潜在狄利克雷分配
Latent Dirichlet Allocation

西安交通大学管理学院
信息管理与电子商务系
智能决策与机器学习研究中心
刘佳鹏

简介

- 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，于2002年由Blei等提出
- LDA在文本数据挖掘、图像处理、生物信息处理、商业和管理(特别是市场营销、信息系统)等领域被广泛使用

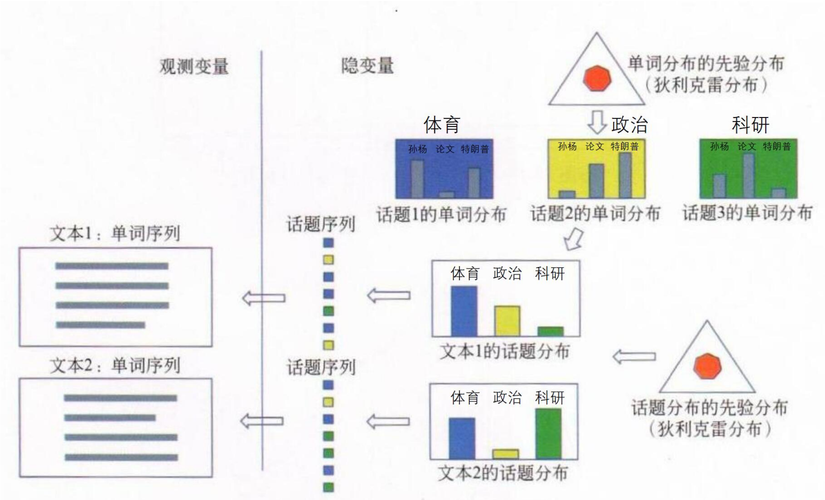
简介

- 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
  - 该假设与概率潜在语义分析的假设相同
- 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
  - 该假设是概率潜在语义分析中没有的
- 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
  - 因为引入了先验分布，所以LDA模型是概率潜在语义分析的贝叶斯扩展

简介

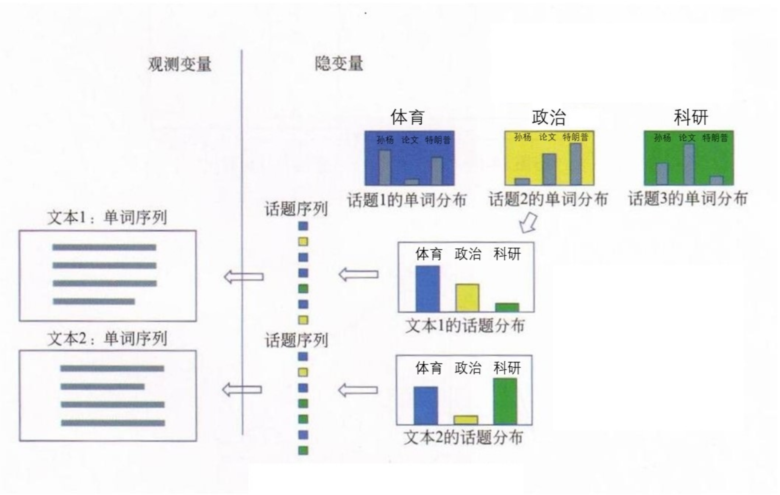
- LDA的文本集合的生成过程如下：
- 首先随机生成一个文本的话题分布，
- 之后在该文本的每个位置，依据该文本的话题分布随机生成一个话题
- 然后在该位置依据该话题的单词分布随机生成一个单词
- 直至文本的最后一个位置，生成整个文本
- 重复以上过程生成所有文本

LDA的文本生成过程



观测变量
隐变量
单词分布的先验分布
（狄利克雷分布）
体育
政治
科研
孙杨 论文 特朗普
孙杨 论文 特朗普
孙杨 论文 特朗普
话题1的单词分布
话题2的单词分布
话题3的单词分布
文本1：单词序列
话题序列
体育 政治 科研
文本1的话题分布
文本2：单词序列
话题序列
体育 政治 科研
文本2的话题分布
话题分布的先验分布
（狄利克雷分布）

对比：概率潜在语义分析的文本生成过程



LDA中的几个概率分布

- 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

LDA中的几个概率分布

- **定义（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

$$P(X_1=n_1,X_2=n_2,\cdots,X_k=n_k)=\frac{n!}{n_1!n_2!\cdots n_k!}p_1^{n_1}p_2^{n_2}\cdots p_k^{n_k}$$
$$=\frac{n!}{\prod_{i=1}^k n_i!}\prod_{i=1}^k p_i^{n_i}$$

