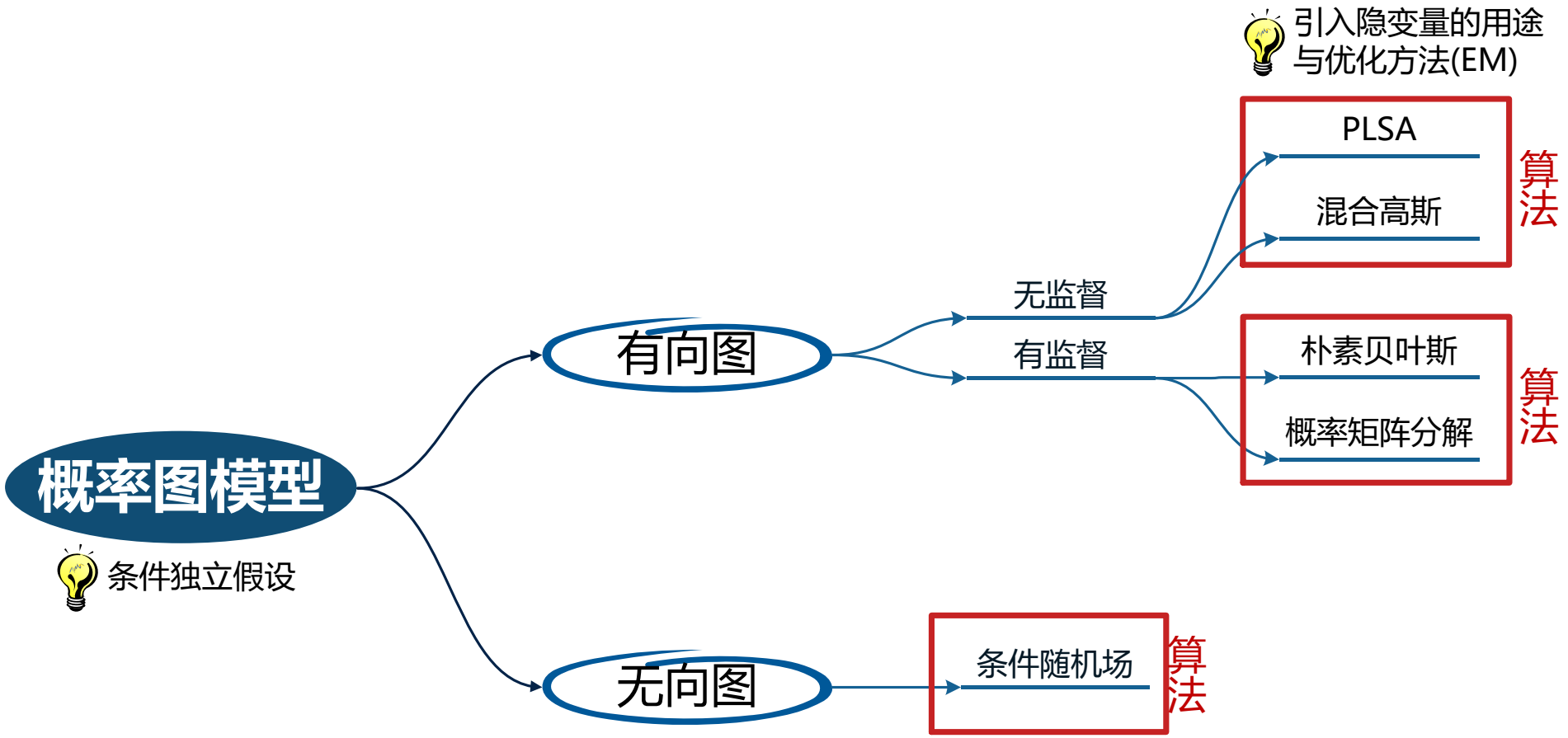
课程专题二：概率图模型方法



内容

- 共现矩阵的一般建模
- 主题模型简介
- 主题模型学习
- 主题模型的扩展
  - **概率图模型方法论**

共现矩阵的一般建模

共现矩阵：数据输入

- 共现矩阵：两个随机变量

第4篇文档中，
第6个词出现了6次

W

D

0	1	0	6	3	5	0	3	1	2	2	1	1	0	0	1
1	0	0	17	0	16	0	2	1	15	1	0	6	0	0	6
0	0	0	0	0	0	0	9	0	0	0	0	0	11	0	0
6	17	0	0	0	6	0	4	0	13	3	0	22	0	0	12
3	0	0	0	0	0	0	0	2	1	0	12	0	0	7	5
5	16	0	6	0	0	0	1	0	59	8	2	2	0	14	7
0	0	0	0	0	0	0	2	0	0	0	0	0	0	1	0
3	2	9	4	0	1	2	0	0	0	1	0	0	5	4	10
1	1	0	0	2	0	0	0	0	1	3	20	0	0	2	1
2	15	0	13	1	59	0	0	1	0	2	2	2	0	5	7
2	1	0	3	0	8	0	1	3	2	0	3	0	0	5	3
1	0	0	0	12	2	0	0	20	2	3	0	1	0	5	4
1	6	0	22	0	2	0	0	0	2	0	1	0	0	0	2
0	0	11	0	0	0	0	5	0	0	0	0	0	0	0	0
0	0	0	0	7	14	1	4	2	5	5	5	0	0	0	6

