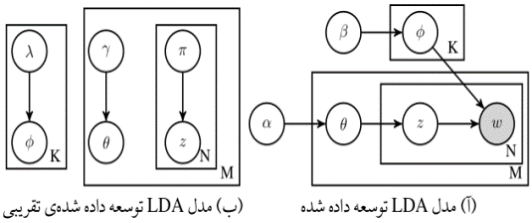
$$\phi_{kj} \propto \sum_{d=1}^{M}\sum_{n=1}^{N_d}\pi_{dnk}^{*} I[\![w_{dn}=j]\!].$$

همچنین، برای به‌روزرسانی شاخص توزیع دیریکله α می‌توانیم یک الگوریتم نیوتن-رافسون بهینه طراحی کنیم که ماتریس هسین[1] آن وارون‌پذیر با مرتبهٔ زمانی خطی است [۲۳].

۳-۳- هموارسازی

الگوی LDA ارائه‌شده در (شکل-۲) دچار یک ضعف بزرگ است و آن مشکلی به نام تُنُک‌بودن[2] است. به دلیل زیاد بودن تعداد واژگان، که مشخصهٔ بیشتر پیکره‌های مستندات است، یک مستند جدید در داده‌های آزمون به احتمال زیاد شامل واژه‌ای است که در هیچ‌یک از مستندات داده‌های آموزشی وجود ندارد. در چنین وضعیتی عناصر متناظر با آن واژگان در ماتریس احتمال ϕ صفر هستند و در نتیجه، احتمال صفر برای مستندات آزمون محاسبه می‌شود. راه‌حل رایج برای رفع این مشکل، هموارسازی[3] شاخص‌های چندجمله‌ای است، به‌این‌صورت‌که به تمام واژگان، چه در داده‌های آموزشی مشاهده شده‌باشند یا نشده‌باشند، یک احتمال مثبت و کوچک اضافه کنیم. به این منظور به‌طور معمول، هموارسازی لاپلاس استفاده می‌شود که با به‌کاربردن توزیع یک‌نواخت دیریکله روی شاخص چندجمله‌ای، میانگین توزیع پسینی را نتیجه می‌دهد.

راه‌حل پیشنهادی در [۷] برای الگوی LDA هموارسازی شده، در شکل (۳-آ) مشاهده می‌شود. اینجا شاخص ϕ را یک ماتریس تصادفی $K\times V$ در نظر می‌گیریم (که هر سطر آن معادل یکی از مؤلفه‌های آمیخته است) و هر سطر آن مستقل از بقیه، می‌تواند از یک توزیع دیریکله نمونه‌برداری شود. اکنون می‌توانیم روند استنباط را توسعه دهیم و ϕ_k را بردار تصادفی، که از توزیع دیریکله به شرط داده‌ها گرفته شده‌است، درنظربگیریم. به‌این‌ترتیب، از برآورد بیز تجربی ارائه‌شده در بخش (۳) فراتر می‌رویم و به یک رویکرد تمام بیزی برای LDA می‌رسیم.



(شکل- ۳): الگوی گرافی LDA هموارسازی‌شده به‌همراه الگوی گرافی تقریبی آن [۱۰]

(Figure- 3): Smoothed version of LDA, and its approximated model [10]

[1] Hessian
[2] Sparsity
[3] Smoothing

برای برآورد شاخص‌های الگوی توسعه‌داده‌شده نیز یک روش بیز تغییراتی درنظرمی‌گیریم که در آن توزیع‌های تغییراتی روی شاخص‌های ϕ، θ و Z عبارت است از:

$$Q(\phi, Z, \theta|\lambda, \pi, \gamma) = \prod_{k=1}^{K} \mathcal{D}ir(\phi_k|\lambda_k) \prod_{d=1}^{M} Q_d(\theta_d, z_d|\pi_d, \gamma_d), \quad (۷)$$

که در آن $Q_d(\theta_d, z_d|\pi_d, \gamma_d)$ همان توزیع تغییراتی است که در بخش ۳ تشریح شد. می‌توان دید که با به‌کاربردن روش بیز تغییراتی به همان روابط (۴) و (۵) برای به‌روزرسانی شاخص‌های توزیع‌های تغییراتی π و γ می‌رسیم. همچنین، برای شاخص جدید نیز به رابطهٔ به‌روزرسانی زیر می‌رسیم:

$$\lambda_{kj} = \beta + \sum_{d=1}^{M} \sum_{n=1}^{N_d} \pi^*_{dnk} I[\![w_{dn} = j]\!]. \quad (۸)$$

با تکرار به‌روزرسانی شاخص‌ها در یک چرخه تا رسیدن به هم‌گرایی می‌توانیم به تقریبی از توزیع پسینی روی شاخص‌های ϕ، θ و Z برسیم.

در پایان، باید ابرشاخص‌های توزیع دیریکله β و α را برآورد کنیم، که مانند روش بخش ۳ آنها را با برآورد بیز تجربی تقریب می‌زنیم. با استفاده از EM-تغییراتی، برآورد بیشینه درست‌نمایی این شاخص‌ها را برای درست‌نمایی حاشیه‌ای به‌دست‌می‌آوریم.

۴- استنباط تکرارشوندهٔ بیز تغییراتی (IVB)

در این بخش یک الگوریتم استنباط تکرارشوندهٔ بیز تغییراتی[1] (IVB) برای تقریب‌زدن تمام متغیرهای پنهان و شاخص‌های الگوی شکل (۳-آ) ارائه می‌دهیم و برای تقریب بیز تغییراتی از الگوی گرافی شکل (۳-ب) استفاده می‌کنیم. نیاز به اشاره است بیشتر منابعی که بررسی شدند، روش EM-تغییراتی را تشریح کرده‌اند. نویسندگان در [۲۵] تمام شاخص‌ها و متغیرهای پنهان را از روش بیشینه‌سازی کران پایین شواهد به‌د[illegible] ولی در زمینهٔ زیست‌داده‌ورزی و بر روی داده[illegible]ژنتیکی است، و بنابراین، در کاربردهای متن‌کاوی [illegible]طور مستقیم، قابل‌استفاده نیست. در [۱۲] یک چار[illegible] یادگیری برخط در کاربرد پردازش متن ارائه شده‌است [illegible]با روشی مشابه EM، شاخص‌ها و متغیرهای پنهان الگ[illegible]با بیشینه‌سازی کران پایین شواهد به‌دست‌می‌آورد. [illegible]ن جهت، شرح روش IVB بر روی داده‌های متنی برای اولین بار در این مقاله ارائه می‌شود.

نخست، الگوریتم استنباط IVB را ارائه می‌دهیم. احتمال توأم مشاهدات، متغیرهای پنهان و شاخص‌ها در الگوی هموارسازی‌شدهٔ LDA را به صورت $P(D, Z, \theta, \phi|\alpha, \beta)$ در نظر می‌گیریم و احتمال توأم الگوی توسعه‌داده‌شدهٔ تقریبی به روش بیز تغییراتی برای شاخص‌ها نیز از رابطهٔ (۷) محاسبه می‌شود.

الگوریتم- ۴: استنباط تکرارشوندهٔ بیز تغییراتی (IVB)

۱. **شاخص‌های توزیع تغییراتی $Q^*(Z)$ را از رابطهٔ زیر به‌روزرسانی کنید:**
$$Q^*(Z) \propto \exp\{\mathbb{E}_{Q(\theta)Q(\phi)}[\![\ln P(D, Z, \theta, \phi)]\!]\}$$
۲. **با استفاده از نتایج گام قبلی شاخص‌های توزیع تغییراتی $Q^*(\theta)$ را از رابطهٔ زیر به‌روزرسانی کنید:**
$$Q^*(\theta) \propto \exp\{\mathbb{E}_{Q(Z)Q(\phi)}[\![\ln P(D, Z, \theta, \phi)]\!]\}$$
۳. **با استفاده از نتایج دو گام قبلی شاخص‌های توزیع تغییراتی $Q^*(\phi)$ را از رابطهٔ زیر به‌روزرسانی کنید:**
$$Q^*(\phi) \propto \exp\{\mathbb{E}_{Q(Z)Q(\theta)}[\![\ln P(D, Z, \theta, \phi)]\!]\}$$

سه گام الگوریتم در یک چرخه تا رسیدن به هم‌گرایی و یا عبور از محدودیت زمان اجرا به‌ترتیب انجام می‌شوند. در پیوست- نشان می‌دهیم که برای شاخص‌های توزیع‌های تغییراتی γ و λ به قواعد به‌روزرسانی (۵) و (۸) می‌رسیم، اما برای شاخص π به رابطهٔ زیر می‌رسیم:

$$\pi_{vk} \propto \exp\{\mathbb{E}_Q[\![\ln \theta_k|\gamma]\!] + \mathbb{E}_Q[\![\ln \phi_{kv}|\lambda]\!]\}, \quad (۹)$$

که در آن $\mathbb{E}_Q[\![\ln \theta_k|\gamma]\!]$ از رابطهٔ (۶) به دست می‌آید و $\mathbb{E}_Q[\![\ln \phi_{kn}|\lambda]\!]$ به‌طور مشابه از رابطهٔ: $= \Psi(\lambda_{kn}) - \Psi(\lambda_{k0})$ که $\lambda_{k0} = \sum_{n=1}^{V} \lambda_{kn}$ به‌دست‌می‌آید. الگوریتم در یک چرخه به‌ترتیب مقادیر روابط (۹)، (۵) و (۸) را محاسبه و شاخص‌های تغییراتی را به‌روزرسانی می‌کند. در پیوست- شرح به‌دست‌آوردن این روابط آمده‌است.

