(Notes) Latent Dirichlet Allocation

shlip@foxmail.com

2019 年 4 月 14 日

Abstract

我们描述了潜在Dirichlet分配（LDA），一种用于离散数据集合（如文本语料库）的生成概率模型。LDA是一个三级分层贝叶斯模型，其中集合中的每个项目都被建模为一组潜在主题的有限混合。同时，每个主题被建模为一组潜在主题概率的无限混合。在文本建模的上下文中，主题概率提供了文档的显式表示。为了对参数求解，又提出了基于变分方法的有效近似推理技术和用于经验贝叶斯参数估计的EM算法。

1 引言

Tf-idf[1] ，选择“单词”或“术语”的基本词汇表，并且对于语料库中的每个文档，由每个单词的出现次数形成计数。在适当的归一化之后，将该术语频率计数与逆文档频率计数进行比较，该逆文档频率计数测量整个语料库中的单词的出现次数（通常在对数标度上，并且再次适当地标准化）。最终结果是逐个文档矩阵X，其列包含语料库中每个文档的tf-idf值。因此，tf-idf方案将任意长度的文档减少到固定长度的数字列表。

LSI (Latent semantic indexing)[2] ，使用X矩阵上的奇异值分解来识别tf-idf特征空间的捕获集合多数变量的线性子空间。

pLSI[3] 将文档中的每个单词建模为混合模型中的样本，其中混合成分是多项随机变量，可以视为“主题”的表示。因此，每个单词都是从单个主题生成的，文档中的不同单词可能从不同的主题生成。每个文档表示为这些混合物组分的混合比例列表，从而减少到一组固定主题的概率分布。

pLSI在文档层面没有提供概率模型。在pLSI中，每个文档都表示为数字列表（主题的混合比例），并且没有这些数字的生成概率模型。这导致了几个问题：（1）模型中的参数数量随着语料库的大小线性增长，这导致过度拟合的严重问题，并且（2）不清楚如何将概率分配给训练集之外的文档。

以上模型都基于“词袋”假设——文档中的词顺序可以忽略。

由de Finetti（1990）提出的经典表示定理确定任何可交换随机变量的集合都可以表示为混合分布- 通常是无限混合。因此，如果我们希望考虑文档和单词的可交换表示，我

们需要考虑捕获单词和文档可交换性的混合模型。这一思路导致我们在本文中提出的潜在Dirichlet分配（LDA）模型。

需要强调的是，可交换性的假设不等于假设随机变量独立同分布。相反，可交换性基本上可以被解释为“条件独立且相同地分布”，其中条件是关于概率分布的潜在隐含参数。

虽然我们在当前论文中讨论的工作侧重于简单的“词袋”模型，这导致单个词（unigrams）的混合分布，但我们的方法也适用于涉及更大结构的混合物的更丰富的模型n-gram或段落等单位。

2 符号与术语

词离散数据的基本单元，定义为词汇表中的一项，索引为$1,\ldots,V$

文档N个词的序列，表示为$W=(w_1,w_2,\ldots,w_N)$

语料M个文档的集合，表示为$D=W_1,W_2,\ldots,W_M$

3 潜在狄利克雷分布

潜在Dirichlet分布（LDA）是语料库的生成概率模型。基本思想是文档被表示为潜在主题的随机混合，其中每个主题的特征在于对单词的分布

LDA假设语料库D中的每个文档W都有以下生成过程：

1.选择$N\sim Possion(\xi)$。

2.选择$\theta\sim Dir(\alpha)$。

3.对于N个单词中的每个单词w_n：

（a）选择一个主题$z_n\sim Multinomial(\theta)$。

（b）从$p(w_n|z_n,\beta)$中选择一个单词w_n，一个以主题z_n为条件的多项式概率。

LDA基本模型做了如下基本假设（将在后边的章节中移除）：

首先，假设Dirichlet分布的维数k（以及主题变量z的维度）已知并且是固定的。

其次，单词概率由$k\times V$矩阵参数化，其中$_{ij}=p(w^j=1|z^i=1)$，现在我们将其视为要估计的固定数量。

最后，泊松假设对于随后的任何事情都不是至关重要的，并且可以根据需要使用更真实的文档长度分布。

此外，请注意N与所有其他数据生成变量(θ和z)无关。

补充，狄利克雷分布：

狄利克雷分布（Dirichlet distribution）或多元Beta分布（multivariate Beta distribution）是一类在实数域以正单纯形（standard simplex）为支撑集（support）的高维连续概率分布，是Beta分布在高维情形的推广。狄利克雷分布是指数族分布之一。其概率密度函数：

