Journal of Machine Learning Research 3 (2003) 993-1022Submetido em 2/02; Publicado em 1/03

Alocação de Dirichlet Latente

David M. Blei BLEI@CS. BERKELEY. EDU
Computer Science Division
University of California
Berkeley, CA 94720, EUA

Andrew Y. Ng ANG@CS. STANFORD. EDU
Departamento de Informática
da Universidade de Stanford
Stanford, CA 94305, EUA

Michael I. Jordan JORDAN@CS. BERKELEY. EDU
Divisão de Ciência da Computação e Departamento de
Estatística da Universidade da Califórnia
Berkeley, CA 94720, EUA

Editor: John Lafferty

Abstrato

Descrevemos a *alocação Dirichlet latente* (LDA), um modelo probabilístico generativo para coleções de dados discretos, como corpora de texto. LDA é um modelo Bayesiano de três níveis hierárquicos, no qual cada item de uma coleção é modelado como uma mistura finita sobre um conjunto de tópicos subjacentes. Cada tópico é, por sua vez, modelado como uma mistura infinita sobre um conjunto subjacente de probabilidades de tópicos. No contexto da modelagem de texto, as probabilidades dos tópicos fornecem uma representação explícita de um documento. Apresentamos técnicas eficientes de inferência aproximada baseadas em métodos variacionais e um algoritmo EM para estimação empírica de parâmetros Bayes. Relatamos resultados na modelagem de documentos, classificação de texto e filtragem colaborativa, comparando com uma mistura de modelo de unigramas e o modelo probabilístico de LSI.

1. Introdução

Neste artigo consideramos o problema de modelagem de corpora de texto e outras coleções de dados discretos. O objetivo é encontrar breves descrições dos membros de uma coleção que permitam o processamento eficiente de grandes coleções, preservando as relações estatísticas essenciais que são úteis para tarefas básicas como classificação, detecção de novidade, sumarização e julgamentos de similaridade e relevância.

Foram feitos progressos significativos neste problema por investigadores na área da recuperação de informação (RI) (Baeza-Yates e Ribeiro-Neto, 1999). A metodologia básica proposta pelos pesquisadores de RI para corporações de texto - uma metodologia implantada com sucesso nos modernos mecanismos de busca da Internet - reduz cada documento no corpus a um vetor de números reais, cada um dos quais repreende rácios de contagens. No popular esquema *tf-idf* (Salton and McGill, 1983), é escolhido um vocabulário básico de "palavras" ou "termos", e, para cada documento do corpus, é formada uma contagem do número de ocorrências de cada palavra. Após uma normalização adequada, este termo contagem de frequência é comparado a uma contagem de frequência inversa do documento, que mede o número de ocorrências de um

palavra em todo o corpus (geralmente em uma escala de log, e novamente devidamente normalizada). O resultado final é uma matriz X termo por documento cujas colunas contêm os valores *tf-idf* para cada um dos documentos do corpus. Assim, o esquema *tf-idf* reduz documentos de comprimento arbitrário a listas de números de comprimento fixo.

Embora a redução do *tf-idf* tenha algumas características apelativas - como é o caso da identificação básica de conjuntos de palavras que são discriminatórias para os documentos na coleção - a abordagem também proporciona uma redução relativamente pequena no comprimento da descrição e revela pouco na forma de inter- ou intra-estrutura estatística do documento. Para resolver estas deficiências, os pesquisadores de RI propuseram várias outras técnicas de redução de dimensionalidade, principalmente a *indexação semântica latente (LSI)* (Deerwester et al., 1990). O LSI usa uma decomposição de valor singular da matriz X para identificar um subespaço linear no espaço de características *tf-idf* que captura a maior parte da variância na coleção. Esta abordagem pode alcançar uma compressão significativa em grandes coleções. Além disso, Deerwester et al. argumentam que as características derivadas do LSI, que são combinações lineares das características *tf-idf* originais, podem capturar alguns aspectos das noções linguísticas básicas, como sinonímia e polissemia.

Para substanciar as alegações relativas ao LSI, e para estudar seus pontos fortes e fracos relativos, é útil desenvolver um modelo probabilístico generativo de corpora de texto e estudar a capacidade do LSI de recuperar aspectos do modelo generativo a partir de dados (Papadimitriou et al., 1998). Dado um modelo generativo de texto, no entanto, não está claro porque se deve adotar o método LSI - pode-se tentar proceder mais diretamente, adequando o modelo aos dados usando a máxima probabilidade ou métodos Bayesianos. Um passo significativo neste sentido foi dado por Hofmann (1999), que apresentou o modelo *probabilístico LSI (pLSI)*, também conhecido como *modelo de aspecto*, como uma alternativa ao LSI. A abordagem pLSI, que descrevemos em detalhes na Seção 4.3, modela cada palavra em um documento como uma amostra de um modelo de mistura, onde os componentes da mistura são variáveis aleatórias multinomiais que podem ser vistas como representações de "tópicos". Assim, cada palavra é gerada a partir de um único tópico, e diferentes palavras em um documento podem ser geradas a partir de diferentes tópicos. Cada documento é representado como uma lista de proporções de mistura para esses componentes de mistura e, assim, reduzido a uma distribuição de probabilidade sobre um conjunto fixo de tópicos. Esta distribuição é a "descrição reduzida" associada a o documento.

Embora o trabalho de Hofmann seja um passo útil para a modelagem probabilística do texto, ele é incompleto na medida em que não fornece nenhum modelo probabilístico no nível dos documentos. Na pLSI, cada documento é representado como uma lista de números (as proporções de mistura para os tópicos), e não existe um modelo probabilístico generativo para esses números. Isto leva a vários problemas: (1) o número de parâmetros no modelo cresce linearmente com o tamanho do corpus, o que leva a sérios problemas de sobreajuste, e (2) não é claro como atribuir probabilidade a um documento fora do conjunto de treinamento. Para ver como proceder além do pLSI, vamos considerar os pressupostos probabilísticos fundamentais subjacentes à classe de métodos de redução de dimensionalidade que inclui o LSI e o pLSI. Todos estes métodos são baseados no pressuposto de "saco de palavras" - que a ordem das palavras em um documento pode ser negligenciada. Na linguagem da teoria da probabilidade, esta é uma suposição de *permuta* das palavras de um documento (Aldous, 1985). Além disso, embora menos frequentemente afirmados formalmente, estes métodos também assumem que os documentos são permutáveis; a ordenação específica dos documentos em um corpus

também pode ser negligenciado.

Um teorema clássico de representação devido a de Finetti (1990) estabelece que qualquer coleção de variáveis aleatórias ex-mudáveis tem uma representação como uma distribuição de mistura - em geral uma mistura infinita. Assim, se quisermos considerar representações intercambiáveis para documentos e palavras, precisamos considerar modelos de mistura que captem a permutabilidade tanto de palavras como de documentos.

Esta linha de pensamento leva ao modelo de *alocação latente Dirichlet (LDA)* que apresentamos no presente artigo.

É importante salientar que uma suposição de permutabilidade não equivale a uma soma de que as variáveis aleatórias são independentes e distribuídas de forma idêntica. Ao contrário, a capacidade de troca pode ser interpretada essencialmente como significando "*condicionalmente* independente e identicamente des-tribuída", onde o condicionamento é com respeito a um parâmetro latente subjacente de uma distribuição de probabilidade. Condicionalmente, a distribuição conjunta das variáveis aleatórias é simples e fatorizada, enquanto marginalmente sobre o parâmetro latente, a distribuição conjunta pode ser bastante complexa. Assim, enquanto uma suposição de permutabilidade é claramente uma suposição simplificadora importante no domínio da modelagem de texto, e sua principal justificativa é que ela leva a métodos que são computacionalmente eficientes, as suposições de permutabilidade não necessariamente levam a métodos que são restritos a simples contagens de freqüência ou operações lineares. Pretendemos demonstrar no presente trabalho que, levando a sério o teorema de Finetti, podemos captar uma estrutura estatística intra-documental significativa através da distribuição da mistura.

Também vale a pena notar que há um grande número de generalizações da noção básica de permutabilidade, incluindo várias formas de permutabilidade parcial, e que os rems de representação também estão disponíveis para estes casos (Diaconis, 1988). Assim, enquanto o trabalho que discutimos no presente trabalho se concentra em modelos simples de "saco de palavras", que levam a distribuições de misturas para palavras únicas (unigramas), nossos métodos também são aplicáveis a modelos mais ricos que envolvem misturas para unidades estruturais maiores, como *n-gramas* ou parágrafos.

O jornal está organizado da seguinte forma. Na Secção 2 introduzimos a notação e terminologia básica. O modelo LDA é apresentado na Secção 3 e é comparado com modelos de variáveis latentes relacionadas na Secção 4. Discutimos inferência e estimativa de parâmetros para LDA na Seção 5. Um exemplo ilustrativo de adaptação da LDA aos dados é apresentado na Secção 6. Resultados empíricos em modelagem de texto, classificação de texto e filtragem colaborativa são apresentados na Seção 7. Finalmente, a Secção 8 apresenta as nossas conclusões.

2. Notação e terminologia

Utilizamos a linguagem das coleções de texto ao longo do papel, referindo entidades como "palavras", "documentos" e "corpora". Isto é útil na medida em que ajuda a orientar a intuição, particularmente quando introduzimos variáveis latentes que visam captar noções abstractas, tais como tópicos. É importante notar, no entanto, que o modelo LDA não está necessariamente ligado ao texto, e tem aplicações para outros problemas envolvendo coleções de dados, incluindo dados de domínios como a filtragem colaborativa, recuperação de imagens baseadas em conteúdo e bioinformática. De facto, na Secção 7.3, apresentamos resultados experimentais no domínio da filtragem colaborativa.

Formalmente, nós definimos os seguintes termos:

- Uma *palavra* é a unidade básica de dados discretos, definida como sendo um item de um vocabulário indexado por $\{1, \ldots, V\}$. Representamos palavras usando vectores de base unitária que têm um único componente igual a um e todos os outros componentes igual a zero. Assim, usando superescritos para denotar componentes, a palavra *vth* no vocabulário é representada por um V-vector w tal que $w^v = 1$ e $w^u = 0$ para

$u \models v$.

- Um *documento* é uma sequência de N palavras denotadas por $\mathbf{w} = (w_1, w_2, \ldots, w_N)$, onde w_n é a *enésima* palavra na sequência.
- Um *corpus* é uma coleção de documentos M denotados por $D = \{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_M\}$.

Queremos encontrar um modelo probabilístico de um corpus que não só atribua alta probabilidade aos membros do corpus, mas também atribua alta probabilidade a outros documentos "similares".

3. Alocação de Dirichlet Latente

A alocação Dirichlet Latente (LDA) é um modelo probabilístico generativo de um corpus. A idéia básica é que os documentos são representados como misturas aleatórias sobre tópicos latentes, onde cada tópico é terciarizado por uma distribuição por palavras.[1]

A LDA assume o seguinte processo generativo para cada documento **w** em um corpus *D*:

1. Escolha $N \sim$ Poisson().
2. Escolha $\sim$ Dir().
3. Para cada uma das N palavras w_n :
   (a) Escolha um tópico $z_n \sim$ Multinomial().
   (b) Escolha uma palavra w_n de $p(w_n \mid z_n ,)$, uma probabilidade multinomial condicionada sobre o tema z_n .

Neste modelo básico são feitas várias suposições simplificadoras, algumas das quais removemos em seções subseqüentes de quent. Primeiro, a dimensionalidade k da distribuição Dirichlet (e, portanto, a dimensionalidade da variável temática z) é assumida conhecida e fixa. Em segundo lugar, as probabilidades da palavra são parametrizadas por uma matriz $k \times V$ onde$_{ij} = p(w^j = 1 \mid z^i = 1)$, que por enquanto tratamos como uma quantidade fixa que deve ser estimada. Finalmente, a suposição de Poisson não é crítica para qualquer coisa que se siga e distribuições mais realistas de comprimento de documento podem ser usadas conforme necessário. Além disso, note que N é independente de todas as outras variáveis geradoras de dados (e **z**). Portanto, é uma variável auxiliar e geralmente ignoraremos sua aleatoriedade no desenvolvimento subseqüente.