هر دو روش EM-تغییراتی و استنباط IVB به روابط به‌روزرسانی به‌طور تقریبی مشابهی می‌رسند، اما تفاوت‌هایی در تفسیر شاخص‌ها دارند. در روش EM-تغییراتی شاخص ϕ در گام-E ثابت درنظرگرفته‌می‌شود و مقدار آن به‌طور مستقیم، در به‌روزرسانی π استفاده می‌شود، سپس، جداگانه در گام-M برآورد می‌شود. درحالی‌که در IVB، با شاخص ϕ همانند متغیرهای پنهان رفتار می‌شود، در نتیجه، به‌جای‌این‌که به‌طور مستقیم، در به‌روزرسانی π استفاده شود، با نمای امید ریاضی $\ln \phi$ جایگزین می‌شود که در ادامه منجر به تقریب‌زدن مقدار ϕ

[1] Iterative Variational Bayes

با شاخص تغییراتی λ می‌شود. همچنین، در IVB تمام گام‌های الگوریتم، شامل بیشینه‌سازی شاخص‌های یک توزیع تغییراتی تحت امید ریاضی سایر توزیع‌های تغییراتی است.

۵- تحلیل داده‌های متنی

در این بخش با استفاده از داده‌های واقعی به بررسی کارایی LDA در الگوبندی زبانی بر اساس ملاک سرگشتگی و دقت الگو بر حسب درصد دسته‌بندی درست می‌پردازیم. همچنین، کاربرد LDA را در پالایش همکارانه بررسی می‌کنیم.

برای بررسی کارایی و دقت الگوریتم ارائه‌شده در این مقاله از پایگاه اینترنتی خبرگزاری فارس استفاده کردیم. دو زیربخش تحلیل بین‌الملل[1] و اقتصاد کلان و بودجه[2] را چون شامل مقالاتی تحلیلی هستند و نسبت به دیگر زیربخش‌ها متن‌های طولانی‌تری دارند، انتخاب کردیم. مراحل استخراج و آماده‌سازی داده‌ها به ترتیب زیر است:

- بارگیری صفحات فهرست‌شده در هر زیربخش
- استخراج متن مقالات و جداسازی واژگان
- حذف واژگان توقف[3] از مجموعه واژگان هر مقاله
- ایجادکردن بردار واژگان

در این مرحله، پیکرهٔ مستنداتی با ۳۶۸۹ واژهٔ یکتا و دو موضوع ایجاد می‌شود که برای هر موضوع حدود هزار مستند وجود دارد. با تنظیماتی که روی این پیکره اعمال می‌شود، می‌توان کارایی الگوریتم را در شرایط مختلف بررسی کرد.

۵-۱- الگوبندی پیکرهٔ مستندات

در این بخش الگوبندی زبانی روش‌های LDA و PLSA را مقایسه می‌کنیم. به‌طور معمول، برای مقایسهٔ الگوهای زبانی، به‌طور مستقیم، از احتمال استفاده نمی‌شود و برای این منظور از ملاک سرگشتگی[4] استفاده می‌شود که یکی از ملاک‌های رایج برای مقایسهٔ میزان تعمیم‌دهی[5] روش‌های الگوبندی زبانی است.

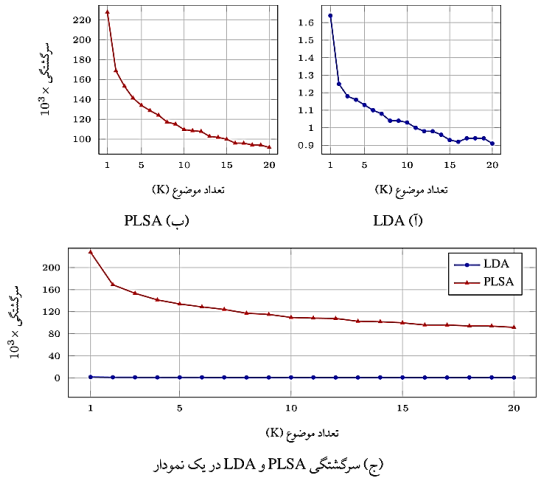
سرگشتگی، معکوس احتمال داده‌های آزمون است که نسبت به تعداد داده‌های آزمون نرمالیزه شده‌است و به شکل زیر تعریف می‌شود [۱۷]:

که در آن W_d مستند d-ام از مجموعه‌داده‌های آزمون، N_d تعداد واژگان درون آن و M تعداد مستندات آزمون است. هرچه مقدار سرگشتگی برای یک الگوی کوچک‌تر باشد، آن الگو کارایی بهتری دارد.

$$PP(D_{test}) = P(W_1, \cdots, W_M)^{-\frac{1}{\sum_{d=1}^{M} N_d}} = \left(\prod_{d=1}^{M} P(W_d)\right)^{-\frac{1}{\sum_{d=1}^{M} N_d}} = \exp\left\{-\frac{\sum_{d=1}^{M} \log P(W_d)}{\sum_{d=1}^{M} N_d}\right\}$$

برای مقایسهٔ الگوی زبانی LDA با PLSA، پیکرهٔ اسناد ایجادشده با روش بیان‌شده در بخش (۳) را به دو قسمت داده‌های آموزشی و داده‌های آزمون تقسیم می‌کنیم و برای مقادیر مختلفی از تعداد موضوعات (K) ملاک سرگشتگی دو روش را با هم مقایسه می‌کنیم.

(شکل- ۴) مقایسهٔ ملاک سرگشتگی را برای LDA و PLSA نشان می‌دهد. ملاحظه می‌شود که برای این مجموعه داده‌های خاص، درحالی‌که مقدار سرگشتگی الگو LDA با افزایش تعداد موضوعات در حدود مقادیر $۱/۶۵ \times ۱۰^۳$ تا $۰/۹۱ \times ۱۰^۳$ تغییر می‌کند، مقدار این ملاک برای الگوی PLSA بسیار بیشتر است و مقادیری حدود $۲۲۷/۸۴ \times ۱۰^۳$ تا $۹۱/۶۳ \times ۱۰^۳$ دارد. روشن است که سرگشتگی الگو LDA کمتر از PLSA است و عملکرد بسیار بهتری در الگوبندی زبانی دارد.



(شکل- ۴) ملاک سرگشتگی برای مجموعه‌داده‌های خبرگزاری فارس

(Figure-4): Perplexity of farsnews data set

[1] https://www.farsnews.ir/world/Analysis-International
[2] https://www.farsnews.ir/economy/macroeconomics
[3] Stop words
[4] Perplexity
[5] Generalization

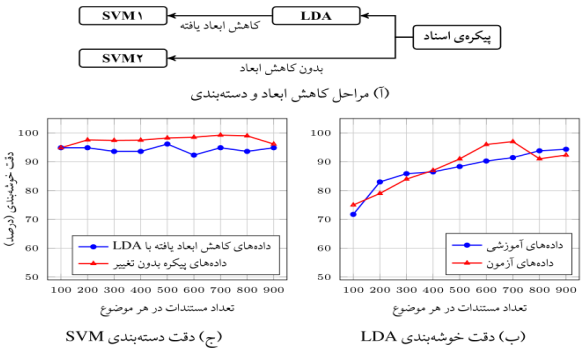
۵-۲- خوشه‌بندی پیکرهٔ مستندات

الگوریتم LDA را روی مجموعه‌داده‌های جمع‌آوری‌شده با روش بخش ۳ اجرا می‌کنیم. هر بار تعدادی از مقالات را به‌عنوان مجموعه‌داده‌های آموزشی درنظرمی‌گیریم و حدود صد مقاله را از باقی‌ماندهٔ مقالات برای مجموعه‌داده‌های آزمون به تصادف انتخاب می‌کنیم و دقت خوشه‌بندی مقالات را برای مجموعه‌داده‌های آموزشی و آزمون به‌دست‌می‌آوریم. به دلیل‌این‌که در زمان جمع‌آوری مجموعه‌داده‌ها، برچسب موضوع هریک از مستندات مشخص است، می‌توانیم دقت خوشه‌بندی را برای داده‌های آموزشی و آزمون به‌دست‌آوریم.