对独立同分布（independent and identically distributed, iid）的连续随机变量$X\in\mathbb{R}_d$ 和支撑集$X\in(0,1),||X||=1$，若X 服从狄利克雷分布，则其概率密度函数$Dir(X|\alpha)$有如下定义：

$$Dir(X|\alpha)=\frac{1}{\mathrm{B}(\alpha)}\prod_{i=1}^{d}X_i^{\alpha_i-1}$$

$$\mathrm{B}(\alpha)=\frac{\prod_{i=1}^{d}\Gamma(\alpha_i)}{\Gamma(\alpha_0)},\quad \alpha_0=\sum_{i=1}^{d}\alpha_i,\quad d\geq 3$$

式中，$\alpha \in \alpha_1, \ldots, \alpha_d$是无量纲的分布参数，$\alpha_0$是分布参数的和，$\beta_\alpha$是多元Beta函数（multivariate beta function），$\Gamma(\alpha)$为Gamma函数。由上述解析形式可知，狄利克雷分布是指数族分布[1] 。

k维Dirichlet随机变量θ可以取$(k-1)-Ub$ 中的值（如果$\theta_i e 0$，$\sum_{i=1}^{k}$，则k向量θ在(k-1)-单纯形中），并且具有以下概率密度：

$$p(\theta|\alpha) = \frac{\Gamma\left(\sum_{i=1}^{k} \alpha_i\right)}{\prod_{i=1}^{k} \Gamma\left(\alpha_i\right)} \theta_1^{\alpha_1 - 1} \cdots \theta_k^{\alpha_k - 1} \quad 1$$

其中参数α是k向量，其中分量$\alpha_i > 0$，而$\Gamma(x)$是Gamma函数。Dirichlet在单纯形上是一个方便的分布- 它在指数族中，具有有限维数的充分统计，并且与多项式分布共轭。

给定参数α和β，主题混合θ的联合分布，N个主题z的集合，N个词的集合由如下方式给出：

$$p(\theta, \mathbf{z}, \mathbf{w}|\alpha, \beta) = p(\theta|\alpha) \prod_{n=1}^{N} p\left(z_n|\theta\right) p\left(w_n|z_n, \beta\right) \quad 2$$

其中，$p(z_n|\theta)$对$z_n^i = 1$的唯一i取θ_i。在θ上求积分并在z上求和，得到文档的边缘分布：

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

最后，对单个文档的边缘概率求乘积，就得到语料的概率：

$$p(D|\alpha, \beta) = \prod_{d=1}^{M} \int p(\theta_d|\alpha) (\prod_{n=1}^{N_d} \sum_{z_{dn}} p(z_{dn}|\theta_d) p(w_{dn}|z_{dn}, \beta)) d\theta_d$$

LDA模型在图1中表示为概率图模型。如图所示，LDA表示有三个级别。参数α 和β 是语料级参数，假设在生成语料库的过程中被采样一次。变量θ_d是文档级变量，每个文档采样一次。最后，变量z_dn和w_dn是字级变量，并且对每个文档中的每个字进行一次采样。

经典聚类模型将涉及两级模型，其中为语料库对Dirichlet进行一次采样，对语料库中的每个文档选择一次多项式聚类变量，并且对于以聚类变量为条件的文档选择一组单词。与许多聚类模型一样，此类模型将文档限制为与单个主题相关联。另一方面，LDA涉及三个级别，特别是主题节点在文档中重复采样。在此模型下，文档可以与多个主题相关联。

3.1 LDA和可交换性

De Finetti的表示定理指出，无限可交换的随机变量序列的联合分布就好像从某个分布中得出随机参数，然后所讨论的随机变量是以参数为条件独立同分布的。

在LDA中，我们假设单词是由主题（通过固定条件分布）生成的，并且这些主题在文档中是无限可交换的。根据de Finetti定理，单词和主题序列的概率必须具有以下形式：

