第十四章：概率图模型

主讲：连德富 特任教授 | 博士生导师
邮箱：liandefu@ustc.edu.cn
手机：13739227137
主页：http://staff.ustc.edu.cn/~liandefu

概率模型

- 机器学习最重要的任务是根据**已观察**到的证据（例如训练样本）对感兴趣的**未知**变量（例如类别标记）进行估计和推测。

- **概率模型**（probabilistic model）提供了一种描述框架，将描述任务归结为**计算变量的概率分布**，在概率模型中，利用**已知**的变量推测**未知**变量的分布称为“推断（inference）”，其**核心**在于基于可观测的变量推测出未知变量的**条件分布**
    - 生成式：计算联合分布 $P(Y, R, O)$
    - 判别式：计算条件分布 $P(Y, R|O)$

- 符号约定
    - Y为关心的变量的集合，O为可观测变量集合，R为其他变量集合

概率模型

直接利用概率求和规则消去变量R的时间和空间复杂度为**指数级别**$O(2^{|Y|+|R|})$，需要一种能够简洁紧凑表达变量间关系的工具

概率图模型

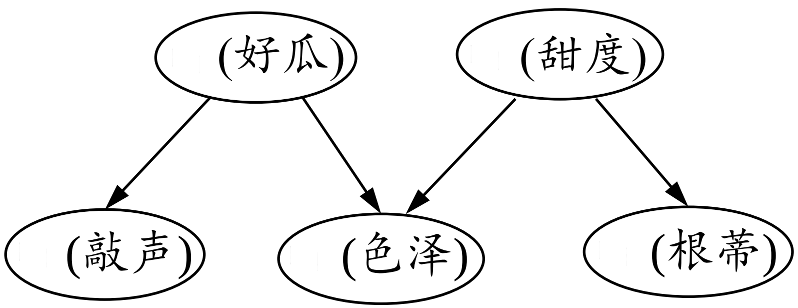
概率图模型

- 概率图模型(probabilistic graphical model)是一类用图来表达**变量相关关系**的概率模型
- 图模型提供了一种**描述框架**，
  - 结点：随机变量（集合）
  - 边：变量之间的依赖关系

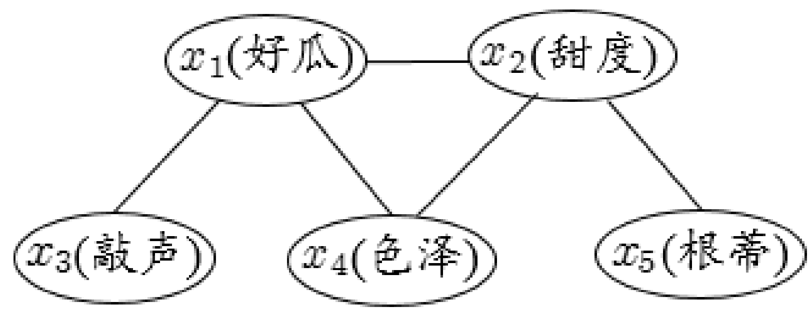
- 分类：
  - **有向图**：贝叶斯网
    - 使用有向无环图表示变量之间的依赖关系
  - **无向图**：马尔可夫网
    - 使用无向图表示变量间的相关关系

概率图模型

- **概率图模型**分类：
  - 有向图：贝叶斯网
  - 无向图：马尔可夫网



有向图



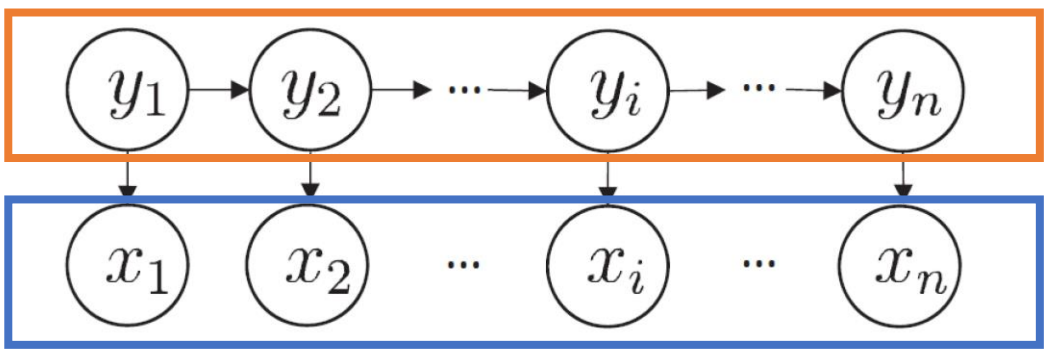
无向图

本章概要

- 隐马尔可夫模型（动态贝叶斯网）
- 马尔可夫随机场 /条件随机场
- 主题模型

隐马尔可夫模型

- 隐马尔可夫模型（Hidden Markov Model，HMM）
- 组成
  - 状态变量: $\{y_1, y_2, \dots, y_n\}$　通常假定是**隐藏**的，不可被观测的
    - 取值范围为$\mathcal{Y}$，通常有N个可能取值的离散空间
  - 观测变量: $\{x_1, x_2, \dots, x_n\}$　表示第i时刻的观测值集合
    - 观测变量可以为离散或连续型，本章中只讨论离散型观测变量，取值范围$\mathcal{X}$为$\{o_1, o_2, \dots, o_M\}$



隐马尔可夫模型

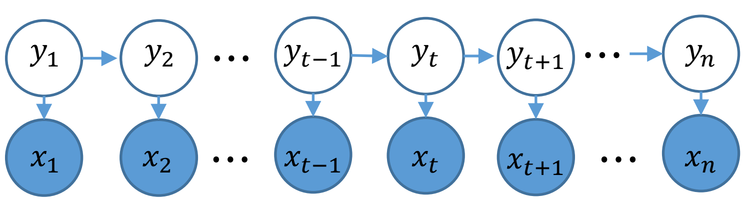
- t时刻的状态x_t仅依赖于$t-1$时刻状态x_{t-1}，与其余n − 2个状态无关

马尔可夫链：

系统下一时刻状态仅由当前状态决定，不依赖于以往的任何状态

$$P(x_1, y_1, \dots, x_n, y_n) = P(y_1)P(x_1|y_1)\prod_{t=2}^{n} P(y_t|y_{t-1})P(x_t|y_t)$$

联合概率



HMM的基本组成

- 确定一个HMM需要**三组参数**
  - 状态**转移概率**：模型在各个状态间转换的概率
  - 表示在任意时刻t，若状态为s_i，下一状态为s_j的概率

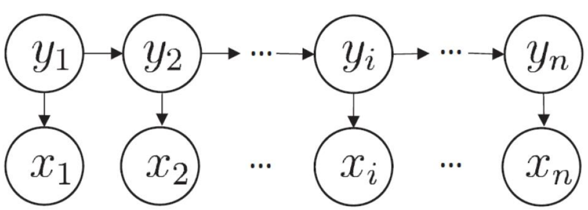
$$A = [a_{i,j}]_{N\times N} \qquad a_{ij} = P(y_{t+1} = s_j | y_t = s_i) \qquad 1 \le i,j \le N$$

  - 输出**观测概率**：模型根据当前状态获 得各个观测值的概率
  - 在任意时刻t，若状态为s_i，则在观测值为o_j的概率

$$B = [b_{i,j}]_{N\times M} \qquad b_{ij} = P(x_t = o_j | y_t = s_i) \qquad 1 \le i \le N \qquad 1 \le j \le M$$

  - **初始**状态概率：模型在初始时刻各个状态出现的概率

$$\pi = [\pi_1, \dots, \pi_n] \qquad \pi_i = P(y_1 = s_i) \qquad 1 \le i \le N$$

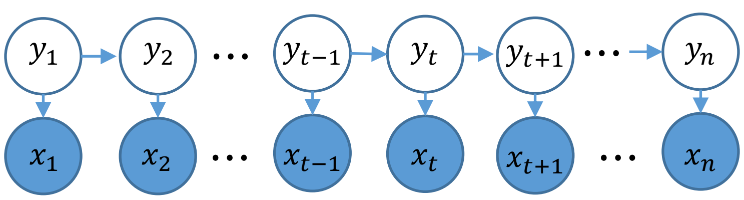


$$P(x_1, y_1, \dots, x_n, y_n) = P(y_1)P(x_1|y_1)\prod_{i=2}^{n} P(y_i|y_{i-1})P(x_i|y_i)$$

HMM的生成过程

- 通过指定状态空间$\mathcal{Y}$，观测空间$\mathcal{X}$和上述三组参数，就能确定一个隐马尔可夫模型。给定$\lambda = [A, B, \pi]$，它按如下过程**生成观察序列**

1. 设置t = 1, 并根据初始状态π选择初始状态y_1
2. 根据状态y_t和输出观测概率B选择观测变量取值x_t
3. 根据状态y_t和状态转移矩阵A转移模型状态，即确定y_{t+1}
4. 若$t < n$, 设置$t = t + 1$，并转到（2）步，否则停止



HMM的基本问题

对于模型 $\lambda = [A, B, \pi]$，给定观测序列 $\boldsymbol{x} = \{x_1, x_2, ..., x_n\}$

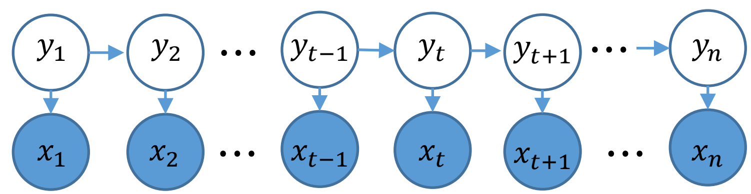
基本问题	具体应用
• **概率计算问题**：评估模型和观测序列间的匹配程度: 有效计算观测序列产生概率$P(\boldsymbol{x}\|\lambda)$	• 根据以往的观序列$\boldsymbol{x} = \{x_1, x_2, ..., x_n\}$预测下一时刻最有可能的观测值$x_{n+1}$
• **预测问题**：根据观测序列"推测"隐藏的模型状态 $\boldsymbol{y} = \{y_1, y_2, ..., y_n\}$	• 语音识别：根据观测的语音信号推测最有可能的状态序列（即：对应的文字）
• **学习问题**：如何调整模型参数 $\lambda = [A, B, \pi]$以使得该序列出现的概率$P(\boldsymbol{x}\|\lambda)$最大	• 通过数据学习参数（模型训练）