به‌منظور بررسی دقت خوشه‌بندی ملاک‌هایی مانند خلوص[1]، شاخص رند[2]، و اطلاعات دوسره[3] وجود دارد [۲۴]. اما در اینجا به‌این‌دلیل‌که در جمع‌آوری مجموعه‌داده‌ها، برچسب هر مستند مشخص است، می‌توانیم از روش ابتکاری زیر استفاده کنیم. پس از خوشه‌بندی می‌توان خوشه‌ها را به هر ترتیبی برچسب‌گذاری کرد، سپس، با ثابت نگه‌داشتن خوشه‌بندی داده‌ها، تمام جای‌گشت‌های برچسب‌گذاری برای خوشه‌ها را تولید می‌کنیم و بر اساس آن تعداد داده‌هایی که برچسب هم‌نام با برچسب خوشه دارند، شمارش و دقت خوشه‌بندی را محاسبه می‌کنیم. جای‌گشتی که بیشینهٔ دقت را داشته‌باشد، به‌عنوان برچسب‌گذاری درست و دقت خوشه‌بندی را بر حسب درصد اعلام می‌کنیم. در این مورد ویژه، به دلیل این که تعداد برچسب‌ها و خوشه‌ها کم است، می‌توان این روش را پیاده‌سازی کرد؛ که از نظر محاسباتی عملی است.

همچنین، برای مقایسهٔ دقت خوشه‌بندی روش LDA با دیگر روش‌های خوشه‌بندی، از ماشین بردار پشتیبان[4] استفاده می‌کنیم. از SVM برای دسته‌بندی دودویی[5] استفاده می‌شود. مجموعه‌داده‌ها را بر اساس موضوع به دو دسته تقسیم و برچسب‌گذاری می‌کنیم. ازآنجاکه می‌توان از LDA برای کاهش ابعاد استفاده کرد و ویژگی‌های یک مستند را در $\gamma^*(W)$ خلاصه‌سازی کرد، یک بار نتایج حاصل از LDA را برای γ^* به‌عنوان بردار ویژگی‌های مستندات درون پیکره به یک SVM می‌دهیم، آن‌گاه بار دیگر بردار واژگان مستندات پیکره را بدون تغییر به‌عنوان بردار ویژگی به SVM دیگری می‌دهیم. سپس، دقت پیش‌بینی برچسب داده‌های آزمون را برای هر دو ماشین به‌دست‌می‌آوریم.

شکل (۵-ب) نشان می‌دهد با افزایش حجم داده‌های آموزشی دقت خوشه‌بندی برای داده‌های آموزشی و آزمون بالا می‌رود. به‌طور مشخص، با افزایش داده‌های آموزشی، دقت خوشه‌بندی برای داده‌های آزمون از حدود 75% به حدود 92.31% افزایش می‌یابد. همچنین، در شکل (۵-ج) مشخص است که دقت استفاده از بردار ویژگی‌های کاهش‌ابعادیافته، نزدیک به حالتی است که از بردار ویژگی‌های بدون تغییر استفاده شود. در حالتی که از بردار کاهش ابعاد یافته برای دسته‌بندی استفاده می‌شود، دقتی به‌طور متوسط در حدود ۹۴.۳٪ داریم و در حالتی که از داده‌های کل پیکره استفاده می‌شود، دقتی به‌طور متوسط در حدود 97/61% داریم.



(شکل- ۵): دقت خوشه‌بندی برای مجموعه‌داده‌های خبرگزاری فارس

(Figure-5): Accuracy of clustering for farsnews data set

پس از خوشه‌بندی، واژه‌ها را بر اساس احتمال تکرار در هر موضوع مرتب و بیست واژه‌ای را که احتمال تکرار بالا دارند، استخراج کردیم و نتیجه را در (جدول نشان دادیم. می‌توان دید که برخی واژگان در هر دو موضوع تکرار شده‌اند؛ ولی روشن است که واژگان هر دسته به موضوع آن دسته ارتباط معنایی دارند.

(جدول- ۱): واژگان با احتمال تکرار بالا در هر موضوع

(Table-1): Most probable words in each topic

اقتصاد کلان و بودجه

دولت	میلیارد	مجلس	ارز
بودجه	درصد	قانون	اقتصاد
کشور	ایران	سازمان	مالی
اقتصادی	افزایش	تومان	لایحه
هزار	منابع	نفت	بانک

تحلیل بین‌الملل

آمریکا	دولت	اسرائیل	نظامی
ایران	سعودی	آمریکایی	امارات
کشور	رژیم	عربستان	چین
ترامپ	صهیونیستی	توافق	سیاسی
عراق	منطقه	یمن	بایدن

[1] Purity
[2] Rand index
[3] Mutual information
[4] Support vector machine (SVM)
[5] Binary classification

۵-۳ -پالایش همکارانه

در تحلیل پایانی از داده‌های MovieLens 1m [۲۲] برای پالایش همکارانه[1] استفاده می‌کنیم. در این مجموعه‌داده، ۶۰۴۰ کاربر به تعدادی از فیلم‌های موردعلاقه‌شان از بین ۳۹۵۳ فیلم موجود در آن، امتیازاتی یک تا پنج داده‌اند. کاربر و فیلم‌های موردعلاقه‌اش به‌ترتیب متناظر مستند و واژگان درون مستند هستند. برای پالایش همکارانه به ترتیب زیر عمل می‌کنیم:

- با داده‌های کاربران مشاهده‌شده، الگویی را آموزش می‌دهیم.
- برای هر کاربر مشاهده‌نشده، تمام فیلم‌های موردعلاقهٔ او را به‌جز یک فیلم به الگو می‌دهیم.
- با استفاده از الگو، امتیاز فیلم‌های بیرون نگه‌داشته‌شده[2] را پیش‌بینی می‌کنیم.

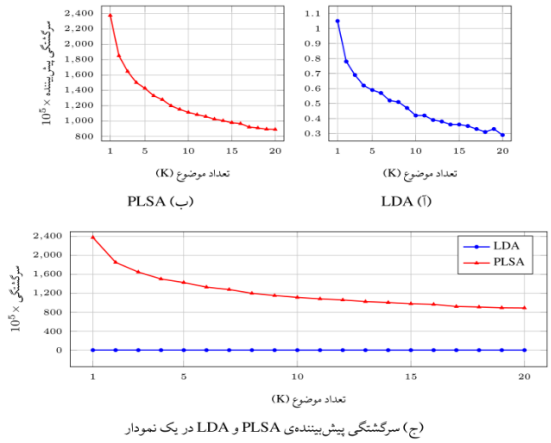
الگوریتم‌های PLSA و LDA روی داده‌های آموزشی ایجادشده اجرا می‌شوند، سپس، لگاریتم درست‌نمایی برای داده‌های بیرون نگه‌داشته‌شده محاسبه می‌شود. برای مقایسهٔ روش‌ها از ملاک سرگشتگی پیش‌بینی‌کننده[3] که به صورت:

$$\mathrm{pred}_{\mathrm{PP}(\mathrm{D_{test}})} = \exp\left\{-\frac{\sum_{d=1}^{M} \ln \mathrm{P}\left(w_{di}\middle|w_{d(-i)}\right)}{M}\right\},$$

تعریف می‌شود، استفاده می‌کنیم که در آن M تعداد کاربرانی

است که به‌عنوان داده‌های آزمون استفاده شده‌اند، w_{di} داده‌ای

است که الگو باید پیش‌بینی کند و $w_{d(-i)}$ همهٔ داده‌ها غیر از داده‌ی i-ام هستند که به الگو نشان داده‌می‌شود.



