北京航空航天大學
BEIHANG UNIVERSITY

自然语言处理

人工智能研究院

主讲教师 沙磊

话题模型

概要

- **Latent Semantic Analysis (LSA)**
- Probabilistic Topic Model
  - Probabilistic Latent Semantic Analysis (pLSA)
  - Latent Dirichlet Allocation (LDA)

Latent Semantic Analysis

- Latent Semantic Analysis (LSA)一般译作潜在语义分析，有时候也称作Latent Semantic Indexing(LSI，潜在语义索 引)
- LSA的提出者是Scott Deerwester、Susan T. Dumais等人，发表的时间是1990年
- LSA的基础是向量空间模型(vector space model)
- LSA的基础理念是，基于词项在文档集合中的共现特性，表达词项的潜在意义
- LSA将文档表示映射到潜在语义空间，从而更好地衡量文本之间的相关性
- LSA也常被视作一种维数缩减技术，因为它把文档从高维的词项空间映射到低维潜在语义空间，去除了噪音

Latent Semantic Analysis

- 在向量空间模型中，一篇文档(document)可以表示为一个向 量，其中每个分量对应一个词项(term)，分量的值是词项在 文档中出现的频率(词项频率，term frequency)或者其它改进 后的词项权值。

$$\vec{d} = \left(tf(t_1), tf(t_2), \ldots, tf(t_M)\right)^{\mathrm{T}}$$

- 因此N篇文档组成的集合可以表示为一个$M \times N$的矩阵C，称作词项-文档矩阵(term-document matrix)，矩阵的行对应 着词项，矩阵的列对应着文档。

C	d_1	d_2	d_3	d_4	d_5	d_6
ship	1	0	1	0	0	0
boat	0	1	0	0	0	0
ocean	1	1	0	0	0	0
wood	1	0	0	1	1	0
tree	0	0	0	1	0	1

Latent Semantic Analysis

- 向量空间模型通过计算文档向量间的相似度来衡量两个 文档之间的相关性，常用的相似度为(夹角)余弦相似度

$$sim(\vec{d}_1, \vec{d}_2) = \frac{\vec{d}_1 \cdot \vec{d}_2}{|\vec{d}_1| \times |\vec{d}_2|}$$

- 例如:

$$sim(\vec{d}_2, \vec{d}_3) = \frac{0 \times 1 + 1 \times 0 + 1 \times 0 + 0 \times 0 + 0 \times 0}{|\vec{d}_2| \times |\vec{d}_3|} = 0$$

- 文档d2和文档d3没有相关性
- 是否合理?
- 建立在词项空间上文档表示没有考虑到同义词关系

Latent Semantic Analysis

- 词项-文档矩阵C不是对称矩阵
- 矩阵CC^T、C^TC是实值对称矩阵
- 矩阵CC^T的含义是任意两个词项的相似度
  - 元素代表着两个词项在文档中的共现次数
- 矩阵C^TC的含义是任意两个文档的相似度
  - 元素代表着两个文档中共同词项的数量
- 按照对称对角化定理，CC^T、C^TC可以分解
  - 令CC^T的正交单位特征向量组成的矩阵为U，则U是$M \times M$的矩阵
  - 令C^TC的正交单位特征向量组成的矩阵为V，则V是$N \times N$的矩阵

Latent Semantic Analysis

- **奇异值分解定理** 若$M \times N$的矩阵C的秩是r，那么可对C进行如下的奇异值分解(Singular Value Decomposition, SVD)：

$$C = U\Sigma V^T$$

其中U和V的含义如前述，且：

(1) CC^T的特征值为$\lambda_1, \lambda_2, \ldots, \lambda_r$，等于$C^TC$的特征值

(2) 对于$1 \le i \le r$，令$\lambda_i \ge \lambda_{i+1}$，$\sigma_i = \sqrt{\lambda_i}$，则对于$1 \le i \le r$,有$\Sigma_{ii} = \sigma_i$，此外$\Sigma$中的其他元素均为0

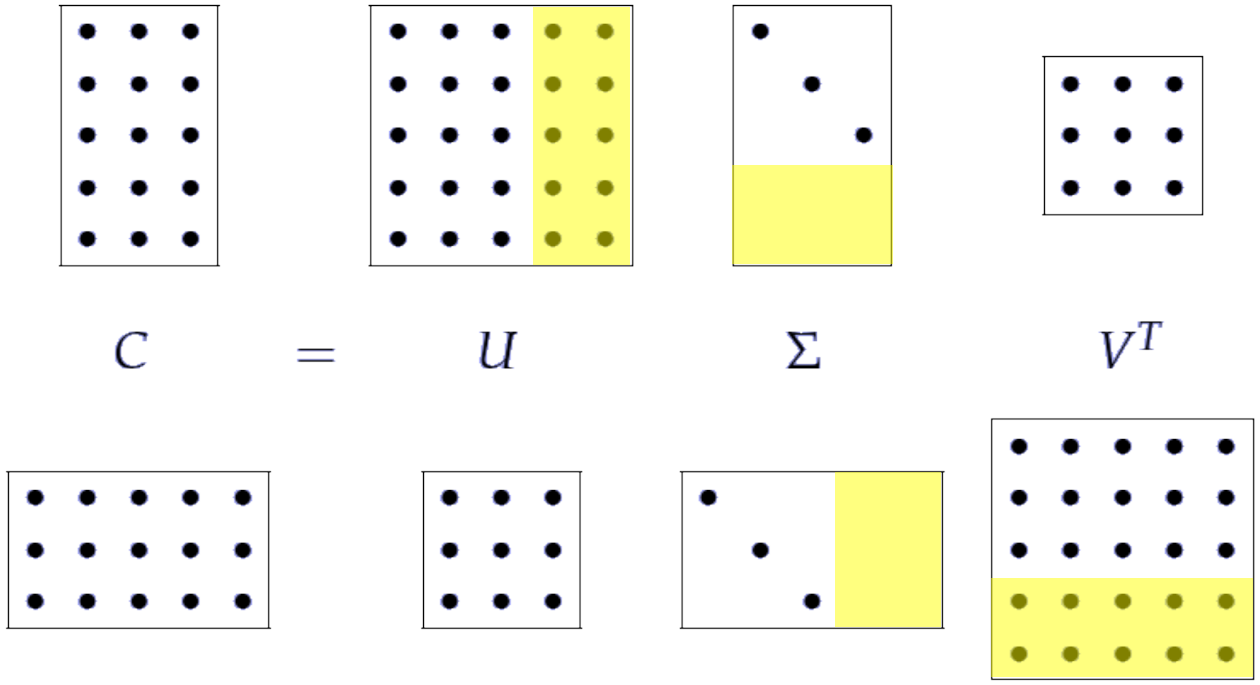
σ_i被称作矩阵C的奇异值

- 依据SVD定理，有

$$CC^{\mathrm{T}} = U\Sigma V^{\mathrm{T}} V\Sigma^{\mathrm{T}} U^{\mathrm{T}} = U\Sigma\Sigma^{\mathrm{T}} U^{\mathrm{T}}$$

$$C^{\mathrm{T}}C = V\Sigma^{\mathrm{T}} U^{\mathrm{T}} U\Sigma V^{\mathrm{T}} = V\Sigma^{\mathrm{T}}\Sigma V^{\mathrm{T}}$$

Latent Semantic Analysis



SVD图示($M > N$及$M < N$)
SVD截断表示

Latent Semantic Analysis

- 例:

$$\begin{bmatrix} 1 & 1 \\ 0 & 1 \\ 1 & 0 \end{bmatrix} = \begin{bmatrix} \sqrt{2/3} & 0 \\ 1/\sqrt{6} & 1/\sqrt{2} \\ 1/\sqrt{6} & -1/\sqrt{2} \end{bmatrix} \times \begin{bmatrix} \sqrt{3} & 0 \\ 0 & 1 \end{bmatrix} \times \begin{bmatrix} 1/\sqrt{2} & -1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix}$$

- **低秩逼近**(low-rank approximation): 寻求矩阵 C 的近似矩阵 C_k，且矩阵C_k的秩为k，并且$k \leq r$
- 所谓C_k逼近C，指的是二者的差矩阵的F范数最小，即下式的值最小，若用C_k代替C误差最小:

$$\|C - C_k\|_F = \sqrt{\sum_{i=1}^{M}\sum_{j=1}^{N}(C_{ij} - C_{kij})^2}$$

Latent Semantic Analysis

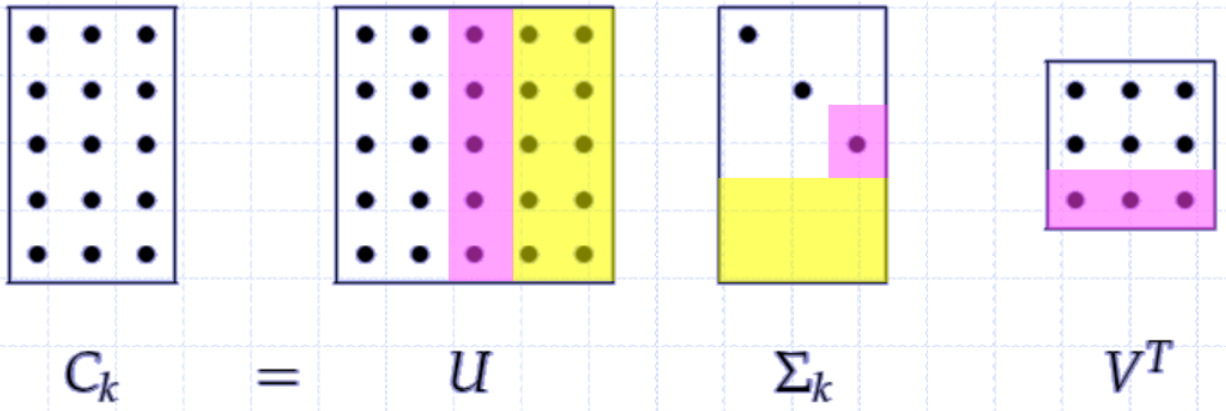
- SVD可以用来解决低秩逼近问题，给定秩r的矩阵C

  (1)构造C的SVD分解，有$C = U\Sigma V^{\mathrm{T}}$

  (2)将Σ中对角线上$r-k$个最小的奇异值置为0，得到Σ_k

  (3)计算$C_k = U\Sigma_k V^{\mathrm{T}}$，将$C_k$作为$C$的近似矩阵

  此时，逼近的误差为: $\sqrt{\Sigma_{i=k+1}^{r}\sigma_i^2}$



Latent Semantic Analysis

- LSA的核心在于将秩r的词项-文档矩阵C进行SVD分解，并寻求词项-文档矩阵的k秩逼近Ck
- 在实际问题中：$r \approx \min(M, N)$，r通常很大，此时可以选择一个较小的k，即$k \ll r$.
- 此时我们可以说，在进行潜在语义分析之前，文档被隐含表示成r维空间中的向量，而在潜在语义分析之后，文档被表示为k维空间中的向量，也就是潜在语义空间中的向量，向量的维数缩减为k维
- 维数k 可以被解释为隐含在文档集合中的话题数量，因此LSA可以被视作一种话题模型