D：表示文档的随机变量，有N个不同文档

W：表示词汇的随机变量，有M个不同词汇

共现矩阵：数据输入

- 共现矩阵：两个随机变量

第6个品牌的商品，
第4个用户买了6次

B

U

0	1	0	6	3	5	0	3	1	2	2	1	1	0	0	1
1	0	0	17	0	16	0	2	1	15	1	0	6	0	0	6
0	0	0	0	0	0	0	9	0	0	0	0	0	11	0	0
6	17	0	0	0	6	0	4	0	13	3	0	22	0	0	12
3	0	0	0	0	0	0	0	2	1	0	12	0	0	7	5
5	16	0	6	0	0	0	1	0	59	8	2	2	0	14	7
0	0	0	0	0	0	0	2	0	0	0	0	0	0	1	0
3	2	9	4	0	1	2	0	0	0	1	0	0	5	4	10
1	1	0	0	2	0	0	0	0	1	3	20	0	0	2	1
2	15	0	13	1	59	0	0	1	0	2	2	2	0	5	7
2	1	0	3	0	8	0	1	3	2	0	3	0	0	5	3
1	0	0	0	12	2	0	0	20	2	3	0	1	0	5	4
1	6	0	22	0	2	0	0	0	2	0	1	0	0	0	2
0	0	11	0	0	0	0	5	0	0	0	0	0	0	0	0
0	0	0	0	7	14	1	4	2	5	5	5	0	0	0	6

U：表示用户的随机变量，有N个不同用户

B：表示商品品牌的随机变量，有M个品牌

一般模型：概率图模型结构

- 可见的随机变量：D，W

- 最大化“所有数据的似然”

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

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

优化问题形式化

$$\underset{P(w_j|d_i),P(d_i)}{\operatorname{argmax}} \prod_{i=1}^{N}\prod_{j=1}^{M} P(d_i, w_j)^{n(d_i,w_j)}$$

$$\text{s.t.} \sum_{i=1}^{N} P(d_i) = 1$$

$$\sum_{j=1}^{M} P(w_j|d_i) = 1 \ \ (i = 1,2,\dots,N)$$

where $P(d_i, w_j) = P(d_i)P(w_j|d_i)$

拉格朗日乘子法

原式先取ln

$$\ln \prod_{i=1}^{N} \prod_{j=1}^{M} P(d_i, w_j)^{n(d_i, w_j)}$$

$$= \sum_{i=1}^{N} \sum_{j=1}^{M} n(d_i, w_j)(\ln P(d_i) + \ln P(w_j|d_i))$$

拉格朗日乘子法

$$L = \sum_{i=1}^{N}\sum_{j=1}^{M} n(d_i, w_j)\left(\ln P(d_i) + \ln P(w_j|d_i)\right)$$

$$+\sum_{i=1}^{N}\lambda_i\left(\sum_{j=1}^{M} P(w_j|d_i) - 1\right) + \lambda_{N+1}\left(\sum_{i=1}^{N} P(d_i) - 1\right)$$

现在，我们求取L极值点位置

求$P(d_i)$

$$\frac{dL}{d\,P(d_i)} = \frac{\sum_{j=1}^{M} n(d_i, w_j)}{P(d_i)} + \lambda_{N+1} = 0$$

于是有

$$P(d_i) = \frac{\sum_{j=1}^{M} n(d_i, w_j)}{-\lambda_{N+1}}$$

又因为

$$\sum_{i=1}^{N} P(d_i) = \sum_{i=1}^{N} \frac{\sum_{j=1}^{M} n(d_i, w_j)}{-\lambda_{N+1}} = 1$$

故而

$$P(d_i) = \frac{\sum_{j=1}^{M} n(d_i, w_j)}{\sum_{i=1}^{N}\sum_{j=1}^{M} n(d_i, w_j)}$$

求$P(w_j|d_i)$

$$\frac{dL}{d\,P(w_j\,|d_i\,)} = \frac{n(d_i, w_j)}{P(w_j\,|d_i\,)} + \lambda_i = 0$$

于是有

$$P(w_j\,|d_i\,) = \frac{n(d_i, w_j)}{-\lambda_i}$$

又因为

$$\sum_{j=1}^{M} P(w_j|d_i) = \sum_{j=1}^{M} \frac{n(d_i, w_j)}{-\lambda_i} = 1$$

故而

$$P(w_j\,|d_i\,) = \frac{n(d_i, w_j)}{\sum_{j=1}^{M} n(d_i, w_j)}$$

主题模型简介

主题模型：数据输入

- 共现矩阵：两个随机变量

W

D

0	1	0	6	3	5	0	3	1	2	2	1	1	0	0	1
1	0	0	17	0	16	0	2	1	15	1	0	6	0	0	6
0	0	0	0	0	0	0	9	0	0	0	0	0	11	0	0
6	17	0	0	0	6	0	4	0	13	3	0	22	0	0	12
3	0	0	0	0	0	0	0	2	1	0	12	0	0	7	5
5	16	0	6	0	0	0	1	0	59	8	2	2	0	14	7
0	0	0	0	0	0	0	2	0	0	0	0	0	0	1	0
3	2	9	4	0	1	2	0	0	0	1	0	0	5	4	10
1	1	0	0	2	0	0	0	0	1	3	20	0	0	2	1
2	15	0	13	1	59	0	0	1	0	2	2	2	0	5	7
2	1	0	3	0	8	0	1	3	2	0	3	0	0	5	3
1	0	0	0	12	2	0	0	20	2	3	0	1	0	5	4
1	6	0	22	0	2	0	0	0	2	0	1	0	0	0	2
0	0	11	0	0	0	0	5	0	0	0	0	0	0	0	0
0	0	0	0	7	14	1	4	2	5	5	5	0	0	0	6