概率计算问题

- 对于模型 $\lambda = [A, B, \pi]$，给定观测序列 $\boldsymbol{x} = \{x_1, x_2, ..., x_n\}$
- 评估模型和观测序列之间的匹配程度: 有效计算观测序列其产生的概率

$$P(\boldsymbol{y}|\lambda) = \pi_{y_1} a_{y_1,y_2} a_{y_2,y_3} \cdots a_{y_{n-1},y_n}$$

$$P(\boldsymbol{x}|\boldsymbol{y},\lambda) = b_{y_1,x_1} b_{y_2,x_2} \cdots b_{y_n,x_n}$$

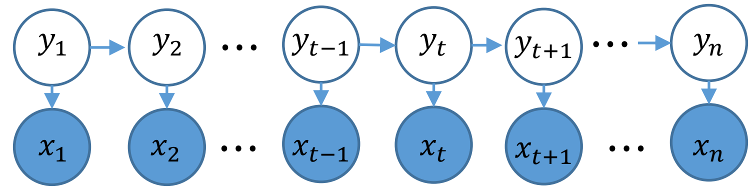


$$P(\boldsymbol{x}|\lambda) = \sum_{\boldsymbol{y}} P(\boldsymbol{x}|\boldsymbol{y},\lambda)\, P(\boldsymbol{y}|\lambda)$$

$$= \sum_{y_1,y_2,\ldots,y_n} \pi_{y_1} b_{y_1,x_1} a_{y_1,y_2} b_{y_2,x_2} a_{y_2,y_3} \cdots a_{y_{n-1},y_n} b_{y_n,x_n}$$

直接计算开销很大，$O(nN^n)$，不可行

概率计算问题



$$P(\boldsymbol{x}|\lambda) = \sum_{y_1,y_2,\ldots,y_n} \pi_{y_1} b_{y_1,x_1} a_{y_1,y_2} b_{y_2,x_2} a_{y_2,y_3} \cdots a_{y_{n-1},y_n} b_{y_n,x_n}$$

$$= \sum_{y_1} \pi_{y_1} b_{y_1,x_1} \sum_{y_2} a_{y_1,y_2} b_{y_2,x_2} \sum_{y_3} a_{y_2,y_3} \cdots \sum_{y_n} a_{y_{n-1},y_n} b_{y_n,x_n}$$

$$= \sum_{y_1} \pi_{y_1} b_{y_1,x_1} \cdots \sum_{y_t} a_{y_{t-1},y_t} b_{y_t,x_t} \sum_{y_{t+1}} a_{y_t,y_{t+1}} b_{y_{t+1},x_{t+1}} \cdots \sum_{y_n} a_{y_{n-1},y_n} b_{y_n,x_n}$$

$$= \sum_{y_t} \left(\sum_{y_1} \pi_{y_1} b_{y_1,x_1} \cdots a_{y_{t-1},y_t} b_{y_t,x_t} \sum_{y_{t+1}} a_{y_t,y_{t+1}} b_{y_{t+1},x_{t+1}} \cdots \sum_{y_n} a_{y_{n-1},y_n} b_{y_n,x_n} \right)$$

$$= \sum_{y_t} P(x_1, x_2, \ldots, x_t, y_t|\lambda) P(x_{t+1}, x_{t+2}, \ldots, x_n|y_t, \lambda)$$

$$= \sum_{y_t} \alpha(y_t)\, \beta(y_t)$$

概率计算问题

$$\alpha(y_t) = P(x_1, x_2, \dots, x_t, y_t|\lambda)$$

$$= \sum_{y_{t-1}} P(x_1, x_2, \dots, x_{t-1}, y_{t-1}|\lambda) P(y_t|y_{t-1}, A) P(x_t|y_t, B)$$

$$= \sum_{y_{t-1}} \alpha(y_{t-1}) P(y_t|y_{t-1}, A) P(x_t|y_t, B)$$

记$\alpha_t(i) = \alpha(y_t = s_i)$，则 $\alpha_t(i) = \sum_{j=1}^{N} \alpha_{t-1}(j) a_{j,i} b_{i,x_t}$

(1) 初值 $\alpha_1(i) = \pi_i b_{i,x_1}$
(2) 递推 $\alpha_t(i) = \sum_{j=1}^{N} \alpha_{t-1}(j) a_{j,i} b_{i,x_t} \quad t = 1,2, \dots, n$
(3) 终止 $P(\boldsymbol{x}|\lambda) = \sum_{i=1}^{N} \alpha_n(i)$

前向算法

概率计算问题

$$\beta(y_t) = P(x_{t+1}, x_{t+2}, \dots, x_n | y_t, \lambda)$$
$$= \sum_{y_{t+1}} P(x_{t+1}, x_{t+2}, \dots, x_n, y_{t+1} | y_t, \lambda)$$
$$= \sum_{y_{t+1}} P(x_{t+2}, \dots, x_n, y_{t+2} | y_{t+1}, \lambda) P(x_{t+1} | y_{t+1}, B) P(y_{t+1} | y_t, A)$$

记$\beta_t(i) = \beta(y_t = s_i)$，则 $\beta_t(i) = \sum_{j=1}^{N} \beta_{t+1}(j) a_{i,j} b_{j,x_{t+1}}$

(1) 初值 $\beta_n(i) = 1$
(2) 递推 $\beta_t(i) = \sum_{j=1}^{N} \beta_{t+1}(j) a_{i,j} b_{j,x_{t+1}}$ $t = n-1, n-2, \dots, 1$
(3) 终止 $P(\boldsymbol{x}|\lambda) = \sum_{i=1}^{N} \beta_1(i) \pi_i b_{i,x_1}$

后向算法

概率计算问题

- 一些概率和期望值计算

给定模型λ和观测序列$\boldsymbol{x}$，在t时刻处于状态s_i的概率

$$\gamma_t(i) = P(y_t = s_i|\boldsymbol{x}, \lambda) = \frac{P(y_t = s_i, \boldsymbol{x}|\lambda)}{P(\boldsymbol{x}|\lambda)} = \frac{\alpha_t(i)\beta_t(i)}{\sum_{i=1}^{N}\alpha_t(i)\beta_t(i)}$$

$$\alpha(y_t) = P(x_1, x_2, \dots, x_t, y_t|\lambda)$$

$$\beta(y_t) = P(x_{t+1}, x_{t+2}, \dots, x_n|y_t, \lambda)$$

$$\alpha(y_t)\beta(y_t) = P(y_t, \boldsymbol{x}|\lambda)$$

给定模型λ和观测序列$\boldsymbol{x}$，在t时刻处于状态s_i且t+1时刻处于状态s_j的概率

$$\xi_t(i,j) = P\left(y_t = s_i, y_{t+1} = s_j \middle| \boldsymbol{x}, \lambda\right) = \frac{P\left(y_t = s_i, y_{t+1} = s_j, \boldsymbol{x} \middle| \lambda\right)}{P(\boldsymbol{x}|\lambda)} = \frac{\alpha_t(i)a_{ij}b_{j,x_{t+1}}\beta_{t+1}(j)}{\sum_{i,j}\alpha_t(i)a_{ij}b_{j,x_{t+1}}\beta_{t+1}(j)}$$

概率计算问题

- 一些概率和期望值计算

给定模型λ和观测序列$\boldsymbol{x}$，在t时刻处于状态s_i的概率 $\gamma_t(i)=\frac{\alpha_t(i)\beta_t(i)}{\sum_{i=1}^{N}\alpha_t(i)\beta_t(i)}$

给定模型λ和观测序列$\boldsymbol{x}$，在t时刻处于状态s_i且t+1时刻处于状态s_j的概率

$$\xi_t(i,j)=\frac{\alpha_t(i)a_{ij}b_{j,x_{t+1}}\beta_{t+1}(j)}{\sum_{i,j}\alpha_t(i)a_{ij}b_{j,x_{t+1}}\beta_{t+1}(j)}$$

给定模型λ和观测序列$\boldsymbol{x}$，状态s_i出现的概率 $\sum_{t=1}^{n}\gamma_t(i)$

给定模型λ和观测序列$\boldsymbol{x}$，状态s_i转移到状态s_j的概率 $\sum_{t=1}^{n-1}\xi_t(i,j)$

预测问题（Viterbi Algorithm）

- 根据观测序列“推测”隐藏的模型状态 $\boldsymbol{y}=\{y_1,y_2,\ldots,y_n\}$

$$\underset{\boldsymbol{y}}{\operatorname{argmax}}\, P(\boldsymbol{y}|\boldsymbol{x},\lambda)=\operatorname{argmax}_{\boldsymbol{y}} P(\boldsymbol{y},\boldsymbol{x}|\lambda)$$

$$\max_{\boldsymbol{y}} P(\boldsymbol{y},\boldsymbol{x}|\lambda) \quad = \max_{y_1,\ldots,y_n} P(y_1)P(x_1|y_1)\prod_{i=2}^{n} P(y_i|y_{i-1})P(x_i|y_i)$$

$$\max(ab,ac)=a\cdot\max(b,c)$$

$$=\underbrace{\max_{y_1} P(y_1)P(x_1|y_1)\max_{y_2} P(y_2|y_1)P(x_2|y_2)\cdots}_{\text{前t项最大化}}\max_{y_n} P(y_n|y_{n-1})P(x_n|y_n)$$

$$\delta(y_{t+1})=\max_{y_1,\ldots,y_t} P(y_{t+1},y_1,\ldots,y_t,x_1,\ldots,x_{t+1})$$

$$\delta(y_2)=\max_{y_1} P(y_2,y_1,x_1,x_2)$$

$$=\max_{y_t} P(y_{t+1}|y_t)P(x_{t+1}|y_{t+1})\max_{y_1,\ldots,y_{t-1}} P(y_t,y_1,\ldots,y_{t-1},x_1,\ldots,x_t)$$

$$=\max_{y_t} P(y_{t+1}|y_t)P(x_{t+1}|y_{t+1})\delta(y_t)$$

预测问题（Viterbi 算法）

概率连乘容易导致溢出，因此引入对数函数

$$\delta(y_{t+1}) = \max_{y_1,\dots,y_t} \log P(y_{t+1}, y_1, \dots, y_t, x_1, \dots, x_{t+1})$$

$$= \log P(x_{t+1}|y_{t+1}) + \max_{y_t} \log P(y_{t+1}|y_t) + \max_{y_1,\dots,y_{t-1}} \log P(y_t, y_1, \dots, y_{t-1}, x_1, \dots, x_t)$$

$$= \log P(x_{t+1}|y_{t+1}) + \max_{y_t} \log P(y_{t+1}|y_t) + \delta(y_t)$$

记$\delta_t(i) = \delta(y_t = s_i)$，则 $\delta_{t+1}(i) = \log b_{i,x_{t+1}} + \max_{1\le j\le N}\left[\log a_{j,i} + \delta_t(j)\right]$

$$\delta_1(i) = \log \pi_i + \log b_{i,x_1}$$

学习问题（Baum-Welch算法）

- 如何调整模型参数 $\lambda = [A, B, \pi]$以使得该序列出现的概率$P(\boldsymbol{x}|\lambda)$最大
- 使用EM算法