  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$X\sim \mathrm{Mult}(n,p)$
- 特别地，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
  - 显然多项分布包含类别分布
  - 实际上LDA模型中的多项分布指的就是类别分布

LDA中的几个概率分布

- 二项分布是多项分布的特殊情况
- 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

$$P(X=m)=\left(\begin{array}{c} n \\ m \end{array}\right) p^m(1-p)^{n-m}, \quad m=0,1,2,\cdots,n$$

  其中 n 和 p $(0 \leqslant p \leqslant 1)$ 是参数
- 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
  - 显然二项分布包含伯努利分布

LDA中的几个概率分布

- 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
- **定义（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

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

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$\theta$服从参数为$\alpha$的狄利克雷分布，记作$\theta \sim \mathrm{Dir}(\alpha)$

LDA中的几个概率分布

- 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

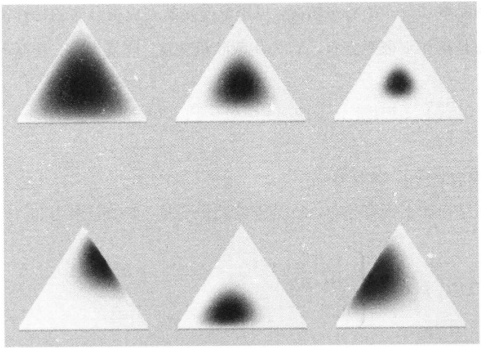
$$\Gamma(s+1) = s\Gamma(s)$$

当 s 是自然数时，有

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

LDA中的几个概率分布

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



- 上图为二维单纯型上的狄利克雷分布，参数分别为$\alpha = (3,3,3), \alpha = (7,7,7), \alpha = (20,20,20), \alpha = (2,6,11), \alpha = (14,9,5), \alpha = (6,2,6)$

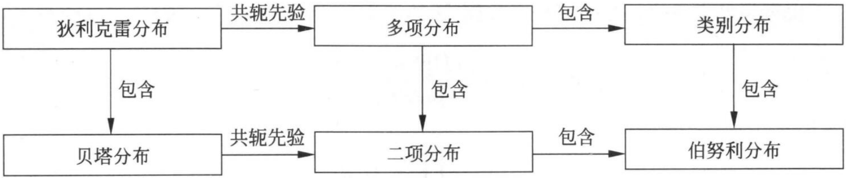
LDA中的几个概率分布

- 贝塔分布是狄利克雷分布的特殊情况
- 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

$$p(x) = \begin{cases} \frac{\Gamma(s+t)}{\Gamma(s)\Gamma(t)} x^{s-1}(1-x)^{t-1}, & 0 \leqslant x \leqslant 1 \\ 0, & \text{其他} \end{cases}$$

其中 $s > 0$ 和 $t > 0$ 是参数

上述几种概率分布之间的关系



共轭先验

- 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轭先验（conjugate prior）
- **共轭先验:** 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轭分布（conjugate distributions）, 先验分布称为似然函数的共轭先验（conjugate prior）
- 例如：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，两者构成共轭分布。作为先验分布的狄利克雷分布的参数又称为超参数
- 使用共轭分布的好处是便于从先验分布计算后验分布

共轭先验

- 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 $\mathcal{W}$ 上的多项分布, $X \sim \mathrm{Mult}(n, \theta)$, 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 $\mathcal{W}$ 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, $i = 1, 2, \cdots, k$
- 参数 θ_i 为 w_i 出现的概率 $(i = 1, 2, \cdots, k)$

共轭先验

- 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 $p(\theta \mid D)$
- 对于给定的样本数据 D, 似然函数是