Uma variável aleatória *k-dimensional* Dirichlet pode assumir valores no $(k - 1)$-simplex (um *k-vector* encontra-se no $(k - 1)$-simplex se$_i \geq 0$, $\sum_{i=1}^{k} {}_i = 1$), e tem a seguinte densidade de probabilidade sobre este simplóri o:

$$p(\mid) = \frac{\Gamma\left(\sum_{i=1}^{k} {}_i\right)}{\prod_{i=1}^{k} \Gamma({}_i)} {}_1^{-1} \cdots {}_k^{-1}, \quad (1)$$

onde o parâmetro é um *k-vector* com componentes$_i > 0$, e onde $\Gamma(x)$ é a função Gama. O Dirichlet é uma distribuição conveniente no simplex - está na família exponencial, tem estatísticas finitas dimensionais suficientes, e é conjugado com a distribuição multinomial. Na Secção 5, estas propriedades facilitarão o desenvolvimento de algoritmos de inferência e estimação de parâmetros para LDA.

Dados os parâmetros e , a distribuição conjunta de uma mistura de tópicos , um conjunto de N tópicos **z**, e um conjunto de N palavras **w** é dado por:

$$p(, \mathbf{z}, \mathbf{w} \mid ,) = p(\mid) \prod_{n}^{N} p(z_n \mid) p(w_n \mid z_n ,), \quad (2)$$

1. Nós nos referimos às variáveis multinomiais latentes no modelo LDA como tópicos, de modo a explorar intuições orientadas ao texto, mas não fazemos reivindicações epistemológicas sobre essas variáveis latentes além de sua utilidade em representar distribuições de probabilidade em conjuntos de palavras.

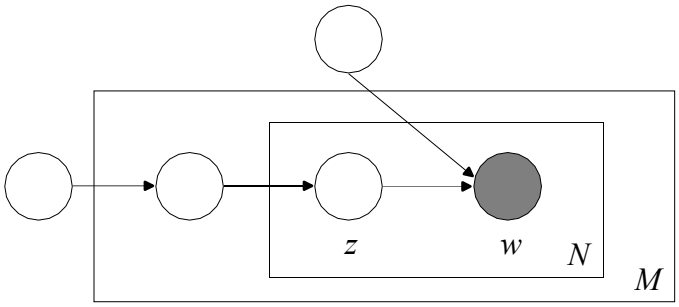

Figura 1: Representação gráfica do modelo de LDA. As caixas são "placas" representando réplicas. A placa externa representa documentos, enquanto a placa interna representa a escolha repetida de tópicos e palavras dentro de um documento.

onde $p(z_n \mid)$ é simplesmente$_i$ para o i único tal que $z^i = 1$. Integrando mais e somando mais z, obtemos a distribuição marginal de um documento:

$$p(\mathbf{w} \mid ,) = \int p(\mid) \left(\prod_{n=1}^{N} \sum_{z_n} p(z_n \mid) p(w_n \mid z_n ,) \right) d. \quad (3)$$

Finalmente, tomando o produto das probabilidades marginais de documentos únicos, obtemos a proba- bilidade de um corpus:

$$p(D \mid ,) = \prod_{d=1}^{M} \int p(_d \mid) \left(\prod_{n=1}^{N_d} \sum_{z_{dn}} p(z_{dn} \mid _d) p(w_{dn} \mid z_{dn} ,) \right) \delta_d .$$

O modelo LDA é representado como um modelo gráfico probabilístico na Figura 1. Como a figura deixa claro, existem três níveis para a representação do LDA. Os parâmetros e são parâmetros de nível de corpus-, supostamente amostrados uma vez no processo de geração de um corpus. As variáveis
$_d$ são variáveis em nível de documento, amostradas uma vez por documento. Finalmente, as variáveis z_{dn} e w_{dn} são variáveis de nível de palavra e são amostradas uma vez para cada palavra em cada documento.

É importante distinguir a LDA de um simples modelo de agrupamento Dirichlet-multinomial. Um modelo clássico de agrupamento envolveria um modelo de dois níveis no qual um Dirichlet é amostrado uma vez para um corpus, uma variável multinomial de agrupamento é selecionada uma vez para cada documento no corpus, e um conjunto de palavras é selecionado para o documento condicional à variável de agrupamento. Como em muitos modelos de clustering, tal modelo restringe um documento a ser associado a um único tópico. O LDA, por outro lado, envolve três níveis, e notavelmente o nó de tópico é amostrado *repetidamente* dentro do documento. Sob este modelo, os documentos podem ser associados a múltiplos tópicos.

Estruturas similares às mostradas na Figura 1 são frequentemente estudadas na modelagem estatística Bayesiana, onde são referidas como *modelos hierárquicos* (Gelman et al., 1995), ou mais precisamente como *modelos hierárquicos con-dialmente independentes* (Kass e Steffey, 1989). Tais modelos também são frequentemente referidos como *modelos empíricos paramétricos Bayes*, um termo que se refere não apenas a uma estrutura particular do modelo, mas também aos métodos usados para estimar parâmetros no modelo (Morris, 1983). De fato,

como discutimos na seção 5, adotamos a abordagem empírica Bayes para estimar parâmetros como e em implementações simples de LDA, mas também consideramos abordagens Bayesianas mais completas.

3.1 LDA e permutabilidade

Um conjunto finito de variáveis aleatórias $\{z_1, \ldots, z_N\}$ é dito ser *permutável* se a distribuição conjunta for invariável à permutação. Se for uma permutação dos números inteiros de 1 a N:

$$p(z_1, \ldots, z_N) = p(z_{\pi(1)}, \ldots, z_{\pi(N)}).$$

Uma sequência infinita de variáveis aleatórias é *infinitamente permutável* se cada subseqüência finita for permutável.

O teorema da representação de De Finetti afirma que a distribuição conjunta de uma sequência infinitamente permutável de variáveis aleatórias é como se um parâmetro aleatório fosse retirado de alguma distribuição e depois as variáveis aleatórias em questão fossem *independentes* e *distribuídas de forma idêntica*, condicionadas a esse parâmetro.

Na LDA, assumimos que as palavras são geradas por tópicos (por distribuições condicionais fixas) e que esses tópicos são infinitamente permutáveis dentro de um documento. Pelo teorema de Finetti, a capacidade de uma sequência de palavras e tópicos deve, portanto, ter a forma:

$$p(\mathbf{w}, \mathbf{z}) = \int p(\theta) \left(\prod_{n=1}^{N} p(z_n \mid \theta) p(w_n \mid z_n) \right) d\theta,$$

onde θ está o parâmetro aleatório de uma multinomial sobre tópicos. Obtemos a distribuição LDA em documentos em Eq. (3) marginalizando as variáveis temáticas e dotando com uma distribuição Dirichlet.

3.2 Uma mistura contínua de unigramas

O modelo LDA mostrado na Figura 1 é um pouco mais elaborado do que os modelos de dois níveis frequentemente estudados na literatura hierárquica clássica Bayesiana. Ao marginalizar sobre a variável temática oculta z, no entanto, podemos entender o LDA como um modelo de dois níveis.

Em particular, vamos formar a palavra distribuição $p(w \mid \beta, \alpha)$:

$$p(w \mid \beta, \alpha) = \sum_z p(w \mid z, \beta) p(z \mid \theta).$$

Note que esta é uma quantidade aleatória, uma vez que depende de θ.

Definimos agora o seguinte processo generativo para um documento $\mathbf{w}$:

1. Escolha $\theta \sim \mathrm{Dir}(\alpha)$.
2. Para cada uma das N palavras w_n:
   (a) Escolha uma palavra w_n de $p(w_n \mid \theta, \beta)$.

Este processo define a distribuição marginal de um documento como uma distribuição contínua da mistura:

$$p(\mathbf{w} \mid \alpha, \beta) = \int p(\theta \mid \alpha) \left(\prod_{n=1}^{N} p(w_n \mid \theta, \beta) \right) d\theta,$$

onde $p(w_n \mid \theta, \beta)$ são os componentes da mistura e $p(\theta \mid \alpha)$ são os pesos da mistura.

A Figura 2 ilustra esta interpretação da LDA. Ela representa a distribuição em $p(w \mid \beta, \alpha)$ que é induzida a partir de uma instância particular de um modelo de LDA. Note que esta distribuição no

$(V - 1)$- simplex é atingida apenas com parâmetros $k + kV$, mas exibe uma estrutura multimodal muito interessante.

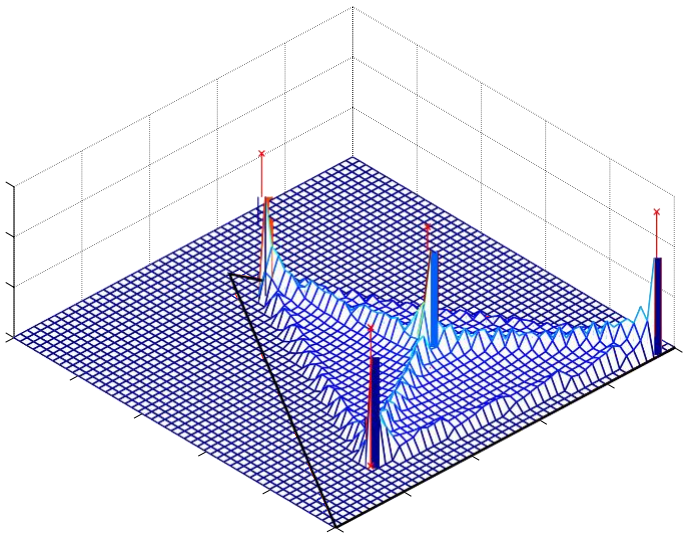

Figura 2: Um exemplo de densidade nas distribuições de unigramas $p(w \mid \ )$ sob LDA para três palavras e quatro tópicos. O triângulo embutido no plano x-y é o simplex 2-D representando todas as distribuições multinomiais possíveis sobre três palavras. Cada um dos vértices do trian- gle corresponde a uma distribuição determinística que atribui probabilidade de uma a uma das palavras; o ponto médio de uma borda dá probabilidade de 0,5 a duas das palavras; e o centróide do triângulo é a distribuição uniforme sobre todas as três palavras. Os quatro pontos marcados com um x são as localizações das distribuições multinomiais $p(w \mid z)$ para cada um dos quatro tópicos, e a superfície mostrada no topo do simplex é um exemplo de uma densidade sobre o $(V - 1)$-simplesx (distribuições multinomiais de palavras) dada pela LDA.

4. Relação com outros modelos de variáveis latentes

Nesta seção comparamos LDA com modelos de variáveis latentes mais simples para texto - o modelo unigramático, uma mistura de unigramas, e o modelo pLSI. Além disso, apresentamos uma interpretação geométrica unificada desses modelos que destaca suas principais diferenças e semelhanças.

4.1 modelo Unigram

Sob o modelo unigramático, as palavras de cada documento são desenhadas independentemente de uma única distribuição multinomial:

$$p(\mathbf{w}) = \prod_{n=1}^{N} p(w_n).$$

Isto está ilustrado no modelo gráfico da Figura 3a.

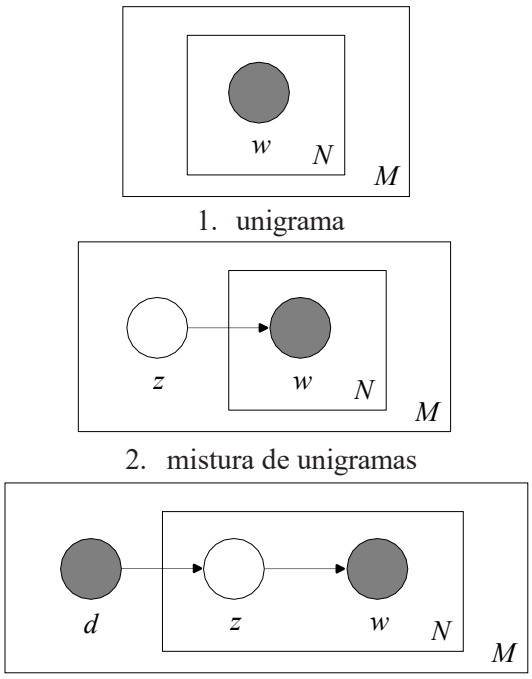

Figura 3: Representação gráfica do modelo de diferentes modelos de dados discretos.