D：表示文档的随机变量，有N个不同文档

W：表示词汇的随机变量，有M个不同词汇

主题模型：用途

- 计算文档之间的相似性：d_i，d_j
  - **原来的方法：计算如下两个向量的相似度**

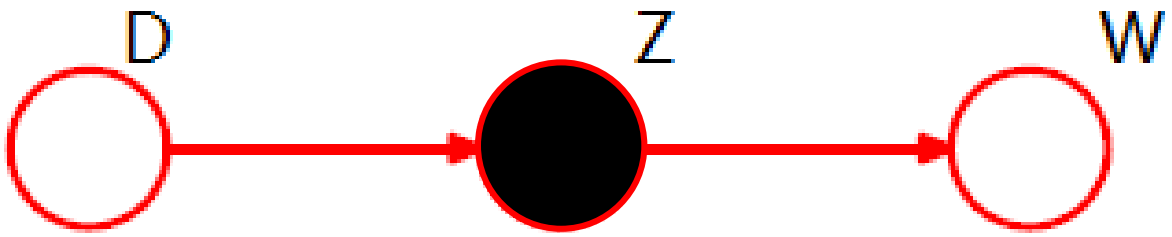
$$(P(w_1|d_i), \cdots, P(w_M|d_i))$$

$$(P(w_1|d_j), \cdots, P(w_M|d_j))$$

- 比较如下两句话：
  - **“我想学习关于编译的知识”**
  - **“我想知道如何把高级编程语言转化为机器语言”**

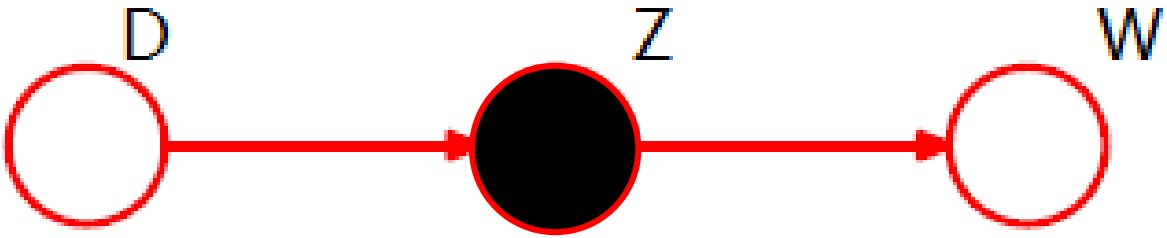
主题模型：概率图模型结构

- 引入一个不可见的随机变量：Z
  - **隐变量Z有K个不同的取值，把任意一个z_k视为一个topic**



主题模型：概率图模型结构

- 主题模型：PLSA（Probabilistic Latent Semantic Analysis）



  - **隐变量Z有K个不同的取值，把任意一个z_i视为一个topic**
  - **假设：D有N个取值，W有M个取值**
  - **该模型需要学习的参数：N+NK+KM**

主题模型：参数的意义

- 某个主题z

$$\left(P(w_1|z), \ldots, P(w_M|z)\right)$$

  - **取概率值最大的前t个词汇**

Aspect 1	Aspect 2	Aspect 3	Aspect 4
plane	space	home	film
airport	shuttle	family	movie
crash	mission	like	music
flight	astronauts	love	new
safety	launch	kids	best
aircraft	station	mother	hollywood
air	crew	life	love
passenger	nasa	happy	actor
board	satellite	friends	entertainment
airline	earth	cnn	star

