清華大學
Tsinghua University

第二十章
潜在狄利克雷分配

潜在狄利克雷分配

- 潜在狄利克雷分配(latent Dirichlet allocation, LDA)，作为基于贝叶斯学习的话题模型，是潜在语义分析、概率潜在语义分析的扩展，

- LDA 在文本数据挖掘、图像处理、生物信息处理等领域被广泛使用

潜在狄利克雷分配

- LDA模型是文本集合的生成概率模型
- 假设每个文本由话题的一个多项分布表示，每个话题由单词的一个多项分布表示
- 特别假设文本的话题分布的先验分布是狄利克雷分布，话题的单词分布的先验分布也是狄利克雷分布
- 先验分布的导入使LDA 能够更好地应对话题模型学习中的过拟合现象

潜在狄利克雷分配

- LDA的文本集合的生成过程如下:

- 首先随机生成一个文本的话题分布
- 之后在该文本的每个位置，依据该文本的话题分布随机生成一个话题
- 然后在该位置依据该话题的单词分布随机生成一个单词，直至文本的最后一个位置，生成整个文本。
- 重复以 上过程生成所有文本。

潜在狄利克雷分配

- LDA模型是含有隐变量的概率图模型
- 模型中，每个话题的单词分布，每个文本的话题分布，文本的每个位置的话题是隐变量
- 文本的每个位置的单词是观测变量
- LDA模型的学习与推理无法直接求解，通常使用吉布斯抽样（Gibbs sampling）和变分EM算法（variational EM algorithm），前者是蒙特卡罗法，而后者是近似算法。

狄利克雷分布

分布定义

- 1. 多项分布

- 多项分布（multinomial distribution)是一种多元离散随机变量的概率分布，是二项分布（binomial distribution）的扩展。

- 假设重复进行n次独立随机试验，每次试验可能出现的结果有k种，第i种结果出现的概率为p_i，第i种结果出现的次数为n_i
- 如果用随机变量 $X = (X_1, X_2, \cdots, X_k)$ 表示试验所有可能结果的次数，其中X_i表示第i种结果出现的次数，那么随机变量x服从多项分布

分布定义

定义 20.1（多项分布） 若多元离散随机变量 $X = (X_1, X_2, \cdots, X_k)$ 的概率质量函数为

$$P(X_1 = n_1, X_2 = n_2, \cdots, X_k = n_k) = \frac{n!}{n_1! n_2! \cdots n_k!} p_1^{n_1} p_2^{n_2} \cdots p_k^{n_k}$$

$$= \frac{n!}{\prod_{i=1}^{k} n_i!} \prod_{i=1}^{k} p_i^{n_i} \tag{20.1}$$

其中 $p = (p_1, p_2, \cdots, p_k)$，$p_i \geqslant 0, i = 1, 2, \cdots, k$，$\sum_{i=1}^{k} p_i = 1$，$\sum_{i=1}^{k} n_i = n$，则称随机变量 X 服从参数为 (n, p) 的多项分布，记作 $X \sim \mathrm{Mult}(n, p)$。

- 当试验的次数n为1时，多项分布变成类别分布（categorical distribution)
- 类别分布表示试验可能出现的k种结果的概率

分布定义

- 2. 狄利克雷分布

- 狄利克雷分布（Dirichlet distribution)是一种多元连续随机变量的概率分布，是贝塔分布（(beta distribution)的扩展

- 在贝叶斯学习中，狄利克雷分布常作为多项分布的先验分布使用

分布定义

定义 20.2（狄利克雷分布） 若多元连续随机变量 $\theta = (\theta_1, \theta_2, \cdots, \theta_k)$ 的概率密度函数为

$$p(\theta|\alpha) = \frac{\Gamma\left(\sum_{i=1}^{k} \alpha_i\right)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \prod_{i=1}^{k} \theta_i^{\alpha_i - 1} \tag{20.2}$$

其中 $\sum_{i=1}^{k} \theta_i = 1$，$\theta_i \geqslant 0$，$\alpha = (\alpha_1, \alpha_2, \cdots, \alpha_k)$，$\alpha_i > 0$，$i = 1, 2, \cdots, k$，则称随机变量 θ 服从参数为 α 的狄利克雷分布，记作 $\theta \sim \mathrm{Dir}(\alpha)$。

分布定义

- 式中 $\Gamma(s)$ 是伽马函数，定义为

$$\Gamma(s)=\int_0^{\infty} x^{s-1}\mathrm{e}^{-x}\mathrm{d}x, \quad s>0$$

- 具有性质

$$\Gamma(s+1)=s\Gamma(s)$$

- 当s是自然数时，有

$$\Gamma(s+1)=s!$$

分布定义

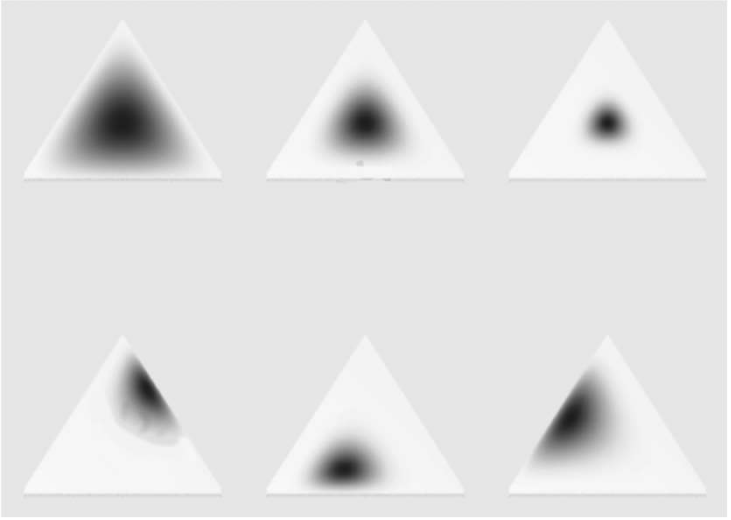
- 由于满足条件 $\theta_i \geqslant 0, \quad \sum_{i=1}^{k} \theta_i = 1$
- 所以狄利克雷分布 θ 存在于 (k1)维单纯形上
- 右图为二维单纯形上的狄利克雷分布

$$\theta_1 + \theta_2 + \theta_3 = 1,\ \theta_1, \theta_2, \theta_3 \geqslant 0$$

- 狄利克雷分布的参数为

$\alpha = (3,3,3), \alpha = (7,7,7), \alpha = (20,20,20),$

$\alpha = (2,6,11), \alpha = (14,9,5), \alpha = (6,2,6).$



分布定义

- 令

$$\mathrm{B}(\alpha) = \frac{\prod_{i=1}^{k}\Gamma(\alpha_i)}{\Gamma\left(\sum_{i=1}^{k}\alpha_i\right)}$$

- 则狄利克雷分布的密度函数可以写成

$$p(\theta|\alpha) = \frac{1}{\mathrm{B}(\alpha)}\prod_{i=1}^{k}\theta_i^{\alpha_i-1}$$

- $\mathrm{B}(\alpha)$ 是规范化因子，称为多元贝塔函数（或扩展的贝塔函数）

分布定义

- 由密度函数的性质

$$\int \frac{\Gamma\left(\sum_{i=1}^{k}\alpha_i\right)}{\prod_{i=1}^{k}\Gamma(\alpha_i)}\prod_{i=1}^{k}\theta_i^{\alpha_i-1}\mathrm{d}\theta = \frac{\Gamma\left(\sum_{i=1}^{k}\alpha_i\right)}{\prod_{i=1}^{k}\Gamma(\alpha_i)}\int\prod_{i=1}^{k}\theta_i^{\alpha_i-1}\mathrm{d}\theta = 1$$

- 得

$$\mathrm{B}(\alpha) = \int\prod_{i=1}^{k}\theta_i^{\alpha_i-1}\mathrm{d}\theta$$

- 即多元贝塔函数的积分表示

分布定义

- 3. 二项分布和贝塔分布

- 二项分布是多项分布的特殊情况，贝塔分布是狄利克雷分布的特殊情况
- 二项分布是指如下概率分布。X为离散随机变量，取值为m，其概率质量函数为

$$P(X=m)=\binom{n}{m}p^m(1-p)^{n-m},\quad m=0,1,2,\cdots,n$$

- 其中n和p (0≤p≤1) 是参数

分布定义

- 贝塔分布是指如下概率分布，X为连续随机变量，取值范围为[[0,1]，其概率密度函数为

$$p(x)=\begin{cases}\dfrac{1}{\mathrm{B}(s,t)}x^{s-1}(1-x)^{t-1}, & 0\leqslant x\leqslant 1\\ 0, & \text{其他}\end{cases}$$

- 其中s>0和t>0是参数，$\mathrm{B}(s,t)=\dfrac{\Gamma(s)\Gamma(t)}{\Gamma(s+t)}$ 是贝塔函数，定义为

