UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação
ISSN 0103-2569

MODELOS PROBABILÍSTICOS DE TÓPICOS: DESVENDANDO O LATENT DIRICHLET ALLOCATION

THIAGO DE PAULO FALEIROS
ALNEU DE ANDRADE LOPES

Nº 409

RELATÓRIOS TÉCNICOS

São Carlos – SP
Abr./2016

Instituto de Ciências Matemáticas e de Computação – ICMC-USP

Modelos probabilísticos de tópicos: desvendando o Latent Dirichlet Allocation

Relatório Técnico

Thiago de Paulo Faleiros
Alneu de Andrade Lopes

USP – São Carlos
Abril de 2016

Agradecimento à Fundação de Amparo à Pesquisa do
Estado de São Paulo (FAPESP), processo nro. 2011/23689-9

RESUMO

FALEIROS, T. P.. **Modelos probabilísticos de tópicos: desvendando o Latent Dirichlet Allocation**. 2016. – Instituto de Ciências Matemáticas e de Computação (ICMC/USP), São Carlos – SP.

Modelos de tópicos têm ganhado bastante atenção e sido alvo de várias pesquisas. A ideia básica dos modelos de tópicos é descobrir, nas relações entre documentos e termos, padrões latentes que sejam significativos para o entendimento dessas relações. Por exemplo, tais modelos podem ranquear um conjunto de termos como importantes para um ou mais temas. Bem como ranquear documentos como tendo relevância para um ou mais temas. Neste trabalho, são descritos os modelos probabilísticos de tópicos, tendo como base o modelo *Latent Dirichlet Allocation* (LDA). O LDA é formalmente descrito por um processo generativo para a criação de documentos textuais dadas distribuições latentes de tópicos. Do ponto de vista computacional, são detalhadamente descritos os dois principais algoritmos de inferência do LDA, o método de Gibbs e o método de inferência variacional. Também é descrita a versão *online* do algoritmo de inferência variacional para o LDA. Além disso, são apresentados os procedimento de avaliação para os modelos probabilísticos de tópicos. Uma outra contribuição desse estudo é a análise comparativa entre o problema estabelecido pela fatoração de matrizes não negativas e o problema de otimização estabelecido pelo método de inferência variacional. Esse trabalho pode servir para auxiliar na exploração, extensão e desenvolvimento de novas abordagens de modelos probabilísticos de tópicos.

Palavras-chave: Latent Dirichlet Allocation, Modelos Probabilísticos de Tópicos, Algoritmos de Inferência, Extração de Tópicos.

SUMÁRIO

CAPÍTULO

1

INTRODUÇÃO

Motivados pela necessidade de técnicas eficientes para extração de informações em texto, uma nova área de pesquisa surgiu em 2003, chamada de modelos probabilísticos de tópicos (*Probabilistic Topic Models*) (BLEI, 2011). O início dessa área se deu basicamente com a apresentação do LDA (*Latent Dirichlet Allocation*) (BLEI; NG; JORDAN, 2003) – LDA é o modelo base. Modelos probabilísticos de tópicos (BLEI; NG; JORDAN, 2003; GRIFFITHS; STEYVERS, 2004; STEYVERS; GRIFFITHS, 2007; HOFMANN, 1999) são um conjunto de algoritmos cujo objetivo é descobrir estruturas temáticas ocultas em grandes coleções de documentos. Inicialmente, esses modelos foram propostos para serem aplicados em documentos textuais, mas logo foram explorados em outros tipos de dados com atributos discretos, como imagens (LI; PERONA, 2005; SIVIC *et al.*, 2005; RUSSELL *et al.*, 2006; CAO; LI, 2007), grafos (HENDERSON; ELIASSI-RAD, 2009; BRONIATOWSKI; MAGEE, 2010; CHANG; BLEI, 2009; MEI *et al.*, 2008) e outros. Como a base do modelo e a descrição teórica é fundamentada em documentos e palavras, neste trabalho não é explorada a aplicação em outros tipos de dados. Porém, a transição para outros tipos de dados é direta uma vez que se entenda como o modelo é aplicado em texto.

A exploração de grandes volumes de dados é simplificada pelos modelos probabilísticos na descoberta dos *tópicos*. Os tópicos são estruturas com valor semântico e que, no contexto de mineração de texto, formam grupos de palavras que frequentemente ocorrem juntas. Esses grupos de palavras quando analisados, dão indícios a um tema ou assunto que ocorre em um subconjunto de documentos. A expressão **tópico** é usada levando-se em conta que o assunto tratado em uma coleção de documentos é extraído automaticamente, ou seja, tópico é definido como um conjunto de palavras que frequentemente ocorrem em documentos semanticamente relacionados. Imagine vários discursos proferido por políticos e transcritos como documentos textuais. Se a cada dia vários documentos são gerados, ao longo dos anos essa coleção de documentos aumentará, inviabilizando o gerenciamento manual. Aplicando uma técnica como o LDA, é possível organizar e agrupar um subconjunto de discursos pelos seus respectivos temas.

Pesquisadores da área de recuperação de informação já propuseram várias técnicas para reduzir o tamanho dos descritores de uma coleção de documentos. Entre as técnicas mais notáveis está o *Latent Semantic Indexing* (LSI) (DEERWESTER *et al.*, 1990; BERRY; DUMAIS; O'BRIEN, 1995). O LSI usa a decomposição em valores singulares de uma matriz documento-termo para identificar um subespaço linear que apresenta uma maior variação no espaço de características. Conhecendo a funcionalidade do LSI, é possível estender essa técnica para um modelo generativo probabilístico. Fazendo isso, Hofmann (1999) propôs o *probabilistic Latent Semantic Indexing* (pLSI). O pLSI é um modelo probabilístico com habilidade de recuperar aspectos de coleções de documentos. No modelo pLSI, cada palavra em um documento é amostrada de uma variável aleatória que representa um tópico. Assim, cada palavra em um documento é gerada por um tópico, e cada documento possui palavras geradas por diferentes tópicos. Isso faz com que um documento possua diferentes proporções de tópicos. Apesar do pLSI ser um modelo probabilístico, ele não é um modelo gerador de documentos completo pois não provê um modelo probabilístico no nível dos documentos. Ou seja, apesar das palavras serem geradas por variáveis aleatórias obedecendo uma distribuição multinomial, os documentos são apenas *bag-of-words*.

O modelo pLSI foi estendido para o modelo LDA (*Latent Dirichlet Allocation*). O LDA é um modelo bayesiano completo e se baseia na geração dos tópicos como distribuições de Dirichlet. Em comparação ao pLSI, o LDA descreve um modelo capaz de classificar documentos não conhecidos (documentos que não foram utilizados no treinamento), e utilizar informações *a priori*. Por essas características, o LDA tem influenciado uma grande quantidade de trabalhos e se tornado a base dos modernos modelos estatísticos de aprendizado de máquina, resultando em uma nova classe de modelos estatísticos chamados *Modelos Probabilísticos de Tópicos*.

O modelo LDA especifica um simples procedimento probabilístico no qual uma coleção de documentos pode ser gerada. Para criar um novo documento, inicialmente, escolhe-se uma distribuição de tópicos. Em seguida, para cada palavra nesse documento, escolhe-se um tópico aleatoriamente de acordo com essa distribuição. A palavra é amostrada de acordo com o tópico escolhido.

O processo inverso da geração de documentos é descobrir a distribuição de tópicos que geraram uma coleção de documentos. Esse processo está relacionada com a inferência do modelo probabilístico. Os algoritmos para inferência de modelos probabilísticos de tópicos são métodos estatísticos que analisam as palavras do texto original para descobrir os tópicos. Esses algoritmos são não supervisionados – os tópicos "emergem" da análise dos textos originais (STEYVERS; GRIFFITHS, 2007).

Neste trabalho é descrito o modelo base e referência para o desenvolvimento de modelos probabilístico de tópicos, o LDA. No Capítulo 2 é descrita a formulação e o processo gerador do modelo LDA. No Capítulo 3.1 são apresentadas as principais técnicas de inferência probabilística desse modelo: método de amostragem de Gibbs e o método de inferência variacional. No Capítulo

4 é apresentada a versão *online* do LDA. No Capítulo 5 são descritas formas de avaliação dos modelos probabilísticos de tópicos. No Capítulo 6 é realizada uma análise comparativa do LDA com o NMF, onde são apontadas similaridades dessas duas técnicas. E por fim, no Capítulo 7 são apresentadas as conclusões deste trabalho.

CAPÍTULO

2

LATENT DIRICHLET ALLOCATION (LDA)

Quando se discute sobre modelos probabilísticos de tópicos, o que se encontra na literatura como estado da arte é o modelo LDA. O LDA é um modelo probabilístico generativo para coleções de dados discretos como corpus de documentos (BLEI; NG; JORDAN, 2003). Um modelo generativo é aquele que aleatoriamente gera os dados a partir das variáveis latentes. Assim, o LDA não é um algoritmo com descrições sequenciais de instruções para encontrar tópicos dada uma coleção de documentos. O LDA é um modelo probabilístico no qual é descrito como os documentos são gerados. Nesse modelo, as variáveis observáveis são os termos de cada documento e as variáveis não observáveis são as distribuições de tópicos. Os parâmetros das distribuições de tópicos, conhecidos como hiper-parâmetros, são dados *a priori* no modelo.

A distribuição utilizada para amostrar a distribuição de tópicos é a distribuição de Dirichlet. No processo generativo, o resultado da amostragem da Dirichlet é usado para alocar as palavras de diferentes tópicos e que preencherão os documentos. Assim, pode-se perceber o significado do nome *Latent Dirichlet Allocation*, que expressa a intenção do modelo de alocar os tópicos latentes que são distribuídos obedecendo a distribuição de Dirichlet.

A função de densidade da distribuição de Dirichlet, denotada como $Dir(z,\alpha)$, é a seguinte:

$$Dir(z,\alpha) = \frac{1}{B(\alpha)} \prod_{k=1}^{K} z_k^{\alpha_k - 1}, \tag{2.1}$$

onde $z = (z_1, \ldots, z_K)$ é uma variável K-dimensional, $0 \leq z \leq 1$ e $\sum_{i=1}^{K} z_i = 1$. Aqui, $\alpha = (\alpha_1, \ldots, \alpha_K)$ são os hiper-parâmetros da distribuição. A função $B(\alpha)$ é a função Beta, na qual pode ser expressa em termos da função gama Γ:

$$B(\alpha) = \frac{\prod_{k=1}^{K} \Gamma(\alpha_k)}{\Gamma\left(\sum_{k=1}^{K} \alpha_k\right)}. \tag{2.2}$$

A distribuição de Dirichlet tem algumas propriedades importantes (BISHOP, 2006), e por isso são comumente usadas em estatística Bayesiana [1].

O processo gerador do LDA é um processo imaginário e, inverso ao que é proposto em uma tarefa computacional de extração de informações, assume-se que os tópicos são especificados antes que qualquer dado seja gerado. Aqui, os tópicos são definidos como distribuições de probabilidade sobre um vocabulário fixo de palavras. Enquanto que os documentos, nada mais do que *bag-of-words*, surgem da escolha aleatória das palavras pertencentes a uma distribuições de tópicos.

Pode-se detalhar o processo gerador do modelo LDA. Para isso, deve-se assumir que um documento d_j é criado da seguinte forma:

1. Crie as distribuições $\phi_k \sim Dir(\phi_k, \beta)$ para todo tópico k, como $0 \leq k \leq K$.
2. Crie uma distribuição $\theta_j \sim Dir(\theta, \alpha)$ para o documento d_j.
3. Para cada posição i das palavras no documento d_j,
   a) Escolha aleatoriamente um tópico $z_{j,i} \sim Multinomial(\theta_j)$.
   b) Escolha aleatoriamente uma palavra $w_{j,i}$ com probabilidade $p(w_{j,i}|\phi_{z_{j,i}})$.

No processo gerador, inicialmente, são utilizadas duas variáveis para representar as distribuições. A variável ϕ é uma variável n-dimensional, onde n é o número de palavras do vocabulário. A variável θ é uma variável K-dimensional, onde o valor de K é o número de tópicos. Essas variáveis descrevem distribuições de probabilidades, logo $\sum_j^n \phi_j = 1$, $\phi_i > 0$, e $\sum_i^K \theta_i = 1$, $\theta_i > 0$. Essas duas variáveis são geradas pela distribuição de Dirichlet (*Dir*) com seus respectivos hiper-parâmetros β e α.

Em seguida, com as distribuições ϕ e θ_j, é gerado o documento d_j. No modelo LDA um documento é simplesmente uma *bag-of-words*, com n_{d_j} termos em um documento d_j. Os termos de uma *bag-of-words* são palavras do vocabulário, e ocasionalmente podem ocorrer repetições de uma mesma palavra. Para cada posição i, das n_{d_j} posições de termos de uma *bag-of-words*, é escolhida uma palavra da distribuição de tópicos. Para isso, deve-se escolher um tópico k dos K tópicos existentes, e associar esse tópico a posição i do documento d_j. A variável $z_{j,i}$ armazenará o tópico escolhido. O tópico é escolhido obedecendo a distribuição θ_j, que informa a participação dos tópicos no documento d_j especificadamente. Em seguida, é escolhido da distribuição ϕ a palavra que irá preencher a posição i. A variável ϕ são K distribuições n-dimensionais, onde cada distribuição k, ϕ_k, corresponde a proporções de palavras que semanticamente descrevem o assunto do qual o tópico k trata. Assim, o termo $w_{j,i}$ deve ser escolhido do tópico $z_{j,i}$, obedecendo a distribuição de palavras $\phi_{z_{j,i}}$.

[1] A distribuição de Dirichlet é a *priori* conjugada da distribuição multinomial

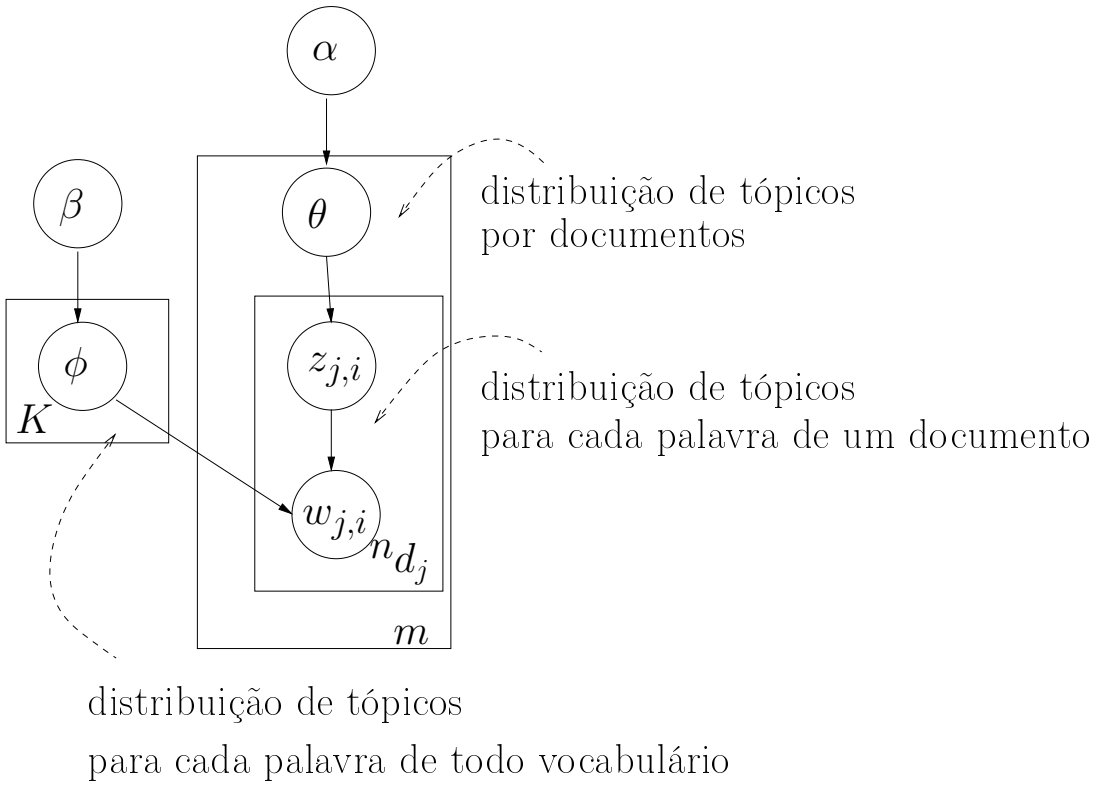
Uma característica importante do LDA é que cada documento possui sua própria distribuição de tópicos θ_j. Assim, um mesmo documento pode estar relacionado com vários tópicos com diferentes proporções de relevâncias. Pode-se perceber isso no modelo generativo pela escolha do tópico atribuído a variável $z_{j,i}$, onde ocasionalmente existirá a chance da escolha de diferentes tópicos segundo a distribuição θ_j.

Todo o processo generativo pode ser representado de forma gráfica por meio de uma rede Bayesiana. Essa rede é ilustrada na Figura 1. Nessa rede, cada vértice corresponde a uma variável e a aresta à relação de dependência. Na notação do modelo gráfico, em vez de ilustrar cada variável repetidas vezes, um retângulo é usado para agrupar variáveis em um subgrafo que se repete. O número de repetições é rotulado na parte inferior de cada retângulo. Os itens abaixo descrevem a notação utilizada na Figura 1.