4.2 Mistura de unigramas

Se aumentarmos o modelo de unigramas com uma discreta variável temática aleatória z (Figura 3b), obtemos uma *mistura de* modelo de *unigramas* (Nigam et al., 2000). Sob este modelo de mistura, cada documento é envelhecido escolhendo primeiro um tópico z e depois gerando N palavras independentemente da multinomial condicional $p(w \mid z)$. A probabilidade de um documento é:

$$p(\mathbf{w}) = \sum_{z} p(z) \prod_{n=1}^{N} p(w_n \mid z).$$

Quando estimado a partir de um corpus, as distribuições de palavras podem ser vistas como representações de tópicos sob a suposição de que cada documento exibe exatamente um tópico. Como os resultados empíricos na Secção 7 ilustram, esta suposição é muitas vezes demasiado limitada para modelar eficazmente uma grande colecção de documentos.

Em contraste, o modelo LDA permite que os documentos exibam múltiplos tópicos em diferentes graus. Isto é conseguido ao custo de apenas um parâmetro adicional: existem $k - 1$ parâmetros associados a $p(z)$ na mistura de unigramas, versus os k parâmetros associados a $p(|)$ no LDA.

4.3 Indexação semântica latente probabilística

A indexação semântica latente probabilística (pLSI) é outro modelo de documento amplamente

utilizado (Hofmann, 1999). O modelo pLSI, ilustrado na Figura 3c, postula que uma etiqueta de documento d e uma palavra w_n são

condicionalmente independente dado um tópico z não observado:

$$p(d, w_n) = p(d) \sum_z p(w_n \mid z) p(z \mid d).$$

O modelo pLSI tenta relaxar a suposição simplificadora feita na mistura de modelos unigramas de que cada documento é gerado a partir de apenas um tópico. De certa forma, ele captura a possibilidade de um documento conter múltiplos tópicos, uma vez que $p(z \mid d)$ serve como o peso da mistura dos tópicos para um determinado documento d. No entanto, é importante notar que d é um índice fictício na lista de documentos do *conjunto de treinamento*. Assim, d é uma variável aleatória multinomial com tantos valores possíveis como os documentos de treinamento e o modelo aprende as misturas de tópicos $p(z \mid d)$ apenas para os documentos nos quais é treinado. Por esta razão, pLSI não é um modelo generativo bem definido de documentos; não há uma forma natural de usá-lo para atribuir probabilidade a um documento previamente não visto.

Uma outra dificuldade com o pLSI, que também deriva do uso de uma distribuição indexada por documentos de treinamento, é que o número de parâmetros que devem ser estimados cresce linearmente com o número de documentos de treinamento. Os parâmetros para um modelo pLSI *k-tópico* são k butions distri- buições multinomiais de tamanho V e M misturas sobre os k tópicos ocultos. Isto dá parâmetros $kV + kM$ e, portanto, crescimento linear em M. O crescimento linear em parâmetros sugere que o modelo é propenso a sobreajustamento e, empiricamente, o sobreajustamento é realmente um problema sério (ver Secção 7.1). Na prática, um heurístico temperado é usado para suavizar os parâmetros do modelo para um desempenho de previsão aceitável. Foi demonstrado, no entanto, que o sobreajuste pode ocorrer mesmo quando se usa a têmpera (Popescul et al., 2001).

O LDA supera esses dois problemas ao tratar os pesos da mistura de tópicos como uma *variável aleatória* oculta do *parâmetro k* em vez de um grande conjunto de parâmetros individuais que estão explicitamente ligados ao conjunto de treinamento. Como descrito na Seção 3, LDA é um modelo generativo bem definido e se generaliza facilmente a novos documentos. Além disso, os parâmetros $k + kV$ em um modelo LDA *k-topic* não crescem com o tamanho do corpus de treinamento. Veremos na Secção 7.1 que o LDA não sofre dos mesmos problemas de sobreajustamento que o pLSI.

4.4 Uma interpretação geométrica

Uma boa maneira de ilustrar as diferenças entre os LDA e os outros modelos temáticos latentes é considerar a geometria do espaço latente, e ver como um documento é representado nessa geometria sob cada modelo.

Todos os quatro modelos descritos acima - unigrama, mistura de unigramas, pLSI e LDA - operam no espaço de distribuições sobre palavras. Cada uma dessas distribuições pode ser vista como um ponto no $(V - 1)$-simplex, que chamamos de simplex.

O modelo de unigrama encontra um único ponto na palavra simplex e postula que todas as palavras no

corpus vem da distribuição correspondente. Os modelos de variáveis latentes consideram k pontos na palavra simplex e formam uma sub-simplex com base nesses pontos, que chamamos de simplex ao tópico. Note que qualquer ponto sobre o tópico simplex é também um ponto sobre a palavra simplex. Os diferentes modelos de variáveis latentes utilizam o tópico simplex de diferentes maneiras para gerar um documento.

- A mistura de unigramas modelo postula que para cada documento, um dos k pontos na palavra simplex (ou seja, um dos cantos do tópico simplex) é escolhido aleatoriamente e todas as palavras do documento são extraídas da distribuição correspondente a esse ponto.

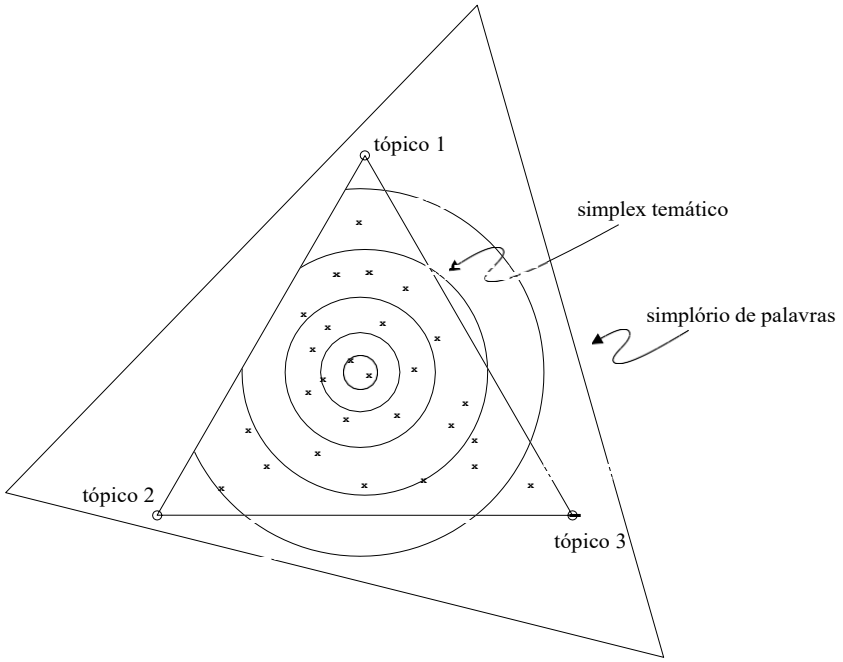

Figura 4: O tópico simplex para três tópicos embutidos na palavra simplex para três palavras. Os cantos da palavra simplex correspondem às três distribuições onde cada palavra (respectivamente) tem uma probabilidade. Os três pontos do tópico simplex correspondem a três distribuições diferentes sobre as palavras. A mistura de unigramas coloca cada documento em um dos cantos do tópico simplex. O modelo pLSI induz uma distribuição empírica sobre o tópico simplex denotado por x. LDA coloca uma distribuição suave sobre o tópico simplex denotado pelas linhas de contorno.

- O modelo pLSI postula que cada palavra de um documento de *treinamento* vem de um tópico escolhido aleatoriamente. Os tópicos são eles próprios extraídos de uma distribuição específica do documento por tópicos, ou seja, um ponto sobre o tópico simplex. Existe uma dessas distribuições para cada documento; o conjunto de documentos de treinamento define assim uma distribuição empírica sobre o tópico simplex.

- LDA postula que cada palavra dos documentos observados e não vistos é gerada por um tópico escolhido aleatoriamente que é extraído de uma distribuição com um parâmetro escolhido aleatoriamente. Este parâmetro é amostrado uma vez por documento a partir de uma distribuição simples sobre o tópico simplex.

Estas diferenças estão destacadas na Figura 4.

5. Inferência e Estimativa de Parâmetros

Descrevemos a motivação por detrás da LDA e ilustramos as suas vantagens conceptuais sobre outros modelos temáticos latentes. Nesta seção, voltamos nossa atenção para os procedimentos de

inferência e estimativa de parâmetros sob LDA.

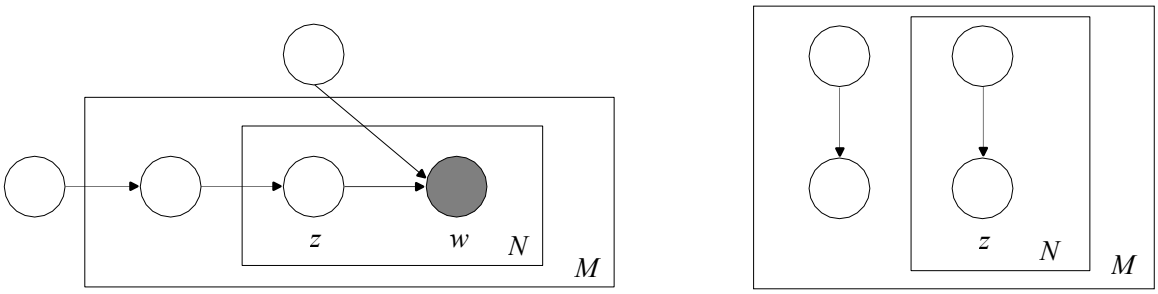

Figura 5: (Esquerda) Representação gráfica do modelo de LDA. (Direita) Representação gráfica do modelo de distribuição variacional utilizado para aproximar o posterior em LDA.

5.1 Inferência

O principal problema inferencial que precisamos resolver para usar LDA é o de calcular a distribuição posterior das variáveis ocultas dado um documento:

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

Infelizmente, esta distribuição é intratável de calcular em geral. Na verdade, para normalizar a distri- buição, marginalizamos sobre as variáveis ocultas e escrevemos Eq. (3) em termos dos parâmetros do modelo:

$$p(\mathbf{w}|\alpha, \beta) = \frac{\Gamma(\sum_i \alpha_i)}{\prod_i \Gamma(\alpha_i)} \int \left(\prod_{i=1}^{k} \theta_i^{\alpha_i - 1} \right) \left(\prod_{n=1}^{N} \sum_{i=1}^{k} \prod_{j=1}^{V} (\theta_i \beta_{ij})^{w_n^j} \right) d\theta,$$

uma função que é intratável devido ao acoplamento entre e na soma sobre tópicos latentes (Dickey, 1983). Dickey mostra que esta função é uma expectativa sob uma extensão particular da distribuição Dirichlet que pode ser representada com funções hipergeométricas especiais. Ela tem sido usada num contexto Bayesiano para dados discretos censurados para representar o posterior no qual, nessa configuração, é um parâmetro aleatório (Dickey et al., 1987).

Embora a distribuição posterior seja intratável para uma inferência exata, uma grande variedade de algoritmos de inferência de aproximadamente mate pode ser considerada para LDA, incluindo a aproximação de Laplace, a aproximação variacional e a cadeia de Markov Monte Carlo (Jordan, 1999). Nesta secção descrevemos um algoritmo simples baseado na convexidade para inferência em LDA, e discutimos algumas das alternativas na Secção 8.

5.2 Inferência variável

A idéia básica da inferência variacional baseada na convexidade é fazer uso da desigualdade de Jensen para ob- tinam um limite inferior ajustável na probabilidade logarítmica (Jordan et al., 1999). Essencialmente, considera-se uma família de limites inferiores, indexada por um conjunto de *parâmetros variacionais*. Os parâmetros variacionais são escolhidos por um procedimento de otimização que tenta encontrar o limite inferior mais apertado possível.

Uma maneira simples de obter uma família rastreável de limites inferiores é considerar modificações simples do modelo gráfico original em que algumas das bordas e nós são

removidos. Considere em particular o modelo LDA mostrado na Figura 5 (esquerda). O acoplamento problemático entre ε

Ao deixar cair estas bordas e os nós **w,** e dotar o modelo gráfico simplificado resultante com parâmetros variacionais livres, obtemos uma família de distribuições sobre as variáveis latentes. Esta família é caracterizada pela seguinte distribuição variacional:

$$q(, \mathbf{z}|,) = q(|) \prod_{n=1}^{N} q(z_n|_n), \tag{4}$$

onde o parâmetro Dirichlet e os parâmetros multinomiais $(_1 , \ldots , _N)$ são os parâmetros variacionais livres.