(ج) سرگشتگی پیش‌بینندهٔ PLSA و LDA در یک نمودار

(شکل-)۶: ملاک سرگشتگی پیش‌بینندهٔ برای مجموعه‌داده‌های MovieLens 1m

(Figure-6): Peredictive perplexity of MovieLens 1M data set

(شکل مقدار ملاک سرگشتگی پیش‌بینندهٔ را برای الگوهای LDA و PLSA نشان می‌دهد. مشاهده می‌شود که برای این مجموعه‌داده میزان سرگشتگی پیش‌بینندهٔ الگوی LDA از حدود ۱۰۵/۰۵ × ۱۰^۵ تا ۰/۲۹ × ۱۰^۵ کاهش می‌یابد و برای PLSA از حدود ۲۳۷۶/۸۷ × ۱۰^۵ تا ۸۹۰/۸۷ × ۱۰^۵ کاهش می‌یابد. سرگشتگی پیش‌بینندهٔ در پالایش همکارانه برای LDA از PLSA کمتر است و در این کاربرد نیز الگوبندی بهتری نسبت به الگوی PLSA دارد.

۶- بحث و نتیجه‌گیری

در این مقاله الگوی LDA را، که یک الگوی احتمالاتی مولد برای الگوبندی داده‌های گسسته است، شرح دادیم. توزیع پسینی الگوی LDA را با روش بیز تغییراتی به‌صورت تقریبی محاسبه کردیم. آنگاه، این رویکرد را بر روی داده‌های واقعی پیاده‌سازی و عملکرد آن را با الگوی PLSA مقایسه کردیم. کاربردهای مختلف آن را به‌طور مختصر بررسی کردیم و نشان دادیم که می‌توان از LDA برای کاهش ابعاد داده‌های حجیم استفاده کرد.

در نتیجه‌گیری کلی، می‌توان نکات زیر را بیان کرد:

- الگوی LDA یک الگوی احتمالاتی مولد برای الگوبندی پیکرهٔ مستندات است. این الگو نخست، به فرایند تولید یا ایجاد یک مستند می‌پردازد. سپس، از روی این فرایند احتمال توأم مستند، شاخص‌ها و متغیرهای پنهان را محاسبه و پس از آن، احتمال کل پیکرهٔ مستندات را حساب می‌کند.
- در حالی که ادعا می‌شود PLSA یک الگوی مولد است، با استفاده از آن نمی‌توان مستند جدید، تولید، یا به‌راحتی احتمال را برای مستندات جدید محاسبه کرد، در نتیجه، PLSA یک الگوی مولد کامل نیست. درصورتی‌که با استفاده از LDA می‌توان مستند جدید تولید یا احتمال مستند به‌تازگی ایجادشده را محاسبه کرد.
- با استفاده از ویژگی‌های اشاره‌شده برای LDA می‌توان داده‌های ساختگی یا مصنوعی با تنظیمات و تعداد دلخواه ایجاد کرد و یک پیکرهٔ مستندات کامل ساخت و از آن برای آزمودن الگوریتم قبل از اجرا روی داده‌های واقعی استفاده کرد.

[1] Collaborative filtering
[2] Held-out
[3] Predictive perplexity

- الگوی LDA یک الگوی احتمالاتی ساده، اما انعطاف‌پذیر است و می‌تواند برای الگوبندی ساختارهای چندسطحی در زمینه‌های مختلف استفاده شود، زیرا یک الگوی احتمالاتی پیمانه‌ای[1] است و می‌تواند در الگوهای احتمالاتی دیگر جاسازی[2] شود. این ویژگی به LDA اجازه می‌دهد تا توسعه‌پذیر باشد و برای ساختن الگوهای پیچیده‌تر و در کاربردهای مختلف استفاده شود، درحالی‌که این قابلیت در الگویی مانند LSA وجود ندارد.
- مزیت روش LDA نسبت به روش‌های دیگر الگوبندی پیکرهٔ مستندات، الگوسازی بهتر موضوعات مربوط به یک مستند است که باعث می‌شود بتوان گردایه‌های حجیم از مستندات را با انعطاف‌پذیری بیشتر الگو کرد. همچنین، این روش خطر بیش‌برازش[3] را کاهش می‌دهد و قدرت تعمیم‌پذیری بیشتری نسبت به روش‌های دیگر دارد.
- مانند بیشتر روش‌های خوشه‌بندی، در LDA فرض می‌شود دانش پیشین کافی دربارهٔ تعداد خوشه‌ها (K) داریم و مقدار آن را از قبل می‌دانیم. البته در واقعیت کمتر پیش می‌آید که چنین شرایطی محقق شود. در این مواقع باید روش‌هایی برای برآورد مقدار K به‌کاربریم. یک روش بیزی می‌تواند برآورد K به شرط داده‌ها باشد. روش ابتکاری دیگر حدس‌زدن حدود K در یک بازه، سپس، اجرای الگوریتم روی تک‌تک مقادیر بازه و انتخاب مقداری از K که بهترین الگو را برای داده‌ها برازش کرده‌است. ملاک انتخاب الگو بهتر می‌تواند درست‌نمایی یا سرگشتگی باشد. برای مثال، در کاربرد الگوبندی پیکرهٔ مستندات از نمودارهای (شکل- ۴) می‌توان مشاهده کرد که شیب نمودار سرگشتگی برای مقادیر ۲ تا ۵ تغییرات زیادی دارد و پس از آن به‌طور تقریبی ثابت می‌شود. انتخاب K از بازه‌ی ۲ تا ۵ می‌تواند مناسب باشد و تا حد زیادی با واقعیت هم تطابق دارد.

در پایان نیاز به اشاره است که پیاده‌سازی انجام‌شده برای LDA و PLSA در مخزن گیتهاب[4] در دسترس است.

[1] Modular
[2] Embed
[3] Overfitting
[4] https://github.com/VahidHeidari/TopicModeling/

۶- مراجع 6- Refrence

[۱] م. رسولی, ب. مینایی‌بیدگلی, ه. فیلی, م. امینیان, "استخراج بی ناظر ظرفیت فعل در زبان فارسی," پردازش علائم و داده‌ها, دوره ۹,شماره ۲, صفحات ۱۲-۳, ۱۳۹۱.

[1] M. S. Rasoli, B. Minaei Bidgoli, H. Faili, and M. Aminian, "Unsupervised Persian Verb Valency Induction," Signal and Data Processing, vol. 9, no. 2, 3-12, 2013.

[۲] ا. عسکریان, م. کاهانی, ش. شریفی, "حس‌نگار: شبکهٔ واژگان فارسی", پردازش علائم و داده‌ها, دوره ۱۵, شمارهٔ ۱, صفحات ۸۶-۷۱, ۱۳۹۷.

[2] E. Asgarian, M. Kahani, and S. Sharifi, "HesNegar: Persian Sentiment WordNet," Signal and Data Processing, vol. 15, no. 1, pp. 71-86, 2018.

[۳] ه. فیلی, "استفاده از تجزیه‌گرهای احتمالاتی زبان طبیعی جهت بهبود ترجمهٔ افعال گروهی انگلیسی به فارسی," پردازش علائم و داده‌ها, دوره ۷, شماره ۱, صفحات ۷۶-۶۵, ۱۳۸۹.

[3] H. Faili, "Phrasal Verb Translation from English to Persian Using Statistical Parsing," Signal and Data Processing, vol. 7, no. 1, pp. 66-76, 2010.

[۴] ه. فیلی, ح. قادر, م. آنالویی, "یک الگوی بیزی برای استخراج با مربی گرامر زبان طبیعی," پردازش علائم و داده‌ها, دوره ۹, شماره ۱, صفحات ۳۴-۱۹, ۱۳۹۱.

[4] H. Faili, H. Ghader, and M. Analoui, "A Bayesian Model for Supervised Grammar Induction," Signal and Data Processing, vol. 9, no. 1, pp. 19-34, 2012.

[۵] ب. مسعودی, س. قوچانی, "رفع ابهام معنایی واژگان مبهم فارسی با مدل موضوعی LDA," پردازش علائم و داده‌ها, دوره ۱۲, شماره ۴, صفحات ۱۲۵-۱۱۷, ۱۳۹۴.

[5] B. Masoudi, and R. G. Saeid, "Farsi Word Sense Disambiguation with LDA Topic Model," Signal and Data Processing, vol. 12, no. 4, pp. 117-125, 2016.

[6] E. Asgari, and J.-C. Chappelier, "Linguistic [1] Analysis of Persian Poems," Proceedings of the Second Workshop on Computational Linguistics for Literature, Atlanta, Georgia, pp. 23-31, 2013.

[7] D. Blei, A. Ng, and J. Michael, "Latent Dirichlet Allocation," Journal of Machine Learning Research, vol. 3, pp. 993-1022, 2003.