E 基于$\boldsymbol{\Theta}^t$推断隐变量$\boldsymbol{Z}$的分布$P(\boldsymbol{Z} \mid \boldsymbol{X}, \boldsymbol{\Theta}^t)$，并计算对数似然$LL(\boldsymbol{\Theta}|\boldsymbol{X}, \boldsymbol{Z})$关于$\boldsymbol{Z}$的期望;

$$\mathcal{L}(P(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\Theta}), \boldsymbol{\Theta}) = \mathbb{E}_{\boldsymbol{Z}\sim P(\boldsymbol{Z} \mid \boldsymbol{X}, \boldsymbol{\Theta}^t)} LL(\boldsymbol{\Theta}|\boldsymbol{X}, \boldsymbol{Z}) = Q(\boldsymbol{\Theta}|\boldsymbol{\Theta}^t)$$

M 寻找参数最大化期望似然;

$$\boldsymbol{\Theta}^{t+1} = \arg\max_{\boldsymbol{\Theta}} Q(\boldsymbol{\Theta}|\boldsymbol{\Theta}^t)$$

学习问题（Baum-Welch算法）

$$Q(\lambda|\lambda^t) = \mathbb{E}_{\boldsymbol{y}\sim P(Y\,|\,\boldsymbol{x},\boldsymbol{\lambda}^t\,)} LL(\lambda|\boldsymbol{x},\boldsymbol{y})$$

$$= \mathbb{E}_{\boldsymbol{y}\sim P(Y\,|\,\boldsymbol{x},\boldsymbol{\lambda}^t\,)} \log P(\boldsymbol{x},\boldsymbol{y}|\lambda)$$

$$= \mathbb{E}_{\boldsymbol{y}\sim P(Y\,|\,\boldsymbol{x},\boldsymbol{\lambda}^t\,)} \log P(y_1)P(x_1|y_1)\prod_{t=2}^{n} P(y_t|y_{t-1})P(x_t|y_t)$$

$$= \mathbb{E}_{\boldsymbol{y}\sim P(Y\,|\,\boldsymbol{x},\boldsymbol{\lambda}^t\,)} \log P(y_1) + \log P(x_1|y_1) + \sum_{t=2}^{n} \log P(y_t|y_{t-1}) + \log P(x_t|y_t)$$

$$= \mathbb{E}_{y_1}[\log P(y_1) + \log P(x_1|y_1)] + \sum_{t=2}^{n} \mathbb{E}_{y_{t-1},y_t}[\log P(y_t|y_{t-1})] + \mathbb{E}_{y_t}[\log P(x_t|y_t)]$$

$$= \sum_{i} \gamma_1(i)\big(\log \pi_i + b_{i,x_1}\big) + \sum_{t=2}^{n}\left(\sum_{i,j} \xi_{t-1}(i,j)\log a_{ij} + \sum_{i} \gamma_t(i) b_{i,x_t}\right)$$

学习问题（Baum-Welch算法）

$$Q(\lambda|\lambda^t) = \sum_i \gamma_1(i)\left(\log \pi_i + b_{i,x_1}\right) + \sum_{t=2}^{n}\left(\sum_{i,j} \xi_{t-1}(i,j) \log a_{ij} + \sum_i \gamma_t(i) \log b_{i,x_t}\right)$$

- 求解初始状态概率

针对约束条件$\sum_i \pi_i = 1$，基于拉格朗日乘子法，求解 $\pi_i = \gamma_1(i)$

- 求解转移概率

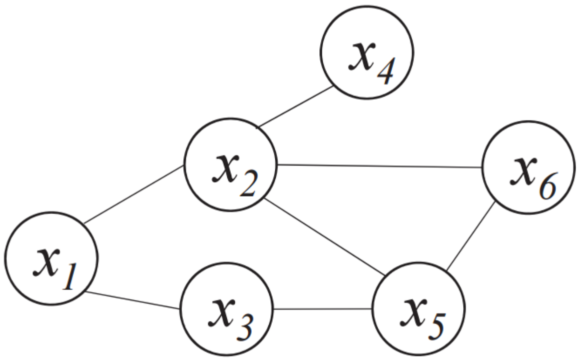
针对约束条件$\sum_j a_{i,j} = 1$，基于拉格朗日乘子法，求解 $a_{i,j} = \frac{\sum_{t=2}^{n} \xi_{t-1}(i,j)}{\sum_j \sum_{t=2}^{n} \xi_{t-1}(i,j)}$

- 求解观测概率

针对约束条件$\sum_j b_{i,j} = 1$，基于拉格朗日乘子法，求解 $b_{i,j} = \frac{\sum_{t=1}^{n} \gamma_t(i)\, \mathbb{I}\left(x_t = o_j\right)}{\sum_{t=1}^{n} \gamma_t(i)}$

马尔可夫随机场

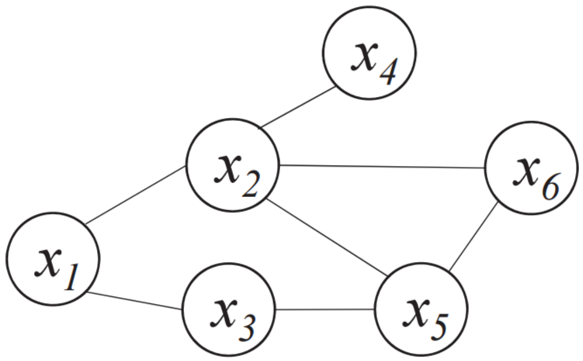
- 马尔可夫随机场（Markov Random Field，MRF）
  - 是典型的马尔可夫网
  - 著名的**无向图**模型



- 图模型表示：
  - 结点表示变量（集），边表示依赖关系
  - 有一组势函数（Potential Functions），亦称“因子”（factor），这是定义在变量子集上的非负实函数，主要用于定义概率分布函数

马尔可夫随机场—分布形式化：

- 使用基于**极大团**的势函数（因子）
  - 对于图中结点的一个子集，若其中任意两结点间都有边连接，则称该结点子集为一个“团”（clique）。若一个团中加入另外任何一个结点都不再形成团，则称该团为“极大团”（maximal clique)
  - 图中$\{x_1, x_2\}, \{x_2, x_6\}, \{x_2, x_5, x_6\}$等为团
  - 图中$\{x_2, x_6\}$不是极大团
  - 每个结点至少出现在一个极大团中



马尔可夫随机场

- 使用基于**极大团**的势函数（因子）
  - 对于n个变量$\boldsymbol{x} = \{x_1, x_2, ... x_n\}$，所有团构成的集合为$\mathcal{C}$，与团$Q \in \mathcal{C}$对应的变量集合记为$\boldsymbol{x}_Q$, 则联合概率定义为：

$$P(\boldsymbol{x}) = \frac{1}{Z}\prod_{Q\in\mathcal{C}}\psi_Q(\boldsymbol{x}_Q)$$

  - 其中，ψ_Q是于团Q对应的势函数，Z为概率的规范化因子
  - 在实际应用中，Z往往很难精确计算。但很多任务中，不需要对Z进行精确计算
  - 若变量问题较多，则团的数目过多，上式的乘积项过多，会给计算带来负担，所以需要考虑极大团

马尔可夫随机场

- 联合概率分布可以使用极大团定义，假设所有极大团构成的集合为$\mathcal{C}^*$

$$P(\boldsymbol{x})=\frac{1}{Z^*}\prod_{Q\in\mathcal{C}^*}\psi_Q(\boldsymbol{x}_Q)$$

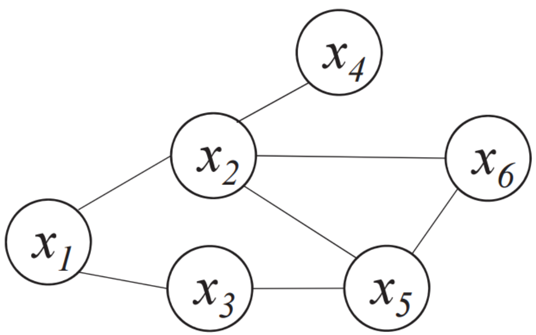
  - 其中，Z^*是规范化因子 $Z^*=\sum_{\boldsymbol{x}}\prod_{Q\in\mathcal{C}^*}\psi_Q(\boldsymbol{x}_Q)$

马尔可夫随机场

- 使用基于**极大团**的势函数（因子）
  - 联合概率分布可以使用极大团定义，假设所有极大团构成的集合为$\mathcal{C}^*$

$$P(\boldsymbol{x}) = \frac{1}{Z^*}\prod_{Q\in\mathcal{C}^*}\psi_Q(\boldsymbol{x}_Q)$$

- 图模型



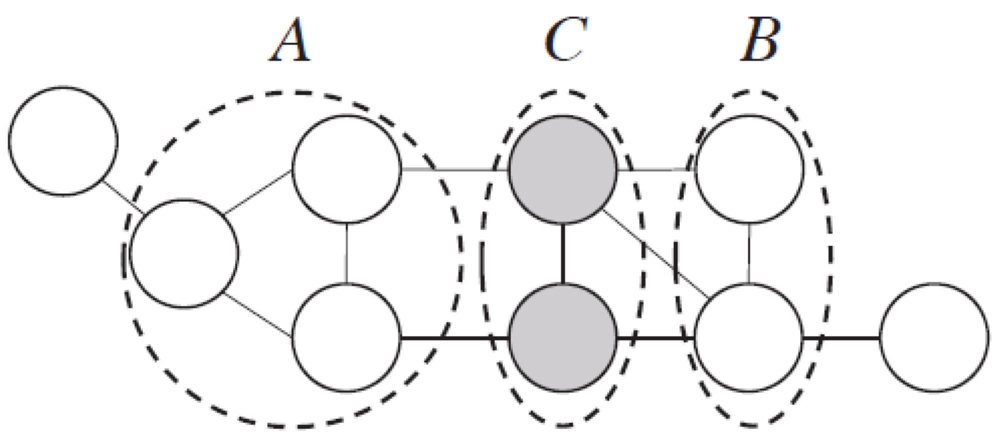
- 联合概率分布

$$P(\boldsymbol{x}) = \frac{1}{Z}\psi_{12}(x_1,x_2)\psi_{13}(x_1,x_3)\psi_{24}(x_2,x_4)\psi_{35}(x_3,x_5)\psi_{256}(x_2,x_5,x_6)$$