更多的例子

主题模型：用途

- 主题：可以把语义上相似的词汇聚到同一个主题z
  - **现在的方法：计算如下两个向量的相似度**

$$(P(z_1|d_i), \cdots, P(z_K|d_i))$$

$$(P(z_1|d_j), \cdots, P(z_K|d_j))$$

- 比较如下两句话：
  - **“我想学习关于编译的知识”**
  - **“我想知道如何把高级编程语言转化为机器语言”**

引入一个隐变量z

倘若

$$P(d_i, w_j) = P(d_i)P(w_j|d_i) = P(d_i)\sum_{k=1}^{K} P(w_j|z_k)P(z_k|d_i)$$

最大化

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

$$\text{s.t.} \sum_{i=1}^{N} P(d_i) = 1$$

$$\sum_{j=1}^{M} P(w_j|z_k) = 1 \ \ (k = 1,2, \dots, K)$$

$$\sum_{k=1}^{K} P(z_k|d_i) = 1 \ \ (i = 1,2, \dots, N)$$

where $P(d_i, w_j) = P(d_i)\sum_{k=1}^{K} P(w_j|z_k)P(z_k|d_i)$

主题模型学习

主题模型学习

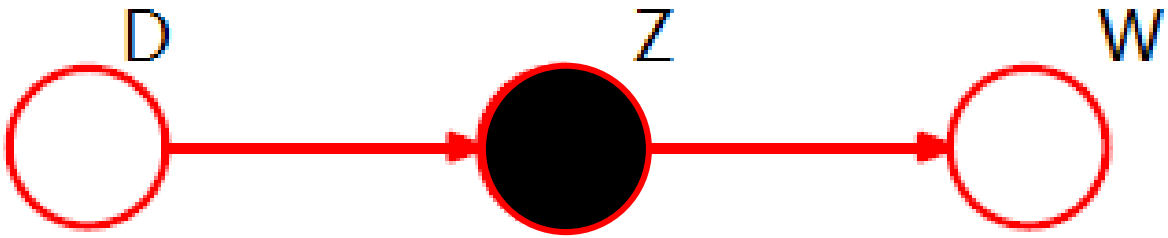
- PLSA模型
  - **使用EM（Expectation Maximization）算法进行优化**

主题模型：EM参数学习

- 最大数据似然

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

$$= \sum_{i=1}^{N} n(d_i)\left[\log P(d_i) + \sum_{j=1}^{M}\frac{n(d_i, w_j)}{n(d_i)}\,\log\sum_{k=1}^{K} P(w_j \mid z_k)P(z_k \mid d_i)\right]$$



$$P(d_i, w_j) = \sum_{k=1}^{K} P(d_i, w_j, z_k) = \sum_{k=1}^{K} P(d_i)P(z_k|d_i)P(w_j|z_k)$$

$$= P(d_i)\sum_{k=1}^{K} P(z_k|d_i)P(w_j|z_k)$$

主题模型：EM参数学习

- 目标函数

$$\sum_{i=1}^{N}\sum_{j=1}^{M} n(d_i, w_j) log \sum_{k=1}^{K} P(w_j|z_k)P(z_k|d_i)$$

凸函数和凸函数

- 目标函数的下界 （使用Jenson不等式）
- If $p_1, \ldots, p_n$ are positive numbers which sum to 1 and f is a real continuous function that is convex (凸), then

$$f\left(\sum_{i=1}^{n} p_i x_i\right) \leq \sum_{i=1}^{n} p_i f(x_i)$$

- If f is concave (凹), then the inequality reverses, giving

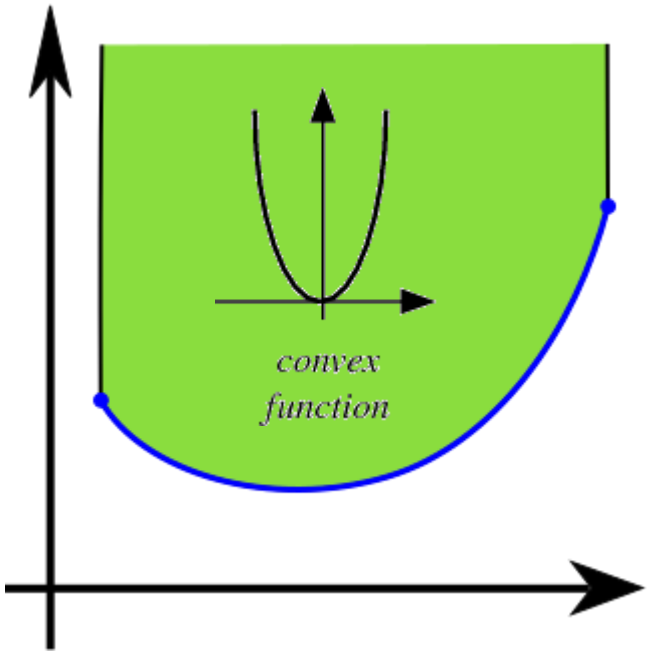
$$f\left(\sum_{i=1}^{n} p_i x_i\right) \geq \sum_{i=1}^{n} p_i f(x_i)$$

凸函数与Jenson不等式

对于每一个凸函数 $f(X)$：

如果 X 是一个随机变量，在 f 的定义域内取值，那么：

$$f\big(E(X)\big) \leq E\big(f(X)\big)$$



一个函数是凸的当且仅当其上境图（在函数图像**上方**的点集）为一个凸集。

而凸集是一个点集合，其中每两点之间的直线点都落在该点集合中

—— Wikipedia

主题模型：参数学习

- 目标函数的下界

$$\sum_{i=1}^{N}\sum_{j=1}^{M} n(d_i, w_j) log \sum_{k=1}^{K} P(w_j|z_k)P(z_k|d_i)$$

$$= \sum_{i=1}^{N}\sum_{j=1}^{M} n(d_i, w_j) log \sum_{k=1}^{K} Q(z_k)\frac{P(w_j|z_k)P(z_k|d_i)}{Q(z_k)}$$

$$\geq \sum_{i=1}^{N}\sum_{j=1}^{M} n(d_i, w_j) \sum_{k=1}^{K} Q(z_k) log \frac{P(w_j|z_k)P(z_k|d_i)}{Q(z_k)}$$

$$= \sum_{i=1}^{N}\sum_{j=1}^{M} n(d_i, w_j) \sum_{k=1}^{K} Q(z_k) log P(w_j|z_k)P(z_k|d_i) - constant$$

where $\sum_{k=1}^{K} Q(z_k) = 1, Q(z_k) \geq 0$

主题模型：参数学习

- 优化目标函数的下界

$$L = \sum_{i=1}^{N} \sum_{j=1}^{M} n(d_i, w_j) \sum_{k=1}^{K} Q(z_k) log P(w_j|z_k) P(z_k|d_i)$$