Latent Semantic Analysis

- 对如下的词项-文档矩阵进行潜在语义分析

C	d_1	d_2	d_3	d_4	d_5	d_6
ship	1	0	1	0	0	0
boat	0	1	0	0	0	0
ocean	1	1	0	0	0	0
wood	1	0	0	1	1	0
tree	0	0	0	1	0	1

$C = U\Sigma V^T$

- 矩阵U(SVD词项矩阵)

	1	2	3	4	5
ship	−0.44	−0.30	0.57	0.58	0.25
boat	−0.13	−0.33	−0.59	0.00	0.73
ocean	−0.48	−0.51	−0.37	0.00	−0.61
voyage	−0.70	0.35	0.15	−0.58	0.16
trip	−0.26	0.65	−0.41	0.58	−0.09

Latent Semantic Analysis

- 矩阵Σ(奇异值矩阵)

2.16	0.00	0.00	0.00	0.00
0.00	1.59	0.00	0.00	0.00
0.00	0.00	1.28	0.00	0.00
0.00	0.00	0.00	1.00	0.00
0.00	0.00	0.00	0.00	0.39

- 矩阵V^T(SVD文档矩阵)

	d_1	d_2	d_3	d_4	d_5	d_6
1	−0.75	−0.28	−0.20	−0.45	−0.33	−0.12
2	−0.29	−0.53	−0.19	0.63	0.22	0.41
3	0.28	−0.75	0.45	−0.20	0.12	−0.33
4	0.00	0.00	0.58	0.00	−0.58	0.58
5	−0.53	0.29	0.63	0.19	0.41	−0.22

Latent Semantic Analysis

- 计算词项-文档矩阵的k秩逼近(令$k = 2$)

2.16	0.00	0.00	0.00	0.00
0.00	1.59	0.00	0.00	0.00
0.00	0.00	1.28	0.00	0.00
0.00	0.00	0.00	1.00	0.00
0.00	0.00	0.00	0.00	0.39

2.16	0.00	0.00	0.00	0.00
0.00	1.59	0.00	0.00	0.00
0.00	0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00	0.00

- 得到如下的$C2$ (在2维话题空间中的词项-文档矩阵)

C_2	d_1	d_2	d_3	d_4	d_5	d_6
ship	0.85	0.52	0.28	0.13	0.21	−0.08
boat	0.36	0.36	0.16	−0.20	−0.02	−0.18
ocean	1.01	0.72	0.36	−0.04	0.16	−0.21
wood	0.97	0.12	0.20	1.03	0.62	0.41
tree	0.12	−0.39	−0.08	0.90	0.41	0.49

Latent Semantic Analysis

C	d_1	d_2	d_3	d_4	d_5	d_6
ship	1	0	1	0	0	0
boat	0	1	0	0	0	0
ocean	1	1	0	0	0	0
wood	1	0	0	1	1	0
tree	0	0	0	1	0	1

C_2	d_1	d_2	d_3	d_4	d_5	d_6
ship	0.85	0.52	0.28	0.13	0.21	−0.08
boat	0.36	0.36	0.16	−0.20	−0.02	−0.18
ocean	1.01	0.72	0.36	−0.04	0.16	−0.21
wood	0.97	0.12	0.20	1.03	0.62	0.41
tree	0.12	−0.39	−0.08	0.90	0.41	0.49

- 在原始词项空间中

$$sim(\vec{d}_2, \vec{d}_3) = \frac{0 \times 1 + 1 \times 0 + 1 \times 0 + 0 \times 0 + 0 \times 0}{|\vec{d}_2| \times |\vec{d}_3|} = 0$$

- 在潜在语义空间中

$$sim(\vec{d}_2, \vec{d}_3) = \frac{0.52 \times 0.28 + 0.36 \times 0.16 + 0.72 \times 0.36 + 0.12 \times 0.20 + (-0.39) \times (-0.08)}{|\vec{d}_2| \times |\vec{d}_3|} = 0.939119$$

- $d2$ 和$d3$ 在潜在语义空间中具有很大相似性，这与我们的直观感觉相符，尽管二者中没有出现重叠的词项，但boat和ship是同义词，二者话题是相近的

Latent Semantic Analysis

- 在原始词项空间中，存在下面的问题
  - (1) 词项多义 (polysemy)现象引起文档相似度被高估 $sim_{true}(\vec{d}_1, \vec{d}_2) < sim(\vec{d}_1, \vec{d}_2)$
  - (2) 词项同义 (Synonymy)现象引起文档相似度被低估 $sim_{true}(\vec{d}_1, \vec{d}_2) > sim(\vec{d}_1, \vec{d}_2)$
- LSA通过空间映射，将话题接近的文档映射到相近的位置
- 存在SVD标准算法(如Lanczos算法)，复杂度较高
- LSA缺陷:

  C_k 中的元素缺直观解释，甚至可以是负值

  k的设定是经验性的

  非统计模型，缺统计学基础

概要

- Latent Semantic Analysis (LSA)
- **Probabilistic Topic Model**
  - Probabilistic Latent Semantic Analysis (pLSA)
  - Latent Dirichlet Allocation (LDA)

Probabilistic Topic Model

- LSA要点:
  (1) 基于词(项)-文档矩阵归纳语义信息
  (2) 基于维数缩减归纳语义信息
  (3) 文档和词(项)被视作欧式空间中的点进行计算
- 概率话题模型，(3)不成立
- 概率话题模型是生成式模型(generative model)
- 概率模型是混合模型(mixture model)
  - 混合模型中，分布表示为若干部件分布按照一定的比例进行组合
- 典型的概率话题模型
  - Probabilistic Latent Semantic Analysis
  - Latent Dirichlet Allocation

Probabilistic Topic Model

- 概率话题模型中
  (1) 文档是关于话题的分布, 不同文档拥有不同的话题比例$p(z)$
  (2) 话题是定义在词表上的概率分布$p(w|z)$，不同的话题是定义 在词表上的不同分布，与LSA不同，话题有着直观的物理解释
- 话题模型是生成模型, 文档是话题模型规定的概率过程的产物
  (1) 对每一个文档，首先选择一个话题分布$p(z)$
  (2) 对文档中的每一个词位，按照话题分布$p(z)$选择一个话题
  (3) 按照话题-词分布$p(w|z)$选择一个词
- 在话题模型中，文档中每个词都对应着一个隐含的话题，这些隐含的话题可以通过统计推断的技术从大量的文档集 合中提取得到

Probabilistic Topic Model

- 话题表示及提取示例(Mark Steyvers)
  - 基于TASA corpus (37000 text passages from educational material)
  - 共提取300个话题

Topic 247

word	prob.
DRUGS	.069
DRUG	.060
MEDICINE	.027
EFFECTS	.026
BODY	.023
MEDICINES	.019
PAIN	.016
PERSON	.016
MARIJUANA	.014
LABEL	.012
ALCOHOL	.012
DANGEROUS	.011
ABUSE	.009
EFFECT	.009
KNOWN	.008
PILLS	.008

Topic 5

word	prob.
RED	.202
BLUE	.099
GREEN	.096
YELLOW	.073
WHITE	.048
COLOR	.048
BRIGHT	.030
COLORS	.029
ORANGE	.027
BROWN	.027
PINK	.017
LOOK	.017
BLACK	.016
PURPLE	.015
CROSS	.011
COLORED	.009