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

共轭先验

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

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

共轭先验

- 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

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

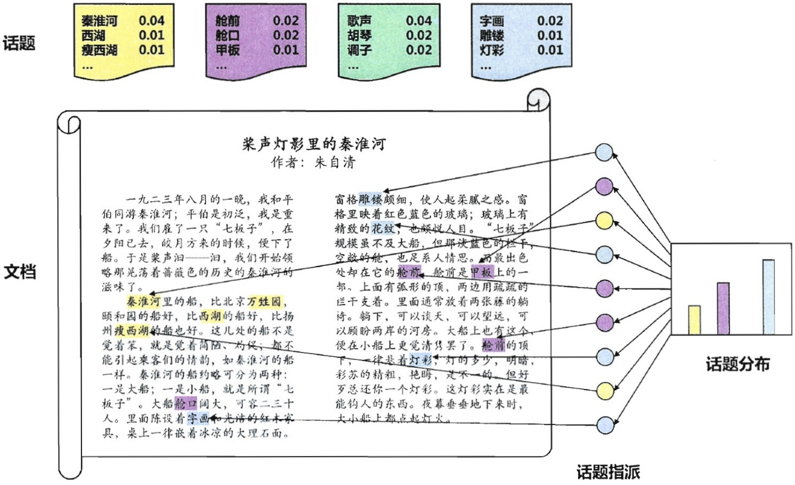
共轭先验

- 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轭先验
- 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计数 $n = (n_1, n_2, \cdots, n_k)$，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）

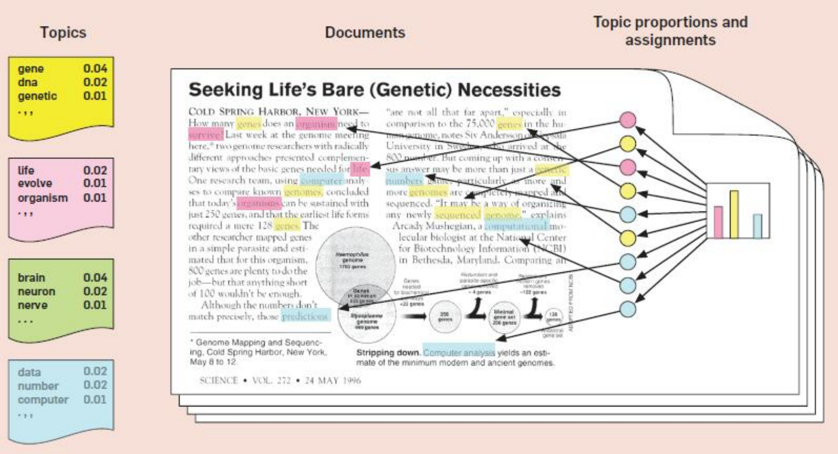
LDA模型——基本想法

- 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
- 严格意义上说，这里的多项分布都是类别分布，在机器学习与自然语言处理中，有时对两者不作严格区分

LDA模型——基本想法


话题
秦淮河 0.04
西湖 0.01
瘦西湖 0.01
...
舱前 0.02
舱口 0.02
甲板 0.01
...
歌声 0.04
胡琴 0.02
调子 0.02
...
字画 0.02
雕镂 0.01
灯彩 0.01
...
文档
桨声灯影里的秦淮河
作者：朱自清
一九二三年八月的一晚，我和平伯同游秦淮河；平伯是初泛，我是重来了。我们雇了一只“七板子”，在夕阳已去，皎月方来的时候，便下了船。于是桨声汩——汩，我们开始领略那晃荡着蔷薇色的历史的秦淮河的滋味了。
秦淮河里的船，比北京万甡园，颐和园的船好，比西湖的船好，比扬州瘦西湖的船也好。这几处的船不是觉着笨，就是觉着简陋、局促；都不能引起乘客们的情韵，如秦淮河的船一样。秦淮河的船约略可分为两种：一是大船；一是小船，就是所谓“七板子”。大船舱口阔大，可容二三十人。里面陈设着字画和光洁的红木家具，桌上一律嵌着冰凉的大理石面。窗格雕镂颇细，使人起柔腻之感。窗格里映着红色蓝色的玻璃；玻璃上有精致的花纹，也颇悦人目。“七板子”规模虽不及大船，但那淡蓝色的栏干，空敞的舱，也足系人情思。而最出色处却在它的舱前。舱前是甲板上的一部。上面有弧形的顶，两边用疏疏的栏干支着。里面通常放着两张藤的躺椅。躺下，可以谈天，可以望远，可以顾盼两岸的河房。大船上也有这个，便在小船上更觉清隽罢了。舱前的顶下，一律悬着灯彩；灯的多少，明暗，彩苏的精粗，艳晦，是不一的。但好歹总还你一个灯彩。这灯彩实在是最能钩人的东西。夜幕垂垂地下来时，大小船上都点起灯火。
话题分布
话题指派