- 限制条件
  - **K条：** $\sum_{j=1}^{M} P(w_j|z_k) = 1$
  - **N条：** $\sum_{k=1}^{K} P(z_k|d_i) = 1$
- 拉格朗日乘数法

主题模型：参数学习

- 拉格朗日乘数法

$$L+\sum_{k=1}^{K}\alpha_k(1-\sum_{j=1}^{M}P(w_j|z_k))+\sum_{i=1}^{N}\beta_i(1-\sum_{k=1}^{K}P(z_k|d_i))$$

- 对自变量求导，各个偏导为0

等式组1：

$$\sum_{i=1}^{N}n(d_i,w_j)Q(z_k)-\alpha_k P(w_j|z_k)=0$$

$$\text{for } 1\le j\le M, 1\le k\le K$$

等式组2：

$$\sum_{j=1}^{M}n(d_i,w_j)Q(z_k)-\beta_i P(z_k|d_i)=0$$

$$\text{for } 1\le i\le N, 1\le k\le K$$

主题模型：参数学习

- 求解等式方程

$$P(w_j|z_k) = \frac{\sum_{i=1}^{N} n(d_i, w_j)Q(z_k)}{\sum_{j=1}^{M}\sum_{i=1}^{N} n(d_i, w_j)Q(z_k)}$$

$$P(z_k|d_i) = \frac{\sum_{j=1}^{M} n(d_i, w_j)Q(z_k)}{n(d_i)}$$

主题模型：参数学习

- 参数学习的迭代框架
- 初始化参数：$P(w_j|z_k)$，$P(z_k|d_i)$
  - **反复迭代直到收敛**
    - 先根据当前的模型参数值，计算$Q(z_k)$

$$P(w_j|z_k) = \frac{\sum_{i=1}^{N} n(d_i, w_j) Q(z_k)}{\sum_{j=1}^{M} \sum_{i=1}^{N} n(d_i, w_j) Q(z_k)}$$