马尔可夫随机场中的分离集

- 马尔可夫随机场中得到“**条件独立性**”

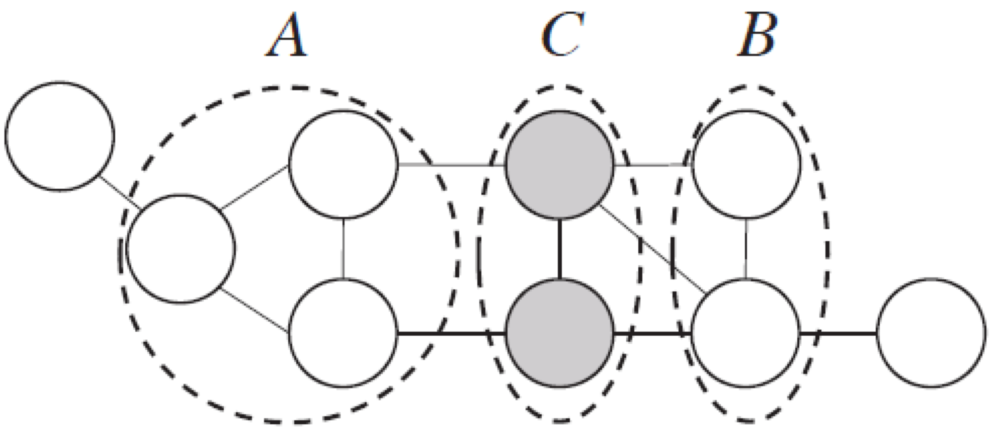
- 借助“**分离**”的概念，若从结点集A中的结点到B中的结点都必须经过结点集C中的结点，则称结点集A，B被结点集C分离，称C为分离集（separating set）



全局马尔可夫性

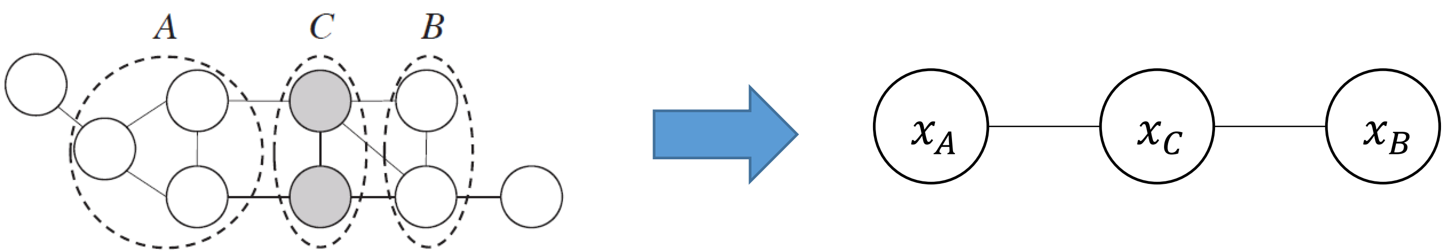
- 借助“分离”的概念，可以得到：
  - **全局马尔可夫性**（global Markov property）：在给定**分离集**的条件下，两个变量子集条件独立

  - 若令A,B,C对应的变量集分别为x_A, x_B, x_C, 则x_A和x_B在x_C给定的条件下独立，记为$x_A \perp x_B \mid x_C$



全局马尔可夫性的验证

- 图模型简化：



- 得到图模型的联合概率为： $P(x_A, x_B, x_C) = \frac{1}{Z}\psi_{AC}(x_A, x_C)\psi_{BC}(x_B, x_C)$

条件概率

$$P(x_A, x_B | x_C) = \frac{P(x_A, x_B, x_C)}{\sum_{x_A, x_B} P(x_A, x_B, x_C)}$$

$$= \frac{\frac{1}{Z}\psi_{AC}(x_A, x_C)\psi_{BC}(x_B, x_C)}{\sum_{x_A, x_B} \frac{1}{Z}\psi_{AC}(x_A, x_C)\psi_{BC}(x_B, x_C)}$$

$$= \frac{\psi_{AC}(x_A, x_C)}{\sum_{x_A} \psi_{AC}(x_A, x_C)} \frac{\psi_{BC}(x_B, x_C)}{\sum_{x_B} \psi_{BC}(x_B, x_C)}$$

$$= P(x_A | x_C) P(x_B | x_C)$$

$$P(x_A | x_C) = \frac{\sum_{x_B} P(x_A, x_B, x_C)}{\sum_{x_A, x_B} P(x_A, x_B, x_C)}$$

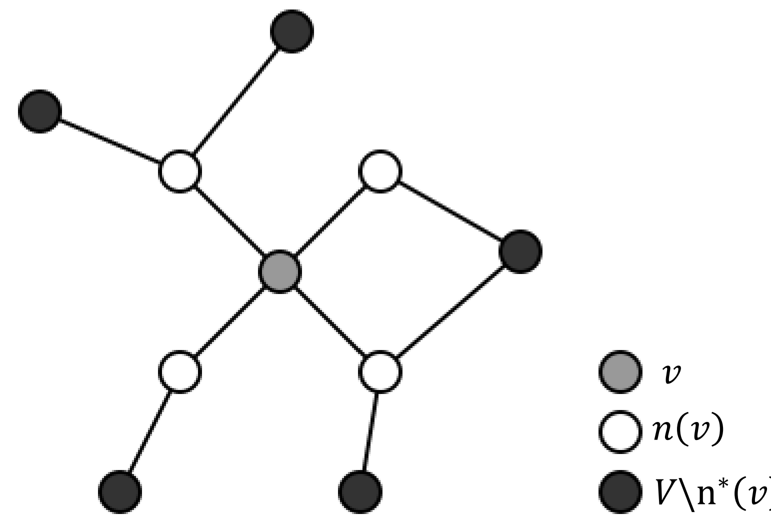
$$= \frac{\sum_{x_B} \frac{1}{Z}\psi_{AC}(x_A, x_C)\psi_{BC}(x_B, x_C)}{\sum_{x_A, x_B} \frac{1}{Z}\psi_{AC}(x_A, x_C)\psi_{BC}(x_B, x_C)}$$

$$= \frac{\psi_{AC}(x_A, x_C)}{\sum_{x_A, x_B} \psi_{AC}(x_A, x_C)}$$

马尔可夫随机场中的条件独立性

全局马尔可夫性（global Markov property）：在给定**分离集**的条件下，两个变量子集条件独立

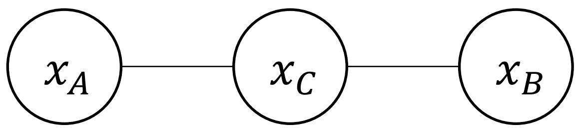
- **局部**马尔可夫性（local Markov property）：在给定**邻接变量**的情况下，一个变量条件独立于其它所有变量
  - 令V为图的结点集，$n(v)$为结点v在图上的邻接节点，$n^*(v) = n(v) \cup \{v\}$, 有$\mathrm{x}_\mathrm{v} \perp \mathrm{x}_{V\backslash \mathrm{n}^*(v)} \mid \mathrm{x}_{n(v)}$



- **成对**马尔可夫性（pairwise Markov property）：在给定**所有其它变量**的情况下，两个非邻接变量条件独立
  - 令V为图的结点集，边集为E，对图中的两个结点u, v，若$\langle u, v\rangle \notin E$，有$\mathrm{x}_\mathrm{u} \perp \mathrm{x}_v \mid \mathrm{x}_{V\backslash\langle u,v\rangle}$

马尔可夫随机场中的势函数

- 势函数$\psi_Q(\boldsymbol{x}_Q)$的作用是定量刻画变量集$\boldsymbol{x}_Q$中变量的相关关系，应为非负函数，且在所偏好的变量取值上有较大的函数值



- 上图中，假定变量均为二值变量，定义势函数：

$$\psi_{AC}(x_A, x_C) = \begin{cases} 1.5, \text{if } x_A = x_C \\ 0.1, \text{otherwise} \end{cases}$$

模型偏好x_A与x_C有相同的取值
x_A与x_C正相关

$$\psi_{BC}(x_B, x_C) = \begin{cases} 0.2, \text{if } x_B = x_C \\ 1.3, \text{otherwise} \end{cases}$$

模型偏好x_B与x_C有不同的取值
x_B与x_C负相关

令x_A与x_C相同且x_B与x_C不同的变量值指派将有较高的联合概率

马尔可夫随机场中的势函数

- 势函数$\psi_Q(\boldsymbol{x}_Q)$的作用是定量刻画变量集$\boldsymbol{x}_Q$中变量的相关关系，应为非负函数，且在所偏好的变量取值上有较大的函数值

- 为了满足**非负性**，指数函数常被用于定义势函数，即：

$$\psi_Q(\boldsymbol{x}_Q) = e^{-H_Q(\boldsymbol{x}_Q)}$$

- 其中，$H_Q(\boldsymbol{x}_Q)$是一个定义在变量 $\boldsymbol{x}_Q$ 上的实值函数，常见形式为：

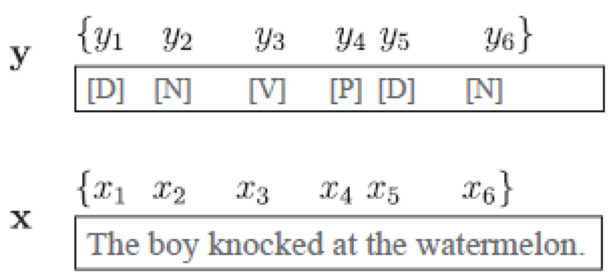
$$H_Q(\boldsymbol{x}_Q) = \sum_{u,v\in Q,u\neq v} \alpha_{uv} x_u\, x_v + \sum_{v\in Q} \beta_v x_v$$