[8] S. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. Harshman, "Indexing by Latent Semantic Analysis," Journal of the

American Society for Information Science, vol. 41, pp. 391-407, 1990.
[9] Y. Du, Y. Yi, X. Li, X. Chen, Y. Fan, and F. Su, “Extracting and Tracking Hot Topics of Micro-blogs Based on Improved Latent Dirichlet Allocation,” Engineering Applications of Artificial Intelligence, vol. 87, pp. 103279, 2020.
[10] C. Geigle, "Inference Methods for Latent Dirichlet Allocation,", Course notes (cs598cxz advanced topics in information retrieval), Department of Computer Science, University of Illinois at Urbana-Champaign, 2016.
[11] Y. Gong, Q. Zhang, and X. Huang, “Hashtag Recommendation for Multimodal Microblog Posts,” Neurocomputing, vol. 272, pp. 170-177, 2018.
[12] M. Hoffman, D. Blei, and F. Bach, "Online Learning for Latent Dirichlet Allocation," Advances in Neural Information Processing Systems. pp. 856-864, 2010.
[13] T. Hofmann, "Probabilistic Latent Semantic Indexing," SIGIR '99. pp. 50–57, 1999.
[14] T. Hofmann, "Probabilistic Latent Semantic Analysis," UAI'99. pp. 289–296, 1999.
[15] T. Hofmann, “Unsupervised Learning by Probabilistic Latent Semantic Analysis,” Machine Learning, vol. 42, pp. 177–196, 2001.
[16] H. Jelodar, Y. Wang, C. Yuan, X. Feng, X. Jiang, Y. Li, and L. Zhao, “Latent Dirichlet Allocation (LDA) and Topic Modeling: Models, Applications, a Survey,” Multimedia Tools Applications, vol. 78, pp. 15169–15211, 2019.
[17] D. Jurafsky, and J. H. Martin, Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, USA: Prentice Hall PTR, 2000.
[18] J. Leskovec, A. Rajaraman, and J. D. Ullman, Mining of Massive Datasets, USA: Cambridge University Press, 2014.
[19] B. Liu, C. Wang, Y. Wang, K. Zhang, and C. Wang, “Microblog Topic Mining Based on FR-DATM,” Chinese Journal of Electronics, vol. 27, pp. 334-341, 2018.
[20] X. Liu, Y. Gao, Z. Cao, and G. Sun, “LDA-based Topic Mining of Microblog Comments,” Journal of Physics: Conference Series, vol. 1757, pp. 012118, 2021.
[21] Y. Lu, Q. Mei, and C. Zhai, “Investigating Task Performance of Probabilistic Topic Models: An Empirical Study of PLSA and LDA,” Information Retrieval, vol. 14, pp. 178–203, 2011.
[22] H. F. Maxwell, and K. Joseph, “The MovieLens Datasets: History and Context,” ACM Transactions on Interactive Intelligent Systems, vol. 5, 2015.
[23] T. Minka, "Estimating a Dirichlet Distribution,", Technical report, M.I.T., 2000.
[24] K. P. Morphy, Machine Learning: A Probabilistic Perspective, London, England: MIT Press, 2012.
[25] A. Raj, M. Stephens, and J. K. Pritchard, “fastSTRUCTURE: Variational Inference of Population Structure in Large SNP Data Sets,” Genetics, vol. 197, pp. 573-589, 2014.
[26] V. Smidl, and A. Quinn, The Variational Bayes Method in Signal Processing, Berlin Heidelberg, Germany: Springer, 2006.

وحید حیدری دانش‌آموختهٔ کارشناسی‌ارشد مهندسی کامپیوتر گرایش الگوریتم‌ها و محاسبات در سال ۱۳۹۹ از دانشگاه تهران است. زمینه‌های پژوهشی موردعلاقهٔ ایشان یادگیری ماشین، زیست‌داده‌ورزی و پردازش زبان‌های طبیعی است.

نشانی رایانامهٔ ایشان عبارت است از:

vahid.heidari@ut.ac.ir

سید محمود طاهری استاد آمار ریاضی دانشکده علوم مهندسی در دانشکدگان فنی دانشگاه تهران است. ایشان دکترای خود را در رشتهٔ آمار ریاضی از دانشگاه شیراز در سال ۱۳۷۶ گرفته‌است. زمینه‌های موردعلاقهٔ ایشان استنباط آماری، ریاضیات و منطق فازی، و آمار و احتمال فازی است.

نشانی رایانامهٔ ایشان عبارت است از:

sm_taheri@ut.ac.ir

سیدمرتضی امینی دانشیار آمار دانشکدهٔ ریاضی، آمار و علوم کامپیوتر دانشکدگان علوم دانشگاه تهران است. ایشان دکترای خود را در رشتهٔ آمار گرایش استنباط آماری در سال ۱۳۹۰ از دانشگاه فردوسی مشهد گرفته‌است. زمینهٔ پژوهشی موردعلاقهٔ ایشان یادگیری ماشین آماری است.

نشانی رایانامهٔ ایشان عبارت است از:

morteza.amini@ut.ac.ir

پیوست‌ها

پیوست آ: روش بیز تغییراتی

در این پیوست، نخست، قضیهٔ بیز تغییراتی را بیان، سپس، رابطه‌های بیان‌شده را در بخش ۳، برای اجرای الگوریتم اثبات می‌کنیم. نیاز به ذکر است در [۷] از روش بهینه‌سازی کران پایین شواهد استفاده شده‌است؛ اما در اینجا از قضیهٔ اول بیز تغییراتی استفاده می‌کنیم و نشان می‌دهیم هر دو روش هم‌ارز هستند و نتایج یکسانی دارند.

پیشنهاد اصلی روش بیز تغییراتی تقریب‌زدن احتمال توأم $P(D,\theta,Z|\alpha,\phi)$ به صورت زیر است:

$$P(D,\theta,Z|\alpha,\phi) = P(\theta,Z|D,\alpha,\phi)\underbrace{P(D|\alpha,\phi)}_{\text{constant}} \propto P(\theta,Z|D,\alpha,\phi)$$

$$\approx Q(\theta,Z|D) = \prod_{d=1}^{M} Q(\theta_d|\gamma_d)\prod_{n=1}^{N} Q(z_{dn}|\pi_{dn}).$$

از توزیع $Q(\theta,Z|D)$ می‌توان به‌عنوان جایگزین توزیع $P(D,\theta,Z|\alpha,\phi)$ استفاده کرد، که در آن γ و π شاخص‌های تغییراتی هستند و آنها را با الگوریتم EM-تغییراتی (الگوریتم) به‌دست‌می‌آوریم. در ادامه، قضیهٔ بیز تغییراتی را بدون اثبات بیان می‌کنیم (برای اثبات به بخش ۳.۳.۱ در [۲۶] مراجعه کنید).

قضیهٔ ۱ (بیز تغییراتی): فرض کنید $P(\theta|X)$ توزیع پسینی با شاخص چندمتغیرهٔ θ باشد که به K زیر بردار به صورت $\theta = (\theta_1,\cdots,\theta_K)$ افراز شده‌باشد و $Q(\theta|X)$ یک توزیع تقریبی محدودشده به توزیع‌های مستقل شرطی برای $\theta_1,\cdots,\theta_K$ به صورت زیر باشد:

$$Q(\theta|X) = Q(\theta_1,\cdots,\theta_K|X) = \prod_{i=1}^{K} Q(\theta_i|X),$$

آنگاه کمینهٔ واگرایی $Q^*(\theta|X) = \arg\min_{Q(\cdot)} KL[Q(\theta|X) \parallel P(\theta|X)]$ از رابطهٔ زیر به‌دست‌می‌آید:

$$Q^*(\theta_i|X) \propto \exp\left\{\mathbb{E}_{Q^*(\theta_{/i}|X)}[\![\ln P(\theta,X)]\!]\right\}, i = 1,\cdots,K,$$

که در آن $\theta_{/i}$ مجموعهٔ شاخص‌های مکمل θ_i در θ و $Q^*(\theta_{/i}|X) = \prod_{j=1,j\neq i}^{K} Q^*(\theta_j|X)$ هستند.

الگوریتم EM-تغییراتی شامل دو گام است که در ادامه مراحل هرکدام شرح داده می‌شود:

گام-E: به‌منظور یافتن تقریب توزیع پسینی متغیرهای پنهان نیاز است گام‌های زیر برداشته‌شوند:

۱. نوشتن احتمال توأم شاخص‌ها و محاسبهٔ لگاریتم آن،

۲. افراز شاخص‌ها و مشخص‌کردن توزیع‌های تغییراتی،

۳. به‌روزرسانی شاخص‌های تغییراتی از توزیع‌های $Q^*(\theta_i|X)$ که طبق **قضیهٔ (۱)** مشخص شده‌اند.