- K – Número de tópicos
- n – Número de palavras do Vocabulário.
- m – Número de documentos.
- n_{d_j} – Número de palavras em um documento d_j, onde $1 \leq j \leq m$.
- θ – Distribuição de tópicos por documentos.
- ϕ – Distribuição dos tópicos sobre as palavras de todo o vocabulário.
- θ_j – Vetor com a proporção dos tópicos para o documento d_j, onde $1 \leq j \leq m$.
- ϕ_k – Vetor com a proporção das palavras do vocabulário para o tópico k, onde $1 \leq k \leq K$.
- α – Priore da distribuição de Dirichlet, relacionada a distribuição documento-termo.
- β – Priore da distribuição de Dirichlet, relacionada a distribuição tópico-palavra.
- w_i – i-ésima palavra do vocabulário, onde $1 \leq i \leq n$.
- $w_{j,i}$ – palavra w_i observada no documento d_j, onde $1 \leq j \leq m$ e $1 \leq i \leq n$.
- $z_{j,i}$ – Distribuição de tópicos associado a palavra $w_{j,i}$ no documento d_j, onde $1 \leq j \leq m$ e $1 \leq i \leq n$.

O modelo Bayesiano do LDA é um modelo hierárquico com três níveis (veja a Figura 1). O primeiro nível representa a distribuição de tópicos em toda a coleção de documentos. No segundo nível, tem-se a distribuição dos tópicos para cada documento. E o último nível, repete-se a distribuição dos tópicos internamente para as palavras em um documento. Com o último nível, tona-se possível representar um documento como uma mistura de tópicos.

Figura 1 – Modelo Gráfico do LDA.



Fonte: Adaptada de Blei, Ng e Jordan (2003).

Utilizando a Figura 1 para esclarecer a representação do modelo, percebe-se que no nível de toda a coleção de documentos estão os hiper-parâmetros α e β. De forma simples, sem o formalismo matemático, pode-se dar uma interpretação para esses parâmetros. Um alto valor de α significa que cada documento provavelmente conterá uma maior mistura de tópicos. Um valor baixo para α indica maior probabilidade dos documentos conterem mistura de poucos tópicos, fazendo uma maior concentração em poucos tópicos. Da mesma forma, um valor alto para β significa que cada tópico terá alta probabilidade de conter misturas de várias palavras. Enquanto que um valor baixo para β indica que o tópico será formado por poucas palavras.

No nível de todo o vocabulário de palavras está a variável ϕ_k, que é amostrada para cada tópico k. Cada vetor ϕ_k forma uma matriz ϕ de tamanho $n \times K$, onde cada linha corresponde às palavras do vocabulário e as colunas aos tópicos. O valor de $\phi_{k,i}$ é a proporção do tópico k para uma palavra w_i.

No nível dos documentos, está a variável θ_j, que é amostrada para cada documento da coleção. Pode-se interpretar essa distribuição de documentos por tópicos como uma matriz θ de tamanho $m \times K$, onde cada linha são os documentos e as colunas os tópicos. Uma linha dessa matriz, referenciada como θ_j, corresponde a proporção de tópicos para um documento d_j da coleção.

No nível das palavras estão as variáveis $z_{j,i}$ e $w_{j,i}$, essas variáveis são amostradas para

cada palavra w_i em cada documento d_j. A variável $z_{j,i}$ é a atribuição de um tópico k ($1 \leq k \leq K$) para uma palavra w_i de um documento d_j.

Aqui, para dar um maior entendimento e também já conhecendo a descrição gráfica do LDA (Figura 1), vamos reescrever o processo generativo, só que dessa vez apresentando os passos para a geração de toda a coleção de documentos. Assim, tem-se o processo generativo do LDA com os seguintes passos:

1. Amostre K multinomiais $\phi_k \sim Dir(\phi_k, \beta)$, um para cada tópico k.
2. Amostre m multinomiais $\theta_j \sim Dir(\theta_j, \alpha)$, um para cada documento d_j.
3. Para cada documento d_j da coleção
    a) Para cada palavra w_i do documento d_j:
        i. Associe um tópico para $z_{j,i}$ amostrado da distribuição de Dirichlet θ_j.
        ii. Amostre uma palavra w_i da distribuição $\phi_{z_{j,i}}$.

Com base no processo generativo, e observando a relação de dependência existente entre as variáveis do modelo, é possível descrever a probabilidade de todas as variáveis latentes do modelo dado as informações *a priori* (BLEI, 2011). Transcrevendo essas probabilidades, tem-se a seguinte distribuição conjunta:

$$p(z, w, \phi, \theta | \alpha, \beta) = \prod_{k=1}^{K} p(\phi_k|\beta) \prod_{j=1}^{M} p(\theta_j|\alpha) \left(\prod_{i=1}^{V} p(z_{j,i}|\vec{\theta}_j) p(w_{i,j}|z_{i,j}, \phi_{z_{j,i}}) \right). \tag{2.3}$$

Essa equação determina uma distribuição de probabilidade com um complexo número de dependências. Por exemplo, a atribuição de tópico $z_{j,i}$ depende da distribuição dos tópicos por documento θ_j, e a palavra observada $w_{j,i}$ depende da atribuição do tópico $z_{j,i}$ e da proporção dessa palavra na distribuição $\phi_{z_{j,i}}$.

Levando em consideração as variáveis observadas e não observadas, almeja-se descobrir as atribuições de tópicos para os documentos e as distribuições de documentos por tópicos e tópicos por termos. Ou seja, o grande problema computacional do LDA é inferir $p(z, \phi, \theta, |w, \alpha, \beta)$, onde w são todas as palavras observadas na coleção de documentos.

Pelo teorema de Bayes, pode-se formular a probabilidade de $p(z, \phi, \theta, |w, \alpha, \beta)$ como o cálculo da *a posteriori* do LDA. Dessa forma, tem-se

$$p(z, \phi, \theta | w, \alpha, \beta) = \frac{p(z, w, \phi, \theta | \alpha, \beta)}{p(w)}. \tag{2.4}$$

O numerador é a distribuição conjunta (Equação 2.3) do modelo e o denominador é a probabilidade marginal dos dados observados.

Logo, o problema computacional central pode ser resolvido inferindo a probabilidade *a posteriori* de todo o modelo, descrito na Equação 2.4. Isso pode ser pensado como o inverso

do processo generativo. Teoricamente, esse cálculo de inferência pode ser feito pela soma da distribuição conjunta de todos os valores possíveis atribuídos as variáveis não observadas (todas as palavras da coleção). Entretanto, o número de atribuições possíveis é exponencialmente grande, fazendo esse cálculo intratável computacionalmente (BLEI, 2011). Apesar disso, existem vários métodos para aproximar a distribuição *a posteriori*. Entre os métodos mais utilizados na literatura para inferência do modelo LDA estão o *Gibbs Sampling* (Amostrador de Gibbs) (GRIFFITHS; STEYVERS, 2004) e *Variational Inference* (Inferência Variacional) (BLEI; NG; JORDAN, 2003).

No próximo capítulo são discutidos os métodos de inferência. Inicialmente, é descrito o Amostrador de Gibbs e como aplicá-lo no caso do LDA. Em seguida é descrito o método de inferência variacional e sua aplicação no LDA.

CAPÍTULO

3

MÉTODOS DE INFERÊNCIA PARA LDA

Como descrito no capítulo anterior, o LDA trata os dados como observações que surgiram de um processo generativo que inclui variáveis ocultas. Para os documentos, essas variáveis ocultas são as estruturas temáticas da coleção. O problema computacional definido pelo LDA está na inferência dessas estruturas temáticas, nas quais correspondem as distribuições de probabilidades relacionadas as relações documentos-tópicos e tópicos-palavras.

Neste capítulo são descritos os algoritmos de inferência do modelo LDA para a descoberta das variáveis ocultas do modelo. São descritos os métodos de inferência de Gibbs e o de inferência variacional. Para ambos métodos são realizadas as derivações completas do modelo LDA e a descrição dos algoritmos. As derivações detalhadas são úteis para alunos que queiram explorar e expandir os modelos probabilísticos de tópicos, pois essas derivações podem servir de base para as derivações de modelos mais específicos ou personalizados. Um leitor que esteja interessado apenas na descrição dos algoritmos e na parte computacional, sugerimos que atentem apenas nas seções relativas ao algoritmo de inferência e ignorem detalhes das derivações devido ao rigoroso embasamento matemático.

3.1 Inferência do LDA via Amostrador de Gibbs

Entre os métodos de inferência, o Amostrador de Gibbs é o mais popular principalmente pela facilidade de implementação e sua aplicação em diversos problemas (GEMAN; GEMAN, 1984). O amostrador de Gibbs é um caso especial da simulação de Monte Carlo em Cadeia de Markov. Métodos de Monte Carlo em Cadeia de Markov podem emular distribuições de probabilidades com alta dimensionalidade por meio do comportamento estacionário da cadeia de Markov.

O processo realizado pelo amostrador de Gibbs se baseia na amostragem de cada dimensão alternadamente, uma de cada vez, condicionada ao valor de todas as outras dimensões.

Suponha que há uma variável K-dimensional não observada, $z = \{z_1, \ldots, z_K\}$, onde z_i corresponde ao valor da i-ésima dimensão do vetor z e $z_{-i} = \{z_1, z_2, \ldots, z_{i-1}, z_{i+1}, \ldots, z_K\}$, suponha também a evidência dada pela variável observada, w. Nesse exemplo, a distribuição a ser inferida é $p(z|w)$. Assim, em vez de fazer amostras em toda a distribuição z para inferir $p(z|w)$, o amostrador de Gibbs faz escolhas separadas para cada dimensão i de z, onde a amostragem de z_i depende das outras dimensões em z_{-i} amostradas até o momento.

Com a intensão de descobrir $p(z|w)$, pode-se apresentar um algoritmo simples descrevendo o processo realizado pelo amostrado de Gibbs. Esse algoritmo está sumarizado no Algoritmo 1.

Algoritmo 1: Amostrador de Gibbs

Entrada : número de iterações T, variável não observada z, variável observada w.

início
  $z^{(0)} \leftarrow \left\{z_1^{(0)}, \ldots, z_K^{(0)}\right\}$;
  para $t = 1$ **to** T **faça**
    para $i = 1$ **to** K **faça**
      $z_i^{(t+1)} \sim P(z_i|z_1^{(t+1)}, \ldots, z_{i-1}^{(t+1)}, z_{i+1}^{(t)}, \ldots, z_K^{(t)}, w)$;
  retorne estimativa de $p(z|w)$;

No Algoritmo 1, a variável não observada no instante inicial, $z^{(0)}$, pode ser iniciada aleatoriamente. Em seguida, é realizado o processo de iteração onde é amostrado cada dimensão de z em relação a todas as outras dimensões amostradas até então. Esse processo é realizado T vezes, tendo no final desse processo a estimativa de $p(z|w)$.

O amostrador de Gibbs gera uma cadeia de Markov de amostras. Com base nisso, é possível demonstrar convergência do algoritmo. Aqui, não serão apresentadas as demonstrações de que o algoritmo alcança um estado estacionário de transições na cadeia de Markov, mas essas demonstrações podem ser encontradas no trabalho de Russell e Norvig (2003).

3.1.1 Integrando o LDA para o Amostrador de Gibbs

No caso do LDA, o amostrador de Gibbs deve fazer amostragem de três variáveis ocultas, z, θ e ϕ. Para simplificar, o método de Gibbs aplicado no LDA é colapsado de forma a amostrar apenas a variável z, e a partir de z encontrar os valores das variáveis θ e ϕ.

O processo de amostragem é realizado por meio de estatísticas obtidas pela contagem das atribuições de palavras para os tópicos e tópicos para documentos feitas após amostragem. Para manter essas estatísticas, serão introduzidas as seguintes variáveis contadoras:

- $c_{j,i,k}$ corresponde ao número de vezes que um termo w_i é atribuída ao tópico k no documento d_j,

- $c_{j,*,k}$ é o número de termos no documento d_j atribuídas ao tópico k,
- $c_{*,i,k}$ é o número de vezes que o termo w_i é atribuída ao tópico k em todos os documento,
- $c_{*,*,k}$ é o número de termos atribuídos ao tópico k considerando toda a coleção de documentos.

O amostrador colapsado de Gibbs computa a probabilidade do tópico $z_{a,b}$ ser atribuído para a posição do termo w_b no documento d_a, dada as atribuições realizadas anteriormente para as outras posições no documento d_b, denotada como $z_{-(a,b)}$,

$$p(z_{a,b}|z_{-(a,b)}, w, \alpha, \beta) \tag{3.1}$$

Pela definição de probabilidade condicional, tem-se

$$= \frac{p(z_{a,b}, z_{-(a,b)}, w|\alpha, \beta)}{p(z_{-(a,b)}, w|\alpha, \beta)}. \tag{3.2}$$

Removendo o denominador, que não depende de $z_{a,b}$, e unindo $z_{a,b}$ e $z_{-(a,b)}$ em z,

$$\propto p(z_{a,b}, z_{-(a,b)}, w|\alpha, \beta) = p(z, w|\alpha, \beta). \tag{3.3}$$

Usando a regra da probabilidade total, integrando a distribuição de documentos por tópicos θ, e a distribuição de tópicos por palavras ϕ,

$$= \int\int p(\theta, \phi, z, w|\alpha, \beta) d\theta d\phi. \tag{3.4}$$

Expandindo a integral dada pela propriedade da distribuição conjunta do LDA (Equação 2.3),

$$= \int p(z|\theta) p(\theta|\alpha) d\theta \times \int p(w|\phi, z) p(\phi|\beta) d\phi, \tag{3.5}$$

e então expandindo os termos,

$$= \int \prod_{j=1}^{m} p(z_j|\theta_j) p(\theta_j|\alpha) d\theta \times \int \prod_{k=1}^{K} p(\phi_k|\beta) \prod_{j=1}^{m} \prod_{i=1}^{n_{d_j}} p(w_{j,i}|\phi_{z_{j,i}}) d\phi \tag{3.6}$$

$$= \prod_{j=1}^{m} \int p(z_j|\theta_j) p(\theta_j|\alpha) d\theta_j \times \prod_{k=1}^{K} \int p(\phi_k|\beta) \prod_{j=1}^{m} \prod_{i=1}^{n_{d_j}} p(w_{j,i}|\phi_{z_{j,i}}) d_{\phi_k} \tag{3.7}$$

Desde que essas probabilidades obedeçam a distribuição de Dirichlet, elas podem ser substituídas pela fórmula usual (Equação 2.1),

$$= \prod_{j=1}^{m} \int \frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} \theta_{j,k}^{\alpha_k - 1} \prod_{i=1}^{n_{d_j}} \theta_{j,z_{j,i}} d\theta_j$$
$$\times \prod_{k=1}^{K} \int \frac{\Gamma(\sum_{l=1}^{n} \beta_l)}{\prod_{l=1}^{n} \Gamma(\beta_l)} \prod_{l=1}^{n} \phi_{l,k}^{\beta_l - 1} \prod_{j=1}^{m} \prod_{i=1}^{n_{d_j}} \phi_{z_{j,i}, w_{j,i}} d\phi_k \tag{3.8}$$

Lembrando que $x^a x^b = x^{a+b}$, pode-se substituir o produto interno das distribuições θ e ϕ,

$$= \prod_{j=1}^{m} \int \frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} \theta_{j,k}^{\alpha_k - 1} \prod_{k=1}^{K} \theta_{j,k}^{c_{j,*,k}} d\theta_j$$

$$\times \prod_{k=1}^{K} \int \frac{\Gamma(\sum_{l=1}^{n} \beta_l)}{\prod_{l=1}^{n} \Gamma(\beta_l)} \prod_{l=1}^{n} \phi_{l,k}^{\beta_l - 1} \prod_{l=1}^{n} \phi_{k,l}^{c_{*,l,k}} d\phi_k \tag{3.9}$$

$$= \prod_{j=1}^{m} \int \frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} \theta_{j,k}^{\alpha_k + c_{j,*,k} - 1} d\theta_j \times \prod_{k=1}^{K} \int \frac{\Gamma(\sum_{l=1}^{n} \beta_l)}{\prod_{l=1}^{n} \Gamma(\beta_l)} \prod_{l=1}^{n} \phi_{l,k}^{\beta_l c_{*,l,k} - 1} d\phi_k. \tag{3.10}$$

Em seguida, multiplicar por uma constante formada por duas frações inversas e com valor igual a um,

$$= \prod_{j=1}^{m} \frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \frac{\prod_{k=1}^{K} \Gamma(c_{j,*,k} + \alpha_k)}{\Gamma(\sum_{k=1}^{K} c_{j,*,k} + \alpha_k)} \int \frac{\Gamma(\sum_{k=1}^{K} c_{j,*,k} + \alpha_k)}{\prod_{k=1}^{K} \Gamma(c_{j,*,k} + \alpha_k)} \prod_{k=1}^{K} \theta_{j,k}^{\alpha_k + c_{j,*,k} - 1} d\theta_j$$

$$\times \prod_{k=1}^{K} \frac{\Gamma(\sum_{l=1}^{n} \beta_l)}{\sum_{l=1}^{n} \Gamma(\beta_l)} \frac{\prod_{l=1}^{n} \Gamma(c_{*,l,k} + \beta_l)}{\Gamma(\sum_{l=1}^{n} c_{*,l,k} + \beta_l)} \int \frac{\Gamma(\sum_{l=1}^{n} c_{*,l,k} + \beta_l)}{\prod_{l=1}^{n} \Gamma(c_{*,l,k} + \beta_l)} \prod_{l=1}^{n} \phi_{l,k}^{\beta_l c_{*,l,k} - 1} d\phi_k. \tag{3.11}$$