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

其中θ 是多项主题的随机参数。我们在方程式(3)中的文档上通过边缘化主题变量并赋予θ一个Dirichlet分布获得LDA分布。

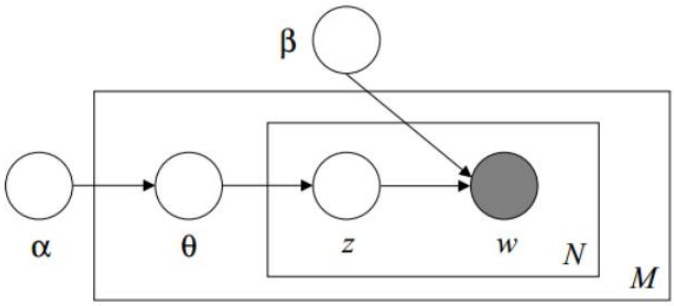

Figure 1: LDA的图形模型表示。盒子是代表重复的“板”。外板表示文档，而内板表示文档中主题和单词的重复选择。

3.2 一元连续混合

图1中所示的LDA模型比经典的分层贝叶斯文献中经常研究的两级模型更精细。然而，通过边缘化隐藏的主题变量z，我们可以将LDA理解为两级模型。形成词概率分布$p(w|\theta,\beta)$：

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

定义文档W的生成过程如下：

1.选择$\theta\sim Dir(\alpha)$。

2.对于N个单词中的每个单词w_n：

(a)从$p(w_n|\theta,\beta)$中选择一个单词w_n。

这个过程将文档的边缘分布定义为连续混合分布：

$$p(W|\alpha,\beta)=\int p(\theta|\alpha)(\prod_{n=1}^{N}(w_n|\theta,\beta))d\theta,$$

其中$p(w_n|\theta,\beta)$为混合组件，$p(\theta|\alpha)$为混合权重。

4 与其他潜在变量模型的关系

4.1 一元模型

每个文档的词独立的从一个单独的多项式分部中抽取：

$$p(W)=\prod_{n=1}^{N}p(w_n)$$

4.2 混合一元模型(Mixture of unigrams)

给一元模型加入离散随机主题变量z，即得到混合一元模型。在混合一元模型下，每个文档的都是通过首先选择主题z然后根据条件多项式$p(w|z)$独立的生成N个单词的过程生成。

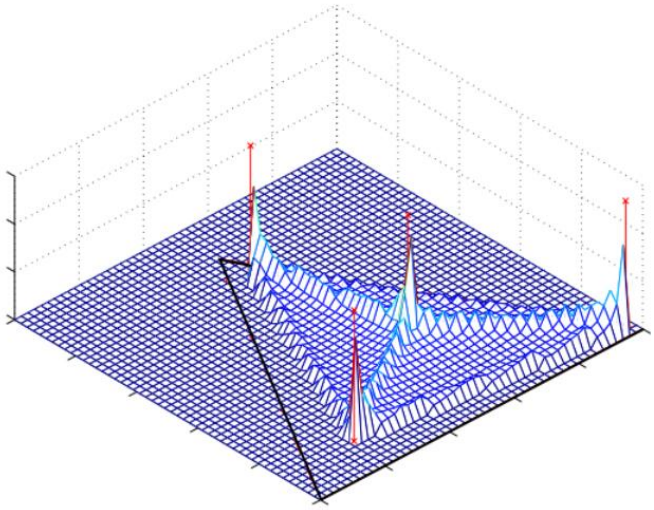

Figure 2: LDA下三个词和四个主题的一元分布$p(w|\theta,\beta)$的密度示例。嵌入在x-y平面中的三角形是2-D单形，表示三个词上的所有可能的多项式分布。三角形的每个顶点对应于确定性分布，将概率1分配给一个单词; 边缘的中点给出概率0.5到2个单词率; 并且三角形的质心是所有三个单词的均匀分布。用x标记的四个点是四个主题中的每个主题的多项分布$p(w|z)$的位置，并且在单形的顶部上显示的表面是LDA给出的(V-1)-单形（词的多项式分布）上的密度的示例。