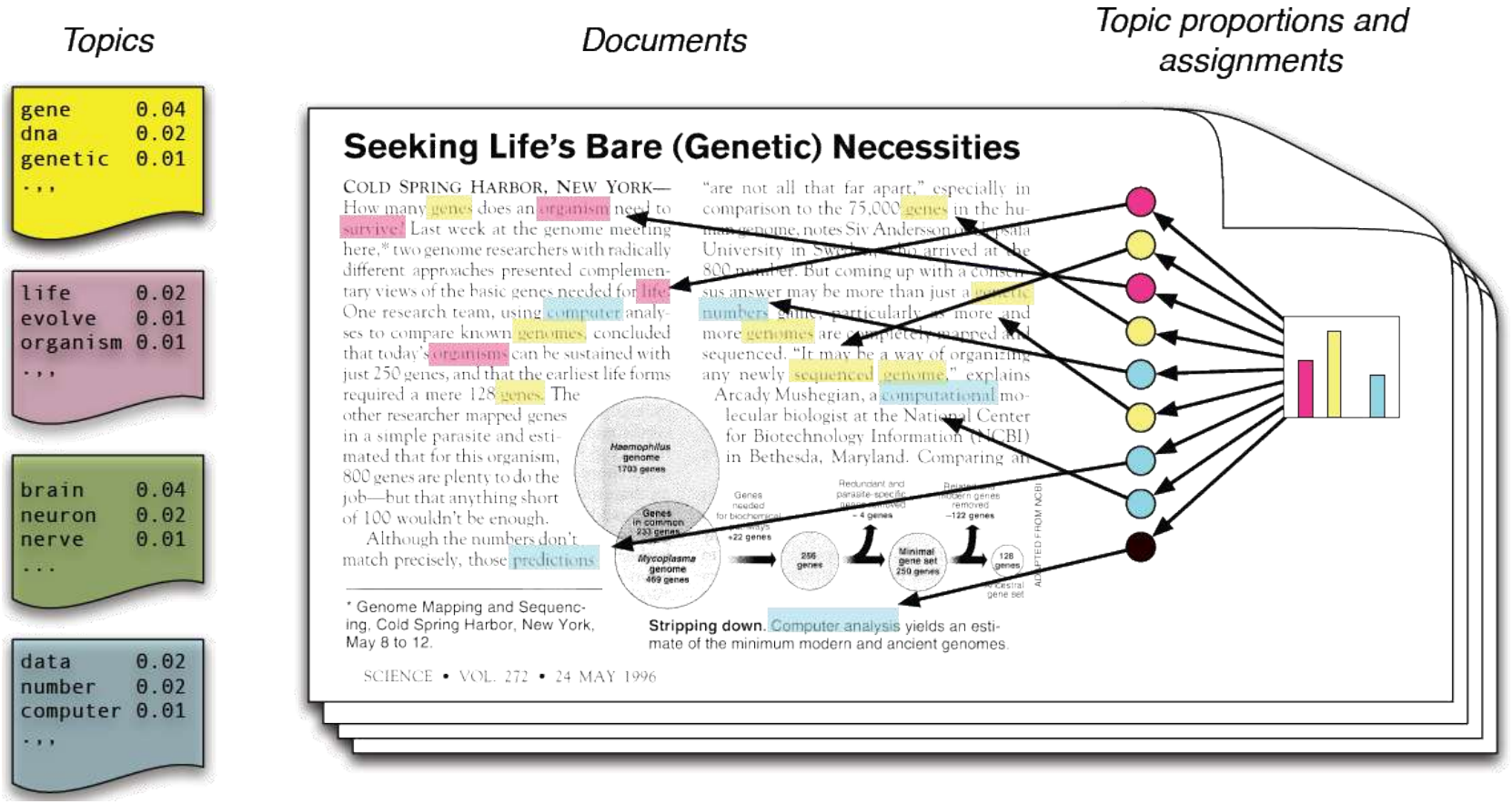
Topic 43

word	prob.
MIND	.081
THOUGHT	.066
REMEMBER	.064
MEMORY	.037
THINKING	.030
PROFESSOR	.028
FELT	.025
REMEMBERED	.022
THOUGHTS	.020
FORGOTTEN	.020
MOMENT	.020
THINK	.019
THING	.016
WONDER	.014
FORGET	.012
RECALL	.012

Topic 56

word	prob.
DOCTOR	.074
DR.	.063
PATIENT	.061
HOSPITAL	.049
CARE	.046
MEDICAL	.042
NURSE	.031
PATIENTS	.029
DOCTORS	.028
HEALTH	.025
MEDICINE	.017
NURSING	.017
DENTAL	.015
NURSES	.013
PHYSICIAN	.012
HOSPITALS	.011

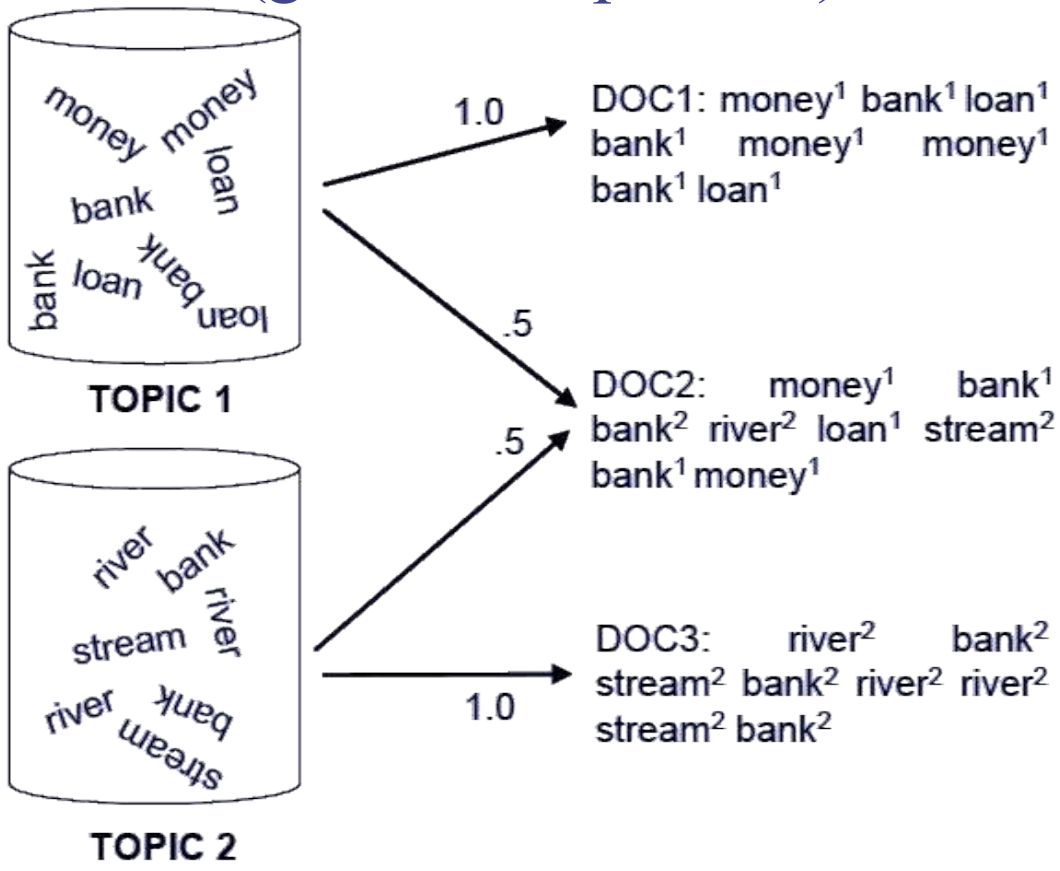
Probabilistic Topic Model



话题模型中的文档与话题

Probabilistic Topic Model

- 基于话题生成文档(generative process)

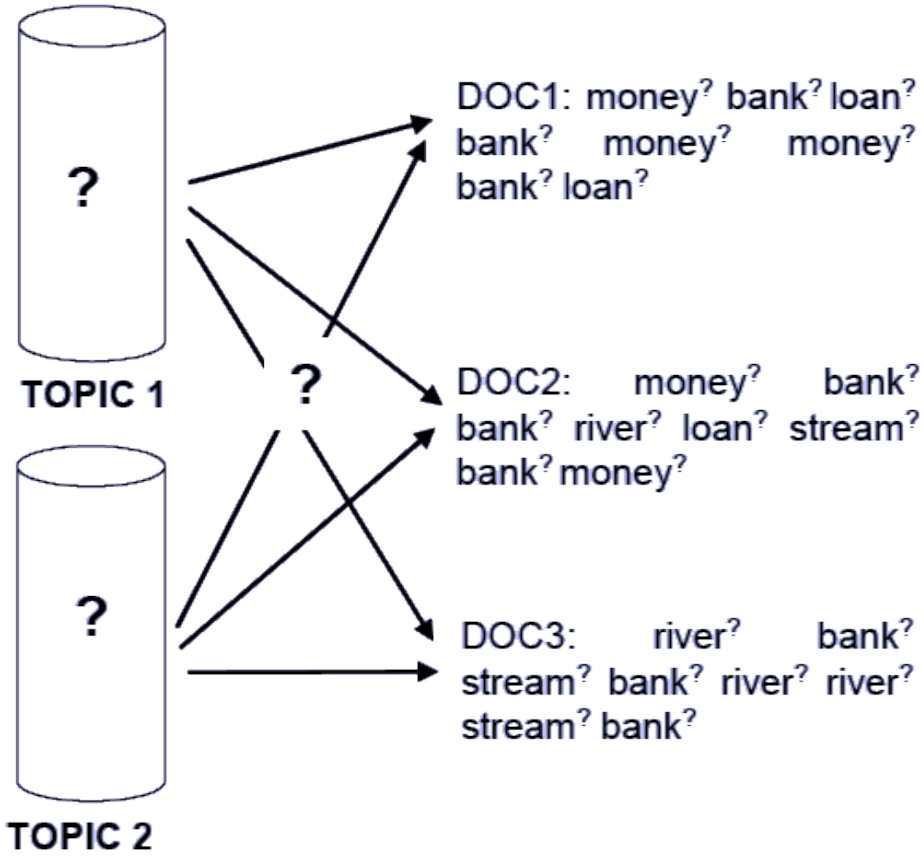


- TOPIC-1 money
- TOPIC-2 river

- DOC-1
  TOPIC-1 1
  TOPIC-2 0
- DOC-2
  TOPIC-1 0.5
  TOPIC-2 0.5
- DOC-3
  TOPIC-1 0
  TOPIC-2 1

Probabilistic Topic Model

- 基于文档推断模型(model inference)



寻求生成文档的最佳模型

- 每个文档中的话题比例?
- 每个话题中的词的分布?
- 生成每个词的隐含话题?