Tendo especificado uma família simplificada de distribuições de probabilidade, o próximo passo é definir um problema de otimização que determine os valores dos parâmetros variacionais e . Como mostramos no Apêndice A, o desiderato de encontrar um limite inferior apertado no log de probabilidade se traduz diretamente no seguinte problema de otimização:

$$(^* , ^*) = \operatorname{argmin}_{(,)} \mathrm{D}(q(, \mathbf{z}|,) \, p(, \mathbf{z}|\mathbf{w}, ,)). \tag{5}$$

Assim, os valores de otimização dos parâmetros variacionais são encontrados minimizando a divergência Kullback- Leibler (KL) entre a distribuição variacional e a verdadeira posterior $p(, \mathbf{z} \mid \mathbf{w}, ,)$. Esta minimização pode ser obtida através de um método iterativo de ponto fixo. Em particular, mostramos no Apêndice A.3 que calculando as derivadas da divergência de KL e fixando-as em zero, obtemos o seguinte par de equações de atualização:

$$_{ni} \propto _{iw_n} \exp\{\mathrm{E}_q [\log(_i) |]\} \tag{6}$$

$$_i = _i + \sum_{n=1}^{N} _{ni} . \tag{7}$$

Como mostramos no Apêndice A.1, a expectativa na atualização multinomial pode ser computada da seguinte forma:

$$\mathrm{E}_q [\log(_i) |] = (_i) - ^k \sum_{j=1} _j, \tag{8}$$

onde é a primeira derivada da função $\lambda o \gamma$ que é computável via Taylor approxima- tions (Abramowitz e Stegun, 1970).

Os Eqs. (6) e (7) têm uma interpretação intuitiva apelativa. A atualização do Dirichlet é um Dirichlet poste- rior dado observações esperadas tomadas sob a distribuição variacional, $\mathrm{E}[z_n|_n]$. A atualização multinomial é semelhante a usar o teorema de Bayes, $p(z_n \mid w_n) \, p(w_n \mid z_n) p(z_n)$, onde $p(z_n)$ é aproximado pelo exponencial do valor esperado do seu logaritmo sob a distribuição variacional, $\mathrm{E}[z \mid w]$, onde p(z) é aproximado pelo exponencial do valor esperado do seu logaritmo sob a distribuição variacional, *E[z | w].*
bução.

É importante notar que a distribuição variacional é realmente uma distribuição condicional, variando em função de **w**. Isto ocorre porque o problema de otimização em Eq. (5) é conduzido para **w** fixo, e assim produz parâmetros de otimização $(^*, ^*)$ que são uma função de **w**. Podemos escrever
a distribuição variacional resultante como $q(, \mathbf{z} \mid ^*(\mathbf{w}), ^*(\mathbf{w}))$, onde explicitamos a dependência de **w**. Assim a distribuição variacional pode ser vista como uma aproximação à distribuição posterior $p(, \mathbf{z}|\mathbf{w}, ,)$.

No idioma do texto, os parâmetros de otimização $(^*(\mathbf{w}), ^*(\mathbf{w}))$ são específicos do documento. Em Em particular, vemos os parâmetros do Dirichlet $^*(\mathbf{w})$ como uma representação de um

documento no tópico simplex.

(1) inicializar $\phi^0_{ni} := 1/k$ para todos i e n
(2) inicializar $\lambda_i := \alpha_i + N/k$ para todos i
(3) **repita**
(4) **para** $n = 1$ **a** N
(5) **para** $i = 1$ **a** k
(6) $\phi^{t+1}_{ni} := w_{iw} \exp(\Psi(\gamma^t_i))$
(7) normalizar ϕ^{t+1}_n para somar a 1.
(8) $\gamma^{t+1} := \alpha + \sum_{n=1}^{N} \phi^{t+1}_n$
(9) **até a** convergência

Figura 6: Algoritmo de inferência variacional para LDA.

Resumimos o procedimento de inferência variacional na Figura 6, com pontos de partida apropriados para ϕ_n e γ. A partir do pseudocódigo é claro que cada iteração de inferência variacional para LDA requer operações de $O((N + 1)k)$. Empiricamente, verificamos que o número de iterações necessárias para um único documento está na ordem do número de palavras do documento. Isto produz um número total de operações aproximadamente na ordem de $N^2 k$.

5.3 Estimativa dos parâmetros

Nesta seção apresentamos um método empírico Bayes para a estimativa de parâmetros α e β no modelo LDA (ver seção 5.4 para uma abordagem Bayesiana mais completa). Em particular, dado um corpus de documentos $D = \{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_M\}$, queremos encontrar parâmetros α e β que maximizem a probabilidade de registo (marginal) dos dados:

$$\ell(\alpha, \beta) = \sum_{d=1}^{M} \log p(\mathbf{w}_d \mid \alpha, \beta).$$

Como descrevemos acima, a quantidade $p(\mathbf{w} \mid \alpha, \beta)$ não pode ser calculada de forma traçável. No entanto, a inferência variacional nos fornece um limite inferior traçável na probabilidade de log, um limite que podemos maximizar em relação a α e β. Assim, podemos encontrar estimativas empíricas aproximadas de Bayes para o modelo LDA através de um procedimento de *EM variacional* alternada que maximiza um limite inferior em relação aos parâmetros variacionais γ e ϕ, e então, para valores fixos dos parâmetros variacionais, maximiza o limite inferior em relação aos parâmetros do modelo α e β.

Nós fornecemos uma derivação detalhada do algoritmo EM variacional para LDA no Apêndice A.4.

A derivação produz o seguinte algoritmo iterativo:

1. (E-step) Para cada documento, encontre os valores de otimização dos parâmetros variacionais $\{\gamma^*_d, \phi^*_d : d \in D\}$. Isto é feito como descrito na secção anterior.

2. (M-step) Maximizar o limite inferior resultante na probabilidade de registo em relação aos

parâmetros do modelo e . Isto corresponde a encontrar estimativas de máxima verosimilhança com estatísticas suficientes esperadas para cada documento sob o posterior aproximado que é calculado no passo E.

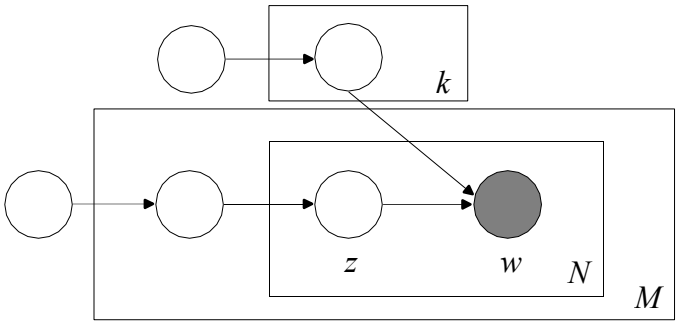

Figura 7: Representação gráfica do modelo alisado de LDA.

Estes dois passos são repetidos até que o limite inferior da probabilidade do log converge.

No Apêndice A.4, mostramos que a atualização dos passos M para o παρ®μετρο condicional multinomial

pode ser escrita de forma analítica:

$$\sum_{d=1}^{M} \sum_{n=1}^{N_d} \phi_{ij}^{*} w_{dni}^{j} w_{dnt}^{i}. \tag{9}$$

Mostramos ainda que o parâmetro M-step update para Dirichlet pode ser implementado usando um eficiente método Newton-Raphson no qual o Hessian é invertido em tempo linear.

5.4 Alisamento

O grande tamanho do vocabulário que é característico de muitos corpora de documentos cria sérios problemas de esparsidade. É muito provável que um novo documento contenha palavras que não apareceram em nenhum dos documentos de um corpus de treinamento. As estimativas de máxima probabilidade dos parâmetros multinomiais atribuem probabilidade zero a tais palavras e, portanto, probabilidade zero a novos documentos. A abordagem padrão para lidar com este problema é "suavizar" os parâmetros multinomiais, atribuindo probabilidade positiva a todos os itens de vocabulário, quer sejam ou não observados no conjunto de treinamento (Jelinek, 1997). O alisamento Laplace é comumente usado; isto essencialmente produz a média da distribuição posterior sob um Dirichlet anterior uniforme sobre os parâmetros multinomiais.

Infelizmente, na configuração do modelo de mistura, o alisamento simples de Laplace já não se justifica como método máximo a posteriori (embora seja frequentemente implementado na prática; cf. Nigam et al., 1999). Na verdade, colocando um Dirichlet prévio no parâmetro multinomial obtemos um posterior intratável no ajuste do modelo de mistura, pela mesma razão que se obtém um posterior intratável no modelo básico de LDA. Nossa solução proposta para este problema é simplesmente aplicar métodos de inferência variacional ao modelo estendido que inclui o alisamento de Dirichlet sobre o parâmetro multinomial.

Na configuração LDA, obtemos o modelo gráfico ampliado mostrado na Figura 7. Tratamos como uma matriz aleatória $k \times V$ (uma linha para cada componente da mistura), onde assumimos que cada linha é extraída independentemente de uma distribuição Dirichlet permutável.[2] Agora estendemos nossos procedimentos de infer- ence para tratar o_i como variáveis aleatórias que são dotadas de uma distribuição posterior,

2. Um Dirichlet permutável é simplesmente uma distribuição Dirichlet com um único parâmetro escalar . A densidade é a mesma de um Dirichlet (Eq. 1) onde $_i =$ para cada componente.

condicionado nos dados. Assim, vamos além do procedimento Bayes empírico do capítulo 5.3 e consideramos uma abordagem Bayesiana mais completa da LDA.

Consideramos uma abordagem variacional da inferência Bayesiana que coloca uma distribuição separável nas variáveis aleatórias , e **z** (Attias, 2000):

$$q(_{1:k}\ ,\ \mathbf{z}_{1:M}\ ,_{1:M}\ |,\ ,\ ) = \prod_{i=1}^{k} \mathrm{Dir}(_i\ |_i\ ) \prod_{d=1}^{M} q_d\ (_d\ ,\ \mathbf{z}_d\ |_d\ ,_d\ ),$$

onde q_d (, **z** |,) é a distribuição variacional definida para LDA em Eq. (4). Como facilmente verificado, o procedimento de inferência variacional resultante produz novamente Eqs. (6) e (7) como as equações de atualização para os parâmetros variacionais e , respectivamente, assim como uma atualização adicional para o novo parâmetro variacional :

$$_{ij} = +* \sum_{d=1}^{M} \sum_{n=1}^{Nd} {}_{dni}\, w_{dn}^{j}.$$

Iterar estas equações para convergência produz uma distribuição posterior aproximada em , e **z**. Ficamos agora com o hiperparâmetro no Dirichlet permutável, bem como o hy- perparâmetro de antes. Nossa abordagem para definir estes hiperparâmetros é novamente (aproximada) empírica Bayes - usamos o EM variacional para encontrar estimativas de máxima verosimilhança destes parâmetros.
com base na probabilidade marginal. Estes procedimentos estão descritos no Apêndice A.4.

6. Exemplo

Nesta seção, fornecemos um exemplo ilustrativo do uso de um modelo LDA em dados reais. Nossos dados são 16.000 documentos de um subconjunto do TREC AP corpus (Harman, 1992). Após remover uma lista padrão de palavras de parada, nós usamos o algoritmo EM descrito na Seção 5.3 para encontrar os parâmetros Dirichlet e multinomial condicional para um modelo LDA de 100 tópicos. As palavras de topo de algumas das distribuições multinomiais resultantes $p(w \mid z)$ estão ilustradas na Figura 8 (topo). Como esperávamos, estas distribuições parecem capturar alguns dos tópicos subjacentes no corpus (e nós os nomeamos de acordo com estes tópicos).

Como enfatizamos na seção 4, uma das vantagens do LDA sobre a variável latente relacionada mod- els é que ele fornece procedimentos de inferência bem definidos para documentos previamente não vistos. Na verdade, podemos ilustrar como a LDA funciona realizando inferências sobre um documento retido e examinando os parâmetros posteriores variacionais resultantes.

A Figura 8 (inferior) é um documento do corpus TREC AP que não foi utilizado para a estimativa de parâmetros. Usando o algoritmo da Secção 5.1, calculámos os parâmetros Dirichlet posteriores variacionais para o artigo e parâmetros multinomiais posteriores variacionais$_n$ para cada palavra no artigo.