$$P(z_k|d_i) = \frac{\sum_{j=1}^{M} n(d_i, w_j) Q(z_k)}{n(d_i)}$$

主题模型：参数学习

- 对 $Q(z_k)$ 的讨论
  - **要求1**

$$\sum_{k=1}^{K} Q(z_k) = 1, Q(z_k) \geq 0$$

  - **要求2：** $Q(z_k)$**可以由** $P(w_j|z_k)$, $P(z_k|d_i)$ **计算**

- $Q(z_k)$的不同选择

$$P(z_k)$$

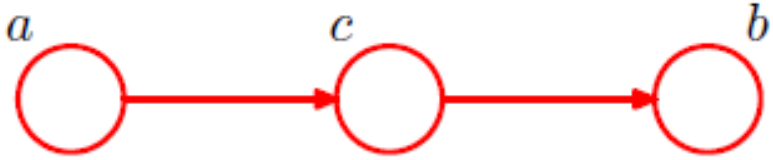
$$P(z_k|d_i)$$

$$P(z_k|d_i, w_j)$$

主题模型：参数学习

- $Q(z_k)$的计算

$$P(z_k|d_i, w_j)$$



$$\boldsymbol{p(a), p(c|a), p(b|c)}$$

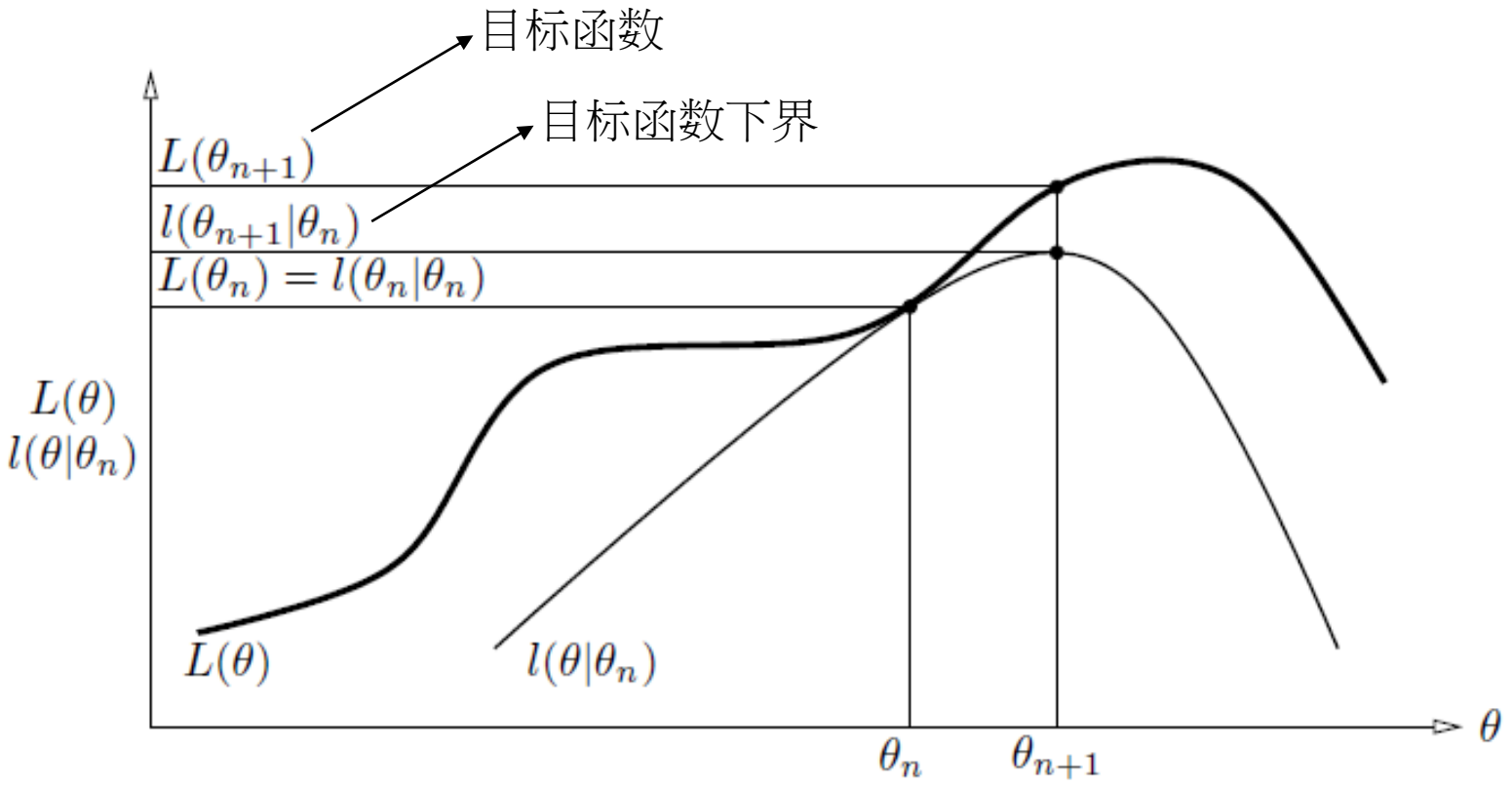
$$p(c|a,b) = \frac{p(a,b,c)}{p(a,b)} = \frac{p(a)p(c|a)p(b|c)}{\sum_c p(a)p(c|a)p(b|c)}$$

$$= \frac{p(c|a)p(b|c)}{\sum_c p(c|a)p(b|c)}$$

$$P(z_k|d_i, w_j) = \frac{P(w_j|z_k)P(z_k|d_i)}{\sum_{k=1}^{K} P(w_j|z_k)P(z_k|d_i)}$$

主题模型：参数学习

- 为什么选择 $P(z_k|d_i, w_j)$
  - **能获得更好的下界**



Sean Borman. The Expectation Maximization Algorithm, a Short Tutorial.

主题模型：参数学习

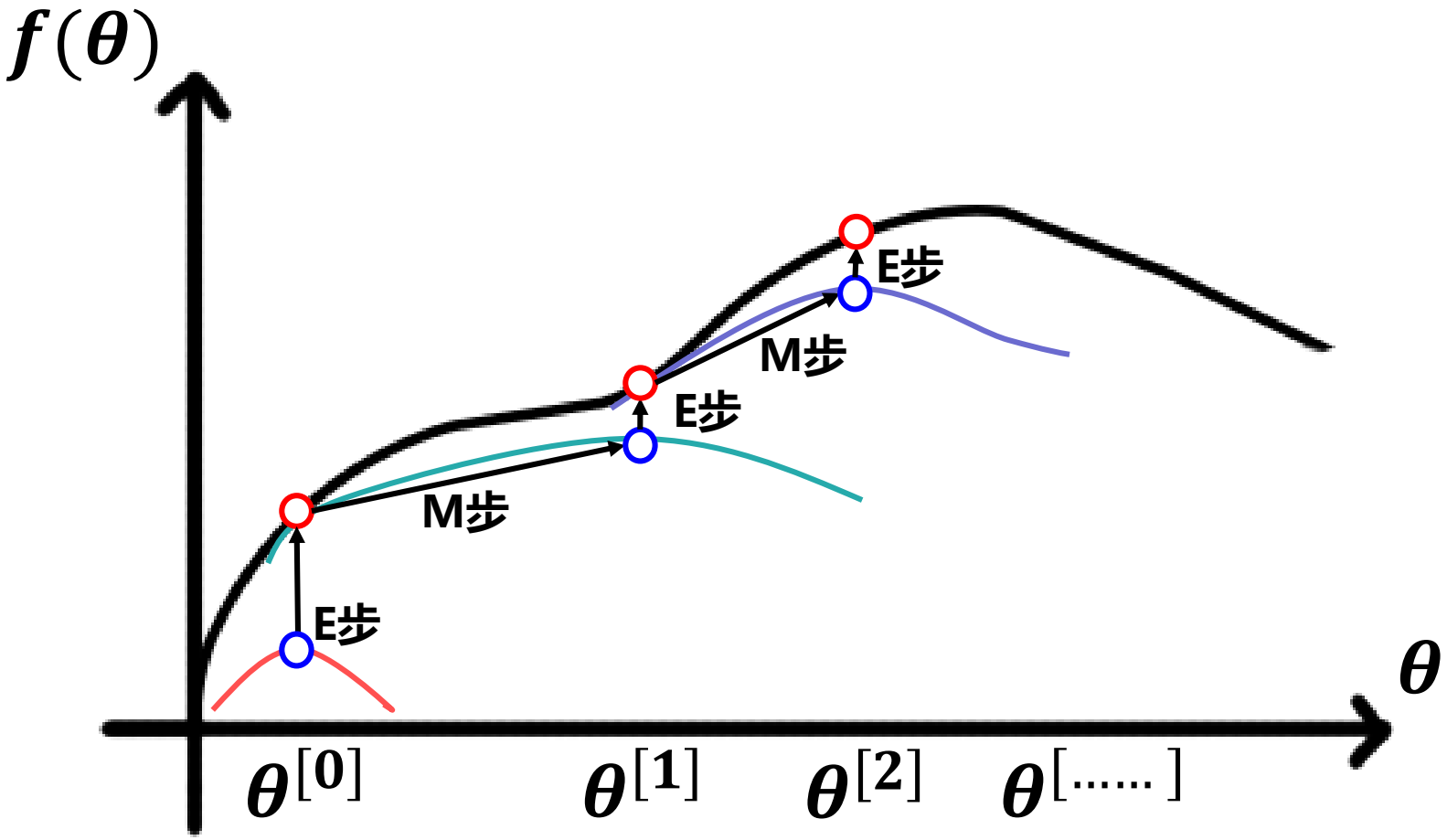
- 完整的学习算法
  - **初始化参数：** $P(w_j|z_k)$，$P(z_k|d_i)$
  - **反复迭代，直到收敛**