Probabilistic Topic Model

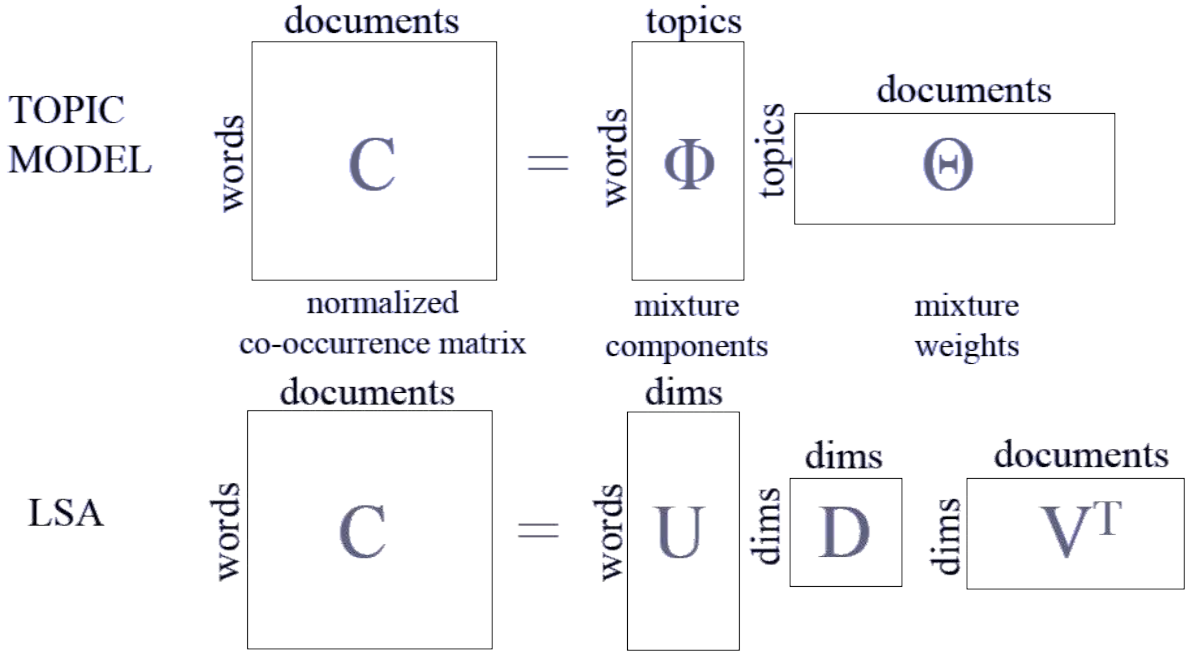
- 给定文档，令文档中的话题分布为$p(z)$，话题-词分布为$p(w|z)$，则有文档中第i个词的概率为：

$$p(w_i) = \sum_{j=1}^{T} p(w_i|z_i = j)p(z_j = j)$$

- 每个文档均可以视作一个关于词的概率分布，从而表示为一个M维向量，向量中的元素代表某词在该文档中的出现概率，若有D个文档，则文档可以表示为一个$M \times D$的矩阵C
- 因为每个话题也是关于词的概率分布，因此每个话题也可以表示为一个M维的向量，所有T个话题也可以表示为一个$M \times T$ 的矩阵Φ
- 按照话题模型，每个文档拥有不同的话题分布，因此文档也是关于话题的分布，文档也可以表示为一个T维话题向量，所有文档也可表示为一个$T \times D$的矩阵Θ

Probabilistic Topic Model

- 与LSA类似，概率话题模型也对应着一个矩阵分解解释



- 这说明，从概念上LSA与概率话题模型有一定相似性，但在概率话题模型，矩阵元素均为非负并拥有直观解释
- 在概率话题中，基于词的文档表示被转换成基于话题的文档表示，也可看作实现了表示维数的缩减

Probabilistic Topic Model

- 在概率话题模型中，同义词因表达话题相近，会出现在同一个话题中，多义词因其表达话题不同而出现在不同的话题-词分布中

Topic 77

word	prob.
MUSIC	.090
DANCE	.034
SONG	.033
PLAY	.030
SING	.026
SINGING	.026
BAND	.026
PLAYED	.023
SANG	.022
SONGS	.021
DANCING	.020
PIANO	.017
PLAYING	.016
RHYTHM	.015
ALBERT	.013
MUSICAL	.013

Topic 82

word	prob.
LITERATURE	.031
POEM	.028
POETRY	.027
POET	.020
PLAYS	.019
POEMS	.019
PLAY	.015
LITERARY	.013
WRITERS	.013
DRAMA	.012
WROTE	.012
POETS	.011
WRITER	.011
SHAKESPEARE	.010
WRITTEN	.009
STAGE	.009

Topic 166

word	prob.
PLAY	.136
BALL	.129
GAME	.065
PLAYING	.042
HIT	.032
PLAYED	.031
BASEBALL	.027
GAMES	.025
BAT	.019
RUN	.019
THROW	.016
BALLS	.015
TENNIS	.011
HOME	.010
CATCH	.010
FIELD	.010

- TOPIC-77 playing music
- TOPIC-82 theater play
- TOPIC-166 playing game

Probabilistic Topic Model

A **Play**[082] is written[082] to be performed[082] on a stage[082] before a live[093] audience[082] or before motion[270] picture[004] or television[004] cameras[004] (for later[054] viewing[004] by large[202] audiences[082]). A Play[082] is written[082] because playwrights[082] have something ...

He was listening[077] to music[077] coming[009] from a passing[043] riverboat. The music[077] had already captured[006] his heart[157] as well as his ear[119]. It was jazz[077]. Bix beiderbecke had already had music[077] lessons[077]. He wanted[268] to play the cornet. And he wanted[268] to play jazz[077]...

Jim[296] **plays**[166] the game[166] Jim[296] likes[081] the game[166] for one. The game[166] book[254] helps[081] jim[296]. Don[180] comes[040] into the house[038]. Don[180] and jim[296] read[254] the game[166] book[254]. The boys[020] see a game[166] for two. The two boys[020] **play**[166] the game[166]....

概要

- Latent Semantic Analysis (LSA)
- Probabilistic Topic Model
  - **Probabilistic Latent Semantic Analysis (pLSA)**
  - Latent Dirichlet Allocation (LDA)

Probabilistic Latent Semantic Analysis

- pLSA的提出者是Thomas Hofmann，发表时间是1999年 pLSA是一种概率话题模型
  - 文档视作话题的混合模型
  - 话题视作词的不同分布
- pLSA是一种生成模型
- pLSA是一种有向图模型
- 模型求解采用EM算法

Probabilistic Latent Semantic Analysis

- N个文档组成的文档集合的生成过程

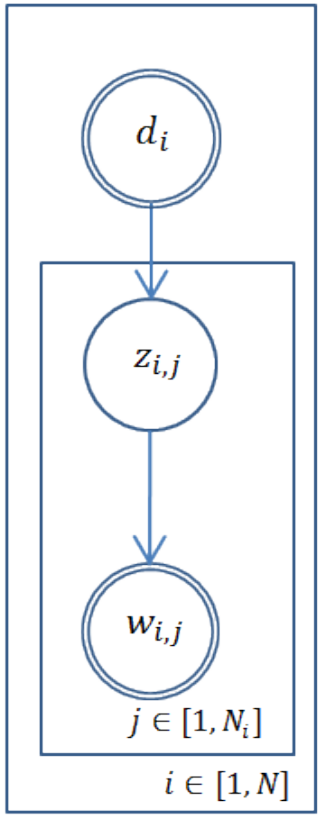
(1) 重复N次

(2) 按照$p(d)$选择一篇文档

重复N_d次

- 按照$p(z|d)$选择一个话题
- 按照$p(w|z)$选择文档中的词

(3) 结束



Probabilistic Latent Semantic Analysis

- 在pLSA中，有：

$$p(d_i, w_j) = p(d_i)p(w_j|d_i)$$

$$p(w_j|d_i) = \sum_{t=1}^{T} p(w_j, z_t|d_i) = \sum_{t=1}^{T} p(w_j|z_t)p(z_t|d_i)$$

- d_i和w_j 关于z_t条件独立
- pLSA遵循词袋假设，(d_i, w_j)的生成互相独立，与顺序无关
- pLSA的任务：基于给定的文档集合，估计参数
  - $p(w_j|z_t)$
  - $p(z_t|d_i)$
- 运用最大似然估计的原则

Probabilistic Latent Semantic Analysis

- 定义似然函数、对数似然函数

$$L(D) = \prod_{i=1}^{N} \prod_{j=1}^{M} p(d_i, w_j)^{n(d_i, w_j)}$$

$$\log L(D) = \sum_{i=1}^{N} \sum_{j=1}^{M} n(d_i, w_j) \log p(d_i, w_j)$$

- $n(d_i, w_j)$ 是w_j在文档d_i中的出现频率
- 根据最大似然估计的原则，参数求解如下：

$$\hat{p} = \underset{p}{\operatorname{argmax}} \log L(D)$$

- 因为存在隐藏变量z，因此可以采用EM算法求解

概要

- Latent Semantic Analysis (LSA)
- Probabilistic Topic Model
  - Probabilistic Latent Semantic Analysis (pLSA)
  - **Latent Dirichlet Allocation (LDA)**

Latent Dirichlet Allocation

- LDA的提出者是David Blei，发表时间是2003年
- LDA可视作是对pLSA的改进，是一种概率话题模型
  - 文档视作话题的混合模型
  - 话题视作词的不同分布
- LDA将贝叶斯估计的思想引入话题模型
- LDA也是有向图模型
- 模型求解通常使用Gibbs采样算法

Latent Dirichlet Allocation

- 参数估计问题，给定观察数据$\mathcal{X}$，求解模型参数ϑ
- 其中

$$p(\vartheta|\mathcal{X}) = \frac{p(\mathcal{X}|\vartheta) \cdot p(\vartheta)}{p(\mathcal{X})}$$

$p(\vartheta)$ – prior　　$p(\mathcal{X}|\vartheta)$ – likelihood

$p(\mathcal{X})$ –evidence　　$p(\vartheta|\mathcal{X})$ – posterior

- 在最大似然估计中，使用了其中$p(\mathcal{X}|\vartheta)$，即

$$\hat{\vartheta}_{\mathrm{ML}} = \underset{\vartheta}{\operatorname{argmax}}\, p(\mathcal{X}|\vartheta)$$

- 贝叶斯分析中，模型参数ϑ被视作随机变量，拥有自己的概率分布，参数估计也就是估计参数的后验分布$p(\vartheta|\mathcal{X})$

Latent Dirichlet Allocation

- 在贝叶斯估计中，允许以$p(\vartheta)$的方式加入参数ϑ的先验分布，表达关于参数ϑ先验知识
- 例如，给定一个硬币抛掷序列，估计硬币正面朝上的概率分布，我们可以加入硬币正面朝上概率的先验分布
- 理论上，先验分布$p(\vartheta)$可以选择任何适用的分布，然而这会引起计算问题，通常的做法是根据观察数据的分布选择相应的共轭分布
- $p(x|\vartheta)$的共轭分布$p(\vartheta)$指的是选择$p(\vartheta)$得到的参数的后验分布$p(\vartheta|X)$ 与$p(\vartheta)$具有同样的分布函数，只是分布参数不同
- 使用共轭先验可以简化参数的估计问题

共轭先验

- In Bayesian probability theory, if the posterior distribution $\mathrm{p}(\theta|x)$ is in the same probability distribution family as the prior probability distribution $p(\theta)$, the prior and posterior are then called **conjugate distributions,** and the prior is called a **conjugate prior** for the likelihood function $p(x|\theta)$
- A conjugate prior is an algebraic convenience, giving a closed-form expression for the posterior; otherwise, numerical integration may be necessary. Further, conjugate priors may give intuition by more transparently showing how a likelihood function updates a prior distribution.