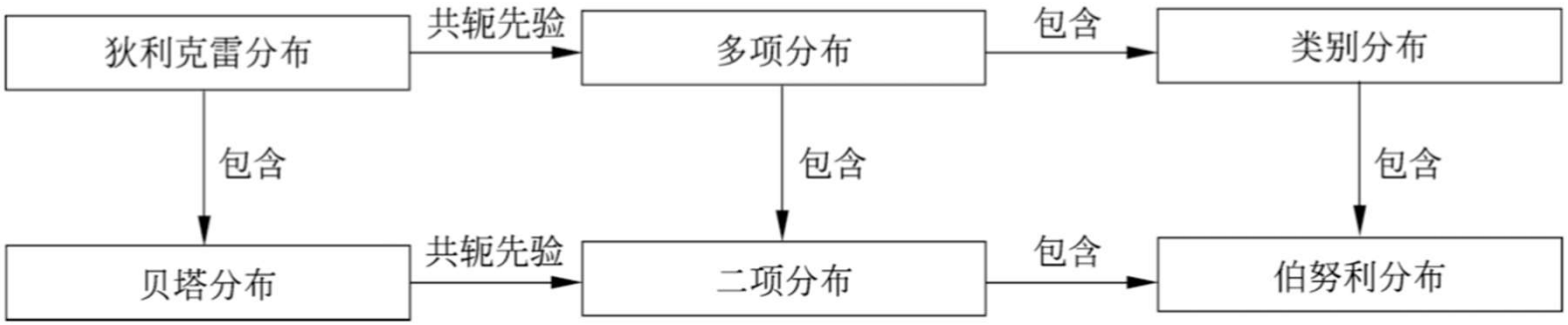
$$\mathrm{B}(s,t)=\int_0^1 x^{s-1}(1-x)^{t-1}\mathrm{d}x$$

- 当然s, t是自然数时，$\mathrm{B}(s,t)=\dfrac{(s-1)!(t-1)!}{(s+t-1)!}$

分布定义

- 当n为1时，二项分布变成伯努利分布（Bernoulli distribution）或0-1分布
- 伯努利分布表示试验可能出现的2种结果的概率

- 下图给出几种概率分布的关系。



共扼先验

- 狄利克雷分布有一些重要性质:
- (1)狄利克雷分布属于指数分布族
- (2)狄利克雷 分布是多项分布的共扼先验（conjugate prior）

共扼先验

- 贝叶斯学习中常使用共扼分布

- 如果后验分布与先验分布属于同类，则先验分布与后验分布称为共扼分布（conjugate distributions），先验分布称为共扼先验（conjugate prior）
- 如果多项分布的先验分布是狄利克雷分布，则其后验分布也为狄利克雷分布，两者构成共扼分布
- 作为先验分布的狄利克雷分布的参数又称为超参数

- 使用共扼分布的好处是便于从先验分布计算后验分布

共扼先验

- 设 $\mathcal{W} = \{w_1, w_2, \cdots, w_k\}$ 是由k个元素组成的集合。随机变量X服从W上的多项分布， $X \sim \text{Mult}(n, \theta)$ ， 其中 $n = (n_1, n_2, \cdots, n_k)$ 和 $\theta = (\theta_1, \theta_2, \cdots, \theta_k)$ 是参数

- 参数n为从W中重复独立抽取样本的次数， n_i为样本中w_i出现的次数 （i θ_i : 1,2,⋯,k)

- 参数 为 w_i 出现的概率 （i = 1,2,⋯,k)

共扼先验

- 将样本数据表示为D，目标是计算在样本数据D给定条件下参数 θ 的后验概率 $p(\theta|D)$ 。
- 对于给定的样本数据D，似然函数是

$$p(D|\theta) = \theta_1^{n_1}\theta_2^{n_2}\cdots\theta_k^{n_k} = \prod_{i=1}^{k}\theta_i^{n_i}$$

- 假设随机变量 θ 服从狄利克雷分布 $p(\theta|\alpha)$ ，其中 $\alpha = (\alpha_1, \alpha_2, \cdots, \alpha_k)$ 为参数。则 θ 的先验分布为

$$p(\theta|\alpha) = \frac{\Gamma\left(\sum_{i=1}^{k}\alpha_i\right)}{\prod_{i=1}^{k}\Gamma(\alpha_i)}\prod_{i=1}^{k}\theta_i^{\alpha_i-1} = \frac{1}{\mathrm{B}(\alpha)}\prod_{i=1}^{k}\theta_i^{\alpha_i-1} = \mathrm{Dir}(\theta|\alpha), \quad \alpha_i > 0$$

共扼先验

- 根据贝叶斯规则，在给定样本数据D和参数α条件下，θ 的后验概率分布是

$$
\begin{aligned}
p(\theta|D,\alpha) &= \frac{p(D|\theta)p(\theta|\alpha)}{p(D|\alpha)} \\
&= \frac{\prod_{i=1}^{k}\theta_i^{n_i}\frac{1}{\mathrm{B}(\alpha)}\theta_i^{\alpha_i-1}}{\int\prod_{i=1}^{k}\theta_i^{n_i}\frac{1}{\mathrm{B}(\alpha)}\theta_i^{\alpha_i-1}\mathrm{d}\theta} \\
&= \frac{1}{\mathrm{B}(\alpha+n)}\prod_{i=1}^{k}\theta_i^{\alpha_i+n_i-1} \\
&= \mathrm{Dir}(\theta|\alpha+n)
\end{aligned}
$$

共扼先验

- 可以看出先验分布和后验分布都是狄利克雷分布
- 两者有不同的参数，所以狄利克雷分布是多项分布的共扼先验
- 狄利克雷后验分布的参数等于狄利克雷先验分布参数 $\alpha = (\alpha_1, \alpha_2, \cdots, \alpha_k)$ 加上多项分布的观测 $n = (n_1, n_2, \cdots, n_k)$ ，好像试验之前就已经观察到计数 $\alpha = (\alpha_1, \alpha_2, \cdots, \alpha_k)$ ，因此也把α叫做先验伪计数（prior pseudo-counts)。

潜在狄利克雷分配模型

基本想法

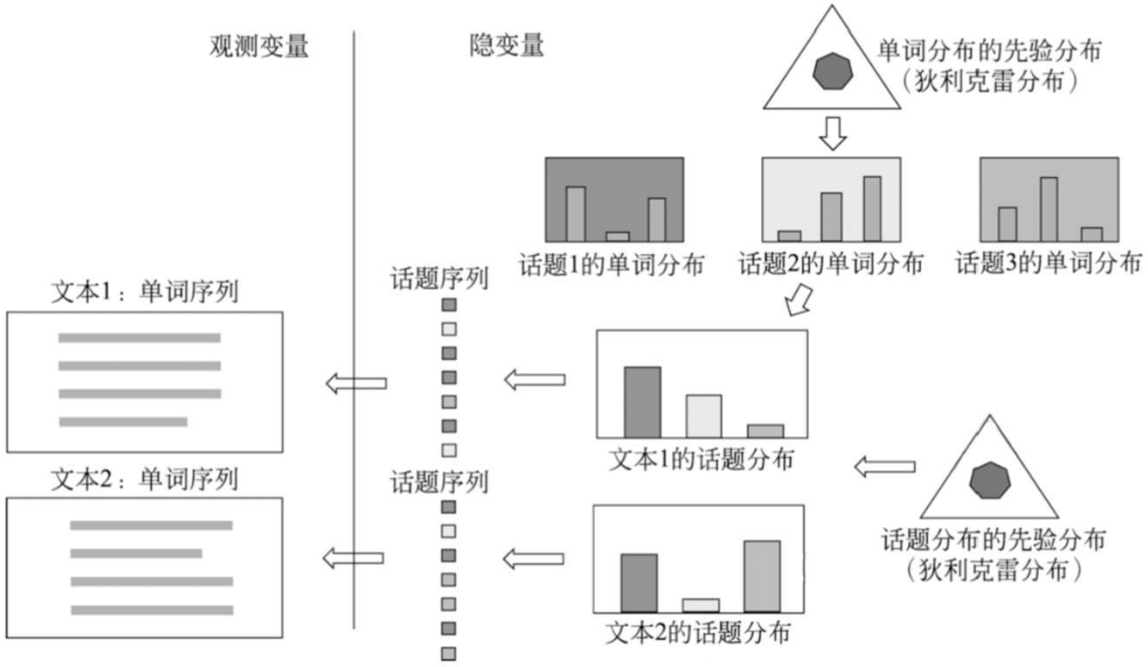
- 潜在狄利克雷分配（LDA）是文本集合的生成概率模型
- 模型假设话题由单词的多项分布表示，文本由话题的多项分布表示，单词分布和话题分布的先验分布都是狄利克雷分布
- 文本内容的不同是由于它们的话题分布不同

基本想法

- LDA模型表示文本集合的自动生成过程:
- 首先，基于单词分布的先验分布（狄利克雷分布）生成多个单词分布，即决定多个话题内容
- 之后，基于话题分布的先验分布（狄利克雷分布）生成多个话题分布，即决定多个文本内容
- 然后，基于每一个话题分布生成话题序列，针对每一个话题，基于话题的单词分布生成单词，整体构成一个单词序列，即生成文本
- 重复这个过程生成所有文本

基本想法

- 文本的单词序列是观测变量，文本的话题序列是隐变量，文本的话题分布和话题的单词分布也是隐变量。



基本想法

- LDA模型是概率图模型，其特点是以狄利克雷分布为多项分布的先验分布
- 学习就是给定文本集合，通过后验概率分布的估计，推断模型的所有参数
- 利用LDA进行 话题分析，就是对给定文本集合，学习到每个文本的话题分布，以及每个话题的单词分布。

基本想法

- 可以认为LDA是PLSA（概率潜在语义分析）的扩展
- 相同点是两者都假设话题是单词的多项分布，文本是话题的多项分布
- 不同点是LDA使用狄利克雷分布作为先验分布，而PLSA不使用先验分布（或者说假设先验分布是均匀分布）
- 学习过程LDA基于贝叶斯学习，而PLSA基于极大似然估计
- LDA的优点是，使用先验概率分布，可以防止学习过程中产生的过拟合（over-fitting）

模型定义

- 1. 模型要素

- 潜在狄利克雷分配（LDA）使用三个集合:
- 单词集合 $W = \{w_1, \cdots, w_v, \cdots, w_V\}$
- 文本集合 $D = \{\mathbf{w}_1, \cdots, \mathbf{w}_m, \cdots, \mathbf{w}_M\}$， 其中 w_m 是一个单词序列 $\mathbf{w}_m = (w_{m1}, \cdots, w_{mn}, \cdots, w_{mN_m})$
- 话题集合 $Z = \{z_1, \cdots, z_k, \cdots, z_K\}$

基本想法

- 每一个话题 z_k 由一个单词的条件概率分布 $p(w|z_k)$ 决定
- 分布 $p(w|z_k)$ 服从多项分布（严格意义上类别分布），其参数为 φ_k
- 参数 φ_k 服从狄利克雷分布（先验分布），其超参数为 β。
- 参数 φ_k 是一个V维向量 $\varphi_k = (\varphi_{k1}, \varphi_{k2}, \cdots, \varphi_{kV})$，其中 φ_{kv} 表示话题 z_k 生成单词 w_v 的概率
- 所有话题的参数向量构成一个 K x V 矩阵 $\varphi = \{\varphi_k\}_{k=1}^{K}$。
- 超参数 β 也是一个V维向量 $\beta = (\beta_1, \beta_2, \cdots, \beta_V)$.