$$Q(z_k) = P(z_k|d_i, w_j) = \frac{P(w_j|z_k)P(z_k|d_i)}{\sum_{k=1}^{K} P(w_j|z_k)P(z_k|d_i)}$$

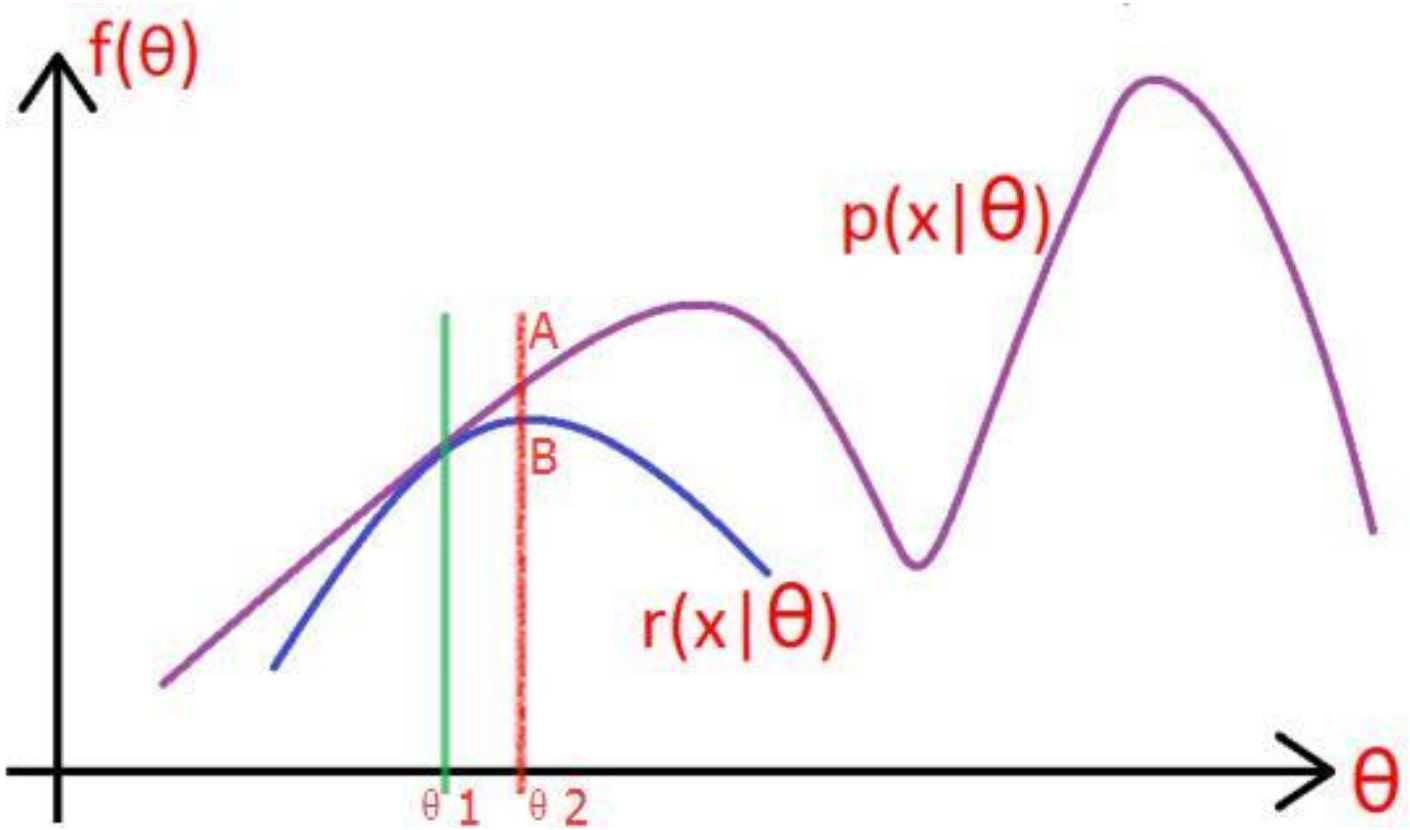
$$P(w_j|z_k) = \frac{\sum_{i=1}^{N} n(d_i, w_j)Q(z_k)}{\sum_{j=1}^{M}\sum_{i=1}^{N} n(d_i, w_j)Q(z_k)}$$