共轭先验

Binominal: $f(x) = \binom{n}{x}\theta^x(1-\theta)^{n-x}$

Beta: $g(\theta) = \frac{1}{B(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$

$$
\begin{aligned}
p(\theta|x) &= \frac{p(x|\theta)p(\theta)}{\int_\theta p(x|\theta)p(\theta)d\theta} \\
&= \frac{\binom{n}{x}\theta^x(1-\theta)^{n-x}\cdot\frac{1}{B(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}}{\int_\theta \binom{n}{x}\theta^x(1-\theta)^{n-x}\cdot\frac{1}{B(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}d\theta} \\
&= \frac{\frac{C_n^x}{B(\alpha,\beta)}\theta^{x+\alpha-1}(1-\theta)^{n-x+\beta-1}}{\frac{C_n^x}{B(\alpha,\beta)}\int_0^1 \theta^{x+\alpha-1}(1-\theta)^{n-x+\beta-1}d\theta} \\
&= Beta(x+\alpha, n-x+\beta)
\end{aligned}
$$

$$\int_0^1 \theta^{x+\alpha-1}(1-\theta)^{n-x+\beta-1}d\theta = B(x+\alpha, n-x+\beta)$$

共轭先验

When the likelihood function is a discrete distribution

Likelihood	Model parameters	Conjugate prior distribution	Prior hyperparameters	Posterior hyperparameters[note 1]
Bernoulli	p (probability)	Beta	$\alpha,\ \beta \in \mathbb{R}$	$\alpha + \sum_{i=1}^{n} x_i,\ \beta + n - \sum_{i=1}^{n} x_i$
Binomial with known number of trials, m	p (probability)	Beta	$\alpha,\ \beta \in \mathbb{R}$	$\alpha + \sum_{i=1}^{n} x_i,\ \beta + \sum_{i=1}^{n} N_i - \sum_{i=1}^{n} x_i$
Negative binomial with known failure number, r	p (probability)	Beta	$\alpha,\ \beta \in \mathbb{R}$	$\alpha + rn,\ \beta + \sum_{i=1}^{n} x_i$
Poisson	λ (rate)	Gamma	$k,\ \theta \in \mathbb{R}$	$k + \sum_{i=1}^{n} x_i,\ \frac{\theta}{n\theta + 1}$
			$\alpha,\ \beta$[note 4]	$\alpha + \sum_{i=1}^{n} x_i,\ \beta + n$
Categorical	$\boldsymbol{p}$ (probability vector), k (number of categories; i.e., size of $\boldsymbol{p}$)	Dirichlet	$\boldsymbol{\alpha} \in \mathbb{R}^k$	$\boldsymbol{\alpha} + (c_1, \ldots, c_k)$, where c_i is the number of observations in category i
Multinomial	$\boldsymbol{p}$ (probability vector), k (number of categories; i.e., size of $\boldsymbol{p}$)	Dirichlet	$\boldsymbol{\alpha} \in \mathbb{R}^k$	$\boldsymbol{\alpha} + \sum_{i=1}^{n} \mathbf{x}_i$

共轭先验

When likelihood function is a continuous distribution [edit]

Likelihood	Model parameters	Conjugate prior distribution	Prior hyperparameters	Posterior hyperparameters[note 1]	Interpretation of hyperparameters	Posterior predictive[note 5]
Normal with known variance σ^2	μ (mean)	Normal	$\mu_0,\ \sigma_0^2$	$\frac{1}{\frac{1}{\sigma_0^2}+\frac{n}{\sigma^2}}\left(\frac{\mu_0}{\sigma_0^2}+\frac{\sum_{i=1}^n x_i}{\sigma^2}\right),\left(\frac{1}{\sigma_0^2}+\frac{n}{\sigma^2}\right)^{-1}$	mean was estimated from observations with total precision (sum of all individual precisions) $1/\sigma_0^2$ and with sample mean μ_0	$\mathcal{N}(\tilde{x}\vert\mu_0',\sigma_0^{2\prime}+\sigma^2)$[4]
Normal with known precision τ	μ (mean)	Normal	$\mu_0,\ \tau_0^{-1}$	$\frac{\tau_0\mu_0+\tau\sum_{i=1}^n x_i}{\tau_0+n\tau},\ (\tau_0+n\tau)^{-1}$	mean was estimated from observations with total precision (sum of all individual precisions) τ_0 and with sample mean μ_0	$\mathcal{N}\left(\tilde{x}\mid\mu_0',\frac{1}{\tau_0'}+\frac{1}{\tau}\right)$[4]
Normal with known mean μ	σ^2 (variance)	Inverse gamma	$\alpha,\ \beta$ [note 6]	$\alpha+\frac{n}{2},\ \beta+\frac{\sum_{i=1}^n(x_i-\mu)^2}{2}$	variance was estimated from 2α observations with sample variance β/α (i.e. with sum of squared deviations 2β, where deviations are from known mean μ)	$t_{2\alpha'}(\tilde{x}\vert\mu,\sigma^2=\beta'/\alpha')$[4]
Normal with known mean μ	σ^2 (variance)	Scaled inverse chi-squared	$\nu,\ \sigma_0^2$	$\nu+n,\ \frac{\nu\sigma_0^2+\sum_{i=1}^n(x_i-\mu)^2}{\nu+n}$	variance was estimated from ν observations with sample variance σ_0^2	$t_{\nu'}(\tilde{x}\vert\mu,\sigma_0^{2\prime})$[4]
Normal with known mean μ	τ (precision)	Gamma	$\alpha,\ \beta$[note 4]	$\alpha+\frac{n}{2},\ \beta+\frac{\sum_{i=1}^n(x_i-\mu)^2}{2}$	precision was estimated from 2α observations with sample variance β/α (i.e. with sum of squared deviations 2β, where deviations are from known mean μ)	$t_{2\alpha'}(\tilde{x}\mid\mu,\sigma^2=\beta'/\alpha')$[4]
Normal[note 7]	μ and σ^2 Assuming exchangeability	Normal-inverse gamma	$\mu_0,\ \nu,\ \alpha,\ \beta$	$\frac{\nu\mu_0+n\bar{x}}{\nu+n},\ \nu+n,\ \alpha+\frac{n}{2},$ $\beta+\frac{1}{2}\sum_{i=1}^n(x_i-\bar{x})^2+\frac{n\nu}{\nu+n}\frac{(\bar{x}-\mu_0)^2}{2}$ • $\bar{x}$ is the sample mean	mean was estimated from ν observations with sample mean μ_0; variance was estimated from 2α observations with sample mean μ_0 and sum of squared deviations 2β	$t_{2\alpha'}\left(\tilde{x}\mid\mu',\frac{\beta'(\nu'+1)}{\nu'\alpha'}\right)$[4]

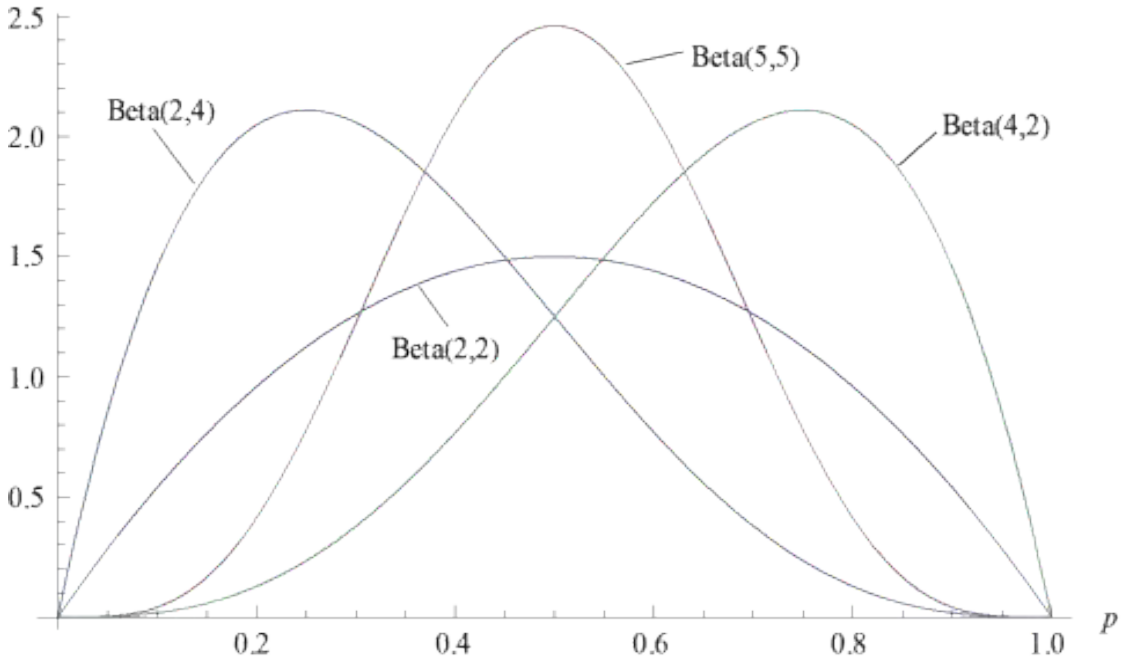
Latent Dirichlet Allocation

- Important prior-likelihood pairs
  - Bernoulli distribution --- Beta distribution
  - Binomial distribution --- Beta distribution
  - General discrete distribution --- Dirichlet distribution
  - Multinomial distribution --- Dirichlet distribution
  - Negative binominal distribution --- Beta distribution
- Beta distribution ---- $\mathrm{Beta}(\mathrm{p}|\alpha,\beta)$
  - 密度函数如下

$$f(p)=\begin{cases}\dfrac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}p^{\alpha-1}(1-p)^{\beta-1} & 0<p<1\\ 0 & others\end{cases}$$