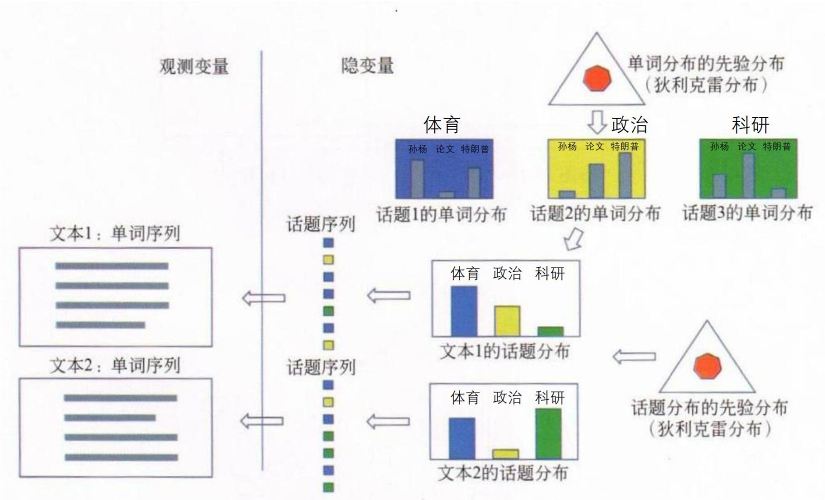
LDA模型——基本想法


Topics
Documents
Topic proportions and assignments
gene 0.04
dna 0.02
genetic 0.01
...
life 0.02
evolve 0.01
organism 0.01
...
brain 0.04
neuron 0.02
nerve 0.01
...
data 0.02
number 0.02
computer 0.01
...
Seeking Life's Bare (Genetic) Necessities
* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.
Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.
SCIENCE • VOL. 272 • 24 MAY 1996

LDA模型——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

LDA模型——基本想法



观测变量
隐变量
单词分布的先验分布
（狄利克雷分布）
体育
政治
科研
孙杨 论文 特朗普
孙杨 论文 特朗普
孙杨 论文 特朗普
话题1的单词分布
话题2的单词分布
话题3的单词分布
文本1：单词序列
话题序列
体育 政治 科研
文本1的话题分布
文本2：单词序列
话题序列
体育 政治 科研
文本2的话题分布
话题分布的先验分布
（狄利克雷分布）

LDA模型——基本想法

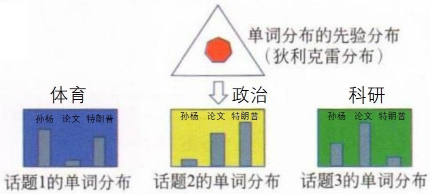
- LDA模型中文本的单词序列是观测变量，文本的话题序列是隐变量，文本的话题分布和话题的单词分布也是隐变量
- 利用LDA进行话题分析，就是对给定文本集合，学习到每个文本的话题分布，以及每个话题的单词分布
- 这就是LDA模型的学习目标：给定文本集合，通过后验概率分布的估计，推断模型的所有参数

LDA模型——模型定义

- 潜在狄利克雷分配（LDA）使用三个集合:
- (1) 单词集合 $W=\{w_1,\cdots,w_v,\cdots,w_V\}$，其中 w_v 是第 v 个单词，$v=1,2,\cdots,V$，V 是单词的个数
- (2) 文本集合 $D=\{\boldsymbol{w}_1,\cdots,\boldsymbol{w}_m,\cdots,\boldsymbol{w}_M\}$，其中 $\boldsymbol{w}_m$ 是第 m 个文本，$m=1,2,\cdots,M$，M 是文本的个数
  - 文本 $\boldsymbol{w}_m$ 是一个单词序列 $\boldsymbol{w}_m=(w_{m1},\cdots,w_{mn},\cdots,w_{mN_m})$，其中 w_{mn} 是文本 $\boldsymbol{w}_m$ 的第 n 个单词，$n=1,2,\cdots,N_m$，N_m 是文本 $\boldsymbol{w}_m$ 中单词的个数
- (3) 话题集合 $Z=\{z_1,\cdots,z_k,\cdots,z_K\}$，其中 z_k 是第 k 个话题，$k=1,2,\cdots,K$，K 是话题的个数