基本想法

- 每一个文本 w_m 由一个话题的条件概率分布 $p(z|w_m)$ 决定
- 分布 $p(z|w_m)$ 服从多项分布（严格意义上类别分布），其参数为 θ_m
- 参数 θ_m 服从狄利克雷分布（先验分布），其超参数为 α
- 参数 θ_m 是一个K维向量 $\theta_m = (\theta_{m1}, \theta_{m2}, \cdots, \theta_{mK})$，其中 θ_{mk} 表示文本 w_m 生成话题 z_k 的概率
- 所有文本的参数向量构成一个 M x K 矩阵 $\boldsymbol{\theta} = \{\theta_m\}_{m=1}^{M}$
- 超参数 α 也是一个K维向量 $\alpha = (\alpha_1, \alpha_2, \cdots, \alpha_K)$
- 每一个文本 w_m 中的每一个单词 w_{mn} 由该文本的话题分布 $p(z|w_m)$ 以及所有话题的单词分布 $p(w|z_k)$ 决定

基本想法

- 2. 生成过程

- LDA文本集合的生成过程如下:

- 给定单词集合W，文本集合D，话题集合Z，狄利克雷分布的超参数 α 和 β

基本想法

- （1）生成话题的单词分布
- 随机生成K个话题的单词分布
- 按照狄利克雷分布Dir(β) 随机生成一个参数向量 φ_k，$\varphi_k \sim \mathrm{Dir}(\beta)$，作为话题 z_k 的单词分布 $p(w|z_k)$

- （2）生成文本的话题分布
- 随机生成M个文本的话题分布
- 按照狄利克雷分布Dir(α) 随机生成一个参数向量 θ_m，$\theta_m \sim \mathrm{Dir}(\alpha)$，作为文本 w_m 的话题分布 $p(z|w_m)$

基本想法

- （3）生成文本的单词序列
- 随机生成M个文本的N_m个单词
- 首先按照多项分布 $\mathrm{Mult}(\theta_m)$ 随机生成一个话题 z_{mn}, $z_{mn} \sim \mathrm{Mult}(\theta_m)$
- 然后按照多项分布 $\mathrm{Mult}(\varphi_{z_{mn}})$ 随机生成一个单词 w_{mn}， $w_{mn} \sim \mathrm{Mult}(\varphi_{z_{mn}})$

- 文本 w_m 本身是单词序列 $\mathbf{w}_m = (w_{m1}, w_{m2}, \cdots, w_{mN_m})$ ， 对应着隐式的话题序列 $\mathbf{z}_m = (z_{m1}, z_{m2}, \cdots, z_{mN_m})$。

LDA的文本生成算法

算法 20.1（LDA 的文本生成算法）

（1）对于话题 z_k（$k = 1, 2, \cdots, K$）：

生成多项分布参数 $\varphi_k \sim \mathrm{Dir}(\beta)$，作为话题的单词分布 $p(w|z_k)$；

（2）对于文本 $\mathbf{w}_m$（$m = 1, 2, \cdots, M$）：

生成多项分布参数 $\theta_m \sim \mathrm{Dir}(\alpha)$，作为文本的话题分布 $p(z|\mathbf{w}_m)$；

（3）对于文本 $\mathbf{w}_m$ 的单词 w_{mn}（$m = 1, 2, \cdots, M$，$n = 1, 2, \cdots, N_m$）：

（a）生成话题 $z_{mn} \sim \mathrm{Mult}(\theta_m)$，作为单词对应的话题；

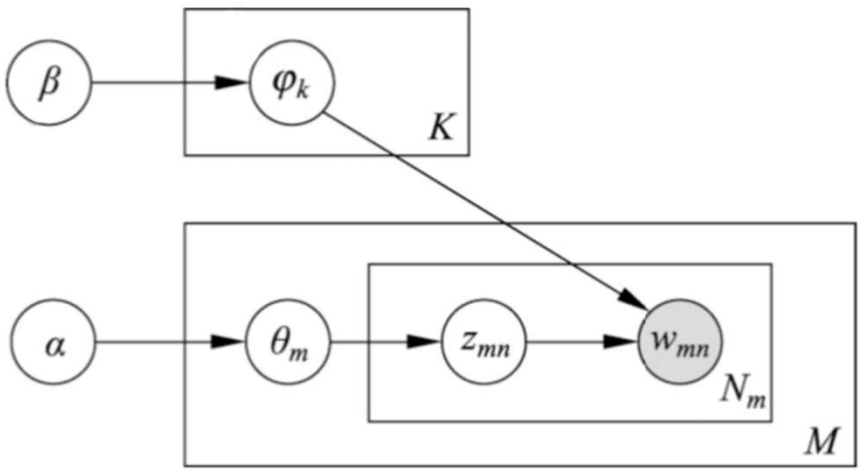
（b）生成单词 $w_{mn} \sim \mathrm{Mult}(\varphi_{z_{mn}})$。 ■

LDA的文本生成算法

- LDA的文本生成过程中，假定话题个数K给定，实际通常通过实验选定
- 狄利 克雷分布的超参数 α 和 β 通常也是事先给定的
- 在没有其他先验知识的情况下，可以假设向量 α 和 β 的所有分量均为1，这时的文本的话题分布 θ_m 是对称的，话题的单词分布 φ_k 也是对称的。

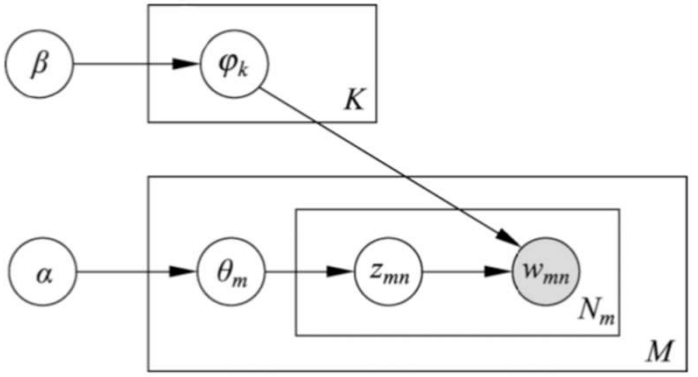
概率图模型

- LDA模型本质是一种概率图模型(probabilistic graphical model)
- 下图为 LDA作为概率图模型的板块表示（plate notation)



- 图中结点表示随机变量
- 实心结点是观测变量
- 空心结点是隐变量
- 有向边表示概率依存关系
- 矩形（板块）表示重复，板块内数字表示重复的次数。

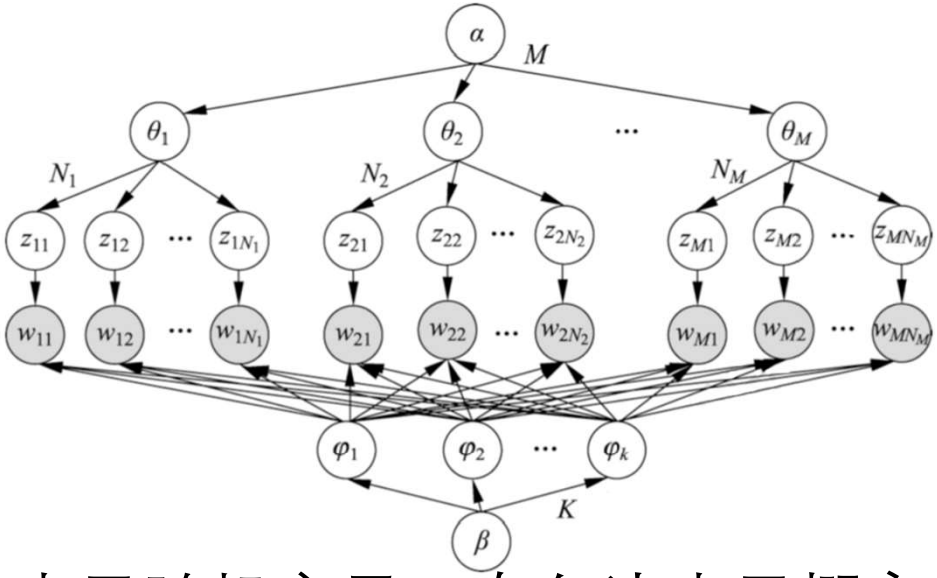
概率图模型



- 图中LDA板块表示，结点 α 和 β 是模型的超参数
- 结点 φ_k 表示话题的单词分布的参数
- 结点 θ_m 表示文本的话题分布的参数
- 结点 z_{mn} 表示话题，结点 v_{mn} 表示单词
- 结点 β 指向结点 φ_k，重复K次，表示根据超参数 β 生成K个话题的单词分布的参数 φ_k
- 结点 α 指向结点 θ_m，重复M次，表示根据超参数 α 生成M个文本的话题分布的参数 θ_m
- 结点 θ_m 指向结点 z_{mn}，重复 N_m 次，表示根据文本的话题分布 θ_m 生成 N_m 个话题 z_{mn}
- 结点 z_{mn} 指向结点 w_{mn}，同时K个结点 φ_k 也指 向结点 w_{mn}，表示根据话题 z_{mn} 以及K个话题的单词分布 φ_k 生成单词 w_{mn} 。