Latent Dirichlet Allocation

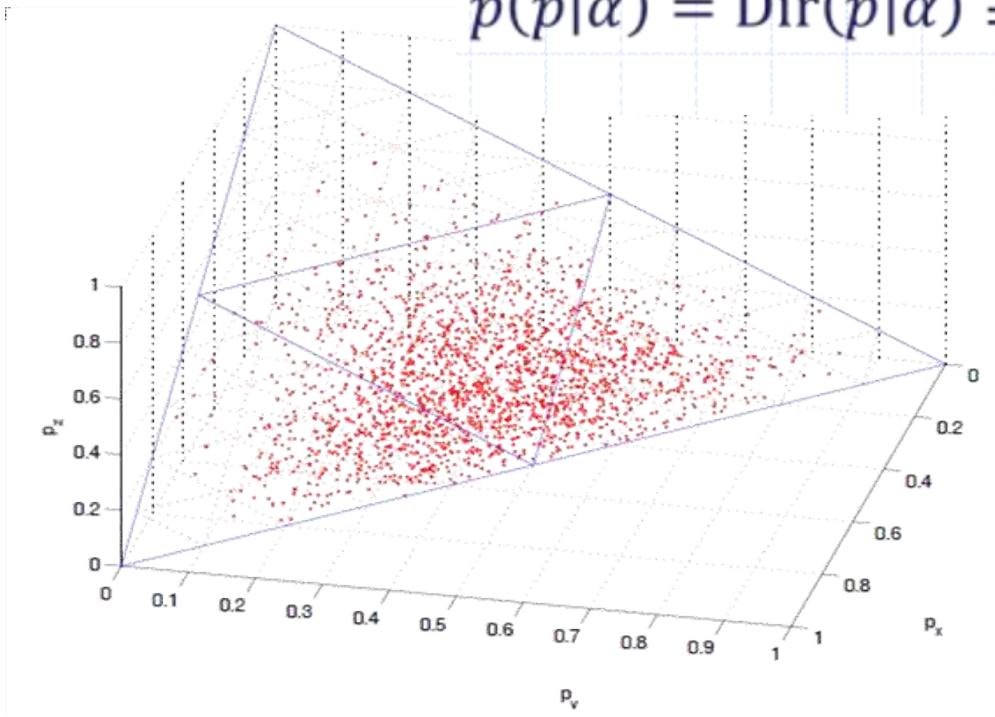
- 当参数 $\alpha = \beta$ 时，Beta分布的密度函数关于0.5对称，当 α 增大时，密度函数越来越集中在0.5附近的区域，当参数$\beta > \alpha$时，密度向左偏，当参数$\beta < \alpha$时，密度向右偏



Latent Dirichlet Allocation

- Dirichlet distribution --- $\mathrm{Dir}(\vec{p}|\vec{\alpha})$
  - Dirichlet distribution generalizes the Beta distribution from 2 to K dimensions

$$p(\vec{p}|\vec{\alpha}) = \mathrm{Dir}(\vec{p}|\vec{\alpha}) \triangleq \frac{\Gamma(\sum_{k=1}^{K}\alpha_k)}{\prod_{k=1}^{K}\Gamma(\alpha_k)}\prod_{k=1}^{K}p_k^{\alpha_k-1}$$



$$E(\vec{p}|\vec{\alpha}) = \frac{\vec{\alpha}}{\sum_i \vec{\alpha}_i}$$

Dir(4,4,2)

Latent Dirichlet Allocation

- 若使用 Beta$(p|\alpha,\beta)$ 作为二项分布 Bin$(n|p)$ 的先验分布， 则参数的后验分布是Beta$(p|n1+\alpha, n0+\beta)$
- 若使用Dir$(p\ |\alpha\ )$作为多项分布Mult$(n|p\ )$的先验分布， 则参数的后验分布是Dir$(p|n+\alpha\ )$ *****
- 在概率话题模型中
  - 文档中的话题分布是多项分布
  - 话题-词分布也是多项分布
- 采用贝叶斯估计，分别为上述两个分布引入相应的共轭先验分布—Dirichlet分布

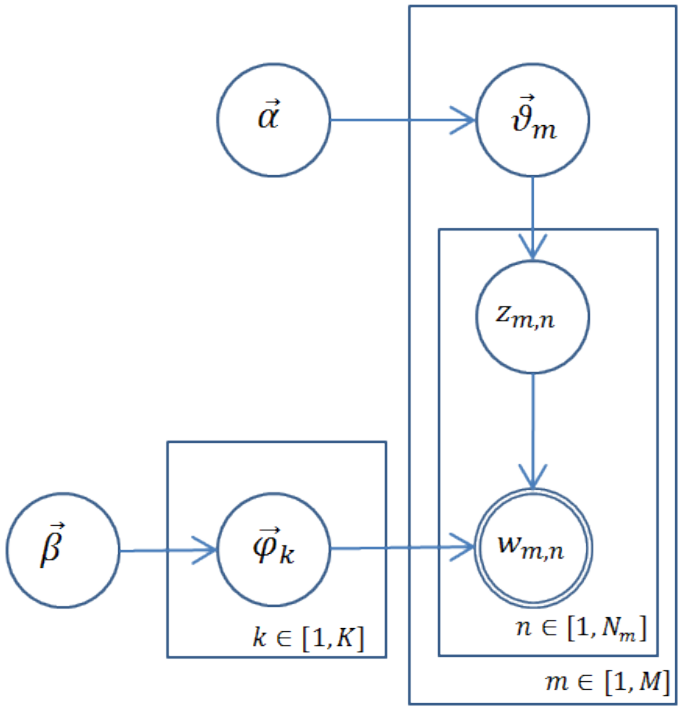
Latent Dirichlet Allocation

- 令$p(w|z=k)=\overrightarrow{\varphi_k}$代表第$k$个话题-词分布，$p(z|d=m)=\overrightarrow{\theta_m}$代表第$m$个文档中的话题分布，LDA参数可以表示为

$$\underline{\Phi}=\{\vec{\varphi}_k\}_{k=1}^{K}$$

$$\underline{\Theta}=\{\vec{\vartheta}_m\}_{m=1}^{M}$$

- 在LDA中，$\overrightarrow{\varphi_k}, \overrightarrow{\theta_m}$都是隐变量



Latent Dirichlet Allocation

- 基于LDA模型的文档生成过程
- //topic plate
- **for** all topics $k \in [1, K]$ **do**
  - sample mixture component $\overrightarrow{\varphi_k} \sim Dir(\vec{\beta})$
- //document plate
- **for** all documents $m \in [1, M]$ **do**
  - sample mixture proportion $\overrightarrow{\theta_m} \sim Dir(\vec{\alpha})$
  - sample document length $N_m \sim Poiss(\xi)$
  - //word plate
  - **for** all words $n \in [1, Nm]$ in document m **do**
    - sample topic index $z_{m,n} = Multi(\overrightarrow{\theta_m})$
    - sample term for word $w_{m,n} \sim Multi(\vec{\varphi}_{z_{m,n}})$

- 基于LDA模型的文档生成过程
- //topic plate
- **for** all topics $k \in [1, K]$ **do**
  - sample mixture component $\overrightarrow{\varphi_k} \sim Dir(\vec{\beta})$
- //document plate
- **for** all documents $m \in [1, M]$ **do**
  - sample mixture proportion $\overrightarrow{\theta_m} \sim Dir(\vec{\alpha})$
  - sample document length $N_m \sim Poiss(\xi)$
  - //word plate
  - **for** all words $n \in [1, Nm]$ in document m **do**
    - sample topic index $z_{m,n} = Multi(\overrightarrow{\theta_m})$
    - sample term for word $w_{m,n} \sim Multi(\vec{\varphi}_{z_{m,n}})$

Latent Dirichlet Allocation

- 对LDA而言，推断问题体现为计算如下的后验分布 $p(\vec{z}, \underline{\Theta}, \underline{\Phi} | \vec{w}, \vec{\alpha}, \vec{\beta})$
- 精确计算该后验分布是困难的
- 利用Gibbs采样算法生成参数的样本
  - Gibbs采样属于Markov Chain Monte Carlo方法
  - 利用马尔可夫链收敛于稳定分布的规律
  - 构造参数的马尔可夫链，基于参数的条件分布对z, Θ, Φ 进行采样或基于下面的公式对z 进行Collapsed采样

$$p(\vec{z} | \vec{w}, \vec{\alpha}, \vec{\beta}) = \iint p(\vec{z}, \underline{\Theta}, \underline{\Phi} | \vec{w}, \vec{\alpha}, \vec{\beta}) d\underline{\Theta} d\underline{\Phi}$$

LDA Gibbs采样公式

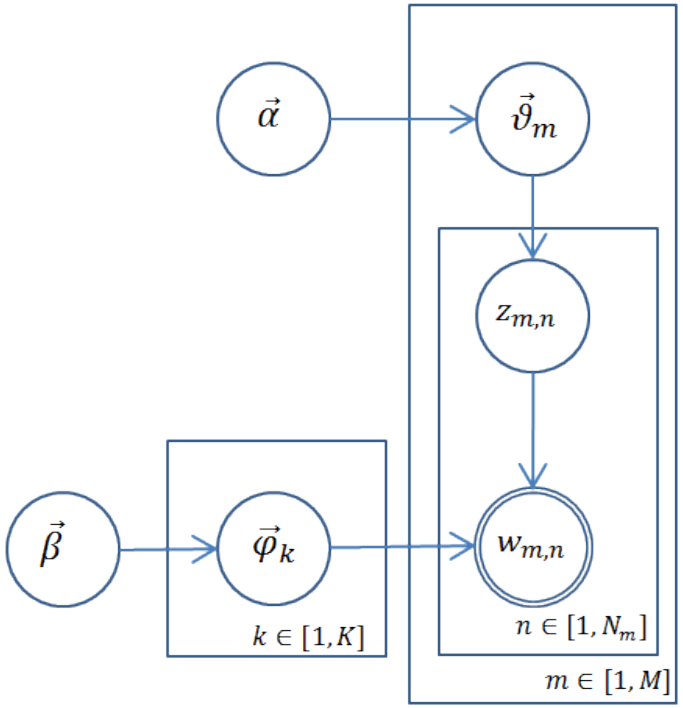
- 语料库中的第i个词对应的主题我们记为z_i，其中$i=(m,n)$是一个二维下标，即语料库中第i个词对应于第m篇文档的第n个词，我们用$\neg i$表示去除下标为i的词。
- 按照 Gibbs Sampling 的要求，我们要求得任一坐标轴i对应的条件分布$p(z_i=k|\vec{z}_{\neg i},\vec{W})$，注意这是一个离散型概率分布的抽样。由贝叶斯法则可得（条件概率正比于联合概率）：
- $p(z_i=k|\vec{z}_{\neg i},\vec{W}) \propto p(z_i=k,w_i=t|\vec{z}_{\neg i},\overrightarrow{W_{\neg i}})$