Lembre-se de que o i-ésimo parâmetro Dirichlet posterior$_i$ é aproximadamente o i-ésimo parâmetro Dirichlet anterior pa- rameter$_i$ mais o número esperado de palavras que foram geradas pelo i-ésimo tópico (ver Eq. 7). Portanto, os parâmetros anteriores do Dirichlet subtraídos dos parâmetros posteriores do Dirichlet indicam o número esperado de palavras que foram atribuídas a cada tópico para um determinado documento. Para o artigo de exemplo da Figura 8 (inferior), a maior parte do$_i$ está perto de$_i$. Quatro tópicos, no entanto, são significativamente maiores (com isso, queremos dizer$_i$ -$_i$ $\geq$ 1). Olhando para as distribuições correspondentes sobre

As palavras identificam os tópicos que se misturam para formar este documento (Figura 8, topo).

Uma visão mais aprofundada vem do exame dos parâmetros$_n$. Estas distribuições aproximam-se de $p(z_n \mid \mathbf{w})$ e tendem a atingir um dos k valores possíveis do tópico. No texto do artigo na Figura 8, as palavras são codificadas por cores de acordo com esses valores$_n$ (ou seja, a i-ésima cor é usada se $q_n\,(z^i = 1) > 0{,}9$). Com esta ilustração, é possível identificar como os diferentes tópicos se misturam no texto do documento.

Embora demonstrando o poder da LDA, a análise posterior também destaca algumas de suas itações de cal. Em particular, o pressuposto do saco de palavras permite que palavras que devem ser geradas pelo mesmo tópico (por exemplo, "William Randolph Hearst Foundation") sejam alocadas a vários top- ics diferentes. Superar esta limitação exigiria alguma forma de extensão do modelo básico da LDA; em particular, poderíamos relaxar a suposição do saco de palavras assumindo a permutabilidade parcial ou Markovianity das seqüências de palavras.

7. Aplicações e Resultados Empíricos

Nesta seção, discutimos nossa avaliação empírica do LDA em vários domínios problemáticos - modelagem de documentos, classificação de documentos e filtragem colaborativa.

Em todos os modelos de mistura, a probabilidade esperada de registro completo dos dados tem max- ima local nos pontos em que todos ou alguns dos componentes da mistura são iguais uns aos outros. Para evitar estes máximos locais, é importante inicializar o algoritmo EM de forma apropriada. Em nossos experimentos, nós inicializamos o EM semeando cada distribuição condicional multinomial com cinco documentos, reduzindo seu comprimento total efetivo para duas palavras, e suavizando todo o vocabulário. Isto é essencialmente uma aproximação ao esquema descrito em Heckerman e Meila (2001).

7.1 Modelagem de documentos

Treinamos uma série de modelos de variáveis latentes, incluindo LDA, em dois corpora de texto para comparar o desempenho de generalização desses modelos. Os documentos nos corpora são tratados como não rotulados; assim, nosso objetivo é a estimativa de densidade - desejamos alcançar alta probabilidade em um conjunto de testes realizados. Em particular, calculamos a *perplexidade* de um conjunto de testes para avaliar os modelos. A perplexidade, usada por convenção na modelagem de linguagem, é monotonicamente decrescente na probabilidade dos dados do teste, e é equivalente algébrica ao inverso da média geométrica por palavra de probabilidade. Uma menor pontuação de perplexidade indica um melhor desempenho na generalização.[3] Mais formalmente, para um conjunto de testes de documentos M, a perplexidade é:

$$\text{perplexidade}(D_{\text{test}}) = \exp\left\{-\frac{\sum_{d=1}^{M}\log p(\mathbf{w}_d)}{\sum_{d=1}^{M} N_d}\right\}.$$

Em nossas experiências, utilizamos um corpus de resumos científicos da comunidade C. Elegans (Av- ery, 2002) contendo 5.225 resumos com 28.414 termos únicos, e um subconjunto do corpus TREC AP contendo 16.333 artigos de notícias com 23.075 termos únicos. Em ambos os casos, foram disponibilizados 10% dos dados para fins de teste e treinados os modelos sobre os 90% restantes. No pré-processamento dos dados,

3. Note que simplesmente usamos a perplexidade como uma figura de mérito para comparar modelos. Os modelos que comparamos são todos modelos unigramáticos ("saco de palavras"), os quais - como discutimos na Introdução - são de interesse no contexto da recuperação de informação. *Não* estamos tentando fazer modelagem de linguagem neste artigo - uma empresa que exigiria que examinássemos o trigrama ou outros modelos de ordem superior. No entanto, notamos de passagem que extensões da LDA poderiam ser consideradas que envolvem Dirichlet-multinomial sobre trigramas em vez de unigramas. Deixamos a exploração de tais extensões para a modelagem de linguagem para o trabalho futuro.

"Artes".	"Orçamentos".	"Crianças".	"Educação".
NOVO	MILHÃO	CRIANÇAS	ESCOLA
FILM	IMPOSTO	MULHERES	ESTUDANTES
PROJETE	PROGRAMA	PESSOAS	ESCOLAS
MÚSICA	ORÇAMENTO	CRIANÇA	EDUCAÇÃO
MOVIE	BILHÃO	EXERCÍCIOS	PROFESSORES
JOGUE	FEDERAL	FAMILIARES	ELEVADO
MUSICAL	EXERCÍCIO	TRABALHO	PÚBLICO
MELHOR	ESPERANÇA	PARENTES	PROFESSOR
ACTOR	NOVO	SAYS	BENNETT
PRIMEIRO	ESTADO	FAMÍLIA	MANIGAT
YORK	PLANO	WELFARE	NAMPHY
OPERA	DINHEIRO	HOMENS	ESTADO
TEATRO	PROGRAMAS	PERCENTRO	PRESIDENTE
ATUALIZAÇÃO	GOVERNAMENTO	CARE	ELEMENTAR
AMOR	CONGRESSO	LIFE	HAITI

Figura 8: Um artigo de exemplo do corpus AP. Cada cor codifica um fator diferente do qual a palavra é gerada putativamente.

A Fundação William Randolph Hearst doará US$ 1,25 milhão ao Lincoln Center, Metropoli- tan Opera Co., New York Philharmonic e Juilliard School. "Nossa diretoria sentiu que tivemos uma oportunidade real de deixar uma marca no futuro das artes cênicas com essas doações um ato tão importante quanto nossas áreas tradicionais de apoio em saúde, pesquisa médica, educação e serviços sociais", disse o presidente da Hearst Foundation, Randolph A. Hearst, na segunda-feira, ao anunciar as doações. A parte do Lincoln Center será de US$ 200.000 para seu novo prédio, que abrigará jovens artistas e oferecerá novas instalações públicas. A Metropolitan Opera Co. e a Filarmônica de Nova York receberão US$ 400.000 cada uma. A Juilliard School, onde a música e as artes cênicas são ensinadas, receberá US$ 250.000 cada. A Hearst Foundation, um dos principais apoiadores do Lincoln Center Consolidated Corporate Fund, também fará sua habitual doação anual de US$100.000.

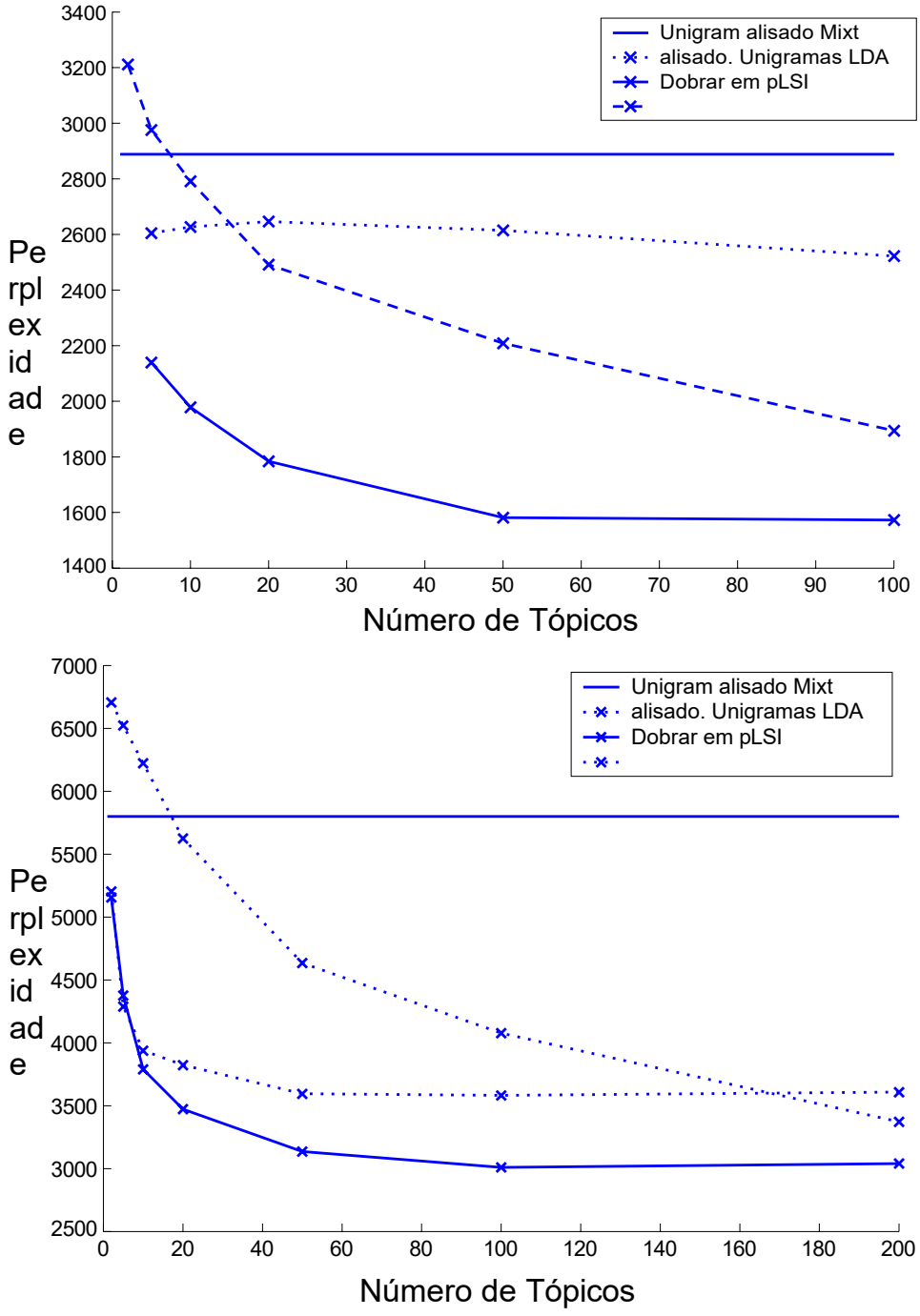

Figura 9: Resultados de perplexidade nos corpos nematódeos (superior) e AP (inferior) para LDA, o modelo de unigrama, mistura de unigramas, e pLSI.

Num. tópicos (k)	Perplexidade (Mult. Mixt.)	Perplexidade (pLSI)
2	22,266	7,052
5	2.20×10^8	17,588
10	1.93×10^{17}	63,800
20	1.20×10^{22}	2.52×10^5
50	4.19×10^{106}	5.04×10^6
100	2.39×10^{150}	1.72×10^7
200	3.51×10^{264}	1.31×10^7

Tabela 1: Sobreposição na mistura de unigramas e modelos pLSI para o corpus AP. Comportamento similar é observado no nematode corpus (não relatado).

removemos uma lista padrão de 50 palavras de paragem de cada corpus. Dos dados do AP, removemos ainda palavras que ocorreram apenas uma vez.

Comparamos o LDA com o unigrama, mistura de unigramas, e modelos pLSI descritos no Sec- tion 4. Nós treinamos todos os modelos de variáveis ocultas usando EM com exatamente o mesmo critério de parada, que a mudança média na probabilidade de log esperada é inferior a 0,001%.