概率图模型

- 板块表示的优点是简洁，板块表示展开之后，成为普通的有向图表示



- 有向图中结点表示随机变量，有向边表示概率依存关系。可以看出LDA是相同随机 变量被重复多次使用的概率图模型。

随机变量序列的可交换性

- 一个有限的随机变量序列是可交换的（exchangeable)，是指随机变量的联合概率 分布对随机变量的排列不变

$$P(x_1, x_2, \cdots, x_N) = P(x_{\pi(1)}, x_{\pi(2)}, \cdots, x_{\pi(N)})$$

- 这里 $\pi(1), \pi(2), \cdots, \pi(N)$ 代表自然数1,2,... ,N的任意一个排列。一个无限的随机变量序列是无限可交换（(infinitely exchangeable）的，是指它的任意一个有限子序列都是可交换的

- 如果一个随机变量序列 $X_1, X_2, \cdots, X_N, \cdots$ 是独立同分布的，那么它们是无限 可交换的。反之不然。

随机变量序列的可交换性

- 随机变量序列可交换的假设在贝叶斯学习中经常使用
- 根据De Finetti定理，任意一个无限可交换的随机变量序列对一个随机参数是条件独立同分布的
- 即任意一个 无限可交换的随机变量序列 $X_1, X_2, \cdots, X_i, \cdots$ 的基于一个随机参数Y的条件概率， 等于基于这个随机参数Y的各个随机变量 $X_1, X_2, \cdots, X_i, \cdots$ 的条件概率的乘积。

$$P(X_1, X_2, \cdots, X_i, \cdots | Y) = P(X_1|Y)P(X_2|Y)\cdots P(X_i|Y)\cdots$$

随机变量序列的可交换性

- LDA假设文本由无限可交换的话题序列组成
- 由De Finetti定理知，实际是假设文本中的话题对一个随机参数是条件独立同分布的
- 所以在参数给定的条件下，文本中的话题的顺序可以忽略
- 作为对比，概率潜在语义模型假设文本中的话题是独立同分布的，文本中的话题的顺序也可以忽略

概率公式

- LDA模型整体是由观测变量和隐变量组成的联合概率分布，可以表为

$$p(\mathbf{w},\mathbf{z},\theta,\varphi|\alpha,\beta)=\prod_{k=1}^{K}p(\varphi_k|\beta)\prod_{m=1}^{M}p(\theta_m|\alpha)\prod_{n=1}^{N_m}p(z_{mn}|\theta_m)p(w_{mn}|z_{mn},\varphi)$$

- 观测变量 w 表示所有文本中的单词序列
- 隐变量 z 表示所有文本中的话题序列
- 隐变量 θ 表示所有文本的话题分布的参数
- 隐变量 φ 表示所有话题的单词分布的参数
- α 和 β 是超参数

概率公式

- $p(\varphi_k|\beta)$ 表示超参数 β 给定条件下第k个话题的单词分布的参数 φ_k 的生成概率
- $p(\theta_m|\alpha)$ 表示超参数 α 给定条件下第m个文本的话题分布的参数 θ_m 的生成概率,
- $p(z_{mn}|\theta_m)$ 表示第m个文本的话题分布 θ_m 给定条件下文本的第n个位置的话题 z_{mn} 的生成概率
- $p(w_{mn}|z_{mn},\varphi)$ 表示在第m个文本的第n个位置的话题 z_{mn} 及所有话题的单词分布的参数 φ 给定条件下第m个文本的第n个位置的单词 w_{mn} 的生成概率

概率公式

- 第m个文本的联合概率分布可以表为

$$p(\mathbf{w}_m, \mathbf{z}_m, \theta_m, \varphi|\alpha, \beta) = \prod_{k=1}^{K} p(\varphi_k|\beta)p(\theta_m|\alpha) \prod_{n=1}^{N_m} p(z_{mn}|\theta_m)p(w_{mn}|z_{mn}, \varphi)$$

- 其中 w_m 表示该文本中的单词序列，z_m 表示该文本的话题序列，θ_m 表示该文本的话 题分布参数。

- LDA模型的联合分布含有隐变量，对隐变量进行积分得到边缘分布

概率公式

- 参数 θ_m 和 φ 给定条件下第m个文本的生成概率是

$$p(\mathbf{w}_m|\theta_m,\varphi)=\prod_{n=1}^{N_m}\left[\sum_{k=1}^{K}p(z_{mn}=k|\theta_m)p(w_{mn}|\varphi_k)\right]$$

- 超参数 α 和 β 给定条件下第m个文本的生成概率是

$$p(\mathbf{w}_m|\alpha,\beta)=\prod_{k=1}^{K}\int p(\varphi_k|\beta)\left[\int p(\theta_m|\alpha)\prod_{n=1}^{N_m}\left[\sum_{l=1}^{K}p(z_{mn}=l|\theta_m)p(w_{mn}|\varphi_l)\right]\mathrm{d}\theta_m\right]\mathrm{d}\varphi_k$$

- 超参数 α 和 β 给定条件下所有文本的生成概率是

$$p(\mathbf{w}|\alpha,\beta)=\prod_{k=1}^{K}\int p(\varphi_k|\beta)\left[\prod_{m=1}^{M}\int p(\theta_m|\alpha)\prod_{n=1}^{N_m}\left[\sum_{l=1}^{K}p(z_{mn}=l|\theta_m)p(w_{mn}|\varphi_l)\right]\mathrm{d}\theta_m\right]\mathrm{d}\varphi_k$$

清華大學
Tsinghua University

LDA的吉布斯抽样算法

LDA的吉布斯抽样算法

- 潜在狄利克雷分配（LDA）的学习（参数估计）是一个复杂的最优化问题，很难精确求解，只能近似求解

- 常用的近似求解方法有吉布斯抽样（Gibbs sampling）和变分 推理（variational inference)

基本想法

- LDA模型的学习，给定文本（单词序列）的集合 $D = \{\mathbf{w}_1, \cdots, \mathbf{w}_m, \cdots, \mathbf{w}_M\}$，

- 目标是要推断:
- (1) 话题序列的集合 $\mathbf{z} = \{\mathbf{z}_1, \cdots, \mathbf{z}_m, \cdots, \mathbf{z}_M\}$ 的后验概率分布
- (2) 参数 $\theta = \{\theta_1, \cdots, \theta_m, \cdots, \theta_M\}$，$\theta_m$ 是文本 w_m 的话题分布的参数
- (3) 参数 $\varphi = \{\varphi_1, \cdots, \varphi_k, \cdots, \varphi_K\}$，$\varphi_k$ 是话题 z_k 的单词分布的参数

- 也就是说，要对联合概率分布 $p(\mathbf{w}, \mathbf{z}, \theta, \varphi | \alpha, \beta)$ 进行估计
- 其中 w 是观测变量，而z，θ，φ 是隐变量。

基本想法

- 为了估计多元随机变量x的联合分布p(x)，吉布斯抽样法选择x的一个分量，固定其他分量，按照其条件概率分布进行随机抽样，依次循环对每一个分量执行这个操作，得到联合分布p(x)的一个随机样本，重复这个过程，在燃烧期之后，得到联合概率分布p(x)的 样本集合

- LDA模型的学习通常采用收缩的吉布斯抽样（collapsed Gibbs sampling）方法

基本想法

- 基本想法是，通过对隐变量 θ 和 φ 积分，得到边缘概率分布 $p(\mathbf{w}, \mathbf{z}|\alpha, \beta)$（也是联合分布）
- 其中变量w是可观测的，变量z是不可观测的
- 对后验概率分布 $p(\mathbf{z}|\mathbf{w}, \alpha, \beta)$ 进行吉布斯抽样，得到分布 $p(\mathbf{z}|\mathbf{w}, \alpha, \beta)$ 的样本集合
- 再利用这个样本集合对参数 θ 和 φ 进行估计，最终得到LDA模型 $p(\mathbf{w}, \mathbf{z}, \theta, \varphi|\alpha, \beta)$ 的所有参数估计

算法的主要部分

- 根据上面的分析，问题转化为对后验概率分布 $p(\mathbf{z}|\mathbf{w},\alpha,\beta)$ 的吉布斯抽样

- 该分布表示在所有文本的单词序列给定条件下所有可能话题序列的条件概率。

抽样分布的表达式

- 首先有关系

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

- 这里变量 w, α 和 β 已知，分母相同，可以不予考虑

- 联合分布域 $p(\mathbf{w},\mathbf{z}|\alpha,\beta)$ 的表达式 可以进一步分解为

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

- 两个因子可以分别处理

抽样分布的表达式

- 推导第一个因子 $p(\mathbf{w}|\mathbf{z},\beta)$ 的表达式。首先

$$p(\mathbf{w}|\mathbf{z},\varphi)=\prod_{k=1}^{K}\prod_{v=1}^{V}\varphi_{kv}^{n_{kv}}$$