LDA Gibbs采样公式

- 由于 $z_i = k, w_i =$ t只涉及到第 m 篇文档和第 k 个主题，所以只会涉及到如下两个 Dirichlet-Multinomial 共轭结构

$$\vec{\alpha} \rightarrow \vec{\Theta}_m \rightarrow \vec{z}_m$$

$$\vec{\beta} \rightarrow \vec{\Phi}_k \rightarrow \vec{w}_k$$



LDA Gibbs采样公式

- 去掉了语料中第i个词对应的(z_i, w_i)并不会改变上面两个共轭结构，只是会减少对应的计数。所以$\vec{\theta}_m, \vec{\varphi}_k$的后验分布都还是狄利克雷分布：

$$p(\vec{\Theta}_m|\vec{Z}_{\neg i}, \vec{W}_{\neg i}) = Dir(\vec{\Theta}_m|\vec{n}_{m,\neg i} + \vec{\alpha})$$
$$p(\vec{\Phi}_k|\vec{Z}_{\neg i}, \vec{W}_{\neg i}) = Dir(\vec{\Phi}_k|\vec{n}_{k,\neg i} + \vec{\beta})$$

LDA Gibbs采样公式

- Gibbs Sampling 公式的推导：

$$
\begin{aligned}
p(z_i = k|\vec{Z}_{\neg i}, \vec{W}) \propto & p(z_i = k, w_i = t|\vec{Z}_{\neg i}, \vec{W}_{\neg i}) \\
= & \int p(z_i = k, w_i = t, \vec{\theta}_m, \vec{\Phi}_k|\vec{Z}_{\neg i}, \vec{W}_{\neg i})\mathrm{d}\vec{\theta}_m\mathrm{d}\vec{\Phi}_k \\
= & \int p(z_i = k, \vec{\theta}_m|\vec{Z}_{\neg i}, \vec{W}_{\neg i}) \cdot p(w_i = t, \vec{\Phi}_k|\vec{Z}_{\neg i}, \vec{W}_{\neg i})\mathrm{d}\vec{\theta}_m\mathrm{d}\vec{\Phi}_k \\
= & \int p(z_i = k|\vec{\theta}_m)p(\vec{\theta}_m|\vec{Z}_{\neg i}, \vec{W}_{\neg i}) \cdot p(w_i = t|\vec{\Phi}_k)p(\vec{\Phi}_k|\vec{Z}_{\neg i}, \vec{W}_{\neg i})\mathrm{d}\vec{\theta}_m\mathrm{d}\vec{\Phi}_k \\
= & \int p(z_i = k|\vec{\theta}_m)Dir(\vec{\theta}_m|\vec{n}_{m,\neg i} + \vec{\alpha})\mathrm{d}\vec{\theta}_m \cdot \int p(w_i = t|\vec{\Phi}_k)Dir(\vec{\Phi}_k|\vec{n}_{k,\neg i} + \vec{\beta})\mathrm{d}\vec{\Phi}_k \\
= & \int \theta_{mk}Dir(\vec{\theta}_m|\vec{n}_{m,\neg i} + \vec{\alpha})\mathrm{d}\vec{\theta}_m \cdot \int \phi_{kt}Dir(\vec{\Phi}_k|\vec{n}_{k,\neg i} + \vec{\beta})\mathrm{d}\vec{\Phi}_k \\
= & E(\theta_{mk}) \cdot E(\phi_{kt}) \\
= & \hat{\theta}_{mk} \cdot \hat{\phi}_{kt}
\end{aligned}
$$

LDA Gibbs采样公式

- 根据狄利克雷分布的期望公式有：

$$\hat{\theta}_{mk} = \frac{n_{m,\neg i}^{(k)} + \alpha_k}{\sum_{k=1}^{K} n_{m,\neg i}^{(k)} + \alpha_k}$$

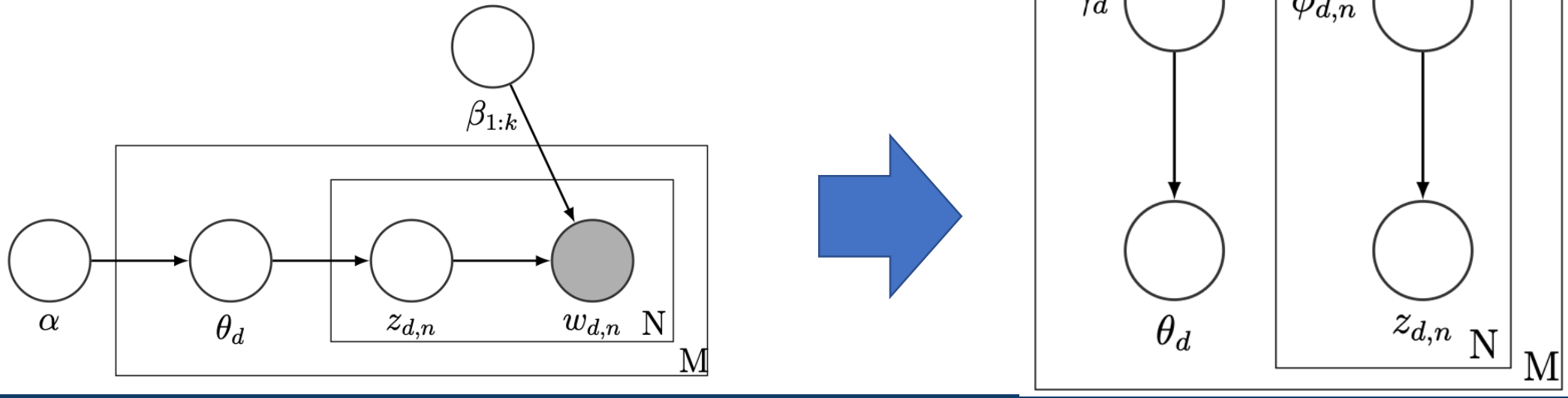
$$\hat{\phi}_{kt} = \frac{n_{k,\neg i}^{(t)} + \beta_t}{\sum_{t=1}^{V} n_{k,\neg i}^{(t)} + \beta_t}$$

- 所以 LDA 模型的采样公式为：

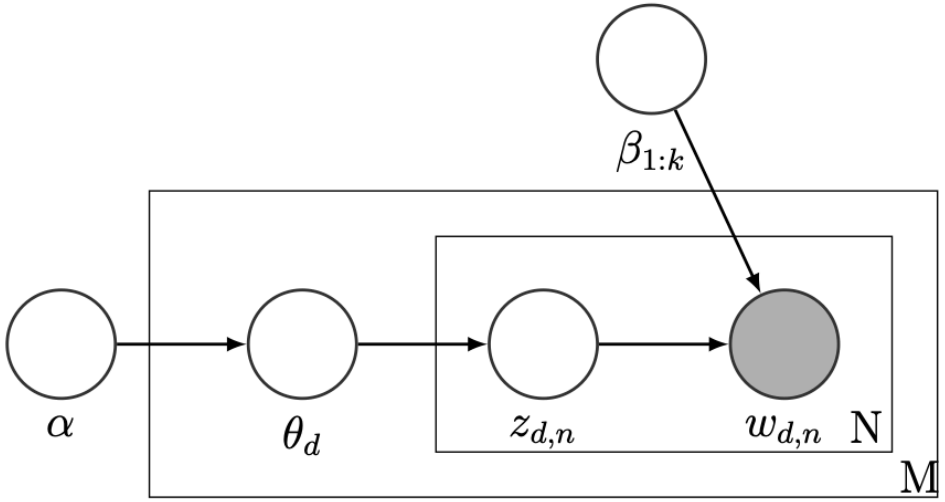
$$p(z_i = k | \vec{Z}_{\neg i}, \vec{W}) \propto \frac{n_{m,\neg i}^{(k)} + \alpha_k}{\sum_{k=1}^{K} n_{m,\neg i}^{(k)} + \alpha_k} \cdot \frac{n_{k,\neg i}^{(t)} + \beta_t}{\sum_{t=1}^{V} n_{k,\neg i}^{(t)} + \beta_t}$$

LDA 变分推断（Variational Inference）

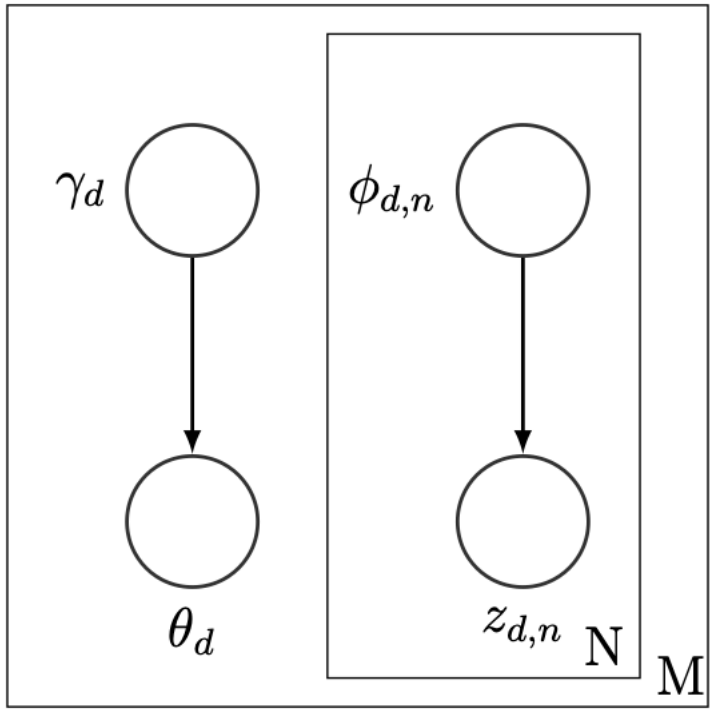
- 变分推断基本思想：用一个简单的凸分布去模拟一个实际分布的对数似然的下界
- **变分参数**描述了用于确定对数似然下界的更简单分布族，并优化以创建最紧密可能的下界。
- 引入变分参数γ，ϕ