Tanto o modelo pLSI como a mistura de unigramas sofrem de sérios problemas de sobreajustamento, embora por diferentes razões. Este fenómeno é ilustrado na Tabela 1. Na mistura do modelo de unigramas, o sobreajustamento é resultado do pico de posteriors no conjunto de treinamento; um fenômeno familiar no cenário super-visado, onde este modelo é conhecido como o modelo Bayes ingênuo (Rennie, 2001). Isto leva a um agrupamento quase determinístico dos documentos de treinamento (no passo E) que é usado para determinar as probabilidades da palavra em cada componente da mistura (no passo M). Um documento anteriormente invisível pode caber melhor em um dos componentes da mistura resultante, mas provavelmente conterá pelo menos uma palavra que não ocorreu nos documentos de treinamento que foram atribuídos a esse componente. Tais palavras terão uma probabilidade muito pequena, o que causa a perplexidade do novo documento a explodir. medida que k aumenta, os documentos do corpo de treinamento são divididos em coleções mais finas e assim induzem mais palavras com pequenas probabilidades.

Na mistura de unigramas, podemos aliviar a sobreposição através da variação do esquema Bayesiano de suavização apresentado na seção 5.4. Isto assegura que todas as palavras terão alguma probabilidade sob cada componente da mistura.

No caso da pLSI, o problema do agrupamento duro é atenuado pelo fato de que cada documento pode exibir uma proporção diferente de tópicos. No entanto, pLSI refere-se apenas ao documento de treinamento - uments e surge um problema de sobreposição diferente que se deve à dimensionalidade do parâmetro $p(z|d)$. Uma abordagem razoável para atribuir probabilidade a um documento previamente invisível é por
marginalizando sobre d:

$$p(\mathbf{w}) = \sum_{d} \prod_{n=1}^{N} p(w_n \mid z) p(z \mid d) p(d).$$

Essencialmente, estamos integrando sobre a distribuição empírica sobre o tema simplex (ver Figura 4).

Este método de inferência, embora teoricamente sólido, faz com que o modelo se ajuste em demasia. A distribuição temática específica do documento tem alguns componentes que são próximos de zero para aqueles tópicos que não aparecem no documento. Assim, certas palavras terão uma probabilidade muito pequena nas estimativas de

cada componente da mistura. Ao determinar a probabilidade de um novo documento por meio de uma digitalização marginal, somente os documentos de treinamento que exibem uma proporção semelhante de tópicos contribuirão para a probabilidade. Para as proporções dos tópicos de um determinado documento de treinamento, qualquer palavra que tenha pequena probabilidade em todos os tópicos constituintes causará a explosão da perplexidade. À medida que k aumenta, a probabilidade de um documento de treinamento exibir tópicos que cobrem todas as palavras do novo documento diminui e, portanto, a perplexidade aumenta. Note que o pLSI não se ajusta tão rapidamente (em relação ao k) como a mistura de unigramas.

Este problema de sobreajuste decorre essencialmente da restrição de que cada documento futuro exiba as mesmas proporções temáticas que foram vistas em um ou mais dos documentos de formação. Dada esta restrição, não somos livres de escolher as proporções mais prováveis de tópicos para o novo documento. Uma abordagem alternativa é a heurística "folding-in" sugerida por Hofmann (1999), onde se ignora
os parâmetros $p(z|d)$ e reajusta $p(z|d_{\text{new}})$. Note que isso dá ao modelo pLSI uma vantagem injusta ao permitir que ele reajuste $k - 1$ parâmetros para os dados de teste.

A LDA não sofre de nenhum destes problemas. Como na pLSI, cada documento pode exibir uma proporção diferente dos tópicos subjacentes. Contudo, a LDA pode facilmente atribuir probabilidade a um novo documento; não são necessárias heurísticas para que um novo documento seja dotado de um conjunto diferente de proporções de tópicos do que as associadas aos documentos do corpo de treinamento.

A Figura 9 apresenta a perplexidade de cada modelo em ambos os corpos para diferentes valores de k. O modelo pLSI e a mistura de unigramas são adequadamente corrigidos para o ajuste excessivo. Os modelos de variáveis latentes têm melhor desempenho do que o modelo de unigramas simples. O LDA tem um desempenho consistentemente melhor que os outros modelos.

7.2 Classificação de documentos

No problema de classificação de texto, desejamos classificar um documento em duas ou mais classes ex-clientes mutuamente. Como em qualquer problema de classificação, podemos querer considerar abordagens generativas ou abordagens discriminatórias. Em particular, usando um módulo LDA para cada classe, obtemos um modelo generativo de classificação. Também é interessante usar LDA no quadro discriminatório, e este é o nosso foco nesta seção.

Um aspecto desafiador do problema de classificação de documentos é a escolha das características. Tratar palavras individuais como características produz um conjunto rico mas muito grande de características (Joachims, 1999). Uma forma de reduzir este conjunto de características é usar um modelo LDA para redução da dimensionalidade. Em particular, o LDA reduz qualquer documento a um conjunto fixo de características de valor real - os parâmetros posteriores do Dirichlet
$*(\mathbf{w})$ associado ao documento. É de interesse ver quanta informação discriminatória
perdemos ao reduzir a descrição do documento a estes parâmetros.

Conduzimos duas experiências de classificação binária usando o conjunto de dados Reuters-21578. O conjunto de dados contém 8000 documentos e 15.818 palavras.

Nessas experiências, estimamos os parâmetros de um modelo LDA em todos os documentos, sem referência à sua verdadeira etiqueta de classe. Em seguida, treinamos uma máquina vetorial de suporte (SVM) sobre as representações de baixa dimensão fornecidas pela LDA e comparamos

esta SVM com uma SVM treinada em todas as características da palavra.

Usando o pacote de software SVMLight (Joachims, 1999), comparamos um SVM treinado em todos os recursos da palavra com aqueles treinados em recursos induzidos por um modelo LDA de 50 tópicos. Note que neste caso reduzimos o espaço das funcionalidades em 99,6 por cento.

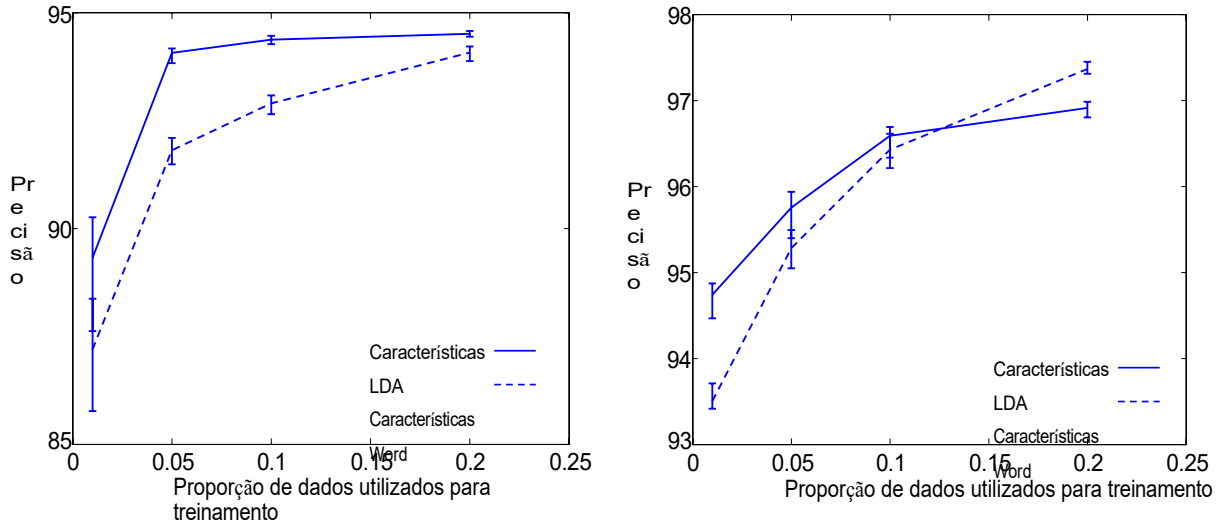

Figura 10: Resultados de classificação em dois problemas de classificação binária do conjunto de dados Reuters-21578 para diferentes proporções de dados de treinamento. O gráfico (a) é EARN VS. NOT EARN. O gráfico (b) é GRAIN VS. NOT GRAIN.

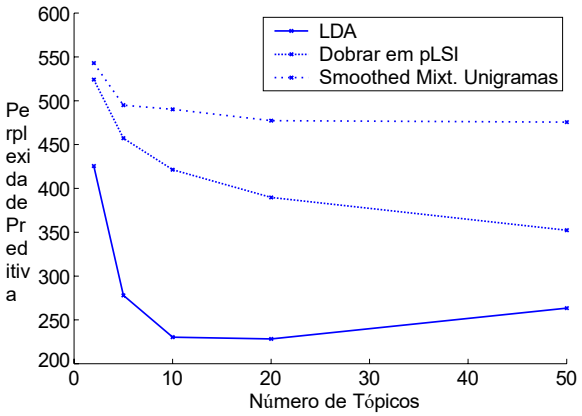

Figura 11: Resultados para filtragem colaborativa nos dados de cadaMovie.

A Figura 10 mostra os nossos resultados. Vemos que há pouca redução no desempenho da classificação no uso das características baseadas em LDA; de fato, em quase todos os casos o desempenho é melhorado com as características de LDA. Embora estes resultados necessitem de uma maior fundamentação, eles sugerem que a representação baseada no tópico fornecida pelo LDA pode ser útil como um algoritmo de filtragem rápida para a seleção de características na

classificação de texto.

7.3 Filtragem colaborativa

Nossa experiência final utiliza os dados de filtragem colaborativa de EachMovie. Neste conjunto de dados, uma coleção de usuários indica suas escolhas preferidas de filmes. Um usuário e os filmes escolhidos são análogos a um documento e as palavras no documento (respectivamente).

A tarefa de filtragem colaborativa é a seguinte. Treinamos um modelo em um conjunto de usuários totalmente observado. Em seguida, para cada usuário não observado, somos exibidos todos os filmes preferidos por esse usuário, com exceção de um, e nos é pedido para prever o que é o filme não observado. Os diferentes algoritmos são avaliados de acordo com a probabilidade que eles atribuem ao filme não observado. Mais precisamente, define-se a perplexidade preditiva dos usuários do teste M a serem:

$$\textit{perplexidade preditiva}(D_{\text{test}}) = \exp\left\{ -\frac{\sum_{d=1}^{M} \log p(w_{d,N_d} \mid \mathbf{wd}_{,1:N_d-1})}{M} \right\}.$$

Nós restringimos o conjunto de dados EachMovie aos usuários que classificaram positivamente pelo menos 100 filmes (uma classificação positiva é de pelo menos quatro em cada cinco estrelas). Dividimos este conjunto de usuários em 3300 usuários de treinamento e 390 usuários de teste.

Sob a mistura de unigramas modelo, a probabilidade de um filme dado um conjunto de filmes observados é obtida a partir da distribuição posterior sobre tópicos:

$$p(w|w_{\text{obs}}) = \sum_z p(w|z)p(z|w_{\text{obs}}).$$

No modelo pLSI, a probabilidade de um filme ser retido é dada pela mesma equação, exceto que $p(z|w_{\text{obs}})$ é computada dobrando nos filmes vistos anteriormente. Por fim, no modelo LDA, a probabilidade de um filme ser retido é dada pela integração sobre o Dirichlet posterior:

$$p(w|w_{\text{obs}}) = \int_z p(w|z)p(z|)p(|\omega_{\text{obs}})d,$$

onde $p(|\omega_{\text{obs}})$ é dado pelo método de inferência variacional descrito na Secção 5.2. Note que esta quantidade é eficiente de calcular. Podemos trocar a soma e o sinal integral, e calcular uma combinação linear das expectativas k Dirichlet.

Com um vocabulário de 1600 filmes, encontramos as perplexidades preditivas ilustradas na Figura 11. Novamente, a mistura do modelo de unigramas e pLSI é corrigida por excesso de ajuste, mas as melhores perplexidades preditivas são obtidas pelo modelo LDA.

8. Discussão

Descrevemos a alocação de Dirichlet latente, um modelo probabilístico generativo flexível para a coleta de dados discretos. O LDA é baseado em um simples pressuposto de permutabilidade para as palavras e tópicos de um documento; ele é, portanto, realizado por uma aplicação direta do teorema de repre- sentação de Finetti. Podemos ver o LDA como uma técnica de redução de dimensionalidade, no espírito do LSI, mas com uma semântica probabilística generativa subjacente adequada que faça sentido para o tipo de dados que ele modela.