- 其中，α_{uv}和β_v是参数，上式第一项考虑每一对结点的关系，第二项考虑单结点

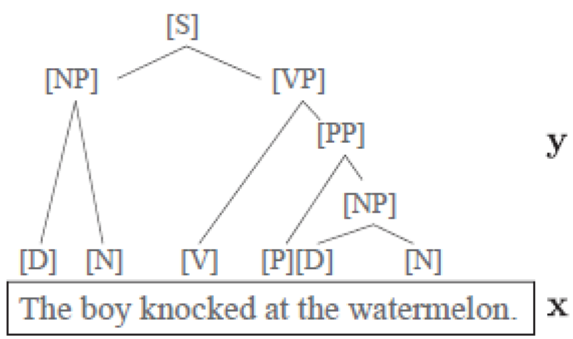
条件随机场

- 条件随机场（Conditional Random Field，CRF)是一种**判别式**无向图模型（可看作给定观测值的MRF），条件随机场对多个变量给定相应**观测值**后的条件概率进行建模
- 若令$\boldsymbol{x}=\{x_1,x_2,\ldots,x_n\}$为观测序列，$\boldsymbol{y}=\{y_1,y_2,\ldots,y_n\}$为对应的标记序列， CRF的目标是构建条件概率模型$P(\boldsymbol{y}|\boldsymbol{x})$

- 标记变量 $\boldsymbol{Y}$ 可以是结构型变量，它各个分量之间具有某种相关性。
  - 自然语言处理的词性标注任务中，观测数据为语句（单词序列），标记为相应的词性序列，具有线性序列结构
  - 在语法分析任务中，输出标记是语法树，具有树形结构



(a) 词性标注



(b) 语法分析

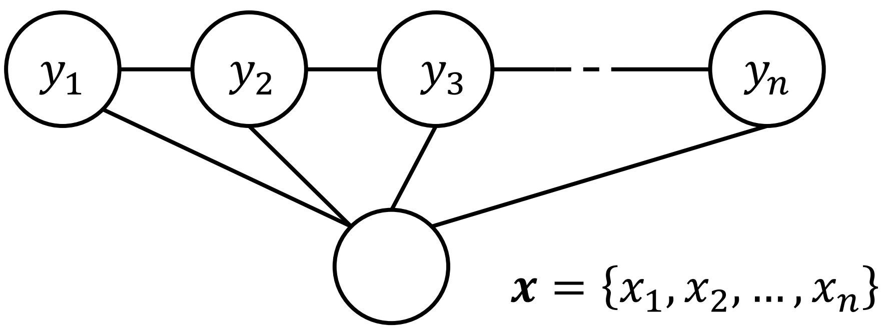
条件随机场

- 令G = $\langle V, E\rangle$表示结点与标记变量$\boldsymbol{y}$中元素一一对应的无向图。无向图中，Y_v表示与节点v对应的标记变量，$n(v)$表示结点v的邻接结点，若图中的每个结点都满足马尔可夫性，

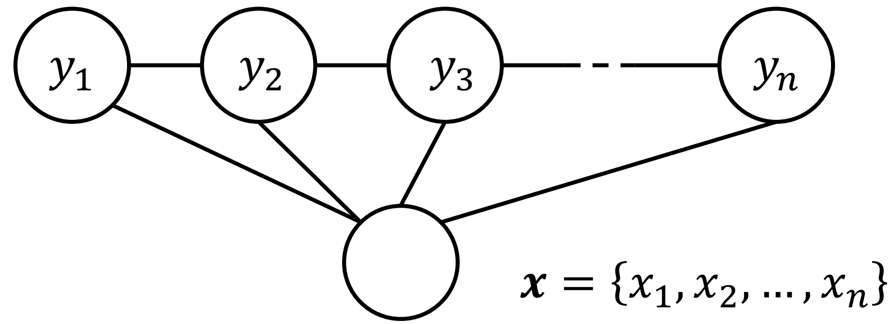
$$P\left(Y_v | X, Y_{V\setminus\{v\}}\right) = P\left(Y_v | X, Y_{n(v)}\right)$$

则(Y, X)构成条件随机场

- CRF使用势函数和图结构上的团来定义$P(\boldsymbol{y}|\boldsymbol{x})$
- 接下来仅考虑**链式**条件随机场（chain-structured CRF）



链式条件随机场



- 包含两种关于标记变量的团：
  - 相邻的标记变量，即 $\{y_{i-1}, y_i\}$；
  - 单个标记变量，$\{y_i\}$；
- 条件概率可被定义为：

$$P(\boldsymbol{y}|\boldsymbol{x}) = \frac{1}{Z}\exp\left(\sum_{j=1}^{K_1}\sum_{i=2}^{n}\lambda_j t_j(y_{i-1}, y_i, \boldsymbol{x}, i) + \sum_{l=1}^{K_2}\sum_{i=1}^{n}\mu_l s_l(y_i, \boldsymbol{x}, i)\right)$$

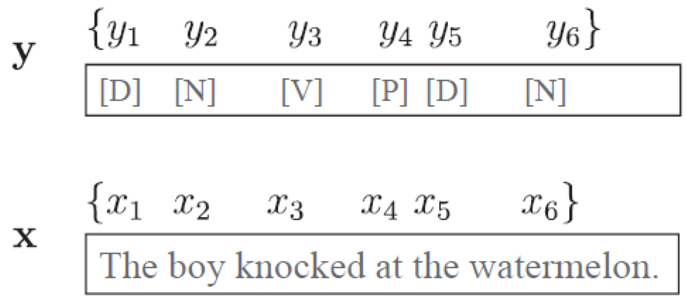
  - $t_j(y_{i+1}, y_i, \boldsymbol{x}, i)$是定义在观测序列的两个相邻标记位置上的转移特征函数（transition feature function），用于刻画相邻标记变量之间的相关关系以及观测序列对它们的影响

  - $s_k(y_i, \boldsymbol{x}, i)$是定义在观测序列的标记位置$i$上的状态特征函数（status feature function），用于刻画观测序列对标记变量的影响

  - λ_j, μ_k为参数，Z为规范化因子

CRF特征函数举例

- **特征函数**通常是实值函数，以刻画数据的一些**很可能成立**或者**期望成立**的经验特性，以词性标注任务为例：

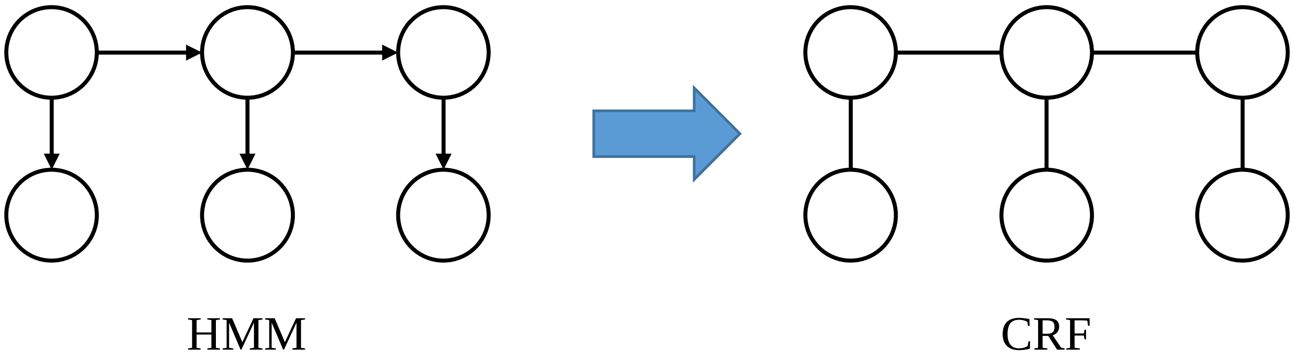


- 采用特征函数：

$$t_j(y_{i+1}, y_i, \boldsymbol{x}, i) = \begin{cases} 1, \text{if } y_{i+1} = [P], y_i = [V], \text{and } x_i = \text{"knock"} \\ 0, \ \text{otherwise} \end{cases}$$

  - 表示第i个观测值x_i为单词'knock'时，相应的标记y_i, y_{i+1}很可能分别为$[V], [P]$.

HMM to CRF



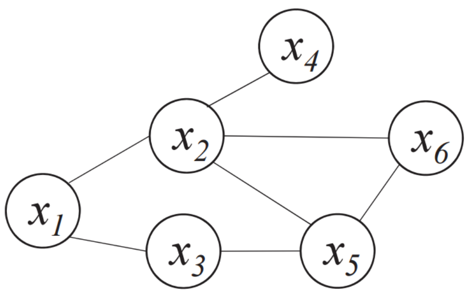
$$P(\boldsymbol{x},\boldsymbol{y}) = P(y_1)P(x_1|y_1)\prod_{t=2}^{n} P(y_t|y_{t-1})P(x_t|y_t)$$

$$= \pi_{y_1} b_{y_1,x_1} a_{y_1,y_2} b_{y_2,x_2} a_{y_2,y_3} \cdots a_{y_{n-1},y_n} b_{y_n,x_n}$$

$$= \exp\left(\sum_{t=1}^{n-1}\sum_{i,j} \mathbb{I}(y_t = s_i, y_{t+1} = s_j)\log a_{ij} + \sum_i \mathbb{I}(y_1 = s_i)\log \pi_i \right.$$
$$\left. + \sum_{t=1}^{n}\sum_{j,k} \mathbb{I}(y_t = s_j, x_t = o_k)\log b_{jk}\right)$$

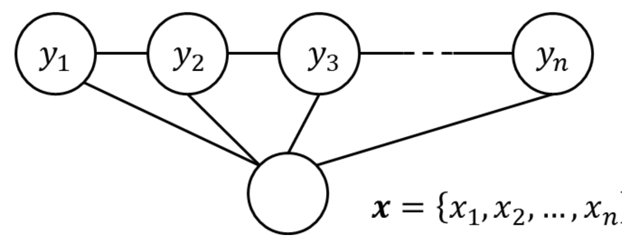
MRF与CRF的对比

- 使用团上的势函数定义概率
- 对**联合概率**建模



$$P(\boldsymbol{x}) = \frac{1}{Z}\psi_{12}(x_1, x_2)\psi_{13}(x_1, x_3)\psi_{24}(x_2, x_4)$$
$$\psi_{35}(x_3, x_5)\psi_{256}(x_2, x_5, x_6)$$

- 使用团上的势函数定义概率
- 有观测变量，对**条件概率**建模



$$P(\boldsymbol{y}|\boldsymbol{x}) = \frac{1}{Z}\exp\left(\sum_{j=1}^{K_1}\sum_{i=2}^{n}\lambda_j t_j(y_{i-1}, y_i, \boldsymbol{x}, i) + \sum_{l=1}^{K_2}\sum_{i=1}^{n}\mu_l s_l(y_i, \boldsymbol{x}, i)\right)$$

CRF基本问题

- **概率计算问题**：评估模型和观测序列间的匹配程度:
  有效计算观测序列产生概率$P(\boldsymbol{x}|\lambda)$

- **预测问题**：根据观测序列“推测”隐藏的模型状态
  $\boldsymbol{y} = \{y_1, y_2, ..., y_n\}$

- **学习问题**：如何调整模型参数 $\lambda = [A, B, \pi]$以使得
  该序列出现的概率$P(\boldsymbol{x}|\lambda)$最大

概率计算问题

$$P(\boldsymbol{y}|\boldsymbol{x}) = \frac{1}{Z}\exp\left(\sum_{j=1}^{K_1}\sum_{i=2}^{n}\lambda_j t_j(y_{i-1}, y_i, \boldsymbol{x}, i) + \sum_{l=1}^{K_2}\sum_{i=1}^{n}\mu_l s_l(y_i, \boldsymbol{x}, i)\right)$$

假设y_0 = "start“且y_{n+1} = "end"

记 $f_k(y_{i-1}, y_i, \boldsymbol{x}, i) = \begin{cases} t_k(y_{i-1}, y_i, \boldsymbol{x}, i), & k = 1, \dots, K_1 \\ s_l(y_i, \boldsymbol{x}, i), & l = K_1 + l; l = 1, \dots, K_2 \end{cases}$

定义 $M_i(\boldsymbol{x}) = [M_i(y_{i-1}, y_i|\boldsymbol{x})] \rightarrow \exp(W_i(y_{i-1}, y_i|\boldsymbol{x})) \rightarrow \sum_{k=1}^{K_1+K_2} w_k f_k(y_{i-1}, y_i, \boldsymbol{x}, i)$

$$P(\boldsymbol{y}|\boldsymbol{x}) = \frac{1}{Z}\prod_{i=1}^{n+1} M_i(y_{i-1}, y_i|\boldsymbol{x}) \qquad Z = \sum_{y_1,\dots,y_n}\prod_{i=1}^{n+1} M_i(y_{i-1}, y_i|\boldsymbol{x})$$

概率计算问题

$$Z=\sum_{y_1,\ldots,y_n}\prod_{i=1}^{n+1}M_i(y_{i-1},y_i|\boldsymbol{x})=\alpha_{n+1}(y_{n+1}|\boldsymbol{x})=[M_1(\boldsymbol{x})M_2(\boldsymbol{x})\cdots M_{n+1}(\boldsymbol{x})]_{start,stop}$$

$$=\underbrace{\sum_{y_1}M_1(y_0,y_1|\boldsymbol{x})\sum_{y_2}M_2(y_1,y_2|\boldsymbol{x})\sum_{y_3}M_3(y_2,y_3|\boldsymbol{x})\ldots\sum_{y_n}}_{\text{前t项求和}}M_n(y_{n-1},y_n|\boldsymbol{x})M_{n+1}(y_n,y_{n+1}|\boldsymbol{x})$$

$$\alpha_{t+1}(y_{t+1}|\boldsymbol{x})=\sum_{y_1,\ldots,y_t}\prod_{i=1}^{t+1}M_i(y_{i-1},y_i|\boldsymbol{x})$$

$$=\sum_{y_t}M_{t+1}(y_t,y_{t+1}|\boldsymbol{x})\left(\sum_{y_1,\ldots,y_{t-1}}\prod_{i=1}^{t}M_i(y_{i-1},y_i|\boldsymbol{x})\right)$$

$$=\sum_{y_t}M_{t+1}(y_t,y_{t+1}|\boldsymbol{x})\,\alpha_t(y_t|\boldsymbol{x})$$

记 $\alpha_0(y_0|\boldsymbol{x})=\begin{cases}1, y_0=start\\ 0, otherwise\end{cases}$

$\Rightarrow$ $\boldsymbol{\alpha}_{t+1}(\boldsymbol{x})=\boldsymbol{\alpha}_t(\boldsymbol{x})M_{t+1}(\boldsymbol{x})$ $\Rightarrow$ $\boldsymbol{\alpha}_{n+1}=\boldsymbol{\alpha}_0(\boldsymbol{x})M_1(\boldsymbol{x})M_2(\boldsymbol{x})\cdots M_{n+1}(\boldsymbol{x})$

概率计算问题

$$\beta_t(y_t|\boldsymbol{x}) = \sum_{y_{t+1},\dots,y_n} \prod_{i=t+1}^{n+1} M_i(y_{i-1}, y_i|\boldsymbol{x})$$

$$= \sum_{y_{t+1}} M_{t+1}(y_t, y_{t+1}|\boldsymbol{x}) \left(\sum_{y_{t+2},\dots,y_{n+1}} \prod_{i=t+2}^{n+1} M_i(y_{i-1}, y_i|\boldsymbol{x}) \right)$$

$$= \sum_{y_{t+1}} M_{t+1}(y_t, y_{t+1}|\boldsymbol{x})\, \beta_{t+1}(y_{t+1}|\boldsymbol{x})$$

⇒ $\boldsymbol{\beta}_t(\boldsymbol{x}) = M_{t+1}(\boldsymbol{x})\boldsymbol{\beta}_{t+1}(\boldsymbol{x})$ ⇒ $\boldsymbol{\beta}_t(\boldsymbol{x}) = M_{t+1}(\boldsymbol{x}) \cdots M_{n+1}(\boldsymbol{x})\boldsymbol{\beta}_{n+1}(\boldsymbol{x})$

记 $\beta_{n+1}(y_{n+1}|\boldsymbol{x}) = \begin{cases} 1, y_{n+1} = \text{end} \\ 0, \text{otherwise} \end{cases}$

概率计算问题

$$P(y_t|\boldsymbol{x}) = \sum_{\boldsymbol{y}_{-t}} P(\boldsymbol{y}|\boldsymbol{x}) \ = \sum_{\boldsymbol{y}_{-t}} \frac{1}{Z} \prod_{i=1}^{n+1} M_i(y_{i-1}, y_i|\boldsymbol{x})$$

$$= \sum_{\boldsymbol{y}_{1:t-1}} \sum_{y_{t+1:n}} \frac{1}{Z} \prod_{i=1}^{n+1} M_i(y_{i-1}, y_i|\boldsymbol{x})$$

$$= \frac{1}{Z} \sum_{\boldsymbol{y}_{1:t-1}} \prod_{i=1}^{t} M_i(y_{i-1}, y_i|\boldsymbol{x}) \sum_{\boldsymbol{y}_{t+1:n}} \prod_{i=t+1}^{n+1} M_i(y_{i-1}, y_i|\boldsymbol{x})$$

$$= \frac{1}{Z} \alpha_t(y_t|\boldsymbol{x}) \beta_t(y_t|\boldsymbol{x})$$

概率计算问题

$$P(y_{t-1}, y_t|\boldsymbol{x}) = \sum_{\boldsymbol{y}_{-t}} P(\boldsymbol{y}|\boldsymbol{x}) \quad = \sum_{\boldsymbol{y}_{-\{t-1,t\}}} \frac{1}{Z} \prod_{i=1}^{n+1} M_i(y_{i-1}, y_i|\boldsymbol{x})$$

$$= \sum_{\boldsymbol{y}_{1:t-2}} \sum_{y_{t+1:n}} \frac{1}{Z} \prod_{i=1}^{n+1} M_i(y_{i-1}, y_i|\boldsymbol{x})$$

$$= \frac{1}{Z} \sum_{\boldsymbol{y}_{1:t-2}} \prod_{i=1}^{t-1} M_i(y_{i-1}, y_i|\boldsymbol{x})\, M_t(y_{t-1}, y_t|\boldsymbol{x}) \sum_{\boldsymbol{y}_{t+1:n}} \prod_{i=t+1}^{n+1} M_i(y_{i-1}, y_i|\boldsymbol{x})$$

$$= \frac{1}{Z} \alpha_{t-1}(y_t|\boldsymbol{x}) M_t(y_{t-1}, y_t|\boldsymbol{x}) \beta_t(y_t|\boldsymbol{x})$$

预测问题（Viterbi Algorithm）

- 作为习题

$$f_k(y_{i-1}, y_i, \boldsymbol{x}, i) = \begin{cases} t_k(y_{i-1}, y_i, \boldsymbol{x}, i), & k = 1, \dots, K_1 \\ s_l(y_i, \boldsymbol{x}, i), & l = K_1 + l; l = 1, \dots, K_2 \end{cases}$$

记 $F_t(y_{t-1}, y_t, \boldsymbol{x}) = \left(f_1(y_{i-1}, y_i, \boldsymbol{x}, i), f_2(y_{i-1}, y_i, \boldsymbol{x}, i), \dots, f_K(y_{i-1}, y_i, \boldsymbol{x}, i)\right)$

$$K = K_1 + K_2$$

$$\boldsymbol{w} = \left(\lambda_1, \dots, \lambda_{K_1}, \mu_1, \dots, \mu_{K_2}\right)$$

$$P(\boldsymbol{y}|\boldsymbol{x}) = \frac{1}{Z}\exp\left(\sum_{j=1}^{K_1}\sum_{i=2}^{n}\lambda_j t_j(y_{i-1}, y_i, \boldsymbol{x}, i) + \sum_{l=1}^{K_2}\sum_{i=1}^{n}\mu_l s_l(y_i, \boldsymbol{x}, i)\right)$$

$$P(\boldsymbol{y}|\boldsymbol{x}) = \frac{1}{Z}\exp\left(\sum_{t=1}^{n}\boldsymbol{w}^{\top}F_t(y_{t-1}, y_t, \boldsymbol{x})\right)$$

学习问题（梯度下降法）

$$f_k(y_{i-1}, y_i, \boldsymbol{x}, i) = \begin{cases} t_k(y_{i-1}, y_i, \boldsymbol{x}, i), & k = 1, \dots, K_1 \\ s_l(y_i, \boldsymbol{x}, i), & l = K_1 + l; l = 1, \dots, K_2 \end{cases}$$

$$K = K_1 + K_2$$

$$f_k(\boldsymbol{y}, \boldsymbol{x}) = \sum_{i=1}^{n} f_k(y_{i-1}, y_i, \boldsymbol{x}, i)$$

$$\boldsymbol{w} = (\lambda_1, \dots, \lambda_{K_1}, \mu_1, \dots, \mu_{K_2})$$

$$P(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{w}) = \frac{1}{Z} \exp\left(\sum_{j=1}^{K_1} \sum_{i=2}^{n} \lambda_j t_j(y_{i-1}, y_i, \boldsymbol{x}, i) + \sum_{l=1}^{K_2} \sum_{i=1}^{n} \mu_l s_l(y_i, \boldsymbol{x}, i) \right)$$

$$= \frac{1}{Z_w(\boldsymbol{x})} \exp\left(\sum_{k=1}^{K} w_k f_k(\boldsymbol{y}, \boldsymbol{x}) \right)$$

$$Z_w(\boldsymbol{x}) = \sum_{\boldsymbol{y}} \exp\left(\sum_{k=1}^{K} w_k f_k(\boldsymbol{y}, \boldsymbol{x}) \right)$$

$$\log P(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{w}) = \sum_{k=1}^{K} w_k f_k(\boldsymbol{y}, \boldsymbol{x}) - \log Z_w(\boldsymbol{x})$$

学习问题（梯度下降法）

记 $F(\boldsymbol{y},\boldsymbol{x}) = \big(f_1(\boldsymbol{y},\boldsymbol{x}), f_2(\boldsymbol{y},\boldsymbol{x}), \dots, f_K(\boldsymbol{y},\boldsymbol{x})\big)$ $\quad f_k(\boldsymbol{y},\boldsymbol{x}) = \sum_{i=1}^{n} f_k(y_{i-1}, y_i, \boldsymbol{x}, i)$

$$\log P(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{w}) = \sum_{k=1}^{K} w_k f_k(\boldsymbol{y},\boldsymbol{x}) - \log Z_w(\boldsymbol{x})$$

$$\nabla_{\boldsymbol{w}} \log P(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{w}) = F(\boldsymbol{y},\boldsymbol{x}) - \boxed{\nabla_w \log Z_w(\boldsymbol{x})} = F(\boldsymbol{y},\boldsymbol{x}) - \mathbb{E}_{\tilde{\boldsymbol{y}}\sim P(\tilde{\boldsymbol{y}}|\boldsymbol{x},\boldsymbol{w})} F(\tilde{\boldsymbol{y}},\boldsymbol{x})$$

$$Z_w(\boldsymbol{x}) = \sum_{\boldsymbol{y}} \exp\left(\sum_{k=1}^{K} w_k f_k(\boldsymbol{y},\boldsymbol{x})\right)$$

$$= \frac{1}{Z_w(\boldsymbol{x})} \nabla_w Z_w(\boldsymbol{x})$$

$$= \frac{1}{Z_w(\boldsymbol{x})} \sum_{\boldsymbol{y}} \exp\left(\sum_{k=1}^{K} w_k f_k(\boldsymbol{y},\boldsymbol{x})\right) F(\boldsymbol{y},\boldsymbol{x})$$

$$= \sum_{\tilde{\boldsymbol{y}}} P(\tilde{\boldsymbol{y}}|\boldsymbol{x},\boldsymbol{w}) F(\tilde{\boldsymbol{y}},\boldsymbol{x}) = \mathbb{E}_{\tilde{\boldsymbol{y}}\sim P(\tilde{\boldsymbol{y}}|\boldsymbol{x},\boldsymbol{w})} F(\tilde{\boldsymbol{y}},\boldsymbol{x})$$

学习问题（梯度下降法）

Repeat

　　从数据集$D = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}$随机采样一个样本对

　　计算梯度信息$\nabla_{\boldsymbol{w}} \log P(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{w}) = F(\boldsymbol{y}, \boldsymbol{x}) - \mathbb{E}_{\tilde{\boldsymbol{y}} \sim P(\tilde{\boldsymbol{y}}|\boldsymbol{x}, \boldsymbol{w})} F(\tilde{\boldsymbol{y}}, \boldsymbol{x})$

　　权重更新 $w^{t+1} = w^t + \nabla_{\boldsymbol{w}} \log P(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{w})$

Until 收敛

计算代价高，如何提升计算效率呢?

习题

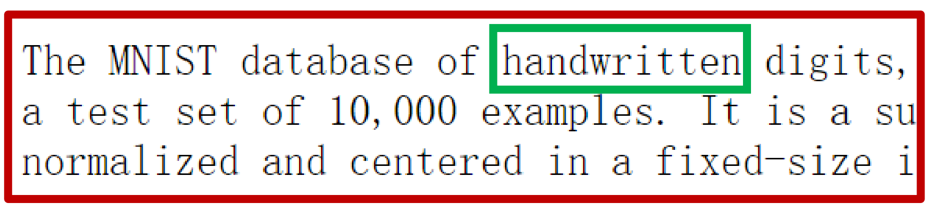
- 1. 在HMM中，求解概率$P(x_{n+1}|x_1, x_2, \ldots, x_n)$.
- 2. PPT 46，给出CRF的预测问题的解法

主题模型

- 主题模型（topic model）是一类生成式**有向图模型**，主要用来处理离散型的数据集合（如文本集合）
- 有效利用海量数据发现文档集合中**隐含的语义**
- 隐狄里克雷分配模型（Latent Dirichlet Allocation, **LDA**）是话题模型的典型代表

隐狄里克雷分配 LDA

- LDA的基本单元
  - **词（word）**：待处理数据中的基本离散单元
  - **文档（document）**：待处理的数据对象，由词组成，词在文档中**不计顺序。**数据对象只要能用“词袋”（bag-of-words）表示就可以使用话题模型
  - **话题（topic）**：表示一个概念，具体表示为一系列相关的词，以及它们在该概念下出现的概率


The MNIST database of handwritten digits,
a test set of 10,000 examples. It is a su
normalized and centered in a fixed-size i

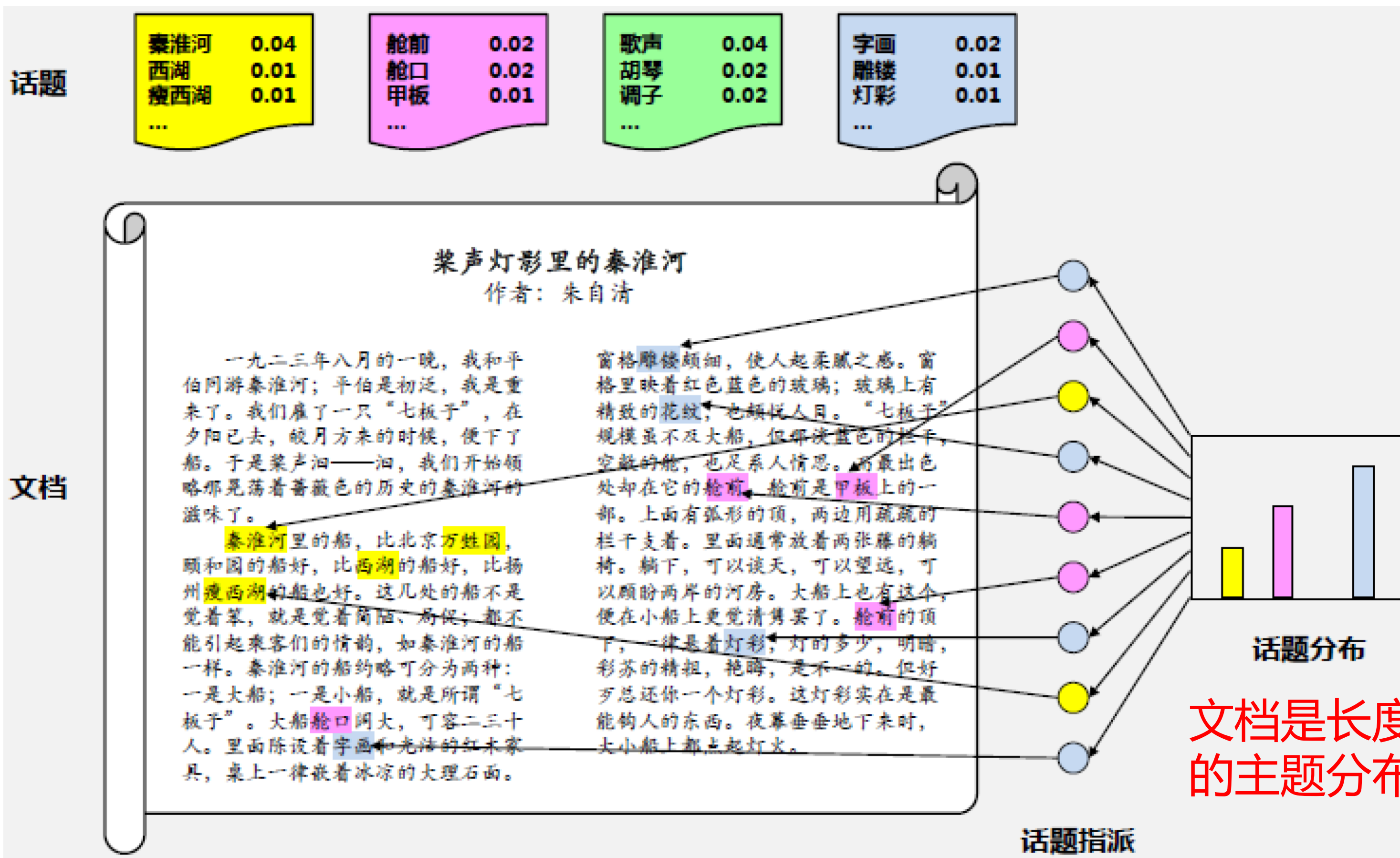
数据　计算机　生物　新闻

建筑　植物　天空

隐狄里克雷分配 LDA

设文档中的词来自一个包含V个词的字典

主题是V维概率词向量

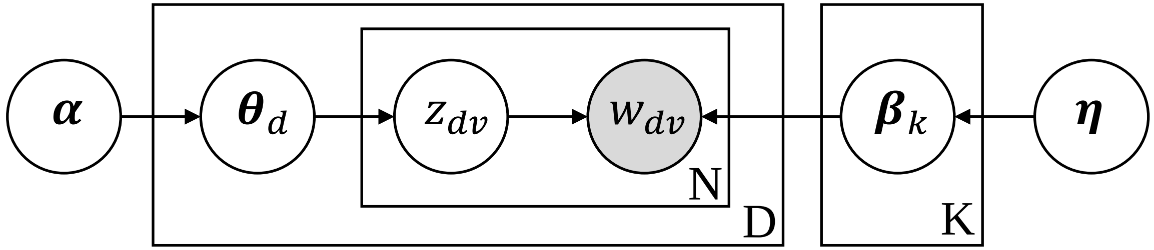


文档是长度为K的主题分布向量

隐狄里克雷分配 LDA

- 假定数据集中共含K个话题和D篇文档，词来自含V个词的字典
- 观测数据：D篇文档，每篇文档用长度为V的词频向量表示
  - $\mathcal{D}=\{\boldsymbol{w}_1,\boldsymbol{w}_2,\dots,\boldsymbol{w}_D\}$ 其中$\boldsymbol{w}_d=\left(w_{d1},\dots,w_{dN_d}\right)$为单词序列
  - 隐变量：K个话题$Z=\{z_1,\dots,z_K\}$，每个话题用长度为N的概率词向量表示
  - 第k话题 $\beta_k\in[0,1]^V$ $\beta_{kv}=P(w_v|z_k)$表示在第k个话题中单词w_v的概率
- 文档表示为话题的分布，由参数$\boldsymbol{\Theta}$ 确定
  - $\boldsymbol{\theta}_t\in[0,1]^K$ $\theta_{mk}=P(z_k|\boldsymbol{w}_m)$

隐狄里克雷分配 LDA

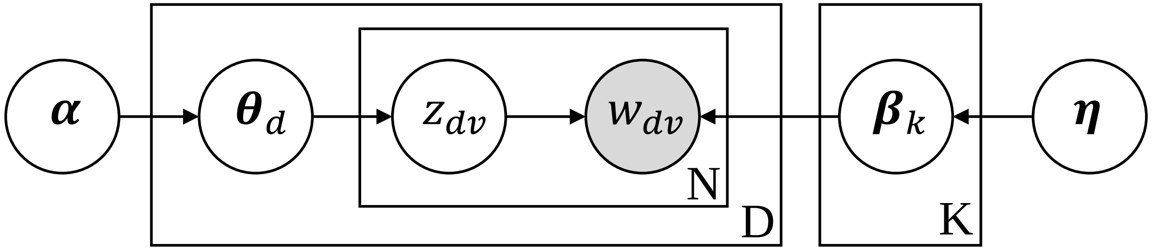


生成文档d过程

- 从以$\boldsymbol{\alpha}$为参数的狄利克雷分布中随机采样一个话题分布$\boldsymbol{\theta}_d$；
- 按如下步骤产生文档中的N_d个词
  - 根据$\boldsymbol{\theta}_d$进行话题指派，得到文档d中词v的话题$z_{d,v}$；
  - 根据指派的话题$z_{d,v}$所对应的的词分布β_k随机采样生成词w_{dv}

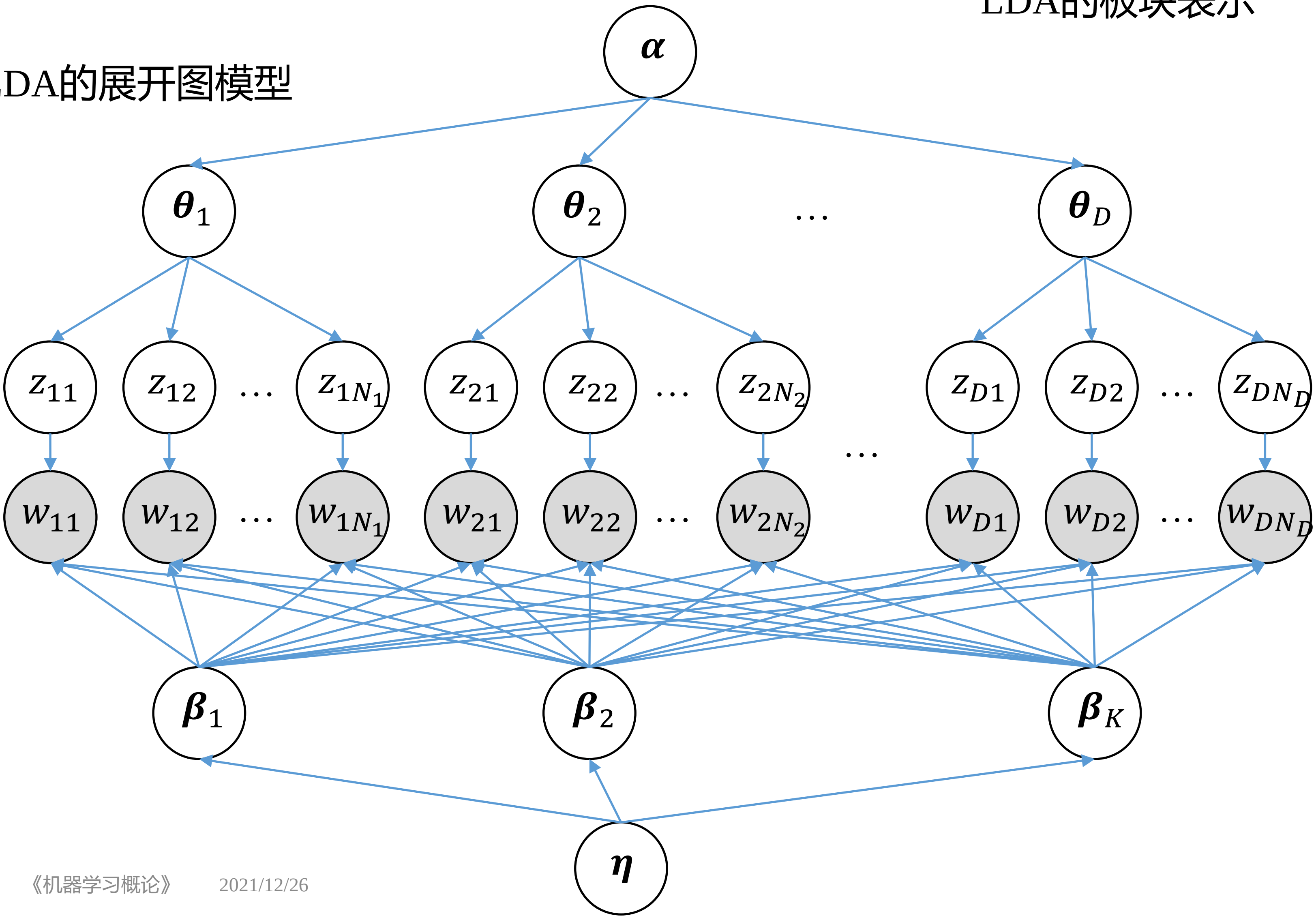
生成主题k过程

- 从以$\boldsymbol{\eta}$为参数的狄利克雷分布中随机采样一个话题分布$\boldsymbol{\beta}_k$；



LDA的板块表示

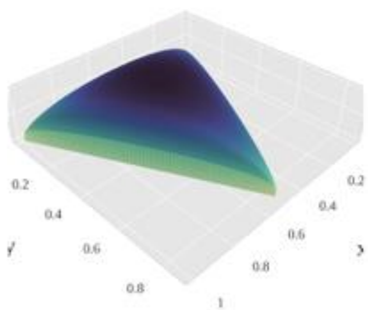
LDA的展开图模型



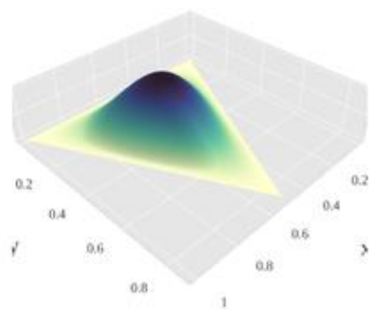
狄里克雷分布

$$\frac{1}{B(\boldsymbol{\alpha})}$$

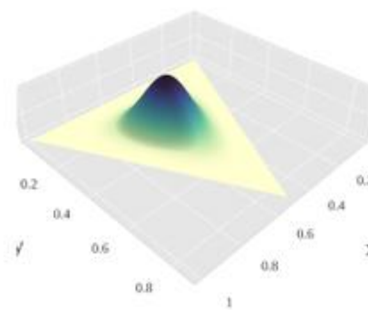
$$f(\boldsymbol{\theta}|\boldsymbol{\alpha}) = \boxed{\frac{\Gamma\left(\sum_{k=1}^{K}\alpha_k\right)}{\prod_{k=1}^{K}\Gamma(\alpha_k)}} \prod_{k=1}^{K}\theta_k^{\alpha_k-1} \qquad \sum_k \theta_k = 1$$



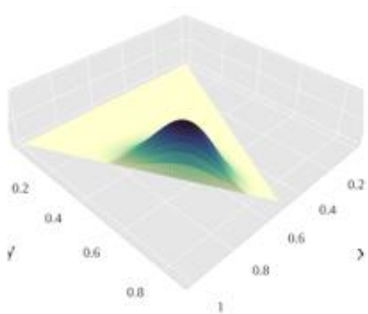
$\boldsymbol{\alpha} = (1.3, 1.3, 1.3)$



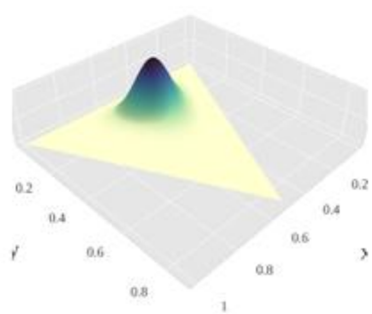
$\boldsymbol{\alpha} = (3,3,3)$



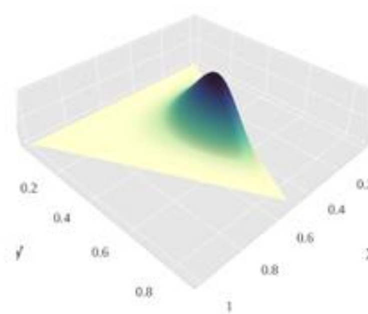
$\boldsymbol{\alpha} = (7,7,7)$



$\boldsymbol{\alpha} = (2,6,