O valor formado pelas integrais correspondem a densidade de uma distribuição de Dirichlet, consequentemente elas tem valor igual a 1 e podem ser removidas,

$$= \prod_{j=1}^{m} \frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \frac{\prod_{k=1}^{K} \Gamma(c_{j,*,k} + \alpha_k)}{\Gamma(\sum_{k=1}^{K} c_{j,*,k} + \alpha_k)} \times \prod_{k=1}^{K} \frac{\Gamma(\sum_{l=1}^{n} \beta_l)}{\sum_{l=1}^{n} \Gamma(\beta_l)} \frac{\prod_{l=1}^{n} \Gamma(c_{*,l,k} + \beta_l)}{\Gamma(\sum_{l=1}^{n} c_{*,l,k} + \beta_l)}. \tag{3.12}$$

Removendo as funções Γ que dependem apenas dos hiper-parâmetros (constantes) α e β, terá a seguinte proporcionalidade

$$\propto \prod_{j=1}^{M} \frac{\prod_{k=1}^{K} \Gamma(c_{j,*,k} + \alpha_k)}{\Gamma(\sum_{k=1}^{K} c_{j,*,k} + \alpha_k)} \times \prod_{k=1}^{K} \frac{\prod_{l=1}^{V} \Gamma(c_{*,l,k} + \beta_l)}{\Gamma(\sum_{l=1}^{V} c_{*,l,k} + \beta_l)}. \tag{3.13}$$

Em seguida, será separado esse produto evidenciando a posição b no documento d_a,

$$= \prod_{j \neq a}^{m} \frac{\prod_{k=1}^{K} \Gamma(c_{j,*,k} + \alpha_k)}{\Gamma(\sum_{k=1}^{K} c_{j,*,k} + \alpha_k)} \times \frac{\prod_{k=1}^{K} \Gamma(c_{a,*,k} + \alpha_k)}{\Gamma(\sum_{k=1}^{K} c_{a,*,k} + \alpha_k)}$$

$$\times \prod_{k=1}^{K} \frac{\prod_{l \neq w_{a,b}}^{n} \Gamma(c_{*,l,k} + \beta_l) \times \Gamma(c_{*,w_{a,b},k})}{\Gamma(\sum_{l=1}^{n} c_{*,l,k} + \beta_l)}. \tag{3.14}$$

Removendo os termos da equação que não dependam da posição (a,b),

$$\propto \frac{\prod_{k=1}^{K} \Gamma(c_{a,*,k} + \alpha_k)}{\Gamma(\sum_{k=1}^{K} c_{a,*,k} + \alpha_k)} \times \prod_{k=1}^{K} \frac{\Gamma(c_{*,w_{a,b},k} + \beta_{w_{a,b}})}{\Gamma(\sum_{l=1}^{n} c_{*,l,k} + \beta_l)}. \tag{3.15}$$

Seja $c^{-(a,b)}$ o valor das contagens feitas como o contador c, mas desconsiderando a contagem da posições (a,b). Note que para contagens que não inclui o documento a, o valor de $c^{(a,b)} = c$, e para as contagens na posição b do documento a, é incrementado 1 mais o valor já contado em $c^{a,b}$,

$$\propto \frac{\prod_{k \neq z_{a,b}}^{K} \Gamma(c_{a,*,k}^{-(a,b)} + \alpha_k) \times \Gamma(c_{a,*,z_{a,b}}^{-(a,b)} + \alpha_{z_{a,b}} + 1)}{\Gamma(1 + \sum_{k=1}^{K} c_{a,*,k}^{-(a,b)} + \alpha_k)}$$

$$\times \prod_{k \neq z_{a,b}}^{K} \frac{\Gamma(c_{*,w_{a,b},k}^{-(a,b)} + \beta_{w_{a,b}})}{\Gamma(\sum_{l=1}^{n} c_{*,l,k} + \beta_l)} \times \frac{\Gamma(c_{*,w_{a,b},z_{a,b}}^{-(a,b)} + \beta_{w_{a,b}} + 1)}{\Gamma(1 + \sum_{l=1}^{n} c_{*,l,z_{a,b}}^{-(a,b)} + \beta_l)}. \tag{3.16}$$

Desde que x é inteiro tem-se que $\Gamma(x+1) = x \times \Gamma(x)$, assim expande-se os termos dependentes da posição (a,b),

$$= \frac{\prod_{k \neq z_{a,b}}^{K} \Gamma(c_{a,*,k}^{-(a,b)} + \alpha_k) \times \Gamma(c_{a,*,z_{a,b}}^{-(a,b)} + \alpha_{z_{a,b}}) \times (c_{a,*,z_{a,b}}^{-(a,b)} + \alpha_{z_{a,b}})}{\Gamma(1 + \sum_{k=1}^{K} c_{a,*,k}^{-(a,b)} + \alpha_k)}$$

$$\times \prod_{k \neq z_{a,b}}^{K} \frac{\Gamma(c_{*,w_{a,b},k}^{-(a,b)} + \beta_{w_{a,b}})}{\Gamma(\sum_{l=1}^{n} c_{*,l,k} + \beta_l)} \times \frac{\Gamma(c_{*,w_{a,b},z_{a,b}}^{-(a,b)} + \beta_{w_{a,b}}) \times (c_{*,w_{a,b},z_{a,b}}^{-(a,b)} + \beta_{w_{a,b}})}{\Gamma(\sum_{l=1}^{n} c_{*,l,z_{a,b}}^{-(a,b)} + \beta_l) \times \sum_{l=1}^{n} c_{*,l,z_{a,b}}^{-(a,b)} + \beta_l}. \tag{3.17}$$

Unindo os produtos de $k \neq z_{a,b}$ e $k = z_{a,b}$,

$$= \frac{\prod_{k=1}^{K} \Gamma(c_{a,*,k}^{-(a,b)} + \alpha_k) \times (c_{a,*,z_{a,b}}^{-(a,b)} + \alpha_{z_{a,b}})}{\Gamma(1 + \sum_{k=1}^{K} c_{a,*,k}^{-(a,b)} + \alpha_k)}$$

$$\times \prod_{k=1}^{K} \frac{\Gamma(c_{*,w_{a,b},k}^{-(a,b)} + \beta_{w_{a,b}})}{\Gamma(\sum_{l=1}^{n} c_{*,l,k} + \beta_l)} \times \frac{(c_{*,w_{a,b},z_{a,b}}^{-(a,b)} + \beta_{w_{a,b}})}{\sum_{l=1}^{n} c_{*,l,z_{a,b}}^{-(a,b)} + \beta_l}. \tag{3.18}$$

Os produtórios indexados por tópicos resultam em valores constantes, logo podem ser removidos,

$$\propto \frac{(c_{a,*,z_{a,b}}^{-(a,b)} + \alpha_{z_{a,b}}) \times (c_{*,w_{a,b},z_{a,b}}^{-(a,b)} + \beta_{w_{a,b}})}{\sum_{l=1}^{V} c_{*,l,z_{a,b}}^{-(a,b)} + \beta_l} \tag{3.19}$$

E por fim, pode-se simplificar o denominador de forma que $\sum_{l=1}^{K} c_{*,l,z_{a,b}}^{-(a,b)} = c_{*,*,k}^{-(a,b)}$,

$$\propto \frac{(c_{a,*,z_{a,b}}^{-(a,b)} + \alpha_{z_{a,b}}) \times (c_{*,w_{a,b},z_{a,b}}^{-(a,b)} + \beta_{w_{a,b}})}{c_{*,*,k}^{-(a,b)} + \sum_{l=1}^{n} \beta_l}. \tag{3.20}$$

Assim, chega-se na equação de amostragem via algoritmo de Gibbs para o modelo LDA:

$$p(z_{a,b} = k | z_{-(a,b)}, w, \alpha, \beta) \propto \frac{(c_{a,*,z_{a,b}}^{-(a,b)} + \alpha_{z_{a,b}}) \times (c_{*,w_{a,b},z_{a,b}}^{-(a,b)} + \beta_{w_{a,b}})}{c_{*,*,k}^{-(a,b)} + \sum_{l=1}^{n} \beta_l}. \tag{3.21}$$

Uma vez estimado z, é possível encontrar os valores para as distribuições θ e ϕ, as respectivas distribuição de documentos por tópicos e tópicos por palavras. Elas podem ser obtidas pelo cálculo

$$\theta_{j,k} = \left(\frac{c_{j,*,k} + \alpha_k}{n_{d_j} + m\alpha_k} \right) \tag{3.22}$$

$$\phi_{k,i} = \left(\frac{c_{*,w_i,k} + \beta_k}{(c_{*,*,k} + n\beta_k)} \right). \tag{3.23}$$

Esses valores correspondem a estatísticas obtidas durante a amostragem e são normalizados levando-se em conta a relação de proporcionalidade.

Com base nas proporções e equações 3.21, 3.22 e 3.23 é apresentada na próxima seção o algoritmo completo para a mostragem.

3.1.2 Algoritmo de inferência via amostrador de Gibbs

O procedimento de amostragem do algoritmo de Gibbs para o LDA pode ser executado usando a Equação 3.21 para a amostragem. O algoritmo é descrito no Algoritmo 3. Ele utiliza apenas quatro grandes estrutura de dados: o contador $c_{j,*,k}$ que é uma matriz de dimensão $m \times K$, onde mantêm o contador do documento d_j para um tópico k; o contador $c_{*,i,k}$ que é uma matriz de dimensão $k \times n$, onde mantêm o contador do tópico k atribuído a uma palavra w_i; o contador $c_{*,*,k}$ que é um vetor K-dimensional, onde mantêm a quantidade de atribuições a um tópico k; e as atribuições z, que é uma matriz $m \times n$ onde mantêm a atribuição de um documento d_j para cada termo w_i. Com isso, o algoritmo executará em dois procedimentos, inicialização (Algoritmo 2) e amostragem (Algoritmo 3).

Algoritmo 2: Inicialização do Amostrador de Gibbs para LDA

Entrada : número de tópicos K, coleção de documentos

início
- todas as variáveis contadoras $c_{j,*,k}$, $c_{*,i,k}$, $c_{*,*,k}$ são iniciados com zero ;
- **para** *documento* d_j *com* $j \in [1,m]$ **faça**
  - **para** *palavra* w_i *com* $i \in [1,n_j]$ *no documento* d_j **faça**
    - amostre o índice do tópico $z_{j,i} \leftarrow Mult(\frac{1}{k})$ para palavra $w_{j,i}$;
    - incremente o contador de documento por tópico: $c_{j,*,z_{j,i}} + +$;
    - incremente o contador de tópico por palavra: $c_{*,i,z_{j,i}} + +$;
    - incremente a soma de palavras do tópico amostrado: $c_{*,*,z_{j,i}} + +$;

A amostragem envolve o cálculo das estatísticas obtidas pelos contadores. Note que na inicialização os contadores são incrementados com tópicos atribuídos aleatoriamente. Em seguida, para atribuir um tópicos para a variável $z_{j,i}$, é necessário decrementar a contagem já atribuída ao termo na posição i do documento d_j, fazer uma nova amostragem (via Equação 3.21), e atualizar os contadores. O novo tópico é atribuído para a variável $z_{j,i}$ e, em seguida, são utilizadas para encontras as distribuições θ e ϕ (via equações 3.22 e 3.23). Veja o Algoritmo 3 para o procedimento completo de amostragem de Gibbs para o LDA.

A convergência desse algoritmo é alcançada quando não existe alterações na distribuição conjunta do modelo (Equação 2.3). Em termos práticos, é definido um número T de iterações.

3.2 Inferência do LDA via método variacional

Nessa seção é apresentado o algoritmo de inferência variacional para o LDA. Esse algoritmo tem uma abordagem diferente do amostrador de Gibbs. O método variacional não se baseia em amostragem, em vez disso, ele transforma o processo de inferência da distribuição *a posteriori* do LDA em um problema de otimização.

Algoritmo 3: Amostrador de Gibbs para o LDA

Entrada : número de tópicos K, coleção de documentos, hiper-parâmetros α e β, número de iterações T

início
  Inicializa os contadores – Algoritmo 2 ;
  enquanto *não terminar o número de iterações* T **faça**
    para *documento* d_j *com* $j \in [1,m]$ **faça**
      para *palavra* w_i *com* $i \in [1,n_j]$ *no documento* d_j **faça**
        $c_{j,*,z_{j,i}} --$, $c_{*,i,z_{j,i}} --$, $c_{*,*,z_{j,i}} --$;
        para *tópico* k *com* $k \in [1,K]$ **faça**
          $p(z_{j,i} = k|\cdot) = \frac{(c^{-(a,b)}_{a,*,z_{a,b}} + \alpha_{z_{a,b}}) \times (c^{-(a,b)}_{*,w_{a,b},z_{a,b}} + \beta_{w_{a,b}})}{c^{-(a,b)}_{*,*,k} + \sum_{l=1}^{n} \beta_l}$
        topico = amostre de $p(z|\cdot)$;
        $z_{j,i}$ = *topico* ;
        $c_{j,*,z_{j,i}} ++$, $c_{*,i,z_{j,i}} ++$, $c_{*,*,z_{j,i}} ++$;
  Atualize o conjunto de parâmetros θ e ϕ de acordo com as equações (3.22) e (3.23) ;

Antes de especificar o método variacional para o LDA, é apresentada a noção básica do método. Para isso, é utilizada uma notação genérica, considerando um modelo onde as variáveis latentes não observadas é denotado por z, e o conjunto de todas as variáveis observáveis e o conhecimento *a priori* é denotado por w. A probabilidade conjunta desse modelo é $p(w,z)$. Aqui, o objetivo é encontrar uma solução para a distribuição *a posteriori* $p(z|w)$, ou seja, descobrir o conhecimento oculto representado pela variável não observada z dada a observação em w. Nessa seção, é introduzido o método variacional (*Variational Bayes Inference*) para inferência da distribuição *a posteriori* $p(z|w)$.

Supõe-se que o cálculo da distribuição $p(z|w)$ seja intratável. Assim, no método variacional, uma solução aproximada para $p(z|w)$ é alcançada por meio de uma outra distribuição $q(z)$. Essa distribuição $q(z)$ é definida por uma família de distribuição mais "fácil" de calcular do que $p(z|w)$. Dessa forma, inicialmente, é necessário definir uma família de distribuições que se aproxime da *a posteriori*. A relação de proximidade entre a distribuição *a posteriori* $p(z|w)$ e a distribuição variacional $q(z)$ é medida pela divergência de Kullback-Leibler (KL) (KULLBACK; LEIBLER, 1951) (veja a Definição 1). Quanto menor a divergência KL entre as distribuições q e a distribuição real p melhor será a aproximação. Logo, o método variacional transforma o problema de inferência em um problema de otimização onde o objetivo é minimizar $KL(q||p)$ (BISHOP, 2006).

Definição 1 (Divergência KL)**.** Para duas distribuições contínuas p e q, a divergência de Kullback-Leibler, ou divergência KL, é calculada da seguinte forma:

$$KL(p||q) = \int_x p(x) log \frac{p(x)}{q(x)} d_x, \tag{3.24}$$

onde x é uma variável aleatória contínua.

Assim como no cálculo da posteriori, minimizar diretamente a divergência $KL(q||p)$ é intratável. Porém, é possível limitar a distribuição marginal do modelo em função apenas da distribuição variacional. Existe uma relação entre as distribuições real p e a variacional q com o logaritmo da probabilidade marginal do modelo. Para alcançar essa relação, será estendido o logaritmo da probabilidade marginal do modelo da seguinte forma:

$$\begin{aligned}\log p(x) &= \log \int_z p(x,z)dz \\ &= \log \int_z p(x,z)\frac{q(z)}{q(z)}dz \\ &= \log \left(E_q \left[\frac{p(x,z)}{q(z)} \right] \right) \\ &\geq E_q \left[p(x,z) \right] - E_q \left[q(z) \right]. \end{aligned} \tag{3.25}$$

O último passo utiliza a desigualdade de Jensen (NEEDHAM, 1993) para encontrar um limite inferior para o logaritmo da probabilidade marginal do modelo.