- 其中 φ_{kv} 是第k个话题生成单词集合第v个单词的概率，n_{kv} 是数据中第k话题生成第v个单词的次数

抽样分布的表达式

- 于是

$$
\begin{aligned}
p(\mathbf{w}|\mathbf{z},\beta) &= \int p(\mathbf{w}|\mathbf{z},\varphi)p(\varphi|\beta)\mathrm{d}\varphi \\
&= \int \prod_{k=1}^{K} \frac{1}{\mathrm{B}(\beta)} \prod_{v=1}^{V} \varphi_{kv}^{n_{kv}+\beta_v-1}\mathrm{d}\varphi \\
&= \prod_{k=1}^{K} \frac{1}{\mathrm{B}(\beta)} \int \prod_{v=1}^{V} \varphi_{kv}^{n_{kv}+\beta_v-1}\mathrm{d}\varphi \\
&= \prod_{k=1}^{K} \frac{\mathrm{B}(n_k+\beta)}{\mathrm{B}(\beta)}
\end{aligned}
$$

- 其中 $n_k = \{n_{k1}, n_{k2}, \cdots, n_{kV}\}$

抽样分布的表达式

- 第二个因子 $p(\mathbf{z}|\alpha)$ 的表达式可以类似推导。首先

$$p(\mathbf{z}|\theta) = \prod_{m=1}^{M} \prod_{k=1}^{K} \theta_{mk}^{n_{mk}}$$

- 其中 θ_{mk} 是第m个文本生成第k个话题的概率，n_{mk} 是数据中第m个文本生成第k 个话题的次数。

抽样分布的表达式

- 于是

$$
\begin{aligned}
p(\mathbf{z}|\alpha) &= \int p(\mathbf{z}|\theta)p(\theta|\alpha)\mathrm{d}\theta \\
&= \int \prod_{m=1}^{M} \frac{1}{\mathrm{B}(\alpha)} \prod_{k=1}^{K} \theta_{mk}^{n_{mk}+\alpha_k-1}\mathrm{d}\theta \\
&= \prod_{m=1}^{M} \frac{1}{\mathrm{B}(\alpha)} \int \prod_{k=1}^{K} \theta_{mk}^{n_{mk}+\alpha_k-1}\mathrm{d}\theta \\
&= \prod_{m=1}^{M} \frac{\mathrm{B}(n_m+\alpha)}{\mathrm{B}(\alpha)}
\end{aligned}
$$

- 其中 $n_m = \{n_{m1}, n_{m2}, \cdots, n_{mK}\}$ ， 可得

$$
p(\mathbf{z}, \mathbf{w}|\alpha, \beta) = \prod_{k=1}^{K} \frac{\mathrm{B}(n_k+\beta)}{\mathrm{B}(\beta)} \cdot \prod_{m=1}^{M} \frac{\mathrm{B}(n_m+\alpha)}{\mathrm{B}(\alpha)}
$$

抽样分布的表达式

- 于是可得收缩的吉布斯抽样分布的公式

$$p(\mathbf{z}|\mathbf{w},\alpha,\beta) \propto \prod_{k=1}^{K} \frac{\mathrm{B}(n_k+\beta)}{\mathrm{B}(\beta)} \cdot \prod_{m=1}^{M} \frac{\mathrm{B}(n_m+\alpha)}{\mathrm{B}(\alpha)}$$

满条件分布的表达式

- 分布 $p(\mathbf{z}|\mathbf{w},\alpha,\beta)$ 的满条件分布可以写成

$$p(z_i|\mathbf{z}_{-i},\mathbf{w},\alpha,\beta)=\frac{1}{Z_{z_i}}p(\mathbf{z}|\mathbf{w},\alpha,\beta)$$

- 即在所有文本单词序列、其他位置话题序列给定条件下第i个位置的话题的条件概率分布。

- w_i 表示所有文本的单词序列的第i个位置的单词
- z_i 表示单词 w_i 对应的话 题
- Z_{z_i} 表示分布 $p(\mathbf{z}|\mathbf{w},\alpha,\beta)$ 对变量 z_i 的边缘化因子。

满条件分布的表达式

- 结合收缩的吉布斯抽样分布的公式，可以推出

$$p(z_i|\mathbf{z}_{-i},\mathbf{w},\alpha,\beta) \propto \frac{n_{kv}+\beta_v}{\sum_{v=1}^{V}(n_{kv}+\beta_v)} \cdot \frac{n_{mk}+\alpha_k}{\sum_{k=1}^{K}(n_{mk}+\alpha_k)}$$

- 第m个文本的第n个位置的单词 w_i 是单词集合的第v个单词
- 其话题 z_i 是话题合集的第k个话题
- n_{kv} 表示第k个话题中第n个单词的计数，但减去当前单词的计数
- n_{mk} 表示第m个文本中第k个话题的计数，但减去当前单词的话题的计数。

算法的后处理

- 通过吉布斯抽样得到的分布 $p(\mathbf{z}|\mathbf{w},\alpha,\beta)$ 的样本，可以得到变量z的分配值，也可以估计变量 θ 和 φ

- 1. 参数 $\theta=\{\theta_m\}$ 的估计
- 根据LDA模型的定义，后验概率满足

$$p(\theta_m|\mathbf{z}_m,\alpha)=\frac{1}{Z_{\theta_m}}\prod_{n=1}^{N_m}p(z_{mn}|\theta_m)p(\theta_m|\alpha)=\mathrm{Dir}(\theta_m|n_m+\alpha)$$

- 这里 $n_m=\{n_{m1},n_{m2},\cdots,n_{mK}\}$ 是第m个文本的话题的计数
- Z_{θ_m} 表示分布 $p(\theta_m,\mathbf{z}_m|\alpha)$ 对变量 θ_m 的边缘化因子

算法的后处理

- 于是得到参数 $\theta = \{\theta_m\}$ 的估计式

$$\theta_{mk} = \frac{n_{mk} + \alpha_k}{\sum_{k=1}^{K} (n_{mk} + \alpha_k)}, \quad m = 1, 2, \cdots, M; \quad k = 1, 2, \cdots, K$$

算法的后处理

- 2. 参数 $\varphi = \{\varphi_k\}$ 的估计

- 后验概率满足

$$p(\varphi_k|\mathbf{w},\mathbf{z},\beta) = \frac{1}{Z_{\varphi_k}} \prod_{i=1}^{I} p(w_i|\varphi_k)p(\varphi_k|\beta) = \mathrm{Dir}(\varphi_k|n_k+\beta)$$

- $n_k = \{n_{k1}, n_{k2}, \cdots, n_{kV}\}$ 是第k个话题的单词的计数
- Z_{φ_k} 表示分布 $p(\varphi_k, \mathbf{w}|\mathbf{z},\beta)$ 对变量 φ_k 的边缘化因子
- I是文本集合单词序列w的单词总数

算法的后处理

- 于是得到参数的估计式

$$\varphi_{kv} = \frac{n_{kv} + \beta_v}{\sum_{v=1}^{V} (n_{kv} + \beta_v)}, \quad k = 1, 2, \cdots, K; \quad v = 1, 2, \cdots, V$$

算法

- 对给定的所有文本的单词序列w，每个位置上随机指派一个话题，整体构成所有 文本的话题序列z。然后循环执行以下操作。
- 在每一个位置上计算在该位置上的话题的满条件概率分布，然后进行随机抽样，得到该位置的新的话题，分派给这个位置。

$$p(z_i|\mathbf{z}_{-i},\mathbf{w},\alpha,\beta) \propto \frac{n_{kv}+\beta_v}{\sum_{v=1}^{V}(n_{kv}+\beta_v)} \cdot \frac{n_{mk}+\alpha_k}{\sum_{k=1}^{K}(n_{mk}+\alpha_k)}$$

算法

- 这个条件概率分布由两个因子组成，第一个因子表示话题生成该位置的单词的概率，第二个因子表示该位置的文本生成话题的概率。
- 整体准备两个计数矩阵:

  话题-单词矩阵 $N_{K\times V} = [n_{kv}]$ 和文本—话题矩阵 $N_{M\times K} = [n_{mk}]$

- 在每一个位置，对两个矩阵中该位置的己有话题的计数减1，计 算满条件概率分布，然后进行抽样，得到该位置的新话题，之后对两个矩阵中该位置的新话题的计数加1。
- 计算移到下一个位置
- 在燃烧期之后得到的所有文本的话题序列就是条件概率分布 $p(\mathbf{z}|\mathbf{w},\alpha,\beta)$ 样本

LDA吉布斯抽样算法

输入：文本的单词序列 $\mathbf{w}=\{\mathbf{w}_1,\cdots,\mathbf{w}_m,\cdots,\mathbf{w}_M\}$，$\mathbf{w}_m=(w_{m1},\cdots,w_{mn},\cdots,w_{mN_m})$；

输出：文本的话题序列 $\mathbf{z}=\{\mathbf{z}_1,\cdots,\mathbf{z}_m,\cdots,\mathbf{z}_M\}$，$\mathbf{z}_m=(z_{m1},\cdots,z_{mn},\cdots,z_{mN_m})$ 的后验概率分布 $p(\mathbf{z}|\mathbf{w},\alpha,\beta)$ 的样本计数，模型的参数 φ 和 θ 的估计值；

参数：超参数 α 和 β，话题个数 K。

（1）设所有计数矩阵的元素 n_{mk}，n_{kv}，计数向量的元素 n_m，n_k 初值为 0；

（2）对所有文本 $\mathbf{w}_m$，$m=1,2,\cdots,M$

对第 m 个文本中的所有单词 w_{mn}，$n=1,2,\cdots,N_m$

（a）抽样话题 $z_{mn}=z_k\sim\text{Mult}\left(\frac{1}{K}\right)$；

增加文本-话题计数 $n_{mk}=n_{mk}+1$，

增加文本-话题和计数 $n_m=n_m+1$，

增加话题-单词计数 $n_{kv}=n_{kv}+1$，

增加话题-单词和计数 $n_k=n_k+1$；

LDA吉布斯抽样算法

（3）循环执行以下操作，直到进入燃烧期

对所有文本 $\mathbf{w}_m$，$m = 1, 2, \cdots, M$

对第 m 个文本中的所有单词 w_{mn}，$n = 1, 2, \cdots, N_m$

（a）当前的单词 w_{mn} 是第 v 个单词，话题指派 z_{mn} 是第 k 个话题；

减少计数 $n_{mk} = n_{mk} - 1$，$n_m = n_m - 1$，$n_{kv} = n_{kv} - 1$ ，$n_k = n_k - 1$；