گام-M: با ثابت درنظرگرفتن متغیرهای پنهان، مقادیر شاخص‌های الگو محاسبه می‌شوند به‌طوری‌که لگاریتم احتمال توأم شاخص‌ها و مشاهدات را بیشینه کنند. در ادامه این مراحل تشریح می‌شوند.

آ-۱- لگاریتم احتمال توأم شاخص‌ها

احتمال توأم شاخص‌ها به صورت زیر است:

$$P(W_d,\theta,Z|\alpha,\phi) = \mathcal{D}ir(\theta_d|\alpha)\prod_{n=1}^{N_d} P(z_{dn}|\theta_{dz_{dn}})\,P(w_{dn}|z_{dn},\phi).$$

از این رابطه لگاریتم طبیعی می‌گیریم، پس داریم:

$$\ln P(W_d,\theta,Z|\alpha,\phi) = \ln\Gamma(\alpha_0) + \sum_{k=1}^{K}(\alpha_k-1)\ln\theta_{dk} - \ln\Gamma(\alpha_k) + \sum_{n=1}^{N_d}\ln\theta_{dz_{dn}} + \ln\phi_{z_{dn}w_n},$$

که در آن $\alpha_0 = \sum_{k=1}^{K}\alpha_k$.

آ-۲- لگاریتم توزیع‌های تغییراتی

توزیع‌های تغییراتی را به صورت زیر تعریف می‌کنیم:

$$Q(\theta_d, Z|\gamma_d, \pi) = \mathcal{D}ir(\theta_d|\gamma_d) \prod_{n=1}^{N_d} Q(z_{dn}|\pi_{z_{dn}w_n}).$$

از عبارت بالا لگاریتم طبیعی می‌گیریم:

$$\ln Q(\theta_d, Z|\gamma_d, \pi) = \ln \Gamma(\gamma_0) + \sum_{k=1}^{K} (\gamma_k - 1) \ln \theta_{dk} - \ln \Gamma(\gamma_k) + \sum_{n=1}^{N_d} \ln \pi_{z_{dn}w_n},$$

که در آن $\gamma_0 = \sum_{k=1}^{K} \gamma_k$.

آ-۳- توزیع تغییراتی $Q^*(Z)$

مطابق قضیهٔ **۱** داریم:

$$\begin{aligned} Q^*(z_{dn}) &\propto \exp\{\mathbb{E}_{Q(\theta)}[\![\ln P(W_d, \theta, Z, |\alpha, \phi)]\!]\} \\ &\propto \exp\{\mathbb{E}_{Q(\theta)}[\![\ln P(z_{dn}|\theta_{dz_{dn}}) + \ln P(w_{dn}|z_{dn}, \phi)]\!]\} \\ &= \exp\{\mathbb{E}_{Q(\theta)}[\![I[\![z_{dn} = k]\!] (\ln \theta_{dk} + \ln \phi_{kw_n})]\!]\} \\ &= (\phi_{kw_n} \exp\{\mathbb{E}_{Q(\theta)}[\![\ln \theta_{dk}]\!]\})^{I[\![z_{dn}=k]\!]}. \end{aligned}$$

هستهٔ توزیع حاصل در آخرین عبارت هم‌خانوادهٔ توزیع $P(z_{dn}|\theta_{dz_{dn}})$ با شاخص زیر است:

$$Q^*(z_{dn}) \propto \pi^*_{z_{dn}w_n}{}^{I[\![z_{dn}=k]\!]},$$

که در آن شاخص $\pi^*_{z_{dn}w_n}$ به صورت:

$$\begin{aligned} \pi^*_{z_{dn}w_n} &\propto \phi_{kw_n} \exp\{\mathbb{E}_{Q(\theta)}[\![\ln \theta_{dk}]\!]\} \\ &= \phi_{kw_n} \exp\{\Psi(\gamma_k) - \Psi(\gamma_0)\}, \end{aligned}$$

محاسبه می‌شود.

آ-۴- توزیع تغییراتی $Q^*(\theta_d)$

با استفاده از **قضیهٔ ۱** می‌نویسیم:

$$\begin{aligned} \ln Q^*(\theta_d) &= \mathbb{E}_{Q(Z)}[\![\ln P(W_d, \theta, \theta|\alpha, \phi)]\!] + C_1 \\ &= \mathbb{E}_{Q(Z)}\left[\!\!\left[\sum_{k=1}^{K} (\alpha_k - 1) \ln \theta_{dk} + \sum_{n=1}^{N_d} \ln \theta_{dz_{dn}}\right]\!\!\right] + C_2 \\ &= \sum_{k=1}^{K} (\alpha_k - 1) \ln \theta_{dk} + \ln \theta_{dk} \sum_{n=1}^{N_d} \mathbb{E}_{Q(Z)}[\![I[\![z_{dn} = k]\!]]\!] + C_2 \\ &= \sum_{k=1}^{K} \left(\alpha_k + \sum_{n=1}^{N_d} \pi_{kw_n} - 1\right) \ln \theta_{dk} + C_2. \end{aligned}$$

از عبارت آخر هستهٔ توزیع دیریکله $\mathcal{D}ir(\theta_d|\gamma)$ در آن مشخص است و شاخص γ_k به صورت:

$$\gamma_k = \alpha_k + \sum_{n=1}^{N_d} \pi_{kw_n},$$

تعریف می‌شود.

پیوست ب: شرح تفضیلی روش برآورد شاخص‌ها

در این بخش گام-M از الگوریتم برای برآورد شاخص ϕ را با روشی متفاوت از [۷] شرح می‌دهیم و در مسئلهٔ برآورد بیز تجربی برای برآورد شاخص α می‌توان به همان منبع مراجعه کرد.

ب-۱- به‌روزرسانی شاخص ϕ

در گام-M می‌خواهیم امید ریاضی لگاریتم احتمال توأم زیر را نسبت به ماتریس احتمال ϕ برای کل مستندات درون‌پیکره بیشینه کنیم:

$$\arg\max_{\phi} \mathbb{E}_{Q(Z)Q(\theta)}[\![\ln P(D,\theta,Z|\alpha,\phi)]\!],$$

که لگاریتم احتمال توأم به صورت زیر است:

$$\ln P(D,\theta,Z|\alpha,\phi) = \sum_{d=1}^{M} \ln \mathcal{D}ir(\theta_d|\alpha) + \sum_{n=1}^{N_d} \ln P(z_{dn}|\theta_d) + \ln P(w_{dn}|z_{dn},\phi)$$

$$= \sum_{d=1}^{M} \ln\Gamma(\alpha_0) + \sum_{k=1}^{K} (\alpha_k - 1)\ln\theta_{dk} - \ln\Gamma(\alpha_k) + \sum_{n=1}^{N_d}\sum_{k=1}^{K} \mathbb{I}[\![z_{dn}=k]\!]\,(\ln\theta_{dk} + \mathbb{I}[\![w_{dn}=v]\!]\ln\phi_{kv}).$$

با ثابت درنظرگرفتن عباراتی که به ϕ ارتباط ندارند، خواهیم داشت:

$$\arg\max_{\phi} \mathbb{E}_{Q(Z)Q(\theta)}\left[\!\!\left[\sum_{d=1}^{M}\sum_{n=1}^{N_d}\sum_{k=1}^{K} \mathbb{I}[\![z_{dn}=k]\!]\,\mathbb{I}[\![w_{dn}=v]\!]\ln\phi_{kv}\right]\!\!\right].$$

هر راه‌حلی برای این مسئله باید دو محدودیت زیر را ارضا کند:

۱. عناصر ماتریس ϕ باید مقادیر مثبت داشته‌باشد؛ یعنی $\forall k \in \{1,\cdots,K\} \& \forall v \in \{1,\cdots,V\} : \phi_{kv} > 0$.

۲. از آنجا که هر سطر ماتریس باید توزیع احتمال باشد، پس باید $\forall k \in \{1,\cdots,K\} : \sum_{v=1}^{V} \phi_{kv} = 1$.