Pela Desigualdade (3.25), tem se que $\log p(x)$ é no mínimo $E_q \left[p(x,z) \right] - E_q \left[q(z) \right]$. Essa expressão, chamada de *Evidence Lower Bound* (ELBO) $\mathscr{L}$, será denotada como

$$\mathscr{L} \triangleq E_q \left[p(x,z) \right] - E_q \left[q(z) \right]. \tag{3.26}$$

Agora, qual a relação do ELBO com a minimização da divergência KL? Para encontrar essa relação, basta calcular a diferença

$$\begin{aligned} & E_q \left[p(x,z) \right] - E_q \left[q(z) \right] - \log p(x) \\ &= \int_z q(z)p(x,z) - \int_z q(z)q(z) - \log p(x) \\ &= \int_z q(z) \log p(x,z) - \int_z q(z) \log q(z) - \int_z q(z) \log p(x) \\ &= \int_z q(z) \log \frac{p(x,z)}{p(x)} - \int_z q(z) \log q(z) \\ &= \int_z q(z) \log p(Z|X) - \int_z q(z) \log q(z) \\ &= \int_z q(z) \log \frac{p(z|x)}{q(z)} \\ &= - \left(\int_z q(z) \log \frac{q(z)}{p(z|x)} \right) \\ &= -KL(q||p) \end{aligned} \tag{3.27}$$

Relembrando que o principal objetivo é minimizar a divergência de Kullback-Leibler, de forma que a distribuição $q(z)$ se aproxime de $p(z|x)$. Com isso, pela Desigualdade (3.25) tem-se que $\log p(x)$ é no máximo $E_q \left[p(x,z) \right] - E_q \left[q(z) \right]$. Já na Equação (3.27) tem-se a expressão para a

divergência KL. Por essas duas equações, nota-se que para minimizar a divergência de Kullback-Leibler é preciso maximizar $E_q[p(x,z)] - E_q[q(z)]$. Assim, foi transformado o problema de inferência em um problema de otimização onde o objetivo é maximizar a Desigualdade 3.25. A Desigualdade 3.25 é o ponto principal no método de inferência variacional, pois todo o processo computacional desse método se baseia em otimizar o ELBO, aqui denotado como $\mathscr{L}$.

3.2.1 Integrando o LDA para o método de inferência variacional

Para utilizar o método de inferência variacional no LDA, inicialmente, é necessário definir uma distribuição q, tal que essa distribuição se aproxime da distribuição *a posteriori* original do LDA. Veja a Figura 1 para a descrição gráfica do LDA e a Equação 2.4 com a descrição da distribuição *a posteriori*. A distribuição do LDA p é denotada nesse texto como a distribuição "original", e a distribuição q é chamada de distribuição variacional. Um modo simples de se obter a família de distribuição variacional q é considerar simples modificações na distribuição original. Removendo as arestas que ligam as variáveis θ, ϕ, z e w, obtêm um modelo simplificado sem a relação de dependência entre essas variáveis. Com as variáveis independentes, o número de combinações de valores atribuídos a elas se tornam computacionalmente viáveis. Cada variável da distribuição variacional, aqui denotada como variáveis variacionais, terão suas correspondentes na distribuição original. Na Figura 2 está o modelo gráfico da distribuição q, com suas variáveis variacionais e suas correspondentes originais. A atribuição dos tópicos $z_{j,i}$ tem como distribuição variacional $q(z_{j,i}|\varphi_{j,i}) = Mult(\varphi_{j,i})$. Note que cada palavra observada w_i terá uma distribuição variacional sobre os tópicos, isso permite que diferentes palavras sejam associadas para diferentes tópicos. A distribuição dos documentos por tópicos, θ_j, tem sua distribuição variacional gerada por uma distribuição de Dirichlet $q(\theta_j) = Dir(\gamma_j, \alpha)$, onde γ_j é um vetor K-dimensional. Existem diferentes distribuições de Dirichlet variacionais para cada documento, permitindo que diferentes documentos sejam atribuídos a diferentes tópicos com diferentes proporções. Por fim, tem-se a distribuição de tópicos por termos, ϕ, que tem distribuição variacional para cada tópicos $q(\phi_k) = Dir(\lambda_k, \beta)$, onde λ_k é um vetor n-dimensional com valores gerados pela distribuição de Dirichlet. Com base nessas simplificações, a família de distribuições q é caracterizada pela seguinte distribuição

$$q(\theta, z, \phi) = \prod_{k=1}^{K} q(\phi_k|\lambda_k) \prod_{j=1}^{m} \left(q(\theta_j|\gamma_j) \prod_{i=1}^{n_{d_j}} q(z_{j,i}|\varphi_{j,i}) \right), \tag{3.28}$$

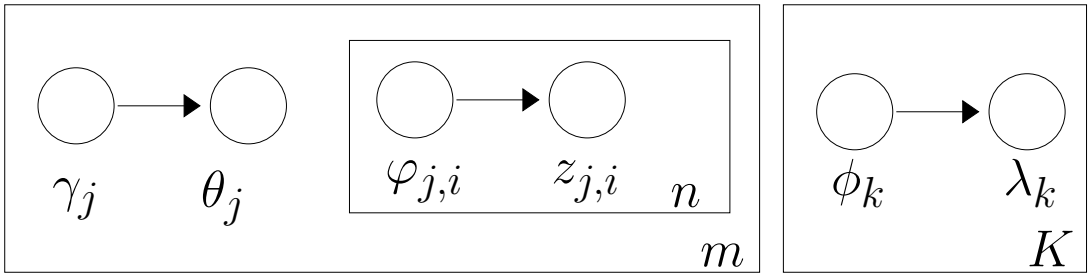
onde φ, γ e λ são as distribuições variacionais.

O método variacional transforma o problema de inferência do LDA em um problema de otimização, onde o objetivo é

$$(\lambda^*, \gamma^*, \varphi^*) = \arg\min_{\lambda^*, \gamma^*, \varphi^*} KL(q(\theta, z, \phi)||p(\theta, z, \phi|w, \alpha, \beta)), \tag{3.29}$$

onde λ^*, γ^* e φ^* são os valores ótimos.

Figura 2 – Distribuição variacional aproximada para o modelo LDA.



Fonte: Adaptada de Blei, Ng e Jordan (2003).

Otimizar a Equação 3.29 diretamente é inviável, mas como foi discutido na Seção 3.2, pode-se otimizar essa equação por meio do ELBO. Para encontrar o ELBO $\mathscr{L}$, o logaritmo da probabilidade marginal do modelo é estendido. Para isso, é descrito a Desigualdade (3.25) em relação a probabilidade marginal do LDA da seguinte forma:

$$\begin{aligned}
\log p(w|\alpha,\beta) &= \log \int \sum_z p(\theta,\phi,z,w|\alpha,\beta)d\theta \\
&= \log \int \sum_z \frac{p(\theta,\phi,z,w|\alpha,\beta)q(\theta,\phi,z)}{q(\theta,\phi,z)}d\theta \\
&\geq \int \sum_Z q(\theta,\phi,z)\log p(\theta,\phi,z,w|\alpha,\beta)d\theta - \int \sum_Z q(\theta,\phi,z)\log q(\theta,\phi,z)d\theta \\
&= E_q[\log p(\theta,\phi,z,w|\alpha,\beta)] - E_q[\log q(\theta,\phi,z)] \\
&\triangleq \mathscr{L}(\gamma,\varphi,\lambda|\alpha,\beta)
\end{aligned} \tag{3.30}$$

Como foi discutido na Seção 3.2, maximizar o ELBO é igual a minimizar a divergência KL entre a distribuição real do LDA e a distribuição variacional (Veja a Equação (3.27)). Os valores encontrados para os parâmetros variacionais γ, φ e λ pela minimização de $\mathscr{L}(\gamma,\varphi,\lambda|\alpha,\beta)$ são aproximações para os parâmetros da distribuição $p(\theta,\phi,z|w,\alpha,\beta)$. Então, o que é feito no método variacional é maximizar o ELBO – $\mathscr{L}(\gamma,\varphi,\lambda|\alpha,\beta)$.

As próximas subseções descrevem a expansão do ELBO ($\mathscr{L}$) segundo o modelo LDA (como descrito na Seção 2), e o procedimento de maximização do ELBO utilizando a técnica de gradiente descendente. Por fim, é descrito o algoritmo de inferência variacional.

3.2.2 Expandindo o ELBO

Relembrando alguns conceitos importantes para auxiliar na expansão do ELBO. Primeiro, resolvendo $E_q[\log p(\theta|\alpha)]$, onde a notação $E_q[\cdot]$ corresponde a esperança em relação a distribuição variacional q, e pela definição da distribuição de Dirichlet (Equação 2.1),

$$p(\theta|\alpha) = \frac{\Gamma(\sum_{k=1}^K \alpha_k)}{\prod_{k=1}^K \Gamma(\alpha_k)} \prod_{k=1}^K \theta_k^{\alpha_k - 1}. \tag{3.31}$$

Aplicando o logaritmo em $p(\theta|\alpha)$,

$$\log p(\theta|\alpha) = \sum_k (\alpha_k - 1)\log\theta_k + \log\Gamma(\sum_k \alpha_k) - \sum_k^K \log\Gamma(\alpha_k). \tag{3.32}$$

Agora, calculando a esperança em relação a distribuição q,

$$E_q[\log p(\theta|\alpha)] = \sum_k (\alpha_k - 1)E_q[\log\theta_k] + \log\Gamma(\sum_k \alpha_k) - \sum_k^K \log\Gamma(\alpha_k). \tag{3.33}$$

A expressão $E[\log\theta]$ corresponde a esperança do logaritmo da distribuição θ e é calculada como

$$E[\log\theta_k] = \Psi(\gamma_i) - \Psi\left(\sum_j \gamma_j\right), \tag{3.34}$$

onde $\Psi(\cdot)$ é a função digama, e γ_i são os parâmetros variacionais da distribuição q correspondente a distribuição original θ (BLEI; NG; JORDAN, 2003).

Com isso, o que é feito agora é reescrever o ELBO. Pela Equação 3.30, tem-se a definição do ELBO

$$\mathscr{L}(\gamma,\varphi,\lambda|\alpha,\beta) \triangleq E_q[\log p(\theta,\phi,z,w|\alpha,\beta)] - E_q[\log q(\theta,\phi,z)] \tag{3.35}$$

Com base na probabilidade conjunta do LDA, descrita na Equação (2.3), pode-se reescrever a Equação (3.35) da seguinte forma:

$$\begin{aligned}
\mathscr{L}(\gamma,\varphi,\lambda|\alpha,\beta) =& E_q[\log p(\phi|\beta)] && (3.36)\\
&+ E_q[\log p(\theta|\alpha)] && (3.37)\\
&+ E_q[\log p(z|\theta)] && (3.38)\\
&+ E_q[\log p(w|z,\phi)] && (3.39)\\
&- E_q[\log q(\theta,\phi,z)] && (3.40)
\end{aligned}$$

O que será feito agora é estender cada um dos termos de $\mathscr{L}(\gamma,\varphi,\lambda|\alpha,\beta)$. Iniciando com o Termo 3.36:

$$\begin{aligned}
E_q[\log p(\phi|\beta)] &= E_q[\sum_{k=1}^K \log p(\phi_k|\beta)]\\
&= E_q\left[\sum_{k=1}^K \left(\sum_{i=1}^n (\beta_i - 1)\log\phi_{k,i} + \log\Gamma(\sum_{i=1}^n \beta_i) - \sum_{i=1}^n \log\Gamma(\beta_i)\right)\right]\\
&= \sum_{k=1}^K \left(\log\Gamma(\sum_{i=1}^n \beta_i) - \sum_{i=1}^n \log\Gamma(\beta_i) + \sum_{i=1}^n (\beta_i - 1)E_q\left[\log\phi_{k,i}\right)\right]\\
&= \sum_{k=1}^K \left(\log\Gamma(\sum_{i=1}^n \beta_i) - \sum_{i=1}^n \log\Gamma(\beta_i) + \sum_{i=1}^n (\beta_i - 1)\left(\Psi(\lambda_{k,i}) - \Psi(\sum_u^n \lambda_{k,u})\right)\right)
\end{aligned}$$

Da mesma forma, para o Termo 3.37 tem-se:

$$
\begin{aligned}
E_q[\log p(\theta|\alpha)] &= E_q[\sum_{j=1}^{m}\left(\log p(\theta_j|\alpha)\right)] \\
&= E_q\left[\sum_{j=1}^{m}\left(\sum_{k=1}^{K}(\alpha_k-1)\log\theta_{j,k}+\log\Gamma(\sum_{i=1}^{K}\alpha_i)-\sum_{k=1}^{K}\log\Gamma(\alpha_k)\right)\right] \\
&= \sum_{j=1}^{m}\left(\log\Gamma(\sum_{k=1}^{K}\alpha_k)-\sum_{k=1}^{K}\log\Gamma(\alpha_i)+\sum_{k=1}^{K}(\alpha_k-1)E_q\left[\log\theta_{j,k}\right]\right) \\
&= \sum_{j=1}^{m}\left(\log\Gamma(\sum_{k=1}^{K}\alpha_k)-\sum_{k=1}^{K}\log\Gamma(\alpha_k)+\sum_{k=1}^{K}(\alpha_k-1)\left(\Psi(\gamma_{j,k})-\Psi(\sum_{k}^{K}\gamma_{j,k})\right)\right)
\end{aligned}
$$

Para expandir o Termo 3.38, é necessário escrever $p(z|\theta)$ da seguinte forma:

$$
\begin{aligned}
p(z|\theta) &= \prod_{j}^{m}\prod_{n}^{n_{d_j}} p(z_{j,i}|\theta_d) \\
&= \prod_{j}^{m}\prod_{n}^{n_{d_j}} \theta_d^{z_{j,i}}.
\end{aligned}
$$

O vetor $z_{j,i}$ contêm a distribuição de tópicos atribuídos a palavra w_i no documento d_j. Quando atribuído a um tópico k, o valor de $z_{j,i,k}=1$, caso contrário, $z_{j,i,k}=0$. Logo, para o Termo 3.39, tem-se

$$
\begin{aligned}
E_q[\log p(z|\theta)] &= E_q\left[\sum_{k}^{K}\sum_{j}^{m}\sum_{i}^{n_{d_j}}\log\theta_{j,k}^{z_{j,i,k}}\right] \\
&= E_q\left[\sum_{k}^{K}\sum_{j}^{m}\sum_{i}^{N_{d_j}} z_{j,i,k}\log\theta_{j,k}\right] \\
&= \sum_{k}^{K}\sum_{j}^{m}\sum_{n}^{n_{d_j}} E_q\left[z_{j,i,k}\right]E_q\left[\log\theta_{j,k}\right] \\
&= \sum_{k}^{K}\sum_{j}^{m}\sum_{n}^{n_{d_j}} \varphi_{j,i,k}\left(\Psi(\gamma_{j,k})-\Psi(\sum_{l}^{K}\gamma_{l,k})\right)
\end{aligned}
\tag{3.41}
$$

Para expandir o Termo 3.39, é necessário escrever $p(w|z,\phi)$ da seguinte forma:

$$
\begin{aligned}
p(w|z,\phi) &= \prod_{k}^{K}\prod_{j}^{m}\prod_{i}^{n_{d_j}} p(w_i|z_{j,i,k},\phi_k) \\
&= \prod_{k}^{K}\prod_{j}^{m}\prod_{i}^{n_{d_j}} \phi_{k,w_i}^{z_{j,i,k}}
\end{aligned}
$$

logo,

$$
\begin{aligned}
E_q[\log p(w|z,\phi)] &= E_q\left[\sum_{k}^{K}\sum_{j}^{m}\sum_{n}^{n_{d_j}}\log \beta_{k,w_i}^{z_{j,i,k}}\right] \\
&= \sum_{k}^{K}\sum_{j}^{m}\sum_{i}^{N_{d_j}} E_q\left[z_{j,i,k}\right] E_q\left[\log \phi_{k,w_i}\right] \\
&= \sum_{k}^{K}\sum_{j}^{m}\sum_{i}^{n_{d_j}} \varphi_{j,i,k}\left(\Psi(\lambda_{k,i}) - \Psi(\sum_{l}^{n}\lambda_{k,l})\right)
\end{aligned}
$$

No Termo 3.40, tem-se o correspondente da distribuição variacional (Equação 3.28)

$$
\begin{aligned}
-E_q[q(\theta,\phi,z)] &= -\int_{k=1}^{K}\int_{j=1}^{m}\sum_{z} q(\theta_d,\phi_k,z)\log q(\theta_j,\phi_k,z)d\theta d\phi dz \\
&= -\int_{k=1}^{K} q(\phi_k)\log q(\phi_k)d\phi - \int_{j=1}^{m} q(\theta_j)\log q(\theta_j)d\theta - \sum_{z} q(z)\log q(z).
\end{aligned}
$$

Note que essa equação corresponde a entropia das distribuições variacionais $q(\theta)$, $q(\phi)$ e $q(z)$. Sabendo disso, basta substituir pela fórmula da entropias das distribuições de Dirichlet θ e ϕ, e distribuição Muntinomial z com parâmetros variacionais. Aplicando a definição de entropia (veja a Definição no trabalho de Frigg e Werndl (2011)) tem-se

$$
\begin{aligned}
-E_q[q(\theta,\phi,z)] &= -\int_{j=1}^{m} q(\theta_j)\log q(\theta_j)d\theta - \sum_{z} q(z)\log q(z) - \int_{k=1}^{K} q(\phi_k)\log q(\phi_k)d\phi \\
&= \sum_{j=1}^{m}\Bigg(-\left(\sum_{k=1}^{K}(\gamma_{j,k}-1)\left(\Psi(\gamma_{j,k}) - \Psi(\sum_{r=1}^{K}\gamma_{j,r})\right)\right) \\
&\quad - \log\Gamma(\sum_{k=1}^{K}\gamma_{j,k}) + \sum_{k=1}^{K}\log\Gamma(\gamma_{j,k}) \\
&\quad - \sum_{i=1}^{n_{d_j}}\sum_{k=1}^{K}\varphi_{j,i,k}\log\varphi_{j,i,k}\Bigg) \\
&\quad + \sum_{k=1}^{K}\Bigg(-\left(\sum_{i=1}^{n}(\lambda_{k,i}-1)\left(\Psi(\lambda_{k,i}) - \Psi(\sum_{u=1}^{n}\lambda_{k,u})\right)\right) \\
&\quad - \log\Gamma(\sum_{i=1}^{n}\lambda_{k,i}) + \sum_{i=1}^{n}\log\Gamma(\lambda_{k,i})\Bigg)
\end{aligned}
$$