文档的概率：

$$p(\mathbf{w}) = \sum_z p(z) \prod_{n=1}^{N} p\left(w_n|z\right)$$

当从语料库中估计时，在假设每个文档恰好展示一个主题的情况下，可以将单词分布视为主题的表示。这种假设过于局限，无法有效地对大量文档建模。

相比之下，LDA模型允许文档在不同程度上展示多个主题。这仅以一个额外参数为代价来实现：在混合一元模型存在与p（z）相关的k-1个参数，而第k个参数与LDA中的$p(|)$相关联。

4.3 概率潜在语义索引

pLSI[3] 假设文档d和词w_n条件独立于未观测到的主题z：

$$p\left(d, w_n\right) = p(d) \sum_z p\left(w_n|z\right) p(z|d)$$

pLSI模型试图放松在混合一元模型对每个文档仅从一个主题生成的假设。在某种意义上，它确实捕获了文档可能包含多个主题的可能性，因为$p(z|d)$用作特定文档d的主题的混合权重。要注意d是训练集中文档列表的虚拟索引。因此，d是具有与训练文档一样多的可能值的多项随机变量，并且模型仅针对训练它的那些文档学习主题混合$p(z|d)$。因此，pLSI不是一个明确定义的文档生成模型，它不能为先前看不见的文档分配概率。

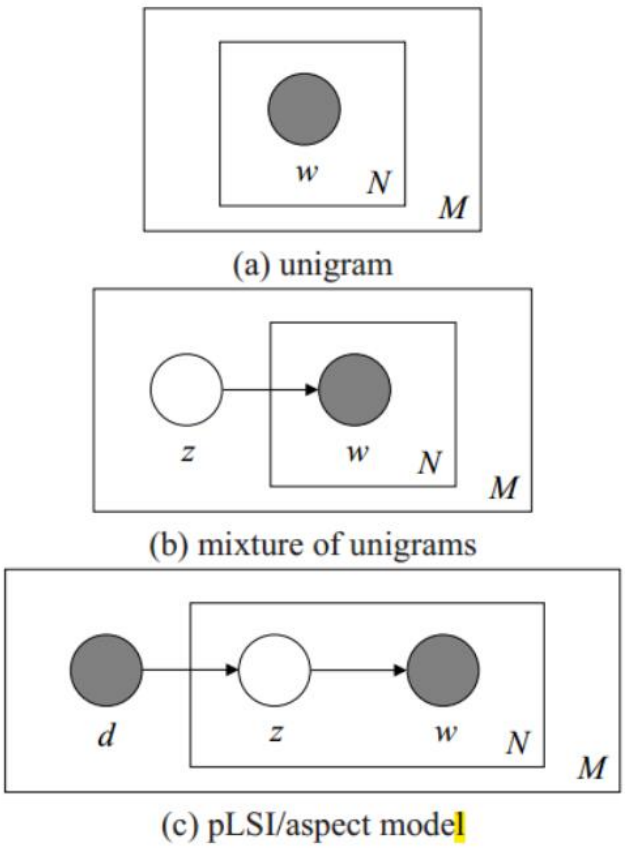

Figure 3: 不同模型在离散数据上的图形化模型表示

pLSI的另一个困难，也是源于使用由训练文档索引的分布，是需要估计的参数的数量随着训练文档的数量线性增长。k-主题pLSI模型的参数是k个大小为V的多项式分布和k个隐藏主题上M混合。这给出了kV + kM个参数，因此在M上线性增长。参数的线性增长表明该模型易于过拟合。在实践中，使用回火启发来平滑模型的参数以获得可接受的预测性能。然而，使用回火也会发生过度拟合。

LDA通过将主题混合权重视为k参数隐藏随机变量而不是明确链接到训练集的大量单个参数来克服这两个问题。LDA是一个定义明确的生成模型，可以很容易地推广到新文档。此外，k-主题LDA模型中的k + kV个参数不随训练语料库的大小而增长。

4.4 几何解释

说明LDA与其他潜在主题模型之间差异的一种方法是考虑潜在空间的几何形状，并观察文档在每个模型下如何在该几何结构中表示。

上面描述的所有四个模型——一元模型，混合一元模型，pLSI和LDA——在词分布空间中运行。每个这样的分布都可以看作(V-1)-simplex上的一个点，我们称之为单纯形式。