LDA模型——模型定义

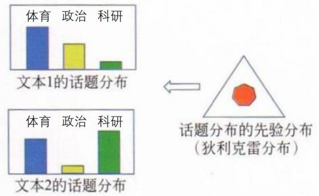
- **话题的单词分布及其先验分布：**



- 每一个话题 z_k 由一个“单词的条件概率分布 $p(w \mid z_k)$” 决定，$w \in W$
- 分布 $p(w \mid z_k)$ 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 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 $K \times V$ 矩阵 $\boldsymbol{\varphi} = \{\varphi_k\}_{k=1}^{K}$
  - 参数 φ_k 服从狄利克雷分布(先验分布)，其超参数为 β
  - 超参数 β 也是一个 V 维向量 $\beta = (\beta_1, \beta_2, \cdots, \beta_V)$

LDA模型——模型定义

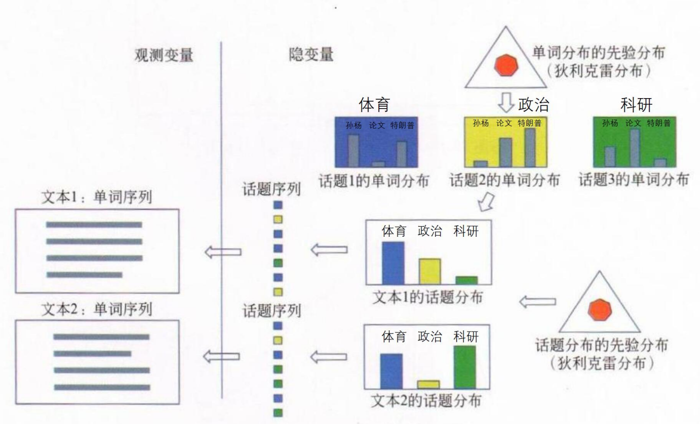
- **文本的话题分布及其先验分布：**



- 每一个文本 $\boldsymbol{w}_m$ 由一个"话题的条件概率分布 $p(z \mid \boldsymbol{w}_m)$" 决定，$z \in Z$
- 分布 $p(z \mid \boldsymbol{w}_m)$ 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 是一个 K 维向量 $\theta_m = (\theta_{m1}, \theta_{m2}, \cdots, \theta_{mK})$，其中 θ_{mk} 表示文本 $\boldsymbol{w}_m$ 生成话题 z_k 的概率
  - 所有文本的参数向量构成一个 $M \times K$ 矩阵 $\boldsymbol{\theta} = \{\theta_m\}_{m=1}^{M}$
  - 参数 θ_m 服从狄利克雷分布(先验分布)，其超参数为 α
  - 超参数 α 也是一个 K 维向量 $\alpha = (\alpha_1, \alpha_2, \cdots, \alpha_K)$

LDA模型——模型定义

- 每一个文本 $\boldsymbol{w}_m$ 中的每一个单词 w_{mn} 由该文本的话题分布 $p(z \mid \boldsymbol{w}_m)$ 以及所有话题的单词分布 $p(w \mid z_k)$ 决定



LDA模型——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
- (1) **生成话题的单词分布:**
  随机生成 K 个话题的单词分布。具体过程如下, 按照狄利克雷分布 $\mathrm{Dir}(\beta)$ 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 $p(w \mid z_k), w \in W, k = 1, 2, \cdots, K$

LDA模型——生成过程

- (2) **生成文本的话题分布:**
  随机生成 M 个文本的话题分布。具体过程如下: 按照狄利克雷分布 $\mathrm{Dir}(\alpha)$ 随机生成一个参数向量 $\theta_m, \theta_m \sim \mathrm{Dir}(\alpha)$, 作为文本 $\boldsymbol{w}_m$ 的话题分布 $p\left(z \mid \boldsymbol{w}_m\right), m=1,2, \cdots, M$

LDA模型——生成过程

- (3) **生成文本的单词序列:**
  随机生成 M 个文本的 N_m 个单词。文本 $\boldsymbol{w}_m(m=1,2,\cdots,M)$ 的单词 $w_{mn}(n=1,2,\cdots,N_m)$ 的生成过程如下:
  (3-1) 首先按照多项分布 $\mathrm{Mult}\left(\theta_m\right)$ 随机生成一个话题 $z_{mn}, z_{mn} \sim \mathrm{Mult}\left(\theta_m\right)$
  (3-2) 然后按照多项分布 $\mathrm{Mult}\left(\varphi_{z_{mn}}\right)$ 随机生成一个单词 $w_{mn}, w_{mn} \sim \mathrm{Mult}\left(\varphi_{z_{mn}}\right)$
- 注: 文本 $\boldsymbol{w}_m$ 本身是单词序列 $\boldsymbol{w}_m=\left(w_{m1}, w_{m2}, \cdots, w_{mN_m}\right)$, 对应着隐式的话题序列$\boldsymbol{z}_m=\left(z_{m1}, z_{m2}, \cdots, z_{mN_m}\right)$