Com as extensões detalhadas de todos os itens, tem-se a formulação expandida do ELBO

$$
\begin{aligned}
\mathscr{L}(\gamma,\varphi,\lambda|\alpha,\beta) &= \sum_{k=1}^{K}\left(\log\Gamma(\sum_{i=1}^{n}\beta_i) - \sum_{i=1}^{n}\log\Gamma(\beta_i) + \sum_{i=1}^{n}(\beta_i-1)\left(\Psi(\lambda_{k,i}) - \Psi(\sum_{u}^{n}\lambda_{k,u})\right)\right) \\
&\quad + \sum_{j=1}^{m}\left(\log\Gamma(\sum_{k=1}^{K}\alpha_k) - \sum_{k=1}^{K}\log\Gamma(\alpha_k) + \sum_{k=1}^{K}(\alpha_k-1)\left(\Psi(\gamma_{j,k}) - \Psi(\sum_{r}^{K}\gamma_{j,r})\right)\right) \\
&\quad + \sum_{k}^{K}\sum_{j}^{m}\sum_{i}^{n_{d_j}}\varphi_{j,i,k}\left(\Psi(\gamma_{j,k}) - \Psi(\sum_{l}^{K}\gamma_{j,l})\right)
\end{aligned}
$$

$$
\begin{aligned}
&+\sum_{k}^{K}\sum_{j}^{m}\sum_{i}^{n_{d_j}}\varphi_{j,i,k}\left(\Psi(\lambda_{k,i})-\Psi(\sum_{v}^{n}\lambda_{k,v})\right)\\
&+\sum_{j=1}^{m}\left(-\left(\sum_{k=1}^{K}(\gamma_{j,k}-1)\left(\Psi(\gamma_{j,k})-\Psi(\sum_{r=1}^{K}\gamma_{j,r})\right)\right)\right.\\
&-\log\Gamma(\sum_{k=1}^{K}\gamma_{j,k})+\sum_{k=1}^{K}\log\Gamma(\gamma_{j,k})\\
&\left.-\sum_{i=1}^{n_{d_j}}\sum_{k=1}^{K}\varphi_{j,i,k}\log\varphi_{j,i,k}\right)\\
&+\sum_{k=1}^{K}\left(-\left(\sum_{i=1}^{n}(\lambda_{k,i}-1)\left(\Psi(\lambda_{k,i})-\Psi(\sum_{u=1}^{n}\lambda_{k,u})\right)\right)\right.\\
&\left.-\log\Gamma(\sum_{i=1}^{n}\lambda_{k,i})+\sum_{i=1}^{n}\log\Gamma(\lambda_{k,i})\right)
\end{aligned}
\tag{3.42}
$$

3.2.3 Otimizando o ELBO

Como discutido na Seção anterior, o objetivo do algoritmo de inferência variacional é encontrar os valores dos parâmetros variacionais resolvendo o problema de otimização na Equação 3.29. A especificação detalhada do ELBO serve para definir a equação a ser otimizada. Assim, deve-se maximizar a Equação (3.42) em relação a cada parâmetro variacional: γ, φ e λ

Primeiro, para maximizar a Equação (3.42) em relação a $\varphi_{d,n,k}$, definida como $\mathscr{L}_{\varphi_{d,n,k}}$, é necessário incluir a restrição $\sum_{k}^{K}\varphi_{d,n,k}=1$. Essa restrição é imposta na equação incorporando o multiplicador de Lagrande $\rho_{j,i}$, tal que

$$
\begin{aligned}
\mathscr{L}_{\varphi}=&\sum_{k}^{K}\sum_{j}^{m}\sum_{i}^{n_{d_j}}\varphi_{j,i,k}\left(\left(\Psi(\gamma_{j,k})-\Psi(\sum_{l}^{K}\gamma_{j,l})\right)\right.\\
&\left.+\left(\Psi(\lambda_{k,i})-\Psi(\sum_{r}^{n}\lambda_{k,r})\right)-\log\varphi_{j,i,k}\right)+\rho_{j,i}\left(\sum_{l}^{K}\varphi_{j,i,l}-1\right)
\end{aligned}
$$

Note que $\mathscr{L}_{\varphi}$ é o ELBO com apenas os termos dependentes de φ.

Para determinar o gradiente de $\mathscr{L}_{\varphi}$, deve-se calcular a derivada de $\mathscr{L}_{\varphi_{j,i,k}}$

$$
\begin{aligned}
\frac{d\mathscr{L}_{\varphi_{j,i,k}}}{d\varphi_{j,i,k}}=&\left(\left(\Psi(\gamma_{j,k})-\Psi(\sum_{r}^{K}\gamma_{j,r})\right)+\left(\Psi(\lambda_{k,i})-\Psi(\sum_{l}^{n}\lambda_{k,l})\right)-\log\varphi_{j,i,k}\right)\\
&-1+\rho_{j,i}
\end{aligned}
$$

Colocando $\mathscr{L}_{\varphi}$ igual a zero e isolando $\varphi_{j,i,k}$, tem-se

$$
\varphi_{j,i,k}=exp\left(\Psi(\gamma_{j,k})-\Psi(\sum_{r}^{K}\gamma_{j,r})+\Psi(\lambda_{k,i})-\Psi(\sum_{l}^{n}\lambda_{k,l})-1+\rho_{j,i}\right)
$$

Não é necessário computar $\rho_{j,i}$ e $\Psi(\sum_r^K \gamma_{j,r})$, pois eles são os mesmos para todo k. Assim,

$$\varphi_{j,i,k} \propto exp\left(\Psi(\gamma_{j,k}) + \Psi(\lambda_{k,i}) - \Psi(\sum_{l}^{n} \lambda_{k,l})\right) \tag{3.43}$$

Em seguida, para maximizar a Equação (3.42) em relação a γ, define-se $\mathscr{L}_\gamma$ como

$$\begin{aligned}\mathscr{L}_\gamma = &\sum_{j=1}^{m}\sum_{k=1}^{K}(\alpha_k - 1)\left(\Psi(\gamma_{j,k}) - \Psi(\sum_{r}^{K}\gamma_{j,r})\right)\\ &+\sum_{j}^{m}\sum_{i}^{n_{d_j}}\sum_{k=1}^{K}\phi_{j,i,k}\left(\Psi(\gamma_{j,k}) - \Psi(\sum_{l}^{K}\gamma_{j,l})\right)\\ &-\sum_{j=1}^{m}\sum_{k=1}^{K}\left((\gamma_{j,k}-1)\left(\Psi(\gamma_{j,k}) - \Psi(\sum_{r=1}^{K}\gamma_{j,r})\right)\right.\\ &\left.-\log\Gamma(\sum_{k=1}^{K}\gamma_{j,k}) + \sum_{k=1}^{K}\log\Gamma(\gamma_{j,k})\right)\end{aligned}$$

Derivando $\mathscr{L}_{\gamma_{j,k}}$,

$$\frac{d\mathscr{L}}{d\gamma_{j,k}} = \left(\Psi'(\gamma_{j,k}) - \Psi'(\sum_{r=1}^{K}\gamma_{j,r})\right)\left(\alpha - 1 + \sum_{i=1}^{n_{d_j}}\phi_{j,i,k} - (\gamma_{j,k}-1)\right)$$

Colocando $\mathscr{L}_{\gamma_{j,k}}$ igual a zero e isolando $\gamma_{j,k}$, tem-se

$$\gamma_{j,k} = \alpha + \sum_{i=1}^{n_{d_j}}\phi_{j,i,k} \tag{3.44}$$

E por fim, maximizar a Equação (3.42) em relação a λ

$$\begin{aligned}\mathscr{L}_\lambda = &\sum_{k=1}^{K}\sum_{i=1}^{n}(\beta_i - 1)\left(\Psi(\lambda_{k,i}) - \Psi(\sum_{l=1}^{n}\lambda_{k,l})\right)\\ &+\sum_{k=1}^{K}\sum_{j=1}^{m}\sum_{i=1}^{n_{d_j}}\varphi_{j,i,k}\left(\Psi(\lambda_{k,i}) - \Psi(\sum_{l=1}^{n}\lambda_{k,l})\right)\\ &-\sum_{k=1}^{K}\left(\left(\sum_{i=1}^{n}(\lambda_{k,i}-1)\left(\Psi(\lambda_{k,i}) - \Psi(\sum_{l=1}^{n}\lambda_{k,l})\right)\right)\right.\\ &\left.-\log\Gamma(\sum_{i=1}^{n}\lambda_{k,i}) + \sum_{i=1}^{n}\log\Gamma(\lambda_{k,i})\right)\end{aligned}$$

Derivando $\mathscr{L}_\lambda$

$$\frac{d\mathscr{L}_{\lambda_{k,i}}}{d\lambda_{k,i}} = \left(\Psi'(\lambda_{k,i}) - \Psi'(\sum_{l=1}^{n}\lambda_{k,l})\right)\left(\beta - 1 + \sum_{j=1}^{m}\sum_{l=1}^{n_{d_j}}1(w_{j,l} = w_i)\varphi_{j,l,k} - (\lambda_{k,i}-1)\right)$$

Colocando $\mathscr{L}_{\lambda_{k,i}}$ igual a zero e isolando $\lambda_{k,i}$, tem-se

$$\lambda_{k,i} = \beta_k + \sum_{j=1}^{m} \sum_{l=1}^{n_{d_j}} 1(w_{j,l} = w_i), \varphi_{j,l,k} \tag{3.45}$$

onde a expressão $1(w_{j,l} = w_i)$ é igual a 1 caso o termo na posição l no documento d_j for igual a palavra w_i, e 0 caso contrário.

Por fim, chegou-se nas operações de atualização do algoritmo – equações 3.45, 3.44 e 3.43. Na próxima seção será descrito o algoritmo de inferência variacional.

3.2.4 Algoritmo de inferência variacional

O método de inferência variacional transforma o problem de inferência probabilística em um problema de otimização. Esse problema de otimização pode ser resolvido iterando em direção do gradiente da função objetiva estabelecida por $\mathscr{L}(\gamma, \varphi, \lambda | \alpha, \beta)$. Derivando o ELBO $\mathscr{L}(\gamma, \varphi, \lambda | \alpha, \beta)$, são obtidas as atualizações referentes as equações 3.45, 3.44 e 3.43 que aproximam o valor da verossimilhança do modelo, $p(w|\alpha, \beta)$. A descrição desse método está no Algoritmo 4.

Algoritmo 4: Algoritmo de inferência variacional para o LDA

Entrada : número de tópicos K, coleção de documentos, hiper-parâmetros α e β, número de iterações T

início
- Inicia $\gamma_j = \alpha + \frac{n}{K}$ para todo documento d_j ;
- Inicia aleatoriamente $\lambda_i = Dir(\beta)$ para toda palavra w_i do vocabulário ;
- $logverossimilhanca = 0$;
- **enquanto** *logverossimilhanca não convergiu* **faça**
  - **para** *documento* d_j *com* $j \in [1, m]$ **faça**
    - **repita**
      - **para** *palavra* w_i *com* $i \in [1, n_{d_j}]$ *no documento* d_j **faça**
        - **para** *tópico* $k \in [1, K]$ **faça**
          - $\varphi_{j,i,k} = \exp\left(\Psi(\gamma_{j,k}) + \Psi(\gamma_{k,i} - \Psi(\sum_{l=1}^{m} \lambda_{k,l}))\right)$;
        - Normaliza o vetor $\varphi_{j,i}$ para somar 1 ;
      - **para** *tópico* $k \in [1, K]$ **faça**
        - $\gamma_{j,k} = \alpha + \sum_{i=1}^{n_{d_j}} \varphi_{j,i,k}$;
    - **até** *convergência do vetor* γ_j;
  - **para** *palavra* w_i *com* $j \in [1, n]$ **faça**
    - **para** *tópico* $k \in [1, K]$ **faça**
      - $\phi_{k,i} = \beta_k + \sum_{j=1}^{m} \sum_{l=1}^{n_{d_j}} 1(w_{j,l} = w_i) \varphi_{j,l,k}$;
    - Normalize o vetor ϕ para somar 1 ;
  - $logverossimilhanca = logverossimilhanca + \mathscr{L}(\gamma, \varphi, \lambda | \alpha, \beta)$;

Algumas considerações devem ser feitas sobre o Algoritmo 4. Primeiro, a versão aqui apresentada não descreve como encontrar os hiperparâmetros α e β, é assumido que eles são simétricos e dado como entrada. O hiperparâmetro é simétrico quando é um valor constante para todas as componentes da distribuição. Em termos práticos, encontra-se na literatura valores padrões de $alpha = 50/K$ e $\beta = 0.01$ (STEYVERS; GRIFFITHS, 2007; WEI; CROFT, 2006), sendo esses valores sugerido para a maioria das aplicações.

CAPÍTULO

4

INFERÊNCIA *ONLINE* PARA O LDA

Com o aumento da quantidade de novos dados no formato textual constantemente publicados, a qualidade dos métodos de mineração de texto podem ser prejudicadas. Quando se trata de notícias, ou documentos textuais disponíveis na Internet, as aplicações devem considerar documentos sobre o fluxo de textos que, em geral, são formados por grandes coleções de documentos, no sentido prático, possivelmente infinitas.

Em muitas aplicações práticas, é essencial transformar de forma rápida esse grande volume de dados textuais em informações e conhecimento útil. Tomando como exemplo o interesse em identificar os tópicos em notícias, deve-se considerar um método que não exija a especificação do número de documentos, ou seja, considera-se infinita a quantidade de documentos que chegam no fluxo. Essa exigência inviabiliza a aplicação de vários métodos encontrados na literatura que percorrem toda a coleção iterativamente e que consomem grande quantidade de memória afim de encontrar as estruturas latentes.

Em um contexto não supervisionado, esses problemas caracterizam o agrupamento em fluxo de dados. Esse problema tem como objetivo agrupar uma sequência X objetos em K grupos distintos. Cada objeto deve ser atribuído a um dos K grupos na ordem que eles chegam no fluxo (CHARIKAR *et al.*, 1997). O problema de extração de tópicos em fluxo de documentos textuais pode ser visto como um caso especial do problema de agrupamento em fluxo de dados. Porém, o agrupamento deve ser realizado tanto para os documentos que chegam no fluxo quanto para as palavras nesses documentos. Com base nisso, pesquisadores desenvolveram alternativas para os modelos probabilísticos de tópicos em fluxo de dados (BANERJEE; BASU, 2007), e também a versão *online* do LDA (HOFFMAN; BLEI; BACH, 2010). Neste capítulo é descrita a versão *online* do algoritmo de inferência variacional para o modelo LDA.

4.1 Aprendizado online

Os algoritmos *online* modernos de aprendizado de máquinas baseiam-se na teoria da aproximação estocástica. Nesta seção é descrito o arcabouço geral dos algoritmos *online* de aprendizado. Inicialmente é descrita a tarefa de aprendizado, com a descrição de uma função geral baseada na minimização do erro e a aplicação do método de gradiente descendente estocástico nessa função geral. Em seguida, é descrita a versão *online* do LDA (oLDA), que utiliza aproximação estocástica para otimizar o problema de otimização estabelecido pelo método de inferência variacional.

4.1.1 Otimização Estocástica

Na tarefa tradicional de aprendizado de máquina, cada exemplo é um par (x,y) composto por um conjunto de atributos x e a informação de classe y. Considere uma função $erro(\hat{y},y)$ que mede a perda ao se estimar uma classe $\hat{y}$ conhecendo a classe real y. O objetivo é encontrar uma função f_v, parametrizada por um vetor de pesos v, que minimiza $Q((x,y),v) = erro(f_v(x),y)$. Assim, dado um conjunto de treino $\{(x_1,y_1),\ldots,(x_n,y_n)\}$, pode-se calcular o risco empírico $E(f_v)$ da seguinte forma

$$E(f_v) = \frac{1}{n}\sum_{i=1}^{n} erro(f_v(x_i),y_i). \tag{4.1}$$

O risco empírico $E(f_v)$ pode ser minimizado utilizando o método de gradiente descendente. Nesse método, em cada iteração t, atualiza-se o vetor de pesos v em direção do gradiente de $E(f_v)$,

$$v^{(t+1)} = v^{(t)} - \rho_t \frac{1}{n}\sum_{i=1}^{n} \nabla_v Q((x_i,y_i),v^{(t)}), \tag{4.2}$$

onde ρ é a taxa com que o erro será considerada na atualização dos vetores de pesos.

Algoritmos de otimização estocástica seguem um processo não determinístico para estimar o gradiente de $E(f_v)$. Em vez de calcular exatamente o gradiente de $E(f_v)$, em cada iteração é estimado esse gradiente com base em um simples exemplo (x_t,y_t) escolhido aleatoriamente (BOTTOU, 1998):

$$v^{(t+1)} = v^{(t)} - \rho_t \nabla_v Q((x_t,y_t),v^{(t)}). \tag{4.3}$$

Com isso, espera-se que a Atualização 4.3 se comporte como a Atualização 4.2, apesar do ruído de gradiente inserido. Algoritmos baseados no gradiente estocástico utilizam estimativas instantâneas do gradiente, o que implica que o vetor com a direção de atualização está sujeito a flutuações aleatórias denominadas ruído de gradiente (BOTTOU, 2010). Por outro lado, esse processo depende apenas do exemplo escolhido aleatoriamente no momento t, não sendo necessário percorrer os exemplos das iterações anteriores. Por esse motivo, o gradiente estimado é mais fácil de computar.