一元模型在单词单形上找到单个点，并假定语料库中的所有单词都来自相应的分布。潜在变量模型考虑单词单形上的k个点并基于这些点形成子单纯形，我们将其称为单形单元。请注意，主题单纯形的任何一点也是单词单纯形上的一个点。不同的潜变量模型以不同的方

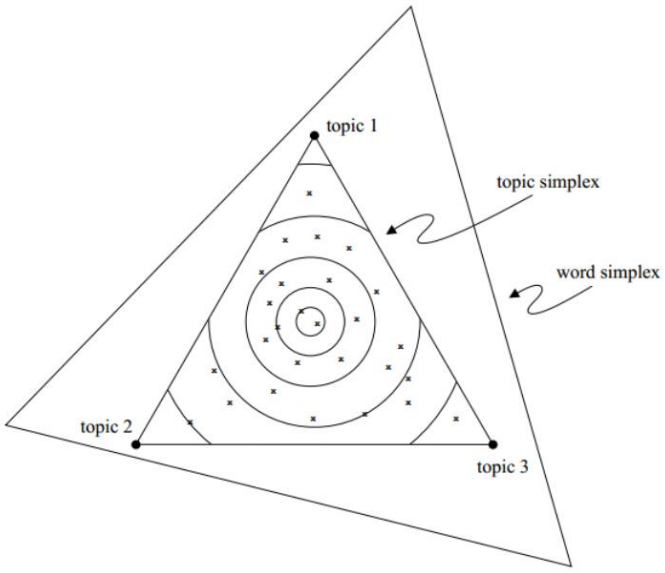

Figure 4: 主题单纯形为三个单词中嵌入单词simplex的三个主题。单词单纯形的角对应于三个分布，其中每个单词（分别）具有概率1。主题单纯形的三个点对应于三个不同的单词分布。混合一元模型将每个文档放在主题单一的角落之一。pLSI模型引起关于由x表示的主题单形的经验分布。LDA在由轮廓线表示的主题单纯形上进行平滑分布。

式使用主题单纯形来生成文档。

- 混合一元模型假设对于每个文档，单词单形上的k个点之一（即，单形主题的一个角）是随机选择的，文档的所有单词都是从对应与那个点的分布中提取的。
- pLSI模型假定训练文档的每个单词都来自随机选择的主题。这些主题本身来自特定文档的主题分布，即主题单形上的一个点。每个文档都有一个这样的分布;因此，训练文档集合定义了主题单形上的经验分布。
- LDA假定观察到的和未看到的文档中的每个单词都是由随机选择的主题生成的，该主题是从带有随机选择参数的分布中提取的。从主题单纯形的平滑分布中，对每个文档进行一次采样仪获取该参数。

5 推导与参数估计

5.1 推导

使用LDA需要解决的推导问题是计算给定文档下隐藏变量的后验分布(假设超参数已知)：

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

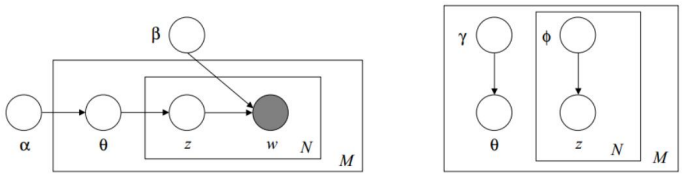

Figure 5: （左）LDA的图形化模型表示。（右）用于近似LDA后验的变分分布的图形模型表示。

后验分布不容易精确推导，因此可以考虑LDA的很多近似推导方法，包括拉普拉斯近似、变分近似、马尔科夫链蒙特卡洛算法。

5.2 变分推导

基于凸性的变分推理的基本思想是利用Jensen不等式来获得对数似然的可调下界。本质上，人们考虑一系列下界，由一组变分参数索引。变分参数由优化程序选择，该程序试图找到最紧密可能下界。

获得易处理的下界族的简单方法是考虑原始图形模型的简单修改，移除其中一些边和节点。考虑图5中的LDA模型。θ，z和w之间的带来的连接问题。通过丢弃这些边和w节点，并赋予所得到的简化图形模型以自由变分参数，就可以得到潜在变量的一系列分布。该族的以如下变量分布为特征：

$$q(\theta, \mathbf{z}|\gamma, \phi) = q(\theta|\gamma) \prod_{n=1}^{N} q\left(z_n|\phi_n\right) \tag{4}$$