LDA模型——生成过程

（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；

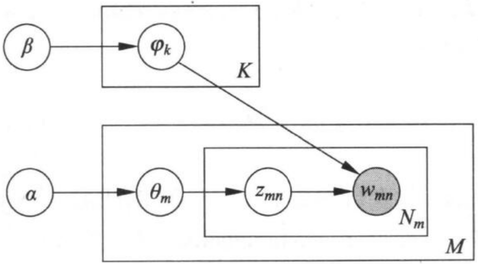
（b）生成单词 $w_{mn} \sim \mathrm{Mult}(\varphi_{z_{mn}})$。

LDA模型——生成过程

- 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

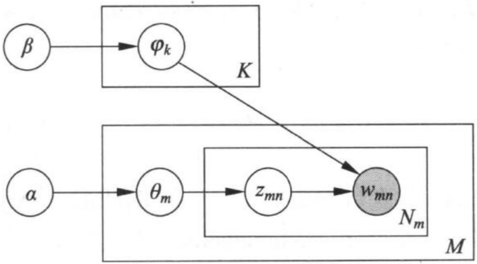
概率图模型

- 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），亦称为盘式记法



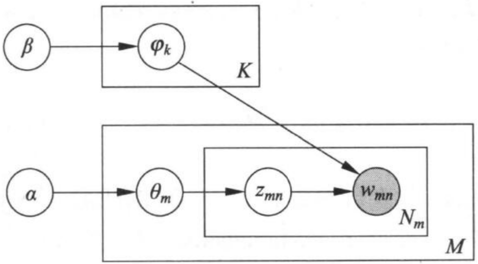
- 结点表示随机变量，实心结点是观测变量，空心结点是隐变量
- 有向边表示概率依存关系
- 矩形（板块）表示重复，板块内数字表示重复的次数

概率图模型



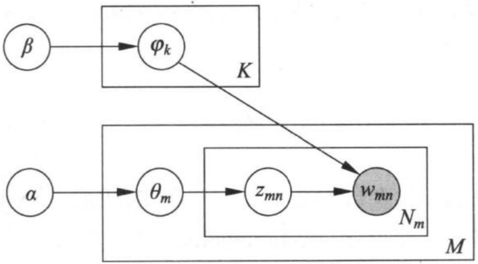
- 结点 α 和 β 是模型的超参数，结点 φ_k 表示话题的单词分布的参数，结点 θ_m 表示文本的话题分布的参数，结点 z_{mn} 表示话题，结点 w_{mn} 表示单词

概率图模型



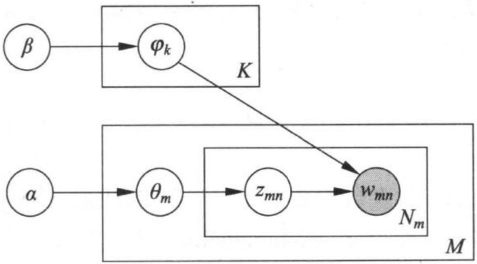
- 结点 β 指向结点 φ_k, 重复 K 次, 表示根据超参数 β 生成 K 个话题的单词分布的参数 φ_k

概率图模型



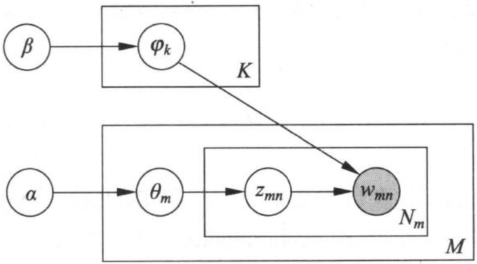
- 结点 α 指向结点 θ_m, 重复 M 次, 表示根据超参数 α 生成 M 个文本的话题分布的参数 θ_m

概率图模型



- 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}

概率图模型



- 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