A convergência do gradiente estocástico foi bastante estudada, principalmente nos trabalhos de Bottou (1998), Bottou (2004), Bottou (2010). Bottou demonstrou que o gradiente estocástico tende a uma solução ótima global v^* de $E(f_{v^*})$, caso a função Q seja uma função convexa, caso contrário, tende a uma solução ótimo local. Essa convergência é garantida desde que a taxa de erro ρ diminua ao longo das iterações, satisfazendo as seguintes condições: $\sum_t \rho_t^2 < \infty$ e $\sum_t \rho_t = \infty$.

A desvantagem do processo estocástico está na velocidade de convergência, que pode se tornar lenta devido ao ruído aplicado pelo cálculo do gradiente aproximado. Além disso, a velocidade com que o valor da taxa de erro ρ diminui influencia na velocidade de convergência.

Uma técnica comum em aprendizado estocástico é utilizar *mini-batches* para atualizar o modelo. Com isso, em vez de utilizar apenas um exemplo por vez, utiliza-se vários exemplos com o objetivo de diminuir o ruído. Utilizando um *minibatch* com b exemplos, a aproximação do gradiente é a seguinte

$$v^{(t+1)} = w^{(t)} - \rho_t \frac{1}{b} \sum_{i=1}^{b} \nabla_v Q((x_i, y_i), v^{(t)}). \tag{4.4}$$

A média do gradiente estocástico dos b exemplos possuem o mesmo valor esperado, logo a aproximação ainda é válida.

4.1.2 *Aprendizado* online *com o LDA*

As técnicas tradicionais de inferência do LDA, como o algoritmo de amostragem de Gibbs e o algoritmo de inferência variacional, requerem uma passagem completa por toda a coleção de documentos em cada iteração. Tanto para o algoritmo de Gibbs quanto para o algoritmo variacional é necessário percorrer todos os termos de cada documento da coleção. Isso claramente pode retardar o processamento em grandes coleções de documentos, e também não é adequado aplicar tais técnicas onde novos documentos estão constantemente chegando. Assim, para resolver esses problemas, Hoffman, Blei e Bach (2010) propôs uma versão *online* do LDA. O algoritmo para inferência *online* do LDA proposto por Hoffman baseia-se na utilização da técnica de otimização estocástica para resolver o problema de otimização estabelecido no método de inferência variacional. Algoritmos de otimização estocástica seguem um processo não determinístico para estimar o gradiente de uma função objetivo. Inferência variacional estocástica provê uma abordagem escalável e muito mais eficiente para aproximar a distribuição *a posteriori* do LDA.

O método de inferência variacional tradicional aplicado no LDA e a notação utilizada nesta seção são descritos com detalhes na Seção 3.2. Na versão *online* em vez de otimizar todo o conjunto de dados, o método de inferência variacional estocástico utiliza apenas uma amostra dos dados escolhidos aleatoriamente (BOTTOU, 2004). Uma amostra pode ser um documento único ou um subconjunto de documentos da coleção. Com apenas as estatísticas obtidas por

uma amostra, o algoritmo ajusta as variáveis variacionais. É importante distinguir as variáveis variacionais entre variáveis locais e globais (Veja na figura 2 a representação das variáveis variacionais do modelo LDA). As variáveis locais mantém estatísticas de cada documento especificadamente e correspondem a distribuição variacional γ e φ. As variáveis globais são as proporções que relacionam tópicos e palavras, correspondente a distribuição λ.

Note que as atualizações das variáveis locais (equações 3.44 e 3.43) utilizam apenas estatísticas do documento que está sendo processados. Assim, na versão online, é amostrado um documento d_j (ou um sub-conjunto de documentos) da coleção e computado os parâmetros γ e φ utilizando a mesma sub-rotina da versão em lote. A variável $\hat{\lambda}$ é criada para manter as estatísticas dos tópicos obtidos pelas variáveis locais,

$$\hat{\lambda} = \eta + n\sum_{i=1}^{N} \varphi_{j,i,k} w_{i,j}, \tag{4.5}$$

onde n é o número de documentos, considerando que n é um número grande para justificar o processamento *online*. Essa equação vem da Equação 3.45 aplicado n vezes para um documento amostrado.

Agora, para atualizar as variáveis globais é necessário aplicar a interpolação da variável $\hat{\lambda}$ com a variável global λ,

$$\lambda_k^{(t+1)} = (1-\rho_t)\lambda_k^{(t)} + \rho_t \hat{\lambda}_k, \tag{4.6}$$

onde ρ_t é o fator de aprendizado.

Note que a atualização da variável global λ em um momento $(t+1)$ utiliza estatística obtidas em um momento anterior (t), e que essa atualização não requer a passagem por toda a coleção de documentos. Em cada momento, novas amostras de documentos são retiradas da coleção (ou do fluxo de documentos) e são atualizadas as estatísticas locais e globais.

No Algoritmo 5 estão descritos os passos para a inferência *online* do LDA. A convergência desse algoritmo é garantida pelas propriedades de otimização estocástica (veja Seção 4.1.1).

A fundamentação da versão online do LDA é baseada na aplicação da otimização estocástica na otimização estabelecida pelo método de inferência variacional do LDA. No caso do método de inferência variacional, objetiva-se otimizar o logaritmo da verossimilhança do modelo. Essa otimização é obtida pela maximização do ELBO, definido pela Equação 3.42. Aqui, é reescrito o ELBO $\mathscr{L}$ em função dos parâmetros variacionais locais, γ e φ, e do parâmetro global λ. Então, para o LDA, pode-se escrever o ELBO como:

$$\mathscr{L}(\gamma, \varphi, \lambda) \triangleq \sum_{d_j \in D} l(\gamma_j, \varphi_j, \lambda), \tag{4.7}$$

onde $l(\gamma_j, \varphi_j, \lambda)$ é a contribuição do documento d_j para o ELBO. Como a ocorrência das palavras para um documento d_j é observado, é possível aplicar o $E-step$ semelhante ao processo em

Algoritmo 5: *online* LDA

Entrada:
Um fluxo de documentos textuais S
Número de tópicos K
hiper-parâmetros α e β

Saída :
estimativa das distribuições documento-tópicos θ e tópico-palavra ϕ

início
  Inicializa $\lambda^{(0)}$ aleatoriamente ;
  repita
    Amostre um documento d_j do fluxo de documents ;
    Inicialize $\gamma_{j,k} = 1$, para $k \in \{1,\ldots,K\}$;
    repita
      para *cada termo* w_i *do documento* d_j **faça**
        para $k = 1$ *até* K **faça**
          $\varphi_{j,i,k} \propto \exp\left(\Psi(\gamma_{j,k}) - \Psi(\sum_{l=1}^{K}\gamma_{j,l}) + \Psi(\lambda_{k,i})\right)$;
      $\gamma_j = \alpha + \sum_{i=1}^{N_{d_j}} \varphi_{j,i}$;
    até *convergência dos parâmetros locais*;
    para $k = 1$ *até* K **faça**
      $\hat{\lambda}_k = \beta + D\sum_{i=1}^{N_{d_j}} \varphi_{j,i,k} w_{j,i}$;
    $\lambda^{(t)} = (1-\rho_t)\lambda^{(t-1)} + \rho_t \hat{\lambda}$;
  até *existir documento no fluxo*;

lote para encontrar os parâmetros locais γ e φ, e mantendo o parâmetro global λ fixo. A variável intermediária $\hat{\lambda}$ mantém uma estimativa para o parâmetro global λ utilizando estatísticas da amostra e das variável locais calculadas. Com isso, é possível atualizar $\lambda^{(t+1)}$ usando uma média ponderada entre seu valor anterior, $\lambda^{(t)}$, e $\hat{\lambda}$,

$$\lambda^{(t+1)} = \lambda^{(t)} + \rho_t D \nabla_\lambda l(\gamma_j, \varphi_j, \lambda), \tag{4.8}$$

onde $\rho \triangleq (\tau_0 + t)^{-\kappa}$, $\kappa \in (0.5, 1]$ controla a taxa de esquecimento dos valores antigos de $\hat{\lambda}$ e $\tau_0 \leq 0$ diminui a importância das iterações iniciais. A condição de que $\kappa \in (0.5, 1]$ é necessária para garantir a convergência. O valor de ρ decresce ao longo do tempo. Note que ρ pode ser considerado como o coeficiente de aprendizagem do parâmetro λ.

O Algoritmo 5 também pode ser justificado pelo trabalho de Neal e Hinton (1998), nos quais foram apresentadas provas que demonstram o porquê de versões *online* dos algoritmos baseados em EM (*Expectation Maximization*) funcionarem. Considerando o contexto no qual um algoritmo tradicional de EM é aplicado, tem-se uma variável aleatória observada Z e uma outra variável aleatória não observada Y. Assume-se que a probabilidade conjunta de Y e Z é parametrizada usando uma distribuição θ, como $p(Z,Y|\theta)$. A probabilidade marginal de Z é então $P(Z|\theta) = \sum_y p(y,z|\theta)$. Com os dados observados, Z, deseja-se encontrar o valor de θ que maximiza o logaritmo da verossimilhança do modelo, $L(\theta) = \log P(z|\theta)$. Para as versões *online*

do algoritmo EM, deseja-se encontrar θ, dado um número independente de dados decompostos como um fluxo do tipo $(Z_1, \ldots)$, assim como as variáveis não observadas podem ser decompostas como $(Y_1, \ldots)$. Utilizando as variáveis decompostas, tem-se uma estimativa $\hat{P}(Y) = \prod_i \hat{P}_i(Y_i)$, na qual uma função F retornará o valor dessa estimativa parametrizada por θ, $F(\hat{P}, \theta) = \sum_i F_i(\hat{P}_i, \theta)$. Então, usando o Teorema 2 apresentado no trabalho de Neal e Hinton (1998), é mostrado que: se $F(\hat{P}, \theta)$ tem máximo local em $\hat{P}^*$ e θ^*, então $L(\theta)$ também terá o máximo em θ^*. Assim, um algoritmo EM pode ser usado para otimizar a decomposições dos dados observados, e também otimizar a verossimilhança do modelo.

CAPÍTULO

5

AVALIAÇÃO DOS MODELOS PROBABILÍSTICOS DE TÓPICOS

A forma mais comum de avaliar os modelos probabilísticos de tópicos é calculando o logaritmo da verossimilhança do modelo (WALLACH *et al.*, 2009). Isso é normalmente realizado separando a coleção de documentos em dois subconjuntos, um para treino e outro para teste. Com os documentos de treino cria-se o modelo. Em seguida, averígua-se o quão bom esse modelo descreve documentos não conhecidos, utilizando os documentos de testes. Quanto maior o valor do logaritmo da verossimilhança melhor é o modelo. Para o LDA, o modelo corresponde a distribuição de tópicos por palavras, ϕ. O algoritmo de inferência é aplicado na coleção de treino para encontrar as distribuições ϕ e θ_{treino}. Já o conjunto de teste corresponde aos documentos não conhecidos pelo modelo. A distribuição de documentos por tópicos, θ_{treino}, não é considerada na avaliação pois descreve apenas aos documentos de treino, logo será necessário computar uma nova distribuição θ_{teste} com apenas os documentos de teste.

Avaliar a efetividade do modelo de extração de tópicos está fortemente relacionada com a correta decomposição do conjunto de documentos em conceitos humanamente interpretáveis. O que se encontra na literatura para avaliar modelos de mistura de tópicos são métricas que verificam a capacidade do modelo aprendido em predizer dados não vistos (CHANG *et al.*, 2009). O modelo que descreve uma coleção de documentos será bom se a distribuição de tópicos por palavras ϕ também corresponder aos tópicos contidos na coleção de treino. Uma métrica bastante utilizada é a medida de perplexidade (*perplexity* measure) (WAAL; BARNARD, 2008). Para aplicar essa medida é necessário dividir todo o conjunto de dados em treinamento e teste. O modelo é criado com o conjunto de treino, então mede-se o quão "perplexo" está o modelo no conjunto de teste, ou seja, é medido o quão bem está a probabilidade das palavras do documento de teste representada pela distribuição de tópicos por palavras obtidas pelo modelo. Quanto menor o valor de perplexidade melhor será o modelo. A perplexidade é calculada da seguinte

forma:

$$perplexidade(w) = \exp\left(-\frac{\log p(w|\alpha,\beta)}{\log \sum_{j=1}^{m} n_{d_j}}\right). \tag{5.1}$$

O logaritmo da verossimilhança do modelo, $p(w|\alpha,\beta)$, é obtido de forma diferente dependendo do algoritmo de inferência empregado. No algoritmo de amostragem de Gibbs é calculado da seguinte forma:

$$p(w|\alpha,\beta) = \sum_{j=1}^{m} \sum_{i=1}^{n_{d_j}} \log \sum_{k=1}^{K} \theta_{j,k}\phi_{k,i}. \tag{5.2}$$

Já no algoritmo de inferência variacional, o logaritmo da verossimilhança do modelo é aproximado pelo ELBO ($\mathscr{L}$ – Equação 3.30).

Essas medidas são boas para comparações entre os modelos probabilísticos, entretanto, os valores obtidos nas medições não necessariamente condizem com a correta relação entre os tópicos encontrados e os assuntos descritos na coleção (CHANG *et al.*, 2009).

Informalmente, avaliações desses modelos podem ser realizadas das seguintes formas: (1) Inspecionar cada tópico, fazendo a busca por palavras de maior probabilidade e verificar se essas palavras são coerentemente relacionadas a algum conceito presente na coleção. (2) Manter um conjunto de documentos escolhidos aleatoriamente e ver se os tópicos encontrados fazem sentido ou não. Baseado nisso, o trabalho de (CHANG *et al.*, 2009) propõe métodos quantitativos para mensurar o significado semântico dos tópicos inferidos pelo modelo. Um método, chamado *Word Intrusion*, mede a coerência desses tópicos. Na tarefa realizada por esse método, um tópico é escolhido aleatoriamente, em seguida, as cinco palavras mais relacionadas com esse tópico é selecionada, uma sexta palavra é escolhida do conjunto de palavras menos relacionadas ao tópico escolhido. Entre essas seis palavras, o usuário deve encontrar aquela que menos se relaciona com todas as outras palavras. Se o tópico não é coerente semanticamente, será difícil apontar qual é a palavra menos relacionada. Outro método proposto é chamado *Topic Intrusion*, nessa tarefa é mostrado o título e partes do texto de um documento. Junto com o documento são apresentados quatro tópicos (cada tópico contém oito palavras com maior probabilidade). Desses tópicos, três são altamente relacionados com o documento. O tópico restante é escolhido aleatoriamente do conjunto de tópicos poucos relacionados. O usuário deve encontrar o tópico que menos se relaciona ao documento apresentado.

Ainda com o objetivo de se obter a avaliação da interpretabilidade, no trabalho de Newman *et al.* (2010) foi proposto um método automático baseado na informação mútua entre pares de palavras que formam o tópico, chamado *Pointwise Mutual Information* (PMI), para simular a avaliação humana sobre a qualidade dos tópicos. No trabalho de Mimno *et al.* (2011) foi avaliado a metodologia para computar ao coerência, substituindo PMI pelo logaritmo da probabilidade condicional dos pares de palavras. Já no trabalho de Musat *et al.* (2011) foi incorporado a hierarquia do WordNet para capturar a relevância dos tópicos. No trabalho de Lau,

Newman e Baldwin (2014), foram utilizadas diferentes técnicas de avaliação com o objetivo de encontrar a melhor, como resultado, a medida *Normalized Pointwise Mutual Information* (NPMI) (veja a Definição 2) foi apontada como a que mais se aproxima da avaliação feita por especialistas, e pode ser utilizada para automatizar a avaliação da coerência do tópico considerando os termos selecionados como descritores e sua coocorrência em relação a uma coleção de referência. A coleção de referência é utilizada para calcular a coocorrência entre os termos relacionados, e usualmente, os documentos da Wikipédia[1] são utilizados como referência externa.

Definição 2 (NPMI – *Normalized Pointwise Mutual Information*)**.** Seja $top_K^L = \{w_1, \ldots, w_L\}$ o conjunto das L palavras com maior probabilidade na distribuição de tópicos. Newman *et al.* (2010) assumem que quanto maior a similaridade média entre os pares das palavras em top_K^L, mais coerente é o tópico. Com isso, pode-se definir a função do NPMI como:

$$NPMI(top_K^L) = \sum_{i=1}^{L} \sum_{j=1}^{L-1} \frac{\log \frac{P(w_i, w_j)}{P(w_i)P(w_j)}}{-\log P(w_i, w_j)} \tag{5.3}$$

[1] <https://www.wikipedia.org/>

CAPÍTULO

6

COMPARANDO LDA COM NMF

O método NMF (*Nonnegative Matrix Factorization*) (PAATERO; TAPPER, 1994; LEE; SEUNG, 1999) fatoriza aproximadamente a matriz com elementos não negativos em duas outras matrizes também com elementos não negativos (Veja a Definição 3).