其中狄利克雷参数γ和多项式参数$(\phi_1, \ldots, \phi_N)$为自由变量参数。

指定了简化的概率分布族后，下一步是设置一个确定变分参数γ和ϕ的值的优化问题。在对数似然上找到紧密下界的需求直接转化为以下优化问题：

$$(\gamma^*, \phi^*) = \arg\min_{(\gamma,\phi)} \mathrm{D}(q(\theta, \mathbf{z}|\gamma, \phi)\|p(\theta, \mathbf{z}|\mathbf{w}, \alpha, \beta)) \tag{5}$$

通过最小化变分分布和真实后验$p(\theta, z|w, \alpha, \beta)$之间的KullbackLeibler（KL）散度来找到变分参数的优化值。这种最小化可以通过迭代定点方法实现。通过计算KL散度的导数并将它们设置为零，我们得到以下更新方程：

$$\phi_{ni} \propto \beta_{iw_n} \exp\left\{\mathrm{E}_q\left[\log\left(\theta_i\right)|\gamma\right]\right\} \tag{6}$$

$$\gamma_i = \alpha_i + \sum_{n=1}^{N} \phi_{ni} \tag{7}$$

多项式更新的期望可以按照如下方式计算：

$$\mathrm{E}_q\left[\log\left(\theta_i\right)|\gamma\right] = \Psi\left(\gamma_i\right) - \Psi\left(\sum_{j=1}^{k} \gamma_j\right) \tag{8}$$

其中，Ψ是$log\Gamma$函数的一阶导数，可通过泰勒近似计算得出。

公式(6)和(7)具有吸引人的直观解释。Dirichlet更新是一个后验Dirichlet，给出了在变分分布$E[z_n|\phi_n]$下的预期观测值。多项式更新类似于使用贝叶斯定理$p(z_n|w_n) \propto p(w_n|z_n)p(z_n)$，其中$p(z_n)$通过变分分布下其对数的期望值的指数来近似。

变分分布实际上是一个条件分布，随w的变化而变化。这是因为方程式(5)中的优化问题在固定w情况下进行，从而产生的优化参数(γ^*, ϕ^*)是W的函数。我们可以将得到的变分分布写为$q(\theta, z|\gamma^*(W), \phi^*(W))$。因此，变分分布可以看作是后验分布$p(\theta, z|W, \alpha, \beta)$的近似值。

在文本语言中，优化参数$(\gamma^*(w), \phi^*(w))$是特定于文档的。我们将Dirichlet参数$\gamma^*(w)$视为提供文档在主题单纯形中的表示。

5.3 参数估计

本节给出LDA模型参数估计的经验贝叶斯方法。给定语料$D = W_1, \ldots, W_M$，我们期望找到最大化数据(边缘)对数似然的参数α和β：

$$\mathscr{L}(\alpha, \beta) = \sum_{d=1}^{M} logp(W_d|\alpha, \beta).$$

如上所述，数量$p(W|\alpha, \beta)$不易计算。然而，变分推断为我们提供了对数似然的易处理的下界，这是我们可以最大化关于α和β的界限。因此，我们可以通过交替变分EM程序找到LDA模型的近似经验贝叶斯估计，该程序最大化关于变分参数γ和ϕ的下界，然后，对于变分参数的固定值，最大化关于模型参数α和β的下界。

LDA变分EM算法的推导产生以下迭代算法：

1.（E-step）对于每个文档，找到变分参数$\gamma_d^*, \phi_d^* : d \in D$的优化值。

2.（M-step）最大化对数似然关于模型参数α和β的结果下限。这对应于找到最大似然估计，其中在E步骤中计算的近似后验下的每个文档具有预期的足够统计量。

重复这两个步骤直到对数似然的下限收敛。

条件多项式参数的M步更新可以分析写出：

$$\beta_{i,j} \propto \sum_{d=1}^{M} \sum_{n=1}^{N_d} \phi_{dni}^* w_{dn}^j. \tag{9}$$