（b）按照满条件分布进行抽样

$$p(z_i|\mathbf{z}_{-i}, \mathbf{w}, \alpha, \beta) \propto \frac{n_{kv} + \beta_v}{\sum_{v=1}^{V}(n_{kv} + \beta_v)} \cdot \frac{n_{mk} + \alpha_k}{\sum_{k=1}^{K}(n_{mk} + \alpha_k)}$$

得到新的第 k' 个话题，分配给 z_{mn}；

（c）增加计数 $n_{mk'} = n_{mk'} + 1$，$n_m = n_m + 1$，$n_{k'v} = n_{k'v} + 1$，$n_{k'} = n_{k'} + 1$；

（d）得到更新的两个计数矩阵 $N_{K\times V} = [n_{kv}]$ 和 $N_{M\times K} = [n_{mk}]$，表示后验概率分布 $p(\mathbf{z}|\mathbf{w}, \alpha, \beta)$ 的样本计数；

LDA吉布斯抽样算法

（4）利用得到的样本计数，计算模型参数

$$\theta_{mk} = \frac{n_{mk} + \alpha_k}{\sum_{k=1}^{K} (n_{mk} + \alpha_k)}$$

$$\varphi_{kv} = \frac{n_{kv} + \beta_v}{\sum_{v=1}^{V} (n_{kv} + \beta_v)}$$

清華大學
Tsinghua University

LDA的变分EM算法

变分推理

- 变分推理（variational inference)是贝叶斯学习中常用的、含有隐变量模型的学习和推理方法
- 变分推理和马尔可夫链蒙特卡罗法（MCMC）属于不同的技巧
- MCMC 通过随机抽样的方法近似地计算模型的后验概率，变分推理则通过解析的方法计算模 型的后验概率的近似值

变分推理

- 变分推理的基本想法如下
- 假设模型是联合概率分布 p(x,z)，其中x是观测变 量（数据），z是隐变量，包括参数
- 目标是学习模型的后验概率分布p(z|x)，用模型 进行概率推
- 但这是一个复杂的分布，直接估计分布的参数很困难

变分推理

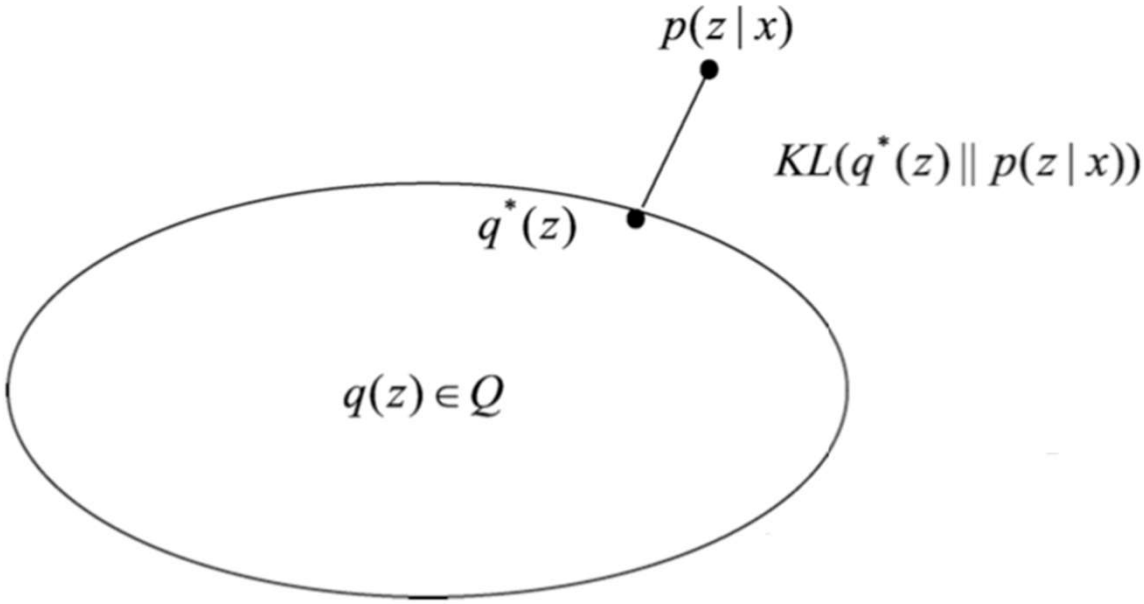
- 考虑用概率分布q(z)近似条件概率分布p(z|x)，用KL散度 D(q(z) || p(z|x))）计算两者的相似度
- q(z）称为变分分布（variational distribution)

- 如果能找到与p(z|x）在KL散度意义下最近的分布$q^*(z)$，则可以用这个分布近似p(z|x)

$$p(z|x) \approx q^*(z)$$

变分推理

- 下图给出了q*(z)与p(z|x)的关系



变分推理

- KL散度可以写成以下形式

$$\begin{aligned} D(q(z)\|p(z|x)) &= E_q\left[\log q(z)\right] - E_q\left[\log p(z|x)\right] \\ &= E_q\left[\log q(z)\right] - E_q\left[\log p(x,z)\right] + \log p(x) \\ &= \log p(x) - \{E_q\left[\log p(x,z)\right] - E_q\left[\log q(z)\right]\} \end{aligned}$$

- 注意到KL散度大于等于零，当且仅当两个分布一致时为零，由此可知右端第一项与第二项满足关系

$$\log p(x) \geqslant E_q\left[\log p(x,z)\right] - E_q\left[\log q(z)\right]$$

- 不等式右端是左端的下界，左端称为证据（(evidence)，右端称为证据下界（(evidence lower bound, ELBO)，证据下界记作

$$L(q) = E_q\left[\log p(x,z)\right] - E_q\left[\log q(z)\right]$$

变分推理

- KL散度的最小化可以通过证据下界的最大化实现，因为目标是求q(z)使KL散度最小化，这时log p(x)是常量

- 因此，变分推理变成求解证据下界最大化的问题

变分推理

- 变分推理目标是通过证据log p(x)的最大化，估计联合概率分布p(x,z)

- 因为含有隐变量z，直接对证据进行最大化困难，转而根据

$$\log p(x) \geqslant E_q\left[\log p(x,z)\right] - E_q\left[\log q(z)\right]$$

- 对证据下界进行最大化。

变分推理

- 对变分分布q(z)要求是具有容易处理的形式，通常假设q(z)对z的所有分量都是互相独立的（实际是条件独立于参数），即满足

$$q(z) = q(z_1)q(z_2)\cdots q(z_n)$$

- 这时的变分分布称为平均场（mean filed)
- KL散度的最小化或证据下界最大化实际是在平均场的集合，即满足独立假设的分布集合

$$Q = \{q(z)|q(z) = \prod_{i=1}^{n} q(z_i)\}$$

- 进行的

变分推理

- 总结起来，变分推理有以下几个步骤：

- 定义变分分布q(z)
- 推导其证据下界表达式
- 用最优化方法对证据下界进行优化，如坐标上升，得到最优分布$q^*(z)$，作为后验分布p(z|x)的近似。

变分EM算法

- 变分推理中，可以通过迭代的方法最大化证据下界，这时算法是EM算法的推广， 称为变分EM算法
- 假设模型是联合概率分布 $p(x, z|\theta)$ ，
- x是观测变量
- z是隐变量
- θ 是参数
- 目标是通过观测数据的概率（证据）$\log p(x|\theta)$ 的最大化，估计模型的参数 θ

变分EM算法

- 使用变分推理，导入平均场 $q(z) = \prod_{i=1}^{n} q(z_i)\}$ 定义证据下界

$$L(q, \theta) = E_q[\log p(x, z|\theta)] - E_q[\log q(z)]$$

- 通过迭代，分别以q和0为变量对证据下界进行最大化，就得到变分EM算法

变分EM算法

算法 20.3（变分 EM 算法）

循环执行以下 E 步和 M 步，直到收敛。

（1）E 步：固定 θ，求 $L(q,\theta)$ 对 q 的最大化。

（2）M 步：固定 q，求 $L(q,\theta)$ 对 θ 的最大化。

给出模型参数 θ 的估计值。 ■

根据变分推理原理，观测数据的概率和证据下界满足

$$\log p(x|\theta) - L(q,\theta) = D(q(z)||p(z|x,\theta)) \geqslant 0 \tag{20.40}$$

变分EM算法

- 变分EM算法的迭代过程中，以下关系成立：

$$\log p(x|\theta^{(t-1)}) = L(q^{(t)}, \theta^{(t-1)}) \leqslant L(q^{(t)}, \theta^{(t)}) \leqslant \log p(x|\theta^{(t)})$$

- 左边的等式基于E步计算和变分推理原理
- 中间的不等式基于M步计算
- 右边的不等式基于变分推理原理

- 说明每次迭代都保证观测数据的概率不递减。因此，变分EM算法一定收敛，但可能收敛到局部最优

变分EM算法

- EM算法实际也是对证据下界进行最大化
- EM算法的推广是求F函数的极大-极大算法，其中的F函数就是证据下界

- EM算法假设q(z) = p(zlx)且p(zx)容易计算，而变分EM算法则考虑一般情况使用容易计算的平均场

$$q(z) = \prod_{i=1}^{n} q(z_i)$$

- 当模型复杂时，EM算法未必可用，但变分EM算法仍然可以使用。

算法推导

- 1. 证据下界的定义

- 为简单起见，一次只考虑一个文本，记作w
- 文本的单词序列$\mathbf{w}=(w_1,\cdots,w_n,\cdots,w_N)$，对应的话题序列$\mathbf{z}=(z_1,\cdots,z_n,\cdots,z_N)$，以及话题分布$\theta$，随机变量w, z和$\theta$的联合分布是

$$p(\theta,\mathbf{z},\mathbf{w}|\alpha,\varphi)=p(\theta|\alpha)\prod_{n=1}^{N}p(z_n|\theta)p(w_n|z_n,\varphi)$$