A inferência exata é intratável para LDA, mas qualquer um de um grande conjunto de inferências aproximadas de algo- rithms pode ser usado para inferência e estimativa de parâmetros dentro da estrutura LDA. Apresentamos uma abordagem simples baseada na

convexidade para a inferência, mostrando que ela produz uma rápida

algoritmo que resulta em um desempenho comparativo razoável em termos de probabilidade do conjunto de testes. Outras abordagens que podem ser consideradas incluem a aproximação de Laplace, técnicas de maior variabilidade de ordem - niques, e métodos Monte Carlo. Em particular, Leisink e Kappen (2002) apresentaram uma metodologia geral para converter limites inferiores de variação de baixa ordem em limites superiores de variação de ordem tional. Também é possível alcançar maior precisão dispensando a exigência de manter um limite, e de fato Minka e Lafferty (2002) mostraram que uma melhor precisão inferencial pode ser obtida para o modelo LDA através de uma técnica de variação de ordem mais alta conhecida como propagação ex pectation. Finalmente, Griffiths e Steyvers (2002) apresentaram um algoritmo de cadeia de Markov Monte Carlo para LDA.

LDA é um modelo simples, e embora o vejamos como um concorrente de métodos como LSI e pLSI no estabelecimento de redução de dimensionalidade para coleções de documentos e outros cor- pora discretos, ele também pretende ser ilustrativo da forma como modelos probabilísticos podem ser escalados para fornecer máquinas inferenciais úteis em domínios que envolvem múltiplos níveis de estrutura. In- Deed, as principais vantagens dos modelos generativos como o LDA incluem a sua modularidade e a sua extensibilidade. Como um módulo probabilístico, o LDA pode ser prontamente incorporado em um modelo mais complexo - uma propriedade que não é possuída pelo LSI. Em trabalhos recentes temos usado pares de módulos LDA para modelar relações entre imagens e suas correspondentes legendas descritivas (Blei e Jordan, 2002). Além disso, existem numerosas extensões possíveis do LDA. Por exemplo, o LDA é facilmente estendido para dados contínuos ou outros dados não-multinomiais. Como é o caso de outros modelos de mistura, incluindo modelos de mistura finita e modelos Markov ocultos, a probabilidade de "emissão" $p(w_n \mid z_n)$ contribui apenas com um valor de probabilidade para os procedimentos de inferência do LDA, e outras probabilidades são prontamente substituídas em seu lugar. Em particular, é simples desenvolver uma variante contínua de LDA na qual os observáveis gaussianos são usados no lugar dos multinomiais. Outra extensão simples da LDA vem de permitir misturas de distribuições de Dirichlet no lugar do Dirichlet único da LDA. Isto permite uma estrutura mais rica no espaço de tópicos latentes e em particular permite uma forma de agrupamento de documentos que é diferente do agrupamento que é alcançado através de tópicos compartilhados. Finalmente, uma variedade de extensões de LDA pode ser considerada na qual as distribuições sobre as variáveis do tópico são elaboradas. Por exemplo, poderíamos organizar os tópicos em uma série temporal, essencialmente relaxando a hipótese de intercambialidade total para uma de intercambialidade parcial. Poderíamos também considerar modelos parcialmente intercambiáveis nos quais condicionamos as variáveis exógenas; assim, por exemplo, a distribuição temática poderia ser condicionada a características como "parágrafo" ou "frase", fornecendo um modelo de texto mais poderoso que faz uso de informações obtidas de um analisador.

Agradecimentos

Este trabalho foi apoiado pela National Science Foundation (NSF grant IIS-9988642) e pelo Programa de Pesquisa Multidisciplinar do Departamento de Defesa (MURI N00014-00-1-0637). Andrew Y. Ng e David M. Blei foram também apoiados por bolsas de estudo da Microsoft Corporation.

Referências

M. Abramowitz e I. Stegun, editores. *Manual de Funções Matemáticas*. Dover, Nova York, 1970.

D. Aldous. Permutabilidade e tópicos relacionados. Em *École d'e'te' de probabilite's de Saint-Flour, XIII- 1983*, páginas 1-198. Springer, Berlim, 1985.

H. Attias. Uma estrutura Bayesiana variável para modelos gráficos. In *Advances in Neural Informa- tion Processing Systems 12*, 2000.

L. Avery. Caenorrhabditis genético centro bibliografia. 2002. URL http://elegans.swmed.edu/wli/cgcbib.

R. Baeza-Yates e B. Ribeiro-Neto. *Modern Information Retrieval*. ACM Press, Nova York, 1999.

D. Blei e M. Jordan. Modelagem de dados anotados. Relatório Técnico UCB//CSD-02-1202, U.C. Berkeley Computer Science Division, 2002.

B. de Finetti. *Teoria da probabilidade. Vol. 1-2*. John Wiley & Sons Ltd., Chichester, 1990. Reimpressão da tradução de 1975.

S. Deerwester, S. Dumais, T. Landauer, G. Furnas, e R. Harshman. Indexação por análise semântica latente. *Journal of the American Society of Information Science*, 41(6):391-407, 1990.

P. Diaconis. Progresso recente nas noções de permutabilidade do de Finetti. Em *Bayesian statistics, 3 (Valencia, 1987)*, páginas 111-125. Oxford Univ. Press, Nova Iorque, 1988.

J. Dickey. Múltiplas funções hipergeométricas: Interpretações probabilísticas e usos estatísticos. *Journal of the American Statistical Association*, 78:628-637, 1983.

J. Dickey, J. Jiang, e J. Kadane. Métodos Bayesianos para dados categóricos censurados. *Journal of the American Statistical Association*, 82:773-781, 1987.

A. Gelman, J. Carlin, H. Stern, e D. Rubin. *Análise de dados Bayesianos*. Chapman & Hall, Londres, 1995.

T. Griffiths e M. Steyvers. Uma abordagem probabilística da representação semântica. Em *Proceedings of the 24th Annual Conference of the Cognitive Science Society*, 2002.

D. Harman. Visão geral da primeira conferência de recuperação de texto (TREC-1). Em *Proceedings of the First Text Retrieval Conference (TREC-1)*, páginas 1-20, 1992.

D. Heckerman e M. Meila. Uma comparação experimental de vários métodos de agrupamento e inicialização. *Machine Learning*, 42:9-29, 2001.

T. Hofmann. Indexação semântica latente probabilística. *Anais da Vigésima Segunda Conferência Anual Internacional SIGIR*, 1999.

F. Jelinek. *Métodos Estatísticos para o Reconhecimento da Fala*. MIT Press, Cambridge, MA, 1997.

T. Joachims. Tornando prática a aprendizagem em larga escala de SVM. Em *Avanços nos Métodos de Kernel - Suporte à Aprendizagem Vetorial*. M.I.T. Press, 1999.

M. Jordan, editor. *Aprendizagem em Modelos Gráficos*. MIT Press, Cambridge, MA, 1999.

M. Jordan, Z. Ghahramani, T. Jaakkola, e L. Saul. Introdução a métodos variacionais para modelos gráficos e físicos. *Machine Learning*, 37:183-233, 1999.

R. Kass e D. Steffey. Inferência Bayesiana aproximada em modelos hierárquicos condicionalmente independentes (modelos empíricos paramétricos Bayes). *Journal of the American Statistical Association*, 84 (407):717-726, 1989.

M. Leisink e H. Kappen. Limites gerais mais baixos baseados em pansions ex de ordem superior gerados por computador. Em *Uncertainty in Artificial Intelligence, Proceedings of the Eighteenth Conference*, 2002.

T. Minka. Estimando uma distribuição Dirichlet. Relatório técnico, M.I.T., 2000.

T. P. Minka e J. Lafferty. Expectativa-propagação para o modelo do aspecto generativo. Em *Uncertainty in Artificial Intelligence (UAI)*, 2002.

C. Morris. Inferência empírica paramétrica de Bayes: Teoria e aplicações. *Journal of the American Statistical Association*, 78(381):47-65, 1983. Com discussão.

K. Nigam, J. Lafferty, e A. McCallum. Usando entropia máxima para a classificação do texto. *IJCAI-99 Workshop on Machine Learning for Information Filtering*, páginas 61-67, 1999.

K. Nigam, A. McCallum, S. Thrun, e T. Mitchell. Classificação de texto de documentos rotulados e não rotulados usando EM. *Machine Learning*, 39(2/3):103-134, 2000.

C. Papadimitriou, H. Tamaki, P. Raghavan, e S. Vempala. Indexação semântica latente: Uma análise proba- bilística. páginas 159-168, 1998.

A.Popescul, L. Ungar, D. Pennock, e S. Lawrence. Modelos probabilísticos para recomendação colaborativa unificada e baseada no conteúdo em ambientes de dados escassos. Em *Uncertainty in Artificial Intelligence, Proceedings of the Seventeenth Conference*, 2001.

J. Rennie. Melhorando a classificação de textos multi-classe com Bayes ingênuo. Relatório Técnico AITR-2001- 004, M.I.T., 2001.

G. Ronning. Estimativa da probabilidade máxima das distribuições de Dirichlet. *Journal of Statistcal Com- putation and Simulation*, 34(4):215-221, 1989.

G. Salton e M. McGill, editores. *Introdução à Recuperação de Informação Moderna*. McGraw-Hill, 1983.

Apêndice A. Inferência e estimação de parâmetros

Neste apêndice, derivamos o procedimento de inferência variacional (Eqs. 6 e 7) e o procedimento de maximização de parâmetros para a multinomial condicional (Eq. 9) e para o Dirichlet. Começamos por derivar uma propriedade útil da distribuição do Dirichlet.

A.1 Informática E[log($_i$ |)]

A necessidade de calcular o valor esperado do log de um único componente de probabilidade sob o Dirichlet surge repetidamente ao derivar os procedimentos de inferência e estimativa de parâmetros para o LDA. Este valor pode ser facilmente calculado a partir da parametrização natural da representação da família exponencial da distribuição Dirichlet.

Recall that a distribution is in the exponential family if it can be written in the form:

$$p(x|) = h(x) \exp^{T} T(x) - A() ,$$

onde está o parâmetro natural, $T(x)$ é a estatística suficiente, e $A()$ é o log do fator de normalização.

Podemos escrever o Dirichlet nesta forma exponenciando o log do Eq. (1):

$$p(|) = \exp^{k} (_i - 1)\log\gamma_i + \log\gamma^{k} \qquad _{i=1} \quad -_i^{k} \quad _{i=1} \quad \log(_i) .$$

Deste formulário, vemos imediatamente que o parâmetro natural do Dirichlet é$_i$ = $_i$ − 1 e a estatística suficiente é $T(_i) = \log\gamma_i$. Além disso, utilizando o fato geral de que a derivada do fator de normalização log em relação ao parâmetro natural é igual à expectativa da estatística suficiente, obtemos:

$$E[\log\gamma_i |] = (_i) - ^{k} \qquad _{j=1} \quad _j$$

onde está a função digamma, a primeira derivada da função gama do log.

A.2 Métodos Newton-Raphson para uma Hessian com estrutura especial

Nesta seção descrevemos um algoritmo linear para o método de otimização geralmente cúbico Newton-Raphson. Este método é usado para a estimativa da máxima verosimilhança da distribuição de Dirichlet (Ron- ning, 1989, Minka, 2000).

A técnica de otimização de Newton-Raphson encontra um ponto estacionário de uma função ao iterar:

$$_{\text{novo}} = _{\text{velho}} - H(_{\text{velho}})^{-1} g(_{\text{velho}})$$

onde $H()$ e $g()$ são a matriz Hessiana e o gradiente, respectivamente, no ponto . Em geral, este algoritmo escalona como $O(N^3)$ devido à inversão da matriz.

Se a matriz Hessiana for da forma:

$$H = \text{diag}(h) + \mathbf{1} z \mathbf{1}^{\mathrm{T}} , \tag{10}$$

onde diag(h) é definido como sendo uma matriz diagonal com os elementos do vector h ao longo da diagonal, então podemos aplicar a matriz de inversão lemma e obter:

$$H^{-1} = \text{diag}(h)^{-1} - \frac{\text{diag}(h)^{-1} \mathbf{1} \mathbf{1}^{\mathrm{T}} \text{diag}(h)^{-1}}{z^{-1} + \sum_{j=1}^{k} h_j^{-1}}$$

Multiplicando pelo gradiente, obtemos o componente i-ésimo:

$$(H^{-1} g)_i = \frac{g_i - c}{h_i}$$

onde

$$c = \frac{\sum_{j=1}^{k} g_j / h_j}{z^{-1} + \sum_{j=1}^{k} h_j^{-1}}.$$

Observe que esta expressão depende apenas dos 2k valores h_i e g_i e assim produz um algoritmo Newton- Raphson que tem complexidade de tempo linear.