محدودیت اول به‌دلیل‌این‌که از لگاریتم استفاده کردیم، پیشتر برآورده شده‌است. برای ارضای محدودیت دوم از ضرایب لاگرانژ استفاده می‌کنیم. برای این منظور باید حاصل جمع تمام محدودیت‌ها را به رابطهٔ بالا اضافه کنیم. در نهایت، به رابطهٔ زیر خواهیم‌رسید که باید آن را بیشینه کنیم:

$$L = \mathbb{E}_{Q(Z)Q(\theta)}\left[\!\!\left[\sum_{d=1}^{M}\sum_{n=1}^{N_d}\sum_{k=1}^{K} \mathbb{I}[\![z_{dn}=k]\!]\,\mathbb{I}[\![w_{dn}=v]\!]\ln\phi_{kv}\right]\!\!\right] + \sum_{k=1}^{K}\lambda_k\left(\sum_{v=1}^{V}\phi_{kv} - 1\right)$$

$$= \sum_{d=1}^{M}\sum_{n=1}^{N_d}\sum_{k=1}^{K} \mathbb{E}_{Q(Z)}[\![\mathbb{I}[\![z_{dn}=k]\!]]\!]\,\mathbb{I}[\![w_{dn}=v]\!]\ln\phi_{kv} + \sum_{k=1}^{K}\lambda_k\left(\sum_{v=1}^{V}\phi_{kv} - 1\right)$$

$$= \sum_{d=1}^{M}\sum_{n=1}^{N_d}\sum_{k=1}^{K} \pi^*_{dkv}\,\mathbb{I}[\![w_{dn}=v]\!]\ln\phi_{kv} + \sum_{k=1}^{K}\lambda_k\left(\sum_{v=1}^{V}\phi_{kv} - 1\right).$$

از L نسبت به ϕ_{kv} مشتق می‌گیریم و برابر صفر قرار می‌دهیم تا مقدار بیشینه به دست آید. پس داریم:

$$\frac{\delta L}{\delta\phi_{kv}} = \frac{\sum_{d=1}^{M}\sum_{n=1}^{N_d}\pi^*_{dkv}\,\mathbb{I}[\![w_{dn}=v]\!]}{\phi_{kv}} + \lambda_k = 0,$$

و نتیجه می‌گیریم:

$$\phi_{kv} = \frac{\sum_{d=1}^{M}\sum_{n=1}^{N_d}\pi^*_{dkv}\,\mathbb{I}[\![w_{dn}=v]\!]}{-\lambda_k},$$

که در آن $-\lambda_k$ همان ثابت نرمال‌ساز است و به صورت $-\lambda_k = \sum_{v=1}^{V}\sum_{d=1}^{M}\sum_{n=1}^{N_d}\pi^*_{dkv}\,\mathbb{I}[\![w_{dn}=v]\!]$ تعریف می‌شود. پس داریم:

$$\phi_{kv} \propto \sum_{d=1}^{M}\sum_{n=1}^{N_d}\pi^*_{dkv}\,\mathbb{I}[\![w_{dn}=v]\!].$$

پیوست- ج: شرح تفضیلی روش تکرارشوندهٔ بیز تغییراتی (IVB)

در این پیوست به مراحل نوشتن الگوریتم تکرارشوندهٔ بیز تغییراتی می‌پردازیم که در بخش ۴ ارائه شد و با استفاده از **قضیهٔ (۱)** قواعد به‌روزرسانی شاخص‌های تغییراتی را به دست می‌آوریم. در الگوریتم IVB به‌طور مشابه با الگوریتم EM-تغییراتی گام‌های زیر نیاز است:

۱. نوشتن احتمال توأم شاخص‌ها و محاسبهٔ لگاریتم آن،
۲. افراز شاخص‌ها و مشخص‌کردن توزیع‌های تغییراتی،
۳. به‌روزرسانی شاخص‌های تغییراتی از توزیع‌های $Q^*(\theta_i|X)$ که طبق **قضیهٔ (۱)** مشخص شده‌اند.

تفاوت IVB با EM-تغییراتی در این است که با شاخص‌های الگو همانند متغیرهای پنهان رفتار می‌شود، در نتیجه، برای آنها توزیع‌های تغییراتی درنظرمی‌گیریم؛ سپس، قواعد به‌روزرسانی شاخص‌های تغییراتی برای آنها محاسبه می‌شوند. به همین دلیل دیگر نیازی به مرحله‌ای جداگانه برای گام-M نداریم. در ادامه، این مراحل را تشریح می‌کنیم.

ج-۱: لگاریتم احتمال توأم شاخص‌ها

احتمال توأم شاخص‌ها به صورت زیر است:

$$P(D,\theta,Z,\phi|\alpha,\beta)=\prod_{k=1}^{K}\mathcal{D}ir(\phi_k|\beta_k)\prod_{d=1}^{M}\mathcal{D}ir(\theta_d|\alpha)\prod_{n=1}^{N_d}P(z_{dn}|\theta_{dz_{dn}})\,P(w_{dn}|z_{dn},\phi_{z_{dn}w_{dn}}).$$

از این رابطه لگاریتم طبیعی می‌گیریم، پس داریم:

$$\begin{aligned}\ln P(D,\theta,Z,\phi|\alpha,\beta)&=\sum_{k=1}^{K}\ln\Gamma(\beta_{k0})+\sum_{v=1}^{V}(\beta_{kv}-1)\ln\phi_{kv}-\ln\Gamma(\beta_{kv})\\&+\sum_{d=1}^{M}\ln\Gamma(\alpha_0)+\sum_{k=1}^{K}(\alpha_k-1)\ln\theta_{dk}-\ln\Gamma(\alpha_k)+\sum_{n=1}^{N_d}\ln\theta_{dz_{dn}}+\ln\phi_{z_{dn}w_n},\end{aligned}$$

که در آن $\alpha_0=\sum_{k=1}^{K}\alpha_k$ و $\beta_{k0}=\sum_{v=1}^{V}\beta_{kv}$.

ج-۲: لگاریتم توزیع‌های تغییراتی

توزیع‌های تغییراتی را به صورت زیر تعریف می‌کنیم:

$$Q(\theta,Z,\phi|\gamma,\pi,\lambda)=\prod_{k=1}^{K}\mathcal{D}ir(\phi_k|\beta_k)\prod_{d=1}^{M}\mathcal{D}ir(\theta_d|\gamma_d)\prod_{n=1}^{N_d}Q(z_{dn}|\pi_{z_{dn}w_n}).$$

از عبارت بالا لگاریتم طبیعی می‌گیریم:

$$\begin{aligned}\ln Q(\theta,Z,\phi|\gamma,\pi,\lambda)&=\sum_{k=1}^{K}\ln\Gamma(\beta_{k0})+\sum_{v=1}^{V}(\beta_{kv}-1)\ln\phi_{kv}-\ln\Gamma(\beta_{kv})\\&+\sum_{d=1}^{M}\ln\Gamma(\gamma_0)+\sum_{k=1}^{K}(\gamma_k-1)\ln\theta_{dk}-\ln\Gamma(\gamma_k)+\sum_{n=1}^{N_d}\ln\pi_{z_{dn}w_n},\end{aligned}$$

که در آن $\gamma_0=\sum_{k=1}^{K}\gamma_k$ و $\beta_{k0}=\sum_{v=1}^{V}\beta_{kv}$.

ج-۳: توزیع تغییراتی $Q^*(Z)$

مانند قبل مطابق **قضیهٔ (۱)** داریم:

$$\begin{aligned}Q^*(z_{dn})&\propto\exp\{\mathbb{E}_{Q(\theta)Q(\phi)}[\![\ln P(D,\theta,Z,\phi|\alpha,\beta)]\!]\}\\&\propto\exp\{\mathbb{E}_{Q(\theta)Q(\phi)}[\![\ln P(z_{dn}|\theta_{dz_{dn}})+\ln P(w_{dn}|z_{dn},\phi)]\!]\}\\&=\exp\{\mathbb{E}_{Q(\theta)Q(\phi)}[\![\mathbb{I}[\![z_{dn}=k]\!]\,(\ln\theta_{dk}+\ln\phi_{kw_n})]\!]\}\\&=\left(\exp\{\mathbb{E}_{Q(\theta)}[\![\ln\theta_{dk}]\!]+\mathbb{I}[\![w_{dn}=v]\!]\,\mathbb{E}_{Q(\phi)}[\![\ln\phi_{kv}]\!]\}\right)^{\mathbb{I}[\![z_{dn}=k]\!]}.\end{aligned}$$

هستهٔ توزیع حاصل در آخرین عبارت هم‌خانوادهٔ توزیع $P(z_{dn}|\theta_{dz_{dn}})$ به شکل زیر است:

$$Q^*(z_{dn})\propto{\pi^*_{kw_{dn}}}^{\mathbb{I}[\![z_{dn}=k]\!]},$$

که در آن شاخص $\pi^*_{kw_{dn}}$ به صورت:

$$\begin{aligned}\pi^*_{kw_{dn}}&\propto\exp\{\mathbb{E}_{Q(\theta)}[\![\ln\theta_{dk}]\!]+\mathbb{I}[\![w_{dn}=v]\!]\,\mathbb{E}_{Q(\phi)}[\![\ln\phi_{kv}]\!]\}\\&=\exp\{\Psi(\gamma_k)-\Psi(\gamma_0)+\mathbb{I}[\![w_{dn}=v]\!](\Psi(\lambda_{kv})-\Psi(\lambda_{k0}))\},\end{aligned}$$

محاسبه می‌شود.