Definição 3 (NMF – *Nonnegative Matrix Factorization*)**.** Dado uma matriz com valores não negativos $F \in \mathbb{R}^{m \times n}$, quando o número de dimensões reduzidas é K, o objetivo do NMF é encontrar duas matrizes $A \in \mathbb{R}^{m \times K}$ e $B \in \mathbb{R}^{K \times n}$ com apenas entradas não negativas tal que

$$F \approx A \cdot B \tag{6.1}$$

Os fatores A e B são obtidos pela minimização de uma função de custo definida por uma medida de "distância". Existem diferentes tipos de funções de custo (LEE; SEUNG, 2001). A função que se relaciona com a formulação dos modelos probabilísticos de tópicos é aquela baseada na divergência KL. Essa função é definida como

$$Q_{NMF-KL} = \min \sum_{j,i} \left(F_{j,i} \log \frac{F_{j,i}}{(AB)_{j,i}} - F_{j,i} + (AB)_{j,i} \right), \tag{6.2}$$

onde $F_{j,i}$ é a entradas da linha j e coluna i matriz F, no caso de uma matriz documento-termo, o valor de $F_{j,i}$ pode ser a frequência do termo w_i no documento d_j. Note que o valor de $-F_{j,i} + (AB)_{j,i}$ será igual a zero caso $F_{j,i} = (AB)_{j,i}$.

A técnica mais simples de resolver a otimização da Equação 6.2 é aplicando o método de gradiente descendente. Fazendo as derivações, chega-se nas seguinte equações de atualização

$$A_{j,k} = A_{j,k} \frac{\sum_i B_{k,i} F_{j,i} / (AB)_{j,k}}{\sum_q B_{k,q}}, \tag{6.3}$$

$$B_{k,i} = B_{k,i} \frac{\sum_j A_{j,k} V_{j,i} / (AB)_{j,i}}{\sum_p A_{p,k}}. \tag{6.4}$$

Assim, interpolando as atualizações das equações 6.3 e 6.4 em várias iterações chega-se nos fatores que aproximam a matriz F. A convergência desse algoritmo não é apropriadamente demonstrada, entretanto, no trabalho de (LEE; SEUNG, 2001) é demonstrado que em cada iteração as atualizações sempre irão diminuir o valor resultante da Equação 6.2.

Apesar de não ser um método probabilístico, o NMF é descrito nesse capítulo pois apresenta similaridades com os modelos de tópicos. Além disso, o NMF e o LDA são duas técnicas popularmente aplicadas no problema de extração de tópicos em coleções de documentos. Nessa seção é realizada uma análise comparativa entre essas duas técnicas, demonstrando que NMF com divergência KL aproxima ao algoritmo de inferência variacional do LDA. Essa análise comparativa é útil para elucidar a implementação do algoritmo de inferência variacional e explorar as relações entre as diferentes técnicas.

A equivalência entre o NMF e o pLSI (*probabilistic Latent Semantic Indexing*) tem sido discutida em vários trabalhos (BUNTINE, 2002; GAUSSIER; GOUTTE, 2005). Ding, Li e Peng (2008) demonstraram que NMF e PLSI otimizam a mesma função objetivo. Apesar do LDA ser a contrapartida com fundamentação em probabilidade Bayesiana do pLSI (GIROLAMI; KABáN, 2003), a equivalência entre NMF e LDA não é bem definida. Entretanto, existem evidências que tal relação intrínseca também exista (J.; LIU; CAO, 2012; GERSHMAN; BLEI, 2012). Nessa seção, o objetivo é esclarecer essa relação em termos de formulação matemática, demonstrando que o NMF com divergência KL aproxima a solução obtida pelo algoritmo de inferência variacional do LDA. Além disso, será demonstrado a relação entre os dois algoritmos.

As correspondências entre NMF com divergência KL e o algoritmo de inferência variacional para o LDA seguem do fato de que ambos tentam minimizar a divergência entre as estatísticas que relacionam a frequência de palavras, documentos por tópicos e tópicos por palavras. Para esclarecer essa relação, o NMF será descrito como uma relaxação do problema estabelecido no método de inferência variacional. A equivalência é alcançada quando as funções $\log\Gamma(\cdot)$ e $\Psi(\cdot)$ são aproximadas e substituídas nas derivações do LDA. Essa relação é demonstrada no Teorema 1.

Teorema 1. A função objetivo do NMF com a divergência KL é uma aproximação do ELBO (*Evidence Lower Bound*) do LDA com *priori* simétricas.

Demonstração. Inicialmente, expande-se o ELBO usando fatoração da distribuição conjunta do LDA, p (Equation 2.3), e a distribuição variacional, q (Equation 3.28):

$$
\begin{aligned}
\mathscr{L} &\triangleq E_q[\log p(\theta,z,w|\alpha,\beta)] - E_q[\log q(\theta,z)] \\
&= E_q[\log p(\theta|\alpha)] + E_q[\log p(z|\theta)] + E_q[\log p(w|z,\beta)] - E_q[\log q(\theta)] - E_q[\log q(z)] \\
&= \prod_{j=1}^{m}\left\{\left[\log\Gamma\left(\sum_{k=1}^{K}\alpha_k\right) - \sum_{k=1}^{K}\log\Gamma(\alpha_k) + \sum_{k=1}^{K}(\alpha_k-1)\left(\Psi\left(\gamma_{j,k}\right) - \Psi\left(\sum_{r=1}^{K}\gamma_{j,r}\right)\right)\right]\right.
\end{aligned}
$$

$$
\begin{aligned}
&+\left[\sum_{i=1}^{n_{d_j}}\sum_{k=1}^{K}\varphi_{j,i,k}\left(\Psi(\gamma_{j,k})-\Psi\left(\sum_{r=1}^{K}\gamma_{j,r}\right)\right)\right]\\
&+\left[\sum_{l=1}^{n_{d_j}}\sum_{k=1}^{K}\sum_{i=1}^{n}\varphi_{j,i}1(w_{j,l}=w_i)\log\beta_{k,i}\right]\\
&+\left[-\log\Gamma\left(\sum_{k=1}^{K}\gamma_{j,k}\right)+\sum_{r=1}^{K}\log\Gamma\left(\gamma_{j,r}\right)-\sum_{k=1}^{K}(\gamma_{j,k}-1)\left(\Psi(\gamma_{j,k})-\Psi\left(\sum_{r=1}^{K}\gamma_{j,r}\right)\right)\right]\\
&+\left[-\sum_{i=1}^{n_{d_j}}\sum_{k=1}^{K}\varphi_{j,i,k}\log\varphi_{j,i,k}\right]\Bigg\}
\end{aligned}
\tag{6.5}
$$

Agora, aproxima-se a Equação 6.5 substituindo as funções Gamma $\Gamma(\cdot)$ e Digamma $\Psi(\cdot)$. A função Gamma é definida como $\Gamma(x)=\int_0^\infty u^{x-1}e^{-u}du$, para $x>0$. Em geral, $\Gamma(x+1)=x\Gamma(x)$, e para argumentos inteiros, $\Gamma(x+1)=x!$. Para propósitos práticos, é considerado a aproximação de Stirlings da função $\Gamma(\cdot)$:

$$
\log\Gamma(x)=\log x!=\sum_{i=1}^{n}\log i\approx\int_{i=1}^{x}\log(i)di\approx x\log x-x. \tag{6.6}
$$

A função Digamma é definida como $\Psi(x)=\frac{d}{dx}\log\Gamma(x)$, e pode ser aproximada por

$$
\Psi(n)\approx\log n-c, \tag{6.7}
$$

onde c é um valor constante (MUQATTASH; YAHDI, 2006).

A distribuição γ_j pode ser relacionada com o vetor A_j associado a cada documento d_j. Da mesma forma, a distribuição β pode ser relacionada a matrix B. Assim, considerando a versão do LDA com hiper-parâmetros α simétricos, é possível reescrever o ELBO usando as correspondentes aproximações para as funções Gamma, Equação 6.6, e Digamma, Equação 6.7:

$$
\begin{aligned}
\mathscr{L}\approx\prod_{j=1}^{m}\Bigg\{&\left[\sum_{k=1}^{k}(\alpha_k-1)\left(\log\frac{A_{j,k}}{\sum_{r=1}^{K}A_{j,r}}\right)\right]\\
&+\left[\sum_{i=1}^{n}\sum_{k=1}^{K}f_{j,i}\varphi_{j,i,k}\left(\log\frac{A_{j,k}}{\sum_{r=1}^{K}A_{j,r}}\right)\right]\\
&+\left[\sum_{i=1}^{n}\sum_{k=1}^{K}F_{j,i}\varphi_{j,i,k}\left(\log\frac{B_{i,k}}{\sum_{l=1}^{n}B_{l,k}}\right)\right]\\
&+\left[\sum_{k=1}^{K}\left(A_{j,k}\left(\log A_{j,k}-1\right)-(A_{j,k}-1)\left(\log\frac{A_{j,k}}{\sum_{r=1}^{K}A_{j,r}}\right)\right)\right]\\
&+\left[\sum_{i=1}^{n}\sum_{k=1}^{K}-F_{j,i}\varphi_{j,i,k}\log\varphi_{j,i,k}\right]\Bigg\}\\
=\sum_{j=1}^{m}\sum_{i=1}^{n}\sum_{k=1}^{K}&\left(F_{j,i}\varphi_{j,i,k}\log\frac{\frac{A_{j,k}}{\sum_{r=1}^{K}A_{j,r}}\frac{B_{i,k}}{\sum_{l=1}^{n}B_{l,k}}}{\varphi_{j,i,k}}\right.
\end{aligned}
$$

$$+(\alpha_k - A_{j,k})\left(\log\frac{A_{j,k}}{\sum_{r=1}^{K} A_{j,r}}\right) - A_{j,k}(\log A_{j,k} - 1)\Bigg) \tag{6.8}$$

Considerando que os vetores A_j e B_i são normalizados de forma que $\sum_{k=1}^{K} A_{j,k} = 1$ e $\sum_{l=1}^{n} B_{i,l} = 1$, e definindo $\mathscr{R}(A_{j,k}, \alpha_k) = (\alpha_k - A_{j,k})(\log A_{j,k}) - A_{j,k}(\log A_{j,k} - 1)$, pode-se reescrever a Equação 6.8 para alcançar o seguinte problema de otimização

$$\begin{aligned}\max \mathscr{L} &\approx \max \sum_{j=1}^{m}\sum_{i=1}^{n}\sum_{k=1}^{K}\left(F_{j,i}\varphi_{j,i,k}\log\frac{A_{j,k}B_{i,k}}{\varphi_{j,i,k}} + \mathscr{R}(A_{j,k},\alpha_k)\right)\\ &\approx \min \sum_{j=1}^{m}\sum_{i=1}^{n}\sum_{k=1}^{K}\left(F_{j,i}\varphi_{j,i,k}\log\frac{\varphi_{j,i,k}}{A_{j,k}B_{i,k}} - \mathscr{R}(A_{j,k},\alpha_k)\right).\end{aligned} \tag{6.9}$$

Sabendo que $\sum_{i=1}^{n} a_i \log\frac{a_i}{b_i} \leq \sum_{i=1}^{n} a_i \log\frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} b_i}$ para qualquer a_i e b_i não negativo, e em seguida adicionando a constante $\sum_{j,i} F_{j,i}\log F_{j,i}$, tem-se

$$\begin{aligned}&\leq \min \sum_{j=1}^{m}\sum_{i=1}^{n}\left(F_{j,i}\sum_{k=1}^{K}\varphi_{j,i,k}\log\frac{\sum_{k=1}^{\mathscr{K}}\varphi_{j,i,k}}{\sum_{k=1}^{K}A_{j,k}B_{i,k}} - \sum_{k=1}^{K}\mathscr{R}(A_{j,k},\alpha_k)\right)\\ &\approx \min \sum_{j}^{m}\sum_{i=1}^{n}\left(F_{j,i}\log\frac{F_{j,i}}{\sum_{k=1}^{K}A_{j,k}B_{i,k}} - \sum_{k=1}^{K}\mathscr{R}(A_{j,k},\alpha_k)\right).\end{aligned} \tag{6.10}$$

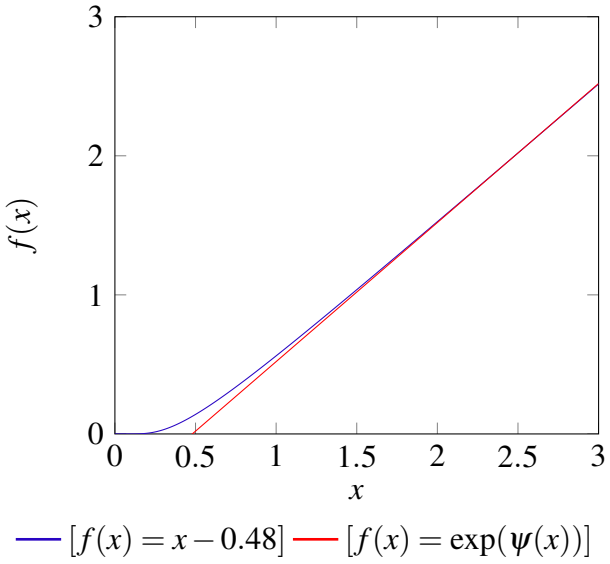
O último termo na Equação 6.10 é equivalente ao NMF (Equação 6.2) menos o termo $\mathscr{R}(A_{j,k}, \alpha_k)$. O termo $\mathscr{R}(A_{j,k}, \alpha_k)$ possui um papel importante no desempenho do LDA, ele corresponde a influência da *priori* e também inclui esparsidade na distribuição de documentos por tópicos. Quando isso é adicionado ao NMF, obtêm-se um termo regularizador que restringe os valores dos vetores A_j. Então, pode-se concluir que maximizar o ELBO do LDA com *priori* simétrica é proporcional a minimizar a função objetiva do NMF com divergência KL desconsiderando o termo regularizador.

■ □

Na teoria, os métodos iterativos aplicados no LDA e NMF são distintos e com diferentes fundamentações. Na prática, existem similaridades nas operações realizadas pelos seus algoritmos. Assim, será indicado essas equivalências e comparado as operações de atualizações do NMF, equações 6.3 e 6.4, e do LDA com inferência variacional, equações 3.44, 3.45 e 3.43.

Na regra de atualização do LDA, a operação de exponenciação sobre a função digama $\Psi(x)$ aproxima uma função linear quando $x > 0.48$ (MUQATTASH; YAHDI, 2006). Para perceber essa aproximação, veja a Figura 3.

Figura 3 – Plote da função linear $f(x) = x - 0.48$ e da função $f(x) = exp(\psi(x))$. Isso indica que a operação exponencial sobre a função digama aproxima uma função linear quando $x > 0.48$, *i.e.* $\exp(\psi(x)) \approx x - 0.48$ se $x > 0.48$



Fonte: Elaborada pelo autor.

Aproveitando a aproximação da função $\exp(\Psi(x))$, é possível aproximar o valor de φ utilizando apenas operações lineares

$$\varphi_{j,i,k} \approx \beta_{k,i} \times \frac{\gamma_{j,k}}{\sum_{k^*=1}^{\mathscr{K}} \gamma_{j,k^*}}. \tag{6.11}$$

Assim, o valor de $\varphi_{j,i}$ aproxima o produto de Hadamard entre o vetor normalizado γ_j e β_k. A matriz resultante, A, é relacionada a distribuição documento-tópicos γ. Da mesma forma, a matriz B é relacionada com a distribuição tópico-palavras β. Considerando essas relações, é possível aproximar a equação de atualização obtidas pelo método de inferência variacional para o parâmetro φ,

$$\varphi_{j,i,k} \propto \left(\frac{A_{j,k} B_{k,i}}{\sum_{k^*=1}^{K} A_{j,k^*} B_{k^*,i}} \right) \tag{6.12}$$

Sem perda de generalidade, será considerada a normalização nas linhas da matriz B, *i.e.* $\sum_i B_{k,i} = 1$. Então, usando a Equação 6.12, é possível reescrever as atualizações de cada posição do fator $A_{j,k}$ na Equação 6.3 como

$$A_{j,k} = \sum_{i=1}^{\mathscr{W}} F_{j,k} \varphi_{j,i,k}. \tag{6.13}$$

Note que a equação de atualização para o fator A_j, Equação 6.13, é similar a atualização da equação dos parâmetros γ_j sem o parâmetros α, Equação 3.44.

A equação de atualização do fator $B_{k,i}$ pode ser reescrita considerando a aproximação φ, Equação 6.12, e o último valor de $A_{j,k}$ obtido na Equação 6.13

$$\begin{aligned} B_{k,i} &= \frac{1}{\sum_j A_{j,k}} \frac{\sum_j F_{j,k} A_{j,k} B_{k,i}}{(AB)_{j,k}} \\ &= \frac{\sum_j F_{j,k} \varphi_{j,i,k}}{\sum_j \sum_i F_{j,k} \varphi_{j,k,i}}. \end{aligned} \tag{6.14}$$

Pela Equação 6.14, pode-se notar que o valor de $B_{k,i}$ é obtido com valor de φ para uma palavra específica w_i e tópico k para cada documento d_j, e normalizado para cada palavra w_i do vocabulário. Isso corresponde a distribuição de tópicos por palavras para um tópico k, representado pela distribuição β_k no LDA.