A.3 Inferência variável

Nesta secção derivamos o algoritmo de inferência variacional descrito na Secção 5.1. Lembre-se de que isto envolve a utilização da seguinte *distribuição variacional*:

$$q(\theta, \mathbf{z} | \gamma, \phi) = q(\theta | \gamma) \prod_{n=1}^{N} q(z_n | \phi_n) \tag{11}$$

como substituto para a distribuição posterior $p(\theta, \mathbf{z}, \mathbf{w} | \alpha, \beta)$, onde os *parâmetros variacionais* γ e ϕ são definidos através de um procedimento de optimização que agora descrevemos.

Depois de Jordan et al. (1999), começamos por limitar a probabilidade logarítmica de um documento usando a desigualdade de Jensen. Omitindo os parâmetros e por simplicidade, temos:

$$\begin{aligned}
\log p(\mathbf{w} | \alpha, \beta) &= \log \int_{\theta} \sum_{\mathbf{z}} p(\theta, \mathbf{z}, \mathbf{w} | \alpha, \beta) d\theta \\
&= \log \int_{\theta} \sum_{\mathbf{z}} \frac{p(\theta, \mathbf{z}, \mathbf{w} | \alpha, \beta) q(\theta, \mathbf{z})}{q(\theta, \mathbf{z})} d\theta \\
&\geq \int_{\theta} \sum_{\mathbf{z}} q(\theta, \mathbf{z}) \log p(\theta, \mathbf{z}, \mathbf{w} | \alpha, \beta) d\theta - \int_{\theta} \sum_{\mathbf{z}} q(\theta, \mathbf{z}) \log q(\theta, \mathbf{z}) d\theta \\
&= \mathrm{E}_q [\log p(\theta, \mathbf{z}, \mathbf{w} | \alpha, \beta)] - \mathrm{E}_q [\log q(\theta, \mathbf{z})].
\end{aligned} \tag{12}$$

Assim, vemos que a desigualdade de Jensen nos proporciona um limite inferior na probabilidade de uma distribuição variacional arbitrária $q(\theta, \mathbf{z} | \gamma, \phi)$.

Pode-se facilmente verificar que a diferença entre o lado esquerdo e o lado direito da Eq. (12) é a divergência KL entre a probabilidade posterior variacional e a probabilidade posterior verdadeira. Ou seja, deixando $L(\gamma, \phi; \alpha, \beta)$ denotar o lado direito da Eq. (12) (onde restabelecemos a dependência dos parâmetros variacionais e em nossa notação), temos:

$$\log p(\mathbf{w} | \alpha, \beta) = L(\gamma, \phi; \alpha, \beta) + \mathrm{D}(q(\theta, \mathbf{z} | \gamma, \phi) \,\|\, p(\theta, \mathbf{z} | \mathbf{w}, \alpha, \beta)). \tag{13}$$

Isto mostra que a maximização do limite inferior $L(\gamma, \phi; \alpha, \beta)$ com respeito e é equivalente à minimização da divergência KL entre a probabilidade posterior variacional e a probabilidade posterior verdadeira, o problema de otimização apresentado anteriormente em Eq. (5).

Agora expandimos o limite inferior usando as factorizações de p e q:

$$\begin{aligned}
L(\gamma, \phi; \alpha, \beta) = {} & \mathrm{E}_q [\log p(\theta | \alpha)] + \mathrm{E}_q [\log p(\mathbf{z} | \theta)] + \mathrm{E}_q [\log p(\mathbf{w} | \mathbf{z}, \beta)] \\
& - \mathrm{E}_q [\log q(\theta)] - \mathrm{E}_q [\log q(\mathbf{z})].
\end{aligned} \tag{14}$$

Finalmente, expandimos Eq. (14) em termos dos parâmetros do modelo $(,)$ e dos parâmetros variacionais
$(,)$. Cada uma das cinco linhas abaixo expande um dos cinco termos no limite:

$$\begin{aligned} L(,;,) = & \log\Gamma^{k} \quad_{j=1} \quad -_{j} \log(_{i}) + (_{i}-1)(_{i}) \quad -^{k} \quad_{j=1} \mathrm{j} \\ & + \quad_{n=1}^{M} \quad_{i=1}^{k} \quad_{ni}(_{i}) \quad -^{k} \quad_{j=1} \mathrm{j} \\ & + \quad_{n=1}^{N} \quad_{i=1}^{k} \quad_{j=1}^{V} \quad_{ni} w^{j} \log\quad_{ij} \quad_{n} \\ & - \log\Gamma^{k} \quad_{j=1\ j} + \log(_{i}) - (_{i}-1)(_{i}) \quad -^{k} \quad_{j=1} \mathrm{j} \\ & + \quad_{n=1}^{N} \quad_{i=1}^{k} \quad_{ni} - \log\quad_{ni} , \end{aligned} \tag{15}$$

onde temos feito uso do Eq. (8).

Nas duas seções seguintes, mostramos como maximizar este limite inferior em relação aos parâmetros variacionais e .

A.3.1 MULTINOMIAL VARIÁVEL

Primeiro maximizamos Eq. (15) com respeito a$_{ni}$, a probabilidade de que a *enésima* palavra seja gerada por
Tópico latente *i*. Observe que esta é uma maximização restrita desde $\sum_{i=1}^{k} {}_{ni} = 1$.

Formamos o Lagrangian isolando os termos que contêm$_{ni}$ e adicionando os multiplicadores de Lagrange apropriados. Que$_{iv}$ seja $p(w_n^v = 1 \mid z^i = 1)$ para o *v*. apropriado (Lembre-se que cada w_n é um vector de tamanho V com exactamente um componente igual a um; podemos seleccionar o v único tal que $w_n^v = 1$):

$$L_{[_{ni}]} = {}_{ni}(_{i}) \quad -^{k} \quad_{j=1\ j} + {}_{ni}\log\Gamma_{iv} - {}_{ni}\log\Gamma_{ni} + \quad_{n}^{k} \quad_{j=1\ ni} - 1 ,$$

onde abandonamos os argumentos de L por simplicidade, e onde o subscrito$_{ni}$ denota que retemos apenas os termos em L que são uma função de$_{ni}$. Tomando derivados com respeito a
$_{ni}$, nós obtemos:

$$\frac{L}{\quad_{ni}} = (_{i}) \quad -^{k} \quad_{j=1\ j} + \log\Gamma_{iv} - \log\Gamma_{ni} - 1 + .$$

Definindo esta derivada como zero produz o valor máximo do parâmetro variacional$_{ni}$ (cf. Eq. 6):

$$\quad_{niiv} \exp(_{i}) \quad -^{k} \quad_{j=1\ j} . \tag{16}$$

A.3.2 DIRICHLET VARIACIONAL

Em seguida, maximizamos Eq. (15) com respeito a_i , o i-ésimo componente do para-éter posterior do Dirichlet. Os termos contendo$_i$ são:

$$L = \sum_{i=1}^{k} \left(\log \Gamma\left(\sum_{j=1}^{k} \alpha_j\right) - \log \Gamma(\alpha_i) + (\alpha_i - 1)\left(\Psi(\gamma_{di}) - \Psi\left(\sum_{j=1}^{k} \gamma_{dj}\right)\right) \right)$$

Isto simplifica:

$$L = \sum_{i=1}^{k} \left(\log \Gamma\left(\sum_{j=1}^{k} \alpha_j\right) - \log \Gamma(\alpha_i) + (\alpha_i - 1)\left(\Psi(\gamma_{di}) - \Psi\left(\sum_{j=1}^{k} \gamma_{dj}\right)\right) \right)$$

Tomamos o derivado com respeito a_i :

$$\frac{\partial L}{\partial \alpha_i} = \Psi\left(\sum_{j=1}^{k} \alpha_j\right) - \Psi(\alpha_i) + \sum_{n=1}^{N} \left(\Psi(\gamma_{ni}) - \Psi\left(\sum_{j=1}^{k} \gamma_{nj}\right) \right).$$

Definindo esta equação como zero produz um rendimento máximo em:

$$\alpha_i = \alpha_i + \sum_{n=1}^{N} \phi_{ni}. \tag{17}$$

Uma vez que Eq. (17) depende da variação multinomial, a inferência variacional completa requer alternância entre Eqs. (16) e (17) até que o limite converge.

A.4 Estimativa dos parâmetros

Nesta seção final, consideramos o problema de obter estimativas empíricas Bayes dos parâmetros do modelo e . Resolvemos este problema usando o limite inferior variacional como um substituto para a probabilidade logarítmica marginal (intratável), com os parâmetros variacionais e fixos aos valores encontrados pela inferência variacional. Em seguida, obtemos estimativas empíricas (aproximadas) de Bayes, maximizando este limite inferior em relação aos parâmetros do modelo.

Até agora, consideramos a probabilidade de um único documento. Dada a nossa suposição de permutabilidade para os documentos, a probabilidade logarítmica global de um corpus $D = \{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_M\}$ é a soma das probabilidades logarítmicas para documentos individuais; além disso, o limite inferior variacional global é a soma dos limites variacionais individuais. No resto desta secção, abusamos
usando L para o limite variacional total, indexando os termos específicos do documento nos limites individuais por d, e somando sobre todos os documentos.

Lembramos da seção 5.3 que nossa abordagem geral para encontrar estimativas empíricas Bayes é baseada em um procedimento EM variável. No passo E variacional, discutido no Apêndice A.3, nós maximizamos o limite $L(\gamma, \phi; \alpha, \beta)$ com respeito aos parâmetros variacionais e . No passo M, que descrevemos nesta seção, nós maximizamos o limite com respeito aos parâmetros do modelo e . O procedimento geral pode assim ser visto como ascensão coordenada

em L .

A.4.1 MULTINÔMIOS CONDICIONAIS

Para maximizar com respeito a , nós isolamos termos e adicionamos multiplicadores Lagrange:

$$L_{[]} = \underset{d=1}{\overset{M}{}} \underset{n=1}{\overset{Nd}{}} \underset{i=1}{\overset{k}{}} \underset{j=1}{\overset{V}{}} {}_{dni}\text{diário} + {}_{ij}\text{ de bordo} \quad {}_{i} \quad {}_{ij} - 1 \quad .$$

Pegamos a derivada em relação a_{ij} , colocamos a zero, e encontramos:

$$\underset{d=1}{\overset{M}{}} \underset{n=1}{\overset{Nd}{}} {}_{ij\,dniwj} \cdot \quad {}_{dn}$$

A.4.2 DIRICHLET

Os termos que contêm são:

$$L_{[]} = \underset{d=1}{\overset{M}{}} \text{tro}^{\,k} \quad {}_{j=1} \quad {}_{j} \text{ - tronco} \quad ( ) + {}_{i} \underset{i=1}{\overset{k}{}} \quad (\quad {}_{i} - 1 \,) \, ( ) \quad {}_{di} \quad - {}^{k} \quad {}_{j=1} \quad {}_{d\,j}$$
$$\text{nc}$$
$$\text{o}$$

Tomando o derivado com respeito a_i dá:

$$\frac{L}{{}_{i}} = \text{M}^{k} \quad {}_{j=1} \qquad ( ) + \quad {}_{d=1} ( ) - {}_{j} \quad {}_{i} \qquad {}^{k}{}_{j=1\,d\,j}$$
$$d_i$$

Esta derivada depende de$_j$, onde $j \neq i$, e por isso devemos usar um método iterativo para encontrar a máxima . Em particular, a Hessian está na forma encontrada em Eq. (10):

$$\frac{L}{{}_{ij}} = (i, j)\text{M}^{0} ({}_{i}) - {}^{0k} \quad {}_{j=1\,j} \,,$$

e assim podemos invocar o algoritmo de Newton-Raphson de tempo linear descrito no Apêndice A.2.

Finalmente, note que podemos usar o mesmo algoritmo para encontrar uma estimativa empírica do ponto Bayes, o parâmetro escalar para o Dirichlet permutável no modelo alisado LDA na seção 5.4.