ج-۴: توزیع تغییراتی $Q^*(\theta_d)$

با استفاده از **قضیهٔ (۱)** می‌نویسیم:

$$\ln Q^*(\theta_d) = \mathbb{E}_{Q(Z)Q(\phi)}[\![\ln P(D,\theta,Z,\phi|\alpha,\beta)]\!] + C_1$$
$$= \mathbb{E}_{Q(Z)Q(\phi)}\left[\!\!\left[\sum_{k=1}^{K}(\alpha_k-1)\ln\theta_{dk} + \sum_{n=1}^{N_d}\ln\theta_{dz_{dn}}\right]\!\!\right] + C_2$$
$$= \sum_{k=1}^{K}(\alpha_k-1)\ln\theta_{dk} + \ln\theta_{dk}\sum_{n=1}^{N_d}\mathbb{E}_{Q(Z)}\left[\!\left[\mathbb{I}[\![z_{dn}=k]\!]\right]\!\right] + C_2$$
$$= \sum_{k=1}^{K}\left(\alpha_k + \sum_{n=1}^{N_d}\pi^*_{kw_n} - 1\right)\ln\theta_{dk} + C_2.$$

در عبارت آخر هستهٔ توزیع دیریکله $\mathcal{D}\mathrm{ir}(\theta_d|\gamma)$ مشخص است و شاخص γ_k به صورت:

$$\gamma_k = \alpha_k + \sum_{n=1}^{N_d}\pi^*_{kw_n},$$

تعریف می‌شود.

ج-۵: توزیع تغییراتی $Q^*(\phi)$

برای به‌دست‌آوردن شاخص توزیع تغییراتی شاخص ϕ با استفاده از **قضیهٔ (۱)** می‌نویسیم:

$$\ln Q^*(\phi_k) = \mathbb{E}_{Q(Z)Q(\theta)}[\![\ln P(D,\theta,Z,\phi|\alpha,\beta)]\!] + C_1$$
$$= \mathbb{E}_{Q(Z)Q(\theta)}\left[\!\!\left[\sum_{v=1}^{V}(\beta_{kv}-1)\ln\phi_{kv} + \sum_{d=1}^{M}\sum_{n=1}^{N_d}\mathbb{I}[\![z_{dn}=k]\!]\mathbb{I}[\![w_{dn}=v]\!]\ln\phi_{kv}\right]\!\!\right] + C_2$$
$$= \left(\beta_{kv} + \sum_{d=1}^{M}\sum_{n=1}^{N_d}\mathbb{E}_{Q(Z)}\left[\!\left[\mathbb{I}[\![z_{dn}=k]\!]\right]\!\right]\mathbb{I}[\![w_{dn}=v]\!] - 1\right)\ln\phi_{kv} + C_2$$
$$= \left(\beta_{kv} + \sum_{d=1}^{M}\sum_{n=1}^{N_d}\pi^*_{kw_{dn}}\mathbb{I}[\![w_{dn}=v]\!] - 1\right)\ln\phi_{kv} + C_2,$$

در عبارت آخر هستهٔ توزیع دیریکله به صورت $\mathcal{D}\mathrm{ir}(\phi_k|\lambda^*_k)$ قابل تشخیص است که شاخص تغییراتی λ^*_{kv} به صورت زیر تعریف می‌شود:

$$\lambda^*_{kv} = \beta_{kv} + \sum_{d=1}^{M}\sum_{n=1}^{N_d}\pi^*_{kw_{dn}}\mathbb{I}[\![w_{dn}=v]\!].$$

پیوست د: کران پایین لگاریتم شواهد

در این پیوست قضیهٔ کران پایین لگاریتم شواهد را شرح می‌دهیم که در [۷] از روش بهینه‌سازی کران پایین لگاریتم شواهد برای به‌دست‌آوردن قواعد به‌روزرسانی شاخص‌های تغییراتی استفاده شده‌است.

قضیهٔ (۲) (کران پایین لگاریتم شواهد). بیشینه‌کردن کران پایین لگاریتم شواهد، معادل کمینه‌کردن واگرایی کولبک-لایبلر بین توزیع‌های تغییراتی و توزیع پسینی متغیرهای پنهان الگوست.

اثبات: واگرایی کولبک-لایبلر را بین توزیع‌های پسینی متغیرهای پنهان می‌نویسیم:

$$
\begin{aligned}
&\mathrm{KL}[Q(\theta, Z|D) \parallel P(\theta, Z|D, \alpha, \phi)] = \\
&= \sum_{Z} \int_{\Theta} Q(\theta, Z|D) \ln \frac{Q(\theta, Z|D)}{P(\theta, Z|D, \alpha, \phi)} d\theta \\
&= \sum_{Z} \int_{\Theta} Q(\theta, Z|D) \ln \frac{Q(\theta, Z|D) P(D|\alpha, \phi)}{P(D, \theta, Z|\alpha, \phi)} d\theta \\
&= \sum_{Z} \int_{\Theta} Q(\theta, Z|D) \ln \frac{Q(\theta, Z|D)}{P(D, \theta, Z|\alpha, \phi)} d\theta + \sum_{Z} \int_{\Theta} Q(\theta, Z|D) \ln P(D|\alpha, \phi)\, d\theta \\
&= \underbrace{\sum_{Z} \int_{\Theta} Q(\theta, Z|D) \ln \frac{Q(\theta, Z|D)}{P(D, \theta, Z|\alpha, \phi)} d\theta}_{-\mathbb{E}_{Q(\theta, Z|D)}\left[\ln \frac{P(D, \theta, Z|\alpha, \phi)}{Q(\theta, Z|D)}\right]} + \ln P(D|\alpha, \phi) \underbrace{\sum_{Z} \int_{\Theta} Q(\theta, Z|D)\, d\theta}_{1},
\end{aligned}
$$

با مرتب کردن داریم:

$$\ln P(D|\alpha, \phi) = \mathrm{KL}[Q(\theta, Z|D) \parallel P(\theta, Z|D, \alpha, \phi)] + \varepsilon[Q(\theta, Z|D); \alpha, \beta],$$

که در آن $\varepsilon[Q(\theta, Z|D); \alpha, \beta] = \mathbb{E}_{Q(\theta, Z|D)}\left[\ln \frac{P(D, \theta, Z|\alpha, \phi)}{Q(\theta, Z|D)}\right]$. از طرفی می‌توان نوشت:

$$
\begin{aligned}
&\ln P(D|\alpha, \phi) = \ln \sum_{Z} \int_{\Theta} P(D, \theta, Z|\alpha, \phi)\, d\theta \\
&= \ln \sum_{Z} \int_{\Theta} Q(\theta, Z|D) \frac{P(D, \theta, Z|\alpha, \phi)}{Q(\theta, Z|D)} d\theta \\
&= \ln \mathbb{E}_{Q(\theta, Z|D)}\left[\frac{P(D, \theta, Z|\alpha, \phi)}{Q(\theta, Z|D)}\right].
\end{aligned}
$$

با استفاده از نابرابری جنسن، لگاریتم طبیعی را وارد امید ریاضی می‌کنیم. پس خواهیم‌داشت:

$$\ln P(D|\alpha, \phi) \geq \mathbb{E}_{Q(\theta, Z|D)}\left[\ln \frac{P(D, \theta, Z|\alpha, \phi)}{Q(\theta, Z|D)}\right] = \varepsilon[Q(\theta, Z|D); \alpha, \beta].$$

در واقع عبارت $\varepsilon[Q(\theta, Z|D); \alpha, \beta]$ یک کران پایین برای لگاریتم شواهد است. پس می‌توان نتیجه گرفت بیشینه‌کردن کران پایین لگاریتم شواهد معادل کمینه‌کردن واگرایی $\mathrm{KL}[Q(\theta, Z|D) \parallel P(\theta, Z|D, \alpha, \phi)]$ است.

در [۱۰] و نیز در بخش پیوست‌های [۷] نشان داده شده‌است که با استفاده از نتیجهٔ به‌دست‌آمده از **قضیهٔ (۲)** و بیشینه‌سازی کران پایین لگاریتم شواهد، یعنی $\varepsilon[Q(\theta, Z|D); \alpha, \beta]$، به همان روابط به‌روزرسانی برای شاخص‌های تغییراتی می‌رسیم که ما در این پیوست با روش دیگری اثبات کردیم.