Dirichlet参数的M步更新可以使用高效的Newton-Raphson方法实现，其中Hessian在线性时间内反转。

5.4 平滑

文档语料库的大词汇量造成严重的稀疏性问题。新文档很可能包含未出现在训练语料库中的任何文档中的单词。多项式参数的最大似然估计为这些单词分配零概率，因此对新文档赋予零概率。解决这个问题的标准方法是“平滑”多项式参数，为所有词汇项目分配正概

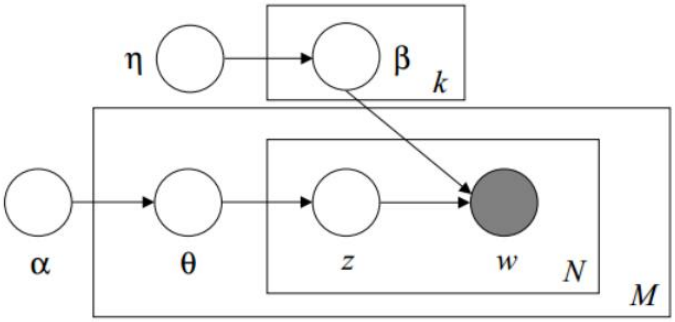

Figure 6: 平滑LDA模型的图形化表示

率，无论它们是否在训练集中被观察到。拉普拉斯平滑是常用平滑方式;这本质上在多项式参数上产生均匀Dirichlet先验下的后验分布的平均值。

然而，在混合模型设置中，简单的拉普拉斯平滑不再是最大后验方法。实际上，通过在多项式参数上放置Dirichlet先验，我们在混合模型设置中获得了难以处理的后验，这与在基本LDA模型中获得难以处理的后验的原因大致相同。我们提出将变分推理方法应用于包含多项式参数上的Dirichlet平滑的扩展模型。

在LDA设置中，我们获得了图6中所示的扩展图形模型。我们将β看作$K \times V$随机矩阵（每个混合分量为一行），我们假设每行是从可交换的Dirichlet分布中独立抽取的。我们现在扩展我们的推理程序，将β_i视为以数据为条件赋予后验分布的随机变量。因此，我们超越了5.3节的经验贝叶斯程序，并考虑采用更完整的贝叶斯方法来研究LDA。我们考虑采用贝叶斯推理的变分方法，在随机变量β，θ和z上放置可分离的分布：

$$q(\beta_{1:k}, z_{1:M}|\lambda\phi, \gamma) = \prod_{i=1}^{k} Dir(\beta_i|\lambda_i) \prod_{d=1}^{M} q_d(\theta_d, z_d|\phi_d, \gamma_d),$$

其中$q_d(\theta, z|\phi, \gamma)$是在方程式(4)中为LDA定义的变分分布。得到的变分推理过程产生方程式(6)和(7)分别作为变分参数γ和ϕ的更新方程式，以及新变分参数λ的附加更新：

$$\lambda_{ij} = \eta + \sum_{d=1}^{M} \sum_{n=1}^{N_d} \phi^*_{dni} w^j_{dn}.$$

将这些方程迭代到收敛会在β，θ和z上产生近似的后验分布。

我们现在留下了可交换Dirichlet的超参数η，以及之前的超参数α。我们设置这些超参数的方法又是（近似的）经验贝叶斯- 我们使用变分EM来基于边际似然找到这些参数的最大似然估计。

6 实例

7 应用与经验结果

7.1 文档建模

习惯在语言建模中使用的困惑困惑度在测试数据的似然中单调递减，并且代数上等效于几何平均每个词似然的倒数。较低的困惑得分表示更好的泛化性能。对于M个文档的测试集，困惑度的形式化表示为：

$$perplexity(D_{test}) = exp\{-\frac{\sum_{d=1}^{M} logp(W_d)}{\sum_{d=1}^{M} N_d}\}$$

8 讨论

References

[1] G. Salton and M. McGill, editors. Introduction to Modern Information Retrieval. McGraw-Hill, 1983.

[2] S. Deerwester, S. Dumais, T. Landauer, G. Furnas, and R. Harshman. Indexing by latent semantic analysis. Journal of the American Society of Information Science, 41(6):391–407, 1990.

[3] T. Hofmann. Probabilistic latent semantic indexing. Proceedings of the Twenty-Second Annual International SIGIR Conference, 1999.