- w是可观测变量，θ和z是隐变量，α和φ是参数

算法推导

- 定义基于平均场的变分分布

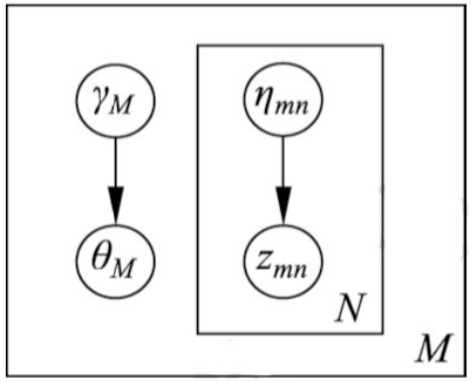
$$q(\theta, \mathbf{z}|\gamma, \eta) = q(\theta|\gamma) \prod_{n=1}^{N} q(z_n|\eta_n)$$

- 其中 γ 是狄利克雷分布参数，$\eta = (\eta_1, \eta_2, \cdots, \eta_n)$ 是多项分布参数，变量 θ 和z的各个分量都是条件独立的

- 目标是求KL散度意义下最相近的变分分布 $q(\theta, \mathbf{z}|\gamma, \eta)$ ， 以近似LDA模型的后验分布 $p(\theta, \mathbf{z}|\mathbf{w}, \alpha, \varphi)$ 。

算法推导

- 下图是变分分布的板块表示。LDA模型中隐变量 θ 和z之间存在依存关系， 变分分布中这些依存关系被去掉，变量 θ 和z条件独立。



算法推导

- 由此得到一个文本的证据下界

$$L(\gamma, \eta, \alpha, \varphi) = E_q[\log p(\theta, \mathbf{z}, \mathbf{w}|\alpha, \varphi)] - E_q[\log q(\theta, \mathbf{z}|\gamma, \eta)]$$

- 其中数学期望是对分布 $q(\theta, \mathbf{z}|\gamma, \eta)$ 定义的，为了方便写作 $E_q[\bullet]$
- γ 和 η 是变分分布的参数，α 和 φ 是LDA模型的参数

- 所有文本的证据下界为

$$L_{\mathbf{w}}(\gamma, \eta, \alpha, \varphi) = \sum_{m=1}^{M} \{E_{q_m}[\log p(\theta_m, \mathbf{z}_m, \mathbf{w}_m|\alpha, \varphi)] - E_{q_m}[\log q(\theta_m, \mathbf{z}_m|\gamma_m, \eta_m)]\}$$

算法推导

- 为求解证据下界 $L(\gamma, \eta, \alpha, \varphi)$ 的最大化，首先写出证据下界的表达式。为此展开证据下界

$$L(\gamma, \eta, \alpha, \varphi) = E_q[\log p(\theta, \mathbf{z}, \mathbf{w}|\alpha, \varphi)] - E_q[\log q(\theta, \mathbf{z}|\gamma, \eta)]$$

$$L(\gamma, \eta, \alpha, \varphi) = E_q[\log p(\theta|\alpha)] + E_q[\log p(\mathbf{z}|\theta)] + E_q[\log p(\mathbf{w}|\mathbf{z}, \varphi)] -$$

$$E_q[\log q(\theta|\gamma)] - E_q[\log q(\mathbf{z}|\eta)]$$

算法推导

- 根据变分参数 γ 和 η ， 模型参数 α 和 φ 继续展开， 并将展开式的每一项写成一行

$$L(\gamma,\eta,\alpha,\varphi)=\log\Gamma\left(\sum_{l=1}^{K}\alpha_l\right)-\sum_{k=1}^{K}\log\Gamma(\alpha_k)+\sum_{k=1}^{K}(\alpha_k-1)\left[\Psi(\gamma_k)-\Psi\left(\sum_{l=1}^{K}\gamma_l\right)\right]+$$

$$\sum_{n=1}^{N}\sum_{k=1}^{K}\eta_{nk}\left[\Psi(\gamma_k)-\Psi\left(\sum_{l=1}^{K}\gamma_l\right)\right]+$$

$$\sum_{n=1}^{N}\sum_{k=1}^{K}\sum_{v=1}^{V}\eta_{nk}w_n^v\log\varphi_{kv}-$$

$$\log\Gamma\left(\sum_{l=1}^{K}\gamma_l\right)+\sum_{k=1}^{K}\log\Gamma(\gamma_k)-\sum_{k=1}^{K}(\gamma_k-1)\left[\Psi(\gamma_k)-\Psi\left(\sum_{l=1}^{K}\gamma_l\right)\right]-\sum_{n=1}^{N}\sum_{k=1}^{K}\eta_{nk}\log\eta_{nk}\qquad(20.47)$$

$$\Psi(\alpha_k)=\frac{\mathrm{d}}{\mathrm{d}\alpha_k}\log\Gamma(\alpha_k)$$

算法推导

第一项推导，求 $E_q[\log p(\theta|\alpha)]$，是关于分布 $q(\theta, \mathbf{z}|\gamma, \eta)$ 的数学期望。

$$E_q[\log p(\theta|\alpha)] = \sum_{k=1}^{K}(\alpha_k - 1)E_q[\log \theta_k] + \log\Gamma\left(\sum_{l=1}^{K}\alpha_l\right) - \sum_{k=1}^{K}\log\Gamma(\alpha_k) \tag{20.49}$$

其中 $\theta \sim \mathrm{Dir}(\theta|\gamma)$，所以利用附录 E 式 (E.7) 有

$$E_{q(\theta|\gamma)}[\log \theta_k] = \Psi(\gamma_k) - \Psi\left(\sum_{l=1}^{K}\gamma_l\right) \tag{20.50}$$

故得

$$E_q[\log p(\theta|\alpha)] = \log\Gamma\left(\sum_{l=1}^{K}\alpha_l\right) - \sum_{k=1}^{K}\log\Gamma(\alpha_k) + \sum_{k=1}^{K}(\alpha_k - 1)\left[\Psi(\gamma_k) - \Psi\left(\sum_{l=1}^{K}\gamma_l\right)\right] \tag{20.51}$$

式中 α_k 和 γ_k 表示第 k 个话题的狄利克雷分布参数。

算法推导

第二项推导，求 $E_q[\log p(\mathbf{z}|\theta)]$，是关于分布 $q(\theta, \mathbf{z}|\gamma, \eta)$ 的数学期望。

$$
\begin{aligned}
E_q(\log p(\mathbf{z}|\theta)) &= \sum_{n=1} E_q\left[\log p(z_n|\theta)\right] \\
&= \sum_{n=1}^{N} E_{q(\theta, z_n|\gamma,\eta)}[\log(z_n|\theta)] \\
&= \sum_{n=1}^{N}\sum_{k=1}^{K} q(z_{nk}|\eta) E_{q(\theta|\gamma)}[\log \theta_k] \\
&= \sum_{n=1}^{N}\sum_{k=1}^{K} \eta_{nk}\left[\Psi(\gamma_k) - \Psi\left(\sum_{l=1}^{K}\gamma_l\right)\right]
\end{aligned}
\tag{20.52}
$$

式中 η_{nk} 表示文档第 n 个位置的单词由第 k 个话题产生的概率，γ_k 表示第 k 个话题的狄利克雷分布参数。最后一步用到附录 E 式 (E.4)。

算法推导

第三项推导，求 $E_q\left[\log p(\mathbf{w}|\mathbf{z},\varphi)\right]$，是关于分布 $q(\theta,\mathbf{z}|\gamma,\eta)$ 的数学期望。

$$
\begin{aligned}
E_q\left[\log p(\mathbf{w}|\mathbf{z},\varphi)\right] &= \sum_{n=1}^{N} E_q\left[\log p(w_n|z_n,\varphi)\right] \\
&= \sum_{n=1}^{N} E_{q(z_n|\eta)}[\log p(w_n|z_n,\varphi)] \\
&= \sum_{n=1}^{N}\sum_{k=1}^{K} q(z_{nk}|\eta)\log p(w_n|z_{nk},\varphi) \\
&= \sum_{n=1}^{N}\sum_{k=1}^{K}\sum_{v=1}^{V} \eta_{nk} w_n^v \log \varphi_{kv}
\end{aligned}
\tag{20.53}
$$

式中 η_{nk} 表示文档第 n 个位置的单词由第 k 个话题产生的概率，w_n^v 在第 n 个位置的单词是单词集合的第 v 个单词时取值为 1，否则取值为 0，φ_{kv} 表示第 k 个话题生成单词集合中第 v 个单词的概率。

算法推导

第四项推导，求 $E_q\left[\log q(\theta|\gamma)\right]$，是关于分布 $q(\theta, \mathbf{z}|\gamma, \eta)$ 的数学期望。由于 $\theta \sim \mathrm{Dir}(\gamma)$，类似式 (20.50) 可以得到

$$E_q[\log q(\theta|\gamma)] = \log\Gamma\left(\sum_{l=1}^{K}\gamma_l\right) - \sum_{k=1}^{K}\log\Gamma(\gamma_k) + \sum_{k=1}^{K}(\gamma_k - 1)\left[\Psi(\gamma_k) - \Psi\left(\sum_{l=1}^{K}\gamma_l\right)\right] \tag{20.54}$$