LDA 变分推断（Variational Inference）



拉近两个分布的距离



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

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

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

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

LDA 变分推断（Variational Inference）

- KL divergence

$$D_{KL}(q||p) = \int q(z) \log \frac{q(z)}{p(z)} dz$$

- 所以变分推断解LDA如下：

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

LDA 变分推断（Variational Inference）

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

$$D(q(\theta, \mathbf{z}|\gamma, \phi)||p(\theta, \mathbf{z}|\mathbf{w}, \alpha, \beta)) = \int_{\theta,\mathbf{z}} q(\theta, \mathbf{z}|\gamma, \phi) \log \frac{q(\theta, \mathbf{z}|\gamma, \phi)}{p(\theta, \mathbf{z}|\mathbf{w}, \alpha, \beta)}$$

$$= \int_{\theta,\mathbf{z}} q(\theta, \mathbf{z}|\gamma, \phi) \log q(\theta, \mathbf{z}|\gamma, \phi) - \int_{\theta,\mathbf{z}} q(\theta, \mathbf{z}|\gamma, \phi) \log p(\theta, \mathbf{z}|\mathbf{w}, \alpha, \beta)$$

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

与待优化的参数无关！

$$\int_{\theta,\mathbf{z}} q(\theta, \mathbf{z}|\gamma, \phi) \log p(\mathbf{w}, \alpha, \beta) = \log p(\mathbf{w}, \alpha, \beta)$$

LDA 变分推断（Variational Inference）

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

$$(\gamma^*, \phi^*) = \arg\min_{\gamma,\phi} \int_{\theta,\mathbf{z}} q(\theta, \mathbf{z}|\gamma, \phi) \log q(\theta, \mathbf{z}|\gamma, \phi) - \int_{\theta,\mathbf{z}} q(\theta, \mathbf{z}|\gamma, \phi) \log p(\theta, \mathbf{z}, \mathbf{w}|\alpha, \beta)$$

$$(\gamma^*, \phi^*) = \arg\max_{\gamma,\phi} E_q[\log p(\theta, \mathbf{z}, \mathbf{w}|\alpha, \beta)] - E_q[\log q(\theta, \mathbf{z}|\gamma, \phi)]$$

$$L(\gamma, \phi; \alpha, \beta)$$

ELBO (Evidence Lower Bound)

LDA 变分推断（Variational Inference）

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

$$\begin{aligned} L(\gamma,\phi;\alpha,\beta) &= \log\Gamma\left(\sum_{j=1}^{k}\alpha_j\right) - \sum_{i=1}^{k}\log\Gamma(\alpha_i) + \sum_{i=1}^{k}(\alpha_i-1)\left(\Psi(\gamma_i) - \Psi\left(\sum_{j=1}^{k}\gamma_j\right)\right) \\ &\quad + \sum_{n=1}^{N}\sum_{i=1}^{k}\phi_{ni}\left(\Psi(\gamma_i) - \Psi\left(\sum_{j=1}^{k}\gamma_j\right)\right) \\ &\quad + \sum_{n=1}^{N}\sum_{i=1}^{k}\sum_{j=1}^{V}\phi_{ni}w_n^j\log\beta_{ij} \\ &\quad - \log\Gamma\left(\sum_{j=1}^{k}\gamma_j\right) + \sum_{i=1}^{k}\log\Gamma(\gamma_i) - \sum_{i=1}^{k}(\gamma_i-1)\left(\Psi(\gamma_i) - \Psi\left(\sum_{j=1}^{k}\gamma_j\right)\right) \\ &\quad - \sum_{n=1}^{N}\sum_{i=1}^{k}\phi_{ni}\log\phi_{ni}, \end{aligned}$$

LDA 变分推断（Variational Inference）

- 拉格朗日乘数法
- ϕ_{ni} 代表话题i生成第n个词的概率，所以存在约束：

$$\sum_{i=1}^{k} \phi_{ni} = 1$$

- 将 $L(\gamma, \phi; \alpha, \beta)$ 中与ϕ_{ni}有关的都拿出来，方便求导：

$$L_{[\phi_{ni}]} = \phi_{ni}\left(\Psi(\gamma_i) - \Psi\left(\sum_{j=1}^{k}\gamma_j\right)\right) + \phi_{ni}\log\beta_{iv} - \phi_{ni}\log\phi_{ni} + \lambda_n\left(\sum_{j=1}^{k}\phi_{ni} - 1\right)$$

LDA 变分推断（Variational Inference）

- 求导：
$$\frac{\partial L}{\partial \phi_{ni}} = \Psi(\gamma_i) - \Psi\left(\sum_{j=1}^{k} \gamma_j\right) + \log \beta_{iv} - \log \phi_{ni} - 1 + \lambda.$$

- 此处Ψ(.)是logΓ函数的一阶导数，近似如下：

$$\Psi(x) \approx \left(\left((0.00416\bar{6}\frac{1}{(x+6)^2} - 0.003968)\frac{1}{(x+6)^2} + 0.008\bar{3}\right)\frac{1}{(x+6)^2} - 0.08\bar{3}\right)\frac{1}{(x+6)^2} + \log(x) - \frac{1}{2x} - \sum_{i=1}^{6}\frac{1}{x-i}$$

- 令 $\frac{\partial L}{\partial \phi_{ni}} = 0$ ，得 $\phi_{ni} \propto \beta_{iv} \exp\left(\Psi(\gamma_i) - \Psi\left(\sum_{j=1}^{k} \gamma_j\right)\right)$

LDA 变分推断（Variational Inference）

- 还有一个参数要求：γ
- 把L中与γ相关的项都拿出来

$$L_{[\gamma]} = \sum_{i=1}^{k}(\alpha_i - 1)\left(\Psi(\gamma_i) - \Psi\left(\sum_{j=1}^{k}\gamma_j\right)\right) + \sum_{n=1}^{N}\phi_{ni}\left(\Psi(\gamma_i) - \Psi\left(\sum_{j=1}^{k}\gamma_j\right)\right)$$

$$-\log\Gamma\left(\sum_{j=1}^{k}\gamma_j\right) + \log\Gamma(\gamma_i) - \sum_{i=1}^{k}(\gamma_i - 1)\left(\Psi(\gamma_i) - \Psi\left(\sum_{j=1}^{k}\gamma_j\right)\right).$$

$$= \sum_{i=1}^{k}\left(\Psi(\gamma_i) - \Psi\left(\sum_{j=1}^{k}\gamma_j\right)\right)\left(\alpha_i + \sum_{n=1}^{N}\phi_{ni} - \gamma_i\right) - \log\Gamma\left(\sum_{j=1}^{k}\gamma_j\right) + \log\Gamma(\gamma_i).$$

- 求导 $\frac{\partial L}{\partial \gamma_i} = \Psi'(\gamma_i)\left(\alpha_i + \sum_{n=1}^{N}\phi_{ni} - \gamma_i\right) - \Psi'\left(\sum_{j=1}^{k}\gamma_j\right)\sum_{j=1}^{k}\left(\alpha_j + \sum_{n=1}^{N}\phi_{nj} - \gamma_j\right)$
- 令 $\frac{\partial L}{\partial \gamma_i} = 0$ ，得 $\gamma_i = \alpha_i + \sum_{n=1}^{N}\phi_{ni}$.

LDA 变分推断（Variational Inference）

- LDA 变分推断流程：$L(\gamma, \phi; \alpha, \beta)$
  - EM算法
    - E步：在上一步的α，β基础上，最小化L求γ，ϕ
    - M步：在上一步的γ，ϕ基础上，最小化L求α，β
- 挑选L中与α，β有关的项并求导

$$L_{[\alpha]} = \sum_{d=1}^{M} \left(\log \Gamma \left(\sum_{j=1}^{k} \alpha_j \right) - \sum_{i=1}^{k} \log \Gamma(\alpha_i) + \sum_{i=1}^{k} \left((\alpha_i - 1) \left(\Psi(\gamma_{di}) - \Psi \left(\sum_{j=1}^{k} \gamma_{dj} \right) \right) \right) \right)$$

$$L_{[\beta]} = \sum_{d=1}^{M} \sum_{n=1}^{N_d} \sum_{i=1}^{k} \sum_{j=1}^{V} \phi_{dni} w_{dn}^{j} \log \beta_{ij} + \sum_{i=1}^{k} \lambda_i \left(\sum_{j=1}^{V} \beta_{ij} - 1 \right)$$

LDA 变分推断（Variational Inference）

- 求导结果 $\beta_{ij} \propto \sum_{d=1}^{M} \sum_{n=1}^{N_d} \phi_{dni} w_{dn}^{j}$

$$\frac{\partial L}{\partial \alpha_i} = M\left(\Psi\left(\sum_{j=1}^{k} \alpha_j\right) - \Psi(\alpha_i)\right) + \sum_{d=1}^{M}\left(\Psi(\gamma_{di}) - \Psi\left(\sum_{j=1}^{k} \gamma_{dj}\right)\right)$$

- $\frac{\partial L}{\partial \alpha_i}$ 的结果取决于$\alpha_i (j \neq i)$，所以需要用牛顿迭代法获得α

$$\log(\alpha^{t+1}) = \log(\alpha^{t}) - \frac{\frac{dL}{d\alpha}}{\frac{d^2 L}{d\alpha^2}\alpha + \frac{dL}{d\alpha}}$$

$$\frac{d^2 L}{d\alpha^2} = M(k^2 \Psi''(k\alpha) - k\Psi''(\alpha))$$

[1] D.M. Blei, A.Y. Ng, and M.I. Jordan. Latent dirichlet allocation. The Journal of Machine Learning Research, 3:993–1022, 2003.
[2] https://people.eecs.berkeley.edu/~cjrd/static/pdfs/lda_tutorial.pdf

反思 LDA

- LDA 并没有做任何词语语义方面的工作，仅仅是做一些数量上的统计，它是怎么挖掘出文档具有语义含义的主题的呢？

总结

- 话题模型在自然语言处理中有大量应用
  - 信息检索
  - 文本聚类
  - 关键词提取
  - 情感分析
  - 社交网络主题分析
  - …
- 基于pLSA、LDA也出现了大量的衍生模型

Thank you!