A indicação da relação entre o NMF com a divergência KL e o LDA com o algoritmo de inferência variacional é importante para entender os procedimentos realizados pelos modelos de tópicos. E com esse objetivo, foi mostrado que o NMF (com divergência KL) é de fato um caso especial do LDA onde é assumido uma distribuição de Dirichlet uniforme, e que o algoritmo de atualizações multiplicativas para resolução do NMF pode ser aproximado para as atualizações estabelecidas pelo algoritmo de inferência variacional do LDA.

CAPÍTULO

7

CONCLUSÃO

A forma como pesquisadores de aprendizado de máquinas modelam textos e outros objetos mudaram após o surgimento dos modelos probabilísticos de tópicos. O arcabouço fornecido pelo LDA serviu como ferramenta para o desenvolvimento de vários outros modelos. Assim, o modelo base LDA foi investigado e detalhado a fim de entender o funcionamento prático dos algoritmos de inferência. O modelo foi descrito, e principalmente, as derivações foram feitas durante os estudos sobre os algoritmos de inferência. Esses estudos servirão como referência para estudos futuros, principalmente para novos alunos que por ventura queiram explorar modelos probabilísticos de tópicos.

Uma limitação deste estudo foi na forma com que a descrição do modelo LDA foi apresentada, na qual se levou pouco em consideração a interpretação Bayesiana do modelo. Em uma descrição que enfatizam o LDA como um modelo Bayesiano completo, as distribuições *priori* deveriam ser melhores descritas. A importância da *priori* no modelo LDA é bem discutida no trabalho (WALLACH; MIMNO; MCCALLUM, 2009). Também foi pouco explorado o LDA como uma Rede Bayesiana, e como estender esse modelo a fim de se obter outros modelos.

O LDA foi formalmente descrito e também apresentado os dois principais algoritmos de inferência, o método de amostragem de Gibbs e o método de inferência variacional. O grande objetivo do estudo apresentado foi registrar detalhadamente o processo de derivação do modelo para a obtenção do algoritmo de inferência. Uma outra contribuição resultante dos estudos descritos neste trabalho está na análise comparativa do modelo LDA com o método de fatoração de matrizes NMF. Uma vez bem definida essa relação, será possível explorar o melhor desses dois métodos, possibilitando o desenvolvimento de novos algoritmos otimizados (FALEIROS; LOPES, 2016).

REFERÊNCIAS

BANERJEE, A.; BASU, S. Topic models over text streams: A study of batch and online unsupervised learning. In: **SDM**. SIAM, 2007. ISBN 978-0-89871-630-6. Disponível em: <http://dblp.uni-trier.de/db/conf/sdm/sdm2007.html#BanerjeeB07>. Citado na página 37.

BERRY, M. W.; DUMAIS, S. T.; O'BRIEN, G. W. Using linear algebra for intelligent information retrieval. **SIAM Rev.**, Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, v. 37, n. 4, p. 573–595, dez. 1995. ISSN 0036-1445. Disponível em: <http://dx.doi.org/10.1137/1037127>. Citado na página 10.

BISHOP, C. M. **Pattern Recognition and Machine Learning (Information Science and Statistics)**. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2006. ISBN 0387310738. Citado 2 vezes nas páginas 14 e 25.

BLEI, D. M. Introduction to probabilistic topic models. **Communications of the ACM**, 2011. Disponível em: <http://www.cs.princeton.edu/~blei/papers/Blei2011.pdf>. Citado 3 vezes nas páginas 9, 17 e 18.

BLEI, D. M.; NG, A. Y.; JORDAN, M. I. Latent dirichlet allocation. **J. Mach. Learn. Res.**, JMLR.org, v. 3, p. 993–1022, mar. 2003. ISSN 1532-4435. Disponível em: <http://dl.acm.org/citation.cfm?id=944919.944937>. Citado 6 vezes nas páginas 9, 13, 16, 18, 28 e 29.

BOTTOU, L. On-line learning in neural networks. In: SAAD, D. (Ed.). New York, NY, USA: Cambridge University Press, 1998. cap. On-line Learning and Stochastic Approximations, p. 9–42. ISBN 0-521-65263-4. Disponível em: <http://dl.acm.org/citation.cfm?id=304710.304720>. Citado 2 vezes nas páginas 38 e 39.

____. Advanced lectures on machine learning: Ml summer schools 2003, canberra, australia, february 2 - 14, 2003, tübingen, germany, august 4 - 16, 2003, revised lectures. In: ____. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004. cap. Stochastic Learning, p. 146–168. ISBN 978-3-540-28650-9. Disponível em: <http://dx.doi.org/10.1007/978-3-540-28650-9_7>. Citado na página 39.

____. Large-Scale Machine Learning with Stochastic Gradient Descent. In: LECHEVALLIER, Y.; SAPORTA, G. (Ed.). **Proceedings of COMPSTAT'2010**. Physica-Verlag HD, 2010. p. 177–186. Disponível em: <http://dx.doi.org/10.1007/978-3-7908-2604-3_16>. Citado 2 vezes nas páginas 38 e 39.

BRONIATOWSKI, D. A.; MAGEE, C. L. Analysis of social dynamics on fda panels using social networks extracted from meeting transcripts. In: **SocCom**. [s.n.], 2010. Disponível em: <http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5591237&tag=1>. Citado na página 9.

BUNTINE, W. Variational extensions to em and multinomial pca. In: **In ECML 2002**. [S.l.]: Springer-Verlag, 2002. p. 23–34. Citado na página 48.

CAO, L.; LI, F.-F. Spatially coherent latent topic model for concurrent segmentation and classification of objects and scenes. In: **ICCV**. IEEE, 2007. p. 1–8. Disponível em: <http://dblp.uni-trier.de/db/conf/iccv/iccv2007.html#CaoF07>. Citado na página 9.

CHANG, J.; BLEI, D. Relational topic models for document networks. In: **AIStats**. [S.l.: s.n.], 2009. Citado na página 9.

CHANG, J.; BOYD-GRABER, J.; WANG, C.; GERRISH, S.; BLEI, D. M. Reading tea leaves: How humans interpret topic models. In: **Neural Information Processing Systems**. [S.l.: s.n.], 2009. Citado 2 vezes nas páginas 43 e 44.

CHARIKAR, M.; CHEKURI, C.; FEDER, T.; MOTWANI, R. Incremental clustering and dynamic information retrieval. In: **Proceedings of the Twenty-Ninth Annual ACM Symposium on Theory of Computing**. [S.l.: s.n.], 1997. p. 626–635. Citado na página 37.

DEERWESTER, S. C.; DUMAIS, S. T.; LANDAUER, T. K.; FURNAS, G. W.; HARSHMAN, R. A. Indexing by latent semantic analysis. **JASIS**, v. 41, n. 6, p. 391–407, 1990. Disponível em: <http://dx.doi.org/10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9>. Citado na página 10.

DING, C.; LI, T.; PENG, W. On the equivalence between non-negative matrix factorization and probabilistic latent semantic indexing. **Comput. Stat. Data Anal.**, Elsevier Science Publishers B. V., Amsterdam, The Netherlands, The Netherlands, v. 52, n. 8, p. 3913–3927, abr. 2008. ISSN 0167-9473. Citado na página 48.

FALEIROS, T. de P.; LOPES, A. de A. On the equivalence between algorithms for non-negative matrix factorization and latent dirichlet allocation. In: **ESANN 2016, 24th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges, Belgium, April 26-29, 2016, Proceedings**. [S.l.: s.n.], 2016. Citado na página 53.

FRIGG, R.; WERNDL, C. **Entropy - A Guide for the Perplexed**. Oxford University Press, 2011. In ?Probabilities in Physics?, Oxford University Press. Disponível em: <http://philsci-archive.pitt.edu/8592/>. Citado na página 31.

GAUSSIER, E.; GOUTTE, C. Relation between plsa and nmf and implications. In: **Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval**. New York, NY, USA: ACM, 2005. (SIGIR '05), p. 601–602. ISBN 1-59593-034-5. Citado na página 48.

GEMAN, S.; GEMAN, D. Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, Taylor & Francis, v. 6, n. 6, p. 721–741, nov. 1984. Disponível em: <http://dx.doi.org/10.1080/02664769300000058>. Citado na página 19.

GERSHMAN, S. J.; BLEI, D. M. A tutorial on Bayesian nonparametric models. **Journal of Mathematical Psychology**, v. 56, n. 1, p. 1–12, fev. 2012. ISSN 00222496. Citado na página 48.

GIROLAMI, M.; KABáN, A. On an equivalence between plsi and lda. In: **Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval**. New York, NY, USA: ACM, 2003. (SIGIR '03), p. 433–434. ISBN 1-58113-646-3. Citado na página 48.

GRIFFITHS, T. L.; STEYVERS, M. Finding scientific topics. **PNAS**, v. 101, n. suppl. 1, p. 5228–5235, 2004. Citado 2 vezes nas páginas 9 e 18.

HENDERSON, K.; ELIASSI-RAD, T. Applying latent dirichlet allocation to group discovery in large graphs. In: **SAC '09: Proceedings of the 2009 ACM symposium on Applied Computing**. New York, NY, USA: ACM, 2009. p. 1456–1461. ISBN 978-1-60558-166-8. Disponível em: <http://portal.acm.org/citation.cfm?id=1529607>. Citado na página 9.

HOFFMAN, M. D.; BLEI, D. M.; BACH, F. R. Online learning for latent dirichlet allocation. In: LAFFERTY, J. D.; WILLIAMS, C. K. I.; SHAWE-TAYLOR, J.; ZEMEL, R. S.; CULOTTA, A. (Ed.). **NIPS**. Curran Associates, Inc., 2010. p. 856–864. Disponível em: <http://dblp.uni-trier.de/db/conf/nips/nips2010.html#HoffmanBB10>. Citado 2 vezes nas páginas 37 e 39.

HOFMANN, T. Probilistic latent semantic analysis. In: **UAI**. [S.l.: s.n.], 1999. Citado 2 vezes nas páginas 9 e 10.

J., Z.; LIU, Z.; CAO, X. Memory-efficient topic modeling. **CoRR**, abs/1206.1147, 2012. Citado na página 48.

KULLBACK, S.; LEIBLER, R. A. On information and sufficiency. **Annals of Mathematical Statistics**, v. 22, p. 49–86, 1951. Citado na página 25.

LAU, J. H.; NEWMAN, D.; BALDWIN, T. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In: BOUMA, G.; PARMENTIER, Y. (Ed.). **Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2014, April 26-30, 2014, Gothenburg, Sweden**. The Association for Computer Linguistics, 2014. p. 530–539. ISBN 978-1-937284-78-7. Disponível em: <http://aclweb.org/anthology/E/E14/E14-1056.pdf>. Citado na página 45.

LEE, D. D.; SEUNG, H. S. Learning the parts of objects by non-negative matrix factorization. **Nature**, Nature Publishing Group, Bell Laboratories, Lucent Technologies, Murray Hill, New Jersey 07974, USA., v. 401, n. 6755, p. 788–791, out. 1999. ISSN 0028-0836. Disponível em: <http://dx.doi.org/10.1038/44565>. Citado na página 47.

____. Algorithms for non-negative matrix factorization. In: LEEN, T. K.; DIETTERICH, T. G.; TRESP, V. (Ed.). **Advances in Neural Information Processing Systems 13**. MIT Press, 2001. p. 556–562. Disponível em: <http://papers.nips.cc/paper/1861-algorithms-for-non-negative-matrix-factorization.pdf>. Citado 2 vezes nas páginas 47 e 48.

LI, F.-F.; PERONA, P. A bayesian hierarchical model for learning natural scene categories. In: **Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 2 - Volume 02**. Washington, DC, USA: IEEE Computer Society, 2005. (CVPR '05), p. 524–531. ISBN 0-7695-2372-2. Disponível em: <http://dx.doi.org/10.1109/CVPR.2005.16>. Citado na página 9.

MEI, Q.; CAI, D.; ZHANG, D.; ZHAI, C. Topic modeling with network regularization. In: **WWW**. [s.n.], 2008. Disponível em: <http://portal.acm.org/citation.cfm?id=1367512>. Citado na página 9.

MIMNO, D.; WALLACH, H. M.; TALLEY, E.; LEENDERS, M.; MCCALLUM, A. Optimizing semantic coherence in topic models. In: **Proceedings of the Conference on Empirical**

Methods in Natural Language Processing. Stroudsburg, PA, USA: Association for Computational Linguistics, 2011. (EMNLP '11), p. 262–272. ISBN 978-1-937284-11-4. Disponível em: <http://dl.acm.org/citation.cfm?id=2145432.2145462>. Citado na página 44.

MUQATTASH, I.; YAHDI, M. Infinite family of approximations of the digamma function. **Mathematical and Computer Modelling**, v. 43, n. 11 - 12, p. 1329 – 1336, 2006. ISSN 0895-7177. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0895717705004735>. Citado 2 vezes nas páginas 49 e 50.

MUSAT, C. C.; VELCIN, J.; TRAUSAN-MATU, S.; RIZOIU, M.-A. Improving topic evaluation using conceptual knowledge. In: **Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence - Volume Volume Three**. AAAI Press, 2011. (IJCAI'11), p. 1866–1871. ISBN 978-1-57735-515-1. Disponível em: <http://dx.doi.org/10.5591/978-1-57735-516-8/IJCAI11-312>. Citado na página 44.

NEAL, R.; HINTON, G. E. A view of the em algorithm that justifies incremental, sparse, and other variants. In: **Learning in Graphical Models**. [S.l.]: Kluwer Academic Publishers, 1998. p. 355–368. Citado 2 vezes nas páginas 41 e 42.

NEEDHAM, T. A visual explanation of jensen's inequality. **The American Mathematical Monthly**, Mathematical Association of America, v. 100, n. 8, p. 768–771, 1993. ISSN 00029890, 19300972. Disponível em: <http://www.jstor.org/stable/2324783>. Citado na página 26.

NEWMAN, D.; LAU, J. H.; GRIESER, K.; BALDWIN, T. Automatic evaluation of topic coherence. In: **Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics**. Stroudsburg, PA, USA: Association for Computational Linguistics, 2010. (HLT '10), p. 100–108. ISBN 1-932432-65-5. Disponível em: <http://dl.acm.org/citation.cfm?id=1857999.1858011>. Citado 2 vezes nas páginas 44 e 45.

PAATERO, P.; TAPPER, U. Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. **Environmetrics**, John Wiley & Sons, Ltd., University of Helsinki, Department of Physics, Siltavuorenpenger 20 D, SF-00170 Helsinki, Finland, v. 5, n. 2, p. 111–126, jun. 1994. Disponível em: <http://dx.doi.org/10.1002/env.3170050203>. Citado na página 47.

RUSSELL, B. C.; FREEMAN, W. T.; EFROS, A. A.; SIVIC, J.; ZISSERMAN, A. Using multiple segmentations to discover objects and their extent in image collections. In: **Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2**. Washington, DC, USA: IEEE Computer Society, 2006. (CVPR '06), p. 1605–1614. ISBN 0-7695-2597-0. Disponível em: <http://dx.doi.org/10.1109/CVPR.2006.326>. Citado na página 9.

RUSSELL, S. J.; NORVIG, P. **Artificial Intelligence: A Modern Approach**. Pearson Education, 2003. ISBN 0137903952. Disponível em: <http://portal.acm.org/citation.cfm?id=773294>. Citado na página 20.

SIVIC, J.; RUSSELL, B. C.; EFROS, A. A.; ZISSERMAN, A.; FREEMAN, W. T. Discovering objects and their location in images. In: **IEEE International Conference on Computer Vision**. [S.l.: s.n.], 2005. Citado na página 9.

STEYVERS, M.; GRIFFITHS, T. Probabilistic Topic Models. In: ____. **Handbook of Latent Semantic Analysis**. Lawrence Erlbaum Associates, 2007. ISBN 1410615340. Disponível em: <http://www.amazon.com/exec/obidos/redirect?tag=citeulike07-20&path=ASIN/1410615340>. Citado 3 vezes nas páginas 9, 10 e 35.

WAAL, A. de; BARNARD, E. Evaluating Topic Models with Stability. 2008. Citado na página 43.

WALLACH, H. M.; MIMNO, D. M.; MCCALLUM, A. Rethinking lda: Why priors matter. In: BENGIO, Y.; SCHUURMANS, D.; LAFFERTY, J. D.; WILLIAMS, C. K. I.; CULOTTA, A. (Ed.). **Advances in Neural Information Processing Systems 22**. Curran Associates, Inc., 2009. p. 1973–1981. Disponível em: <http://papers.nips.cc/paper/3854-rethinking-lda-why-priors-matter.pdf>. Citado na página 53.

WALLACH, H. M.; MURRAY, I.; SALAKHUTDINOV, R.; MIMNO, D. Evaluation methods for topic models. In: **Proceedings of the 26th Annual International Conference on Machine Learning**. New York, NY, USA: ACM, 2009. (ICML '09), p. 1105–1112. ISBN 978-1-60558-516-1. Disponível em: <http://doi.acm.org/10.1145/1553374.1553515>. Citado na página 43.

WEI, X.; CROFT, W. B. Lda-based document models for ad-hoc retrieval. In: **Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval**. New York, NY, USA: ACM, 2006. (SIGIR '06), p. 178–185. ISBN 1-59593-369-7. Disponível em: <http://doi.acm.org/10.1145/1148170.1148204>. Citado na página 35.