式中 γ_k 表示第 k 个话题的狄利克雷分布参数。

算法推导

第五项公式推导，求 $E_q[\log q(\mathbf{z}|\eta)]$，是关于分布 $q(\theta, \mathbf{z}|\gamma, \eta)$ 的数学期望。

$$
\begin{aligned}
E_q[\log q(\mathbf{z}|\eta)] &= \sum_{n=1}^{N} E_q[\log q(z_n|\eta)] \\
&= \sum_{n=1}^{N} E_{q(z_n|\eta)}[\log q(z_n|\eta)] \\
&= \sum_{n=1}^{N}\sum_{k=1}^{K} q(z_{nk}|\eta)\log q(z_{nk}|\eta) \\
&= \sum_{n=1}^{N}\sum_{k=1}^{K} \eta_{nk}\log \eta_{nk} \qquad (20.55)
\end{aligned}
$$

式中 η_{nk} 表示文档第 n 个位置的单词由第 k 个话题产生的概率，γ_k 表示第 k 个话题的狄利克雷分布参数。

算法推导

- 2. 变分参数 γ 和 η 的估计

- 首先通过证据下界最优化估计参数 η 。
- η_{nk} 表示第n个位置的单词是由第k个话题生成的概率。考虑式（20.47）关于 η_{nk} 的最大化，η_{nk} 满足约束条件

$$\sum_{l=1}^{K} \eta_{nl} = 1$$

算法推导

- 包含 η_{nk} 的约束最优化问题拉格朗日函数为

$$L_{[\eta_{nk}]} = \eta_{nk}\left[\Psi(\gamma_k) - \Psi\left(\sum_{l=1}^{K}\gamma_l\right)\right] + \eta_{nk}\log\varphi_{kv} - \eta_{nk}\log\eta_{nk} + \lambda_n\left(\sum_{l=1}^{K}\eta_{nl} - 1\right) \tag{20.56}$$

- 这里 φ_{kv} 是（在第n个位置）由第k个话题生成第v个单词的概率
- 对 η_{nk} 求偏导数得

$$\frac{\partial L}{\partial \eta_{nk}} = \Psi(\gamma_k) - \Psi\left(\sum_{l=1}^{K}\gamma_l\right) + \log\varphi_{kv} - \log\eta_{nk} - 1 + \lambda_n$$

算法推导

- 令偏导数为零，得到参数 η_{nk} 的估计值

$$\eta_{nk} \propto \varphi_{kv} \exp\left(\Psi(\gamma_k) - \Psi\left(\sum_{l=1}^{K} \gamma_l\right)\right)$$

- 接着通过证据下界最优化估计参数 γ 。γ_k 是第k个话题的狄利克雷分布参数。考虑式（20.47）关于 γ_k 的最大化

$$L_{[\gamma_k]} = \sum_{k=1}^{K} (\alpha_k - 1)\left[\Psi(\gamma_k) - \Psi\left(\sum_{l=1}^{K} \gamma_l\right)\right] + \sum_{n=1}^{N}\sum_{k=1}^{K} \eta_{nk}\left[\Psi(\gamma_k) - \Psi\left(\sum_{l=1}^{K} \gamma_l\right)\right] -$$
$$\log\Gamma\left(\sum_{l=1}^{K} \gamma_l\right) + \log\Gamma(\gamma_k) - \sum_{k=1}^{K} (\gamma_k - 1)\left[\Psi(\gamma_k) - \Psi\left(\sum_{l=1}^{K} \gamma_l\right)\right] \qquad (20.59)$$

算法推导

- 简化为

$$L_{[\gamma_k]} = \sum_{k=1}^{K}\left[\Psi(\gamma_k) - \Psi\left(\sum_{l=1}^{K}\gamma_l\right)\right]\left(\alpha_k + \sum_{n=1}^{N}\eta_{nk} - \gamma_k\right) - \log\Gamma\left(\sum_{l=1}^{K}\gamma_l\right) + \log\Gamma(\gamma_k) \tag{20.60}$$

- 对 γ_k 求偏导数得

$$\frac{\partial L}{\partial \gamma_k} = \left[\Psi'(\gamma_k) - \Psi'\left(\sum_{l=1}^{K}\gamma_l\right)\right]\left(\alpha_k + \sum_{n=1}^{N}\eta_{nk} - \gamma_k\right)$$

- 据此，得到由坐标上升算法估计变分参数的方法

LDA的变分参数估计算法

算法 20.4（LDA 的变分参数估计算法）

（1）初始化：对所有 k 和 n，$\eta_{nk}^{(0)} = 1/K$

（2）初始化：对所有 k，$\gamma_k = \alpha_k + N/K$

（3）重复

（4）　　对 $n = 1$ 到 N

（5）　　对 $k = 1$ 到 K

（6）　　$$\eta_{nk}^{(t+1)} = \varphi_{kv} \exp\left[\Psi(\gamma_k^{(t)}) - \Psi\left(\sum_{l=}^{K} \gamma_l^{(t)}\right)\right]$$

（7）　　规范化 $\eta_{nk}^{(t+1)}$ 使其和为 1

（8）　　$$\gamma^{(t+1)} = \alpha + \sum_{n=1}^{N} \eta_n^{(t+1)}$$

（9）直到收敛 ■

算法推导

- 3. 模型参数 α 和 $\varphi_{\cdot}$ 的估计

- 给定一个文本合集 $D = \{\mathbf{w}_1, \cdots, \mathbf{w}_m, \cdots, \mathbf{w}_M\}$, 模型参数估计对所有文本同时进行

算法推导

首先通过证据下界的最大化估计 φ。φ_{kv} 表示第 k 个话题生成单词集合第 v 个单词的概率。将式 (20.47) 扩展到所有文本，并考虑关于 φ 的最大化。满足 K 个约束条件

$$\sum_{v=1}^{V} \varphi_{kv} = 1, \quad k = 1, 2, \cdots, K$$

约束最优化问题的拉格朗日函数为

$$L_{[\beta]} = \sum_{m=1}^{M} \sum_{n=1}^{N_m} \sum_{k=1}^{K} \sum_{v=1}^{V} \eta_{mnk} w_{mn}^{v} \log \varphi_{kv} + \sum_{k=1}^{K} \lambda_k \left(\sum_{v=1}^{V} \varphi_{kv} - 1 \right) \tag{20.63}$$

对 φ_{kv} 求偏导数并令其为零，归一化求解，得到参数 φ_{kv} 的估计值

$$\varphi_{kv} = \sum_{m=1}^{M} \sum_{n=1}^{N_m} \eta_{mnk} w_{mn}^{v} \tag{20.64}$$

其中 η_{mnk} 为第 m 个文本的第 n 个单词属于第 k 个话题的概率，w_{mn}^{v} 在第 m 个文本的第 n 个单词是单词集合的第 v 个单词时取值为 1，否则为 0。

算法推导

接着通过证据下界的最大化估计参数 α。α_k 表示第 k 个话题的狄利克雷分布参数。将式 (20.47) 扩展到所有文本，并考虑关于 α 的最大化

$$L_{[\alpha]}=\sum_{m=1}^{M}\left\{\log\Gamma\left(\sum_{l=1}^{K}\alpha_l\right)-\sum_{k=1}^{K}\log\Gamma(\alpha_k)+\sum_{k=1}^{K}(\alpha_k-1)\left[\Psi(\gamma_{mk})-\Psi\left(\sum_{l=1}^{K}\gamma_{ml}\right)\right]\right\} \tag{20.65}$$

对 α_k 求偏导数得

$$\frac{\partial L}{\partial\alpha_k}=M\left[\Psi\left(\sum_{l=1}^{K}\alpha_l\right)-\Psi(\alpha_k)\right]+\sum_{m=1}^{M}\left[\Psi(\gamma_{mk})-\Psi\left(\sum_{l=1}^{K}\gamma_{ml}\right)\right] \tag{20.66}$$

再对 α_l 求偏导数得

$$\frac{\partial^2 L}{\partial\alpha_k\partial\alpha_l}=M\left[\Psi'\left(\sum_{l=1}^{K}\alpha_l\right)-\delta(k,l)\Psi'(\alpha_k)\right] \tag{20.67}$$

这里 $\delta(k,l)$ 是 delta 函数。

算法推导

式 (20.65) 和式 (20.66) 分别是函数 (20.64) 对变量 α 的梯度 $g(\alpha)$ 和 Hessian 矩阵 $H(\alpha)$。应用牛顿法（又称为牛顿-拉弗森方法）求该函数的最大化[①]。用以下公式迭代，得到参数 α 的估计值。

$$\alpha_{\text{new}} = \alpha_{\text{old}} - H(\alpha_{\text{old}})^{-1} g(\alpha_{\text{old}}) \tag{20.68}$$

据此，得到估计参数 α 的算法。

LDA的变分EM算法

算法 20.5（LDA 的变分 EM 算法）

输入：给定文本集合 $D = \{\mathbf{w}_1, \cdots, \mathbf{w}_m, \cdots, \mathbf{w}_M\}$；

输出：变分参数 γ，η，模型参数 α，φ。

交替迭代 E 步和 M 步，直到收敛。

（1）E 步

固定模型参数 α，φ，通过关于变分参数 γ，η 的证据下界的最大化，估计变分参数 γ，η。具体见算法 20.4。

（2）M 步

固定变分参数 γ，η，通过关于模型参数 α，φ 的证据下界的最大化，估计模型参数 α，φ。具体算法见式 (20.63) 和式 (20.67)。

根据变分参数 (γ, η) 可以估计模型参数 $\theta = (\theta_1, \cdots, \theta_m, \cdots, \theta_M), \mathbf{z} = (\boldsymbol{z}_1, \cdots, \boldsymbol{z}_m, \cdots, \boldsymbol{z}_M)$。 ■