$$P(z_k|d_i) = \frac{\sum_{j=1}^{M} n(d_i, w_j)Q(z_k)}{n(d_i)}$$

EM算法过程展示


$f(\theta)$
E步
M步
E步
M步
E步
θ
$\theta^{[0]}$
$\theta^{[1]}$
$\theta^{[2]}$
$\theta^{[\ldots\ldots]}$

EM算法过程展示

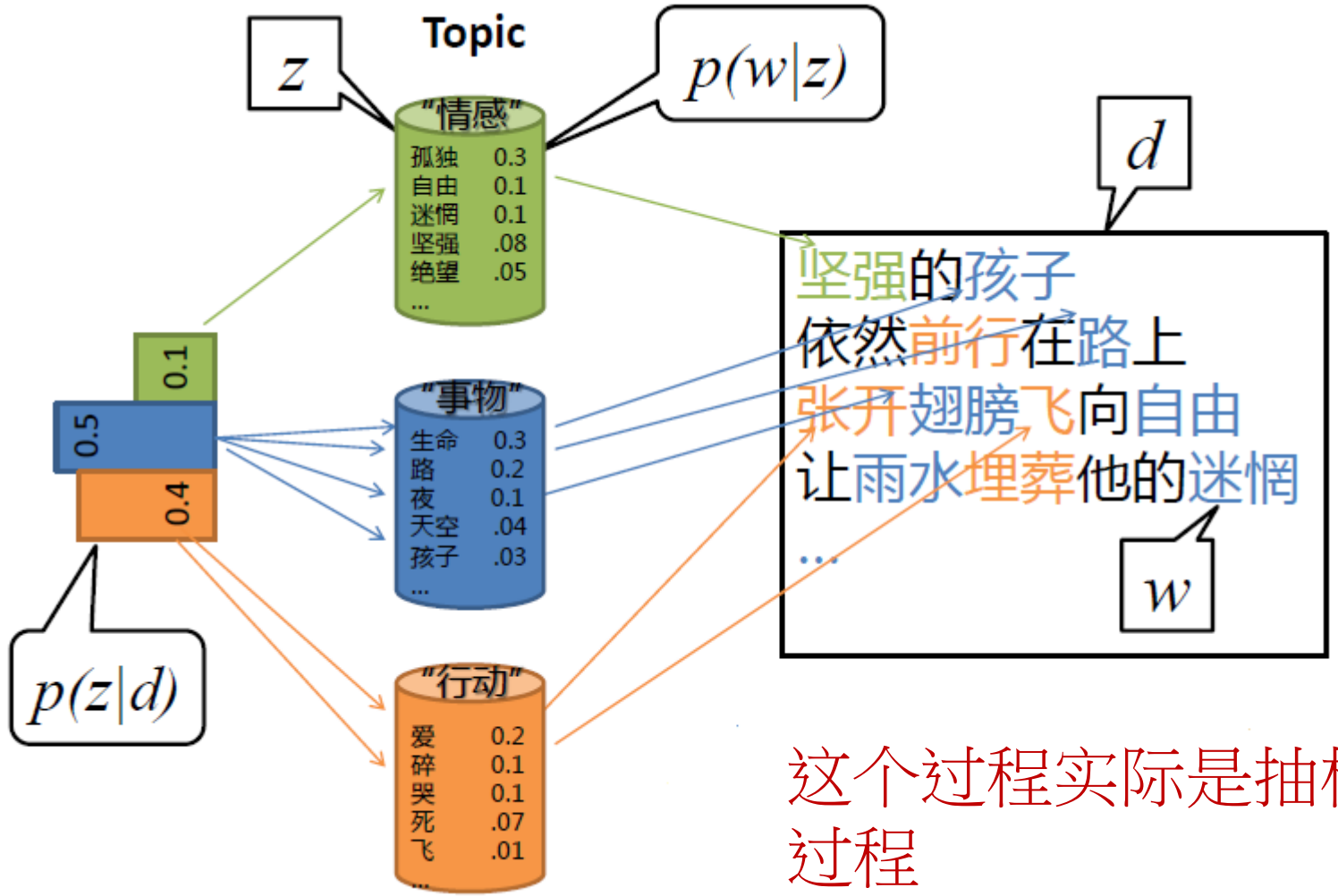


PS：从上图可以看出，EM算法只能求得局部极值点。

http://www.w2bc.com/article/161212

汪峰老师写歌

$$P(w|d) = \sum_{z} P(w|z)P(z|d)$$



这个过程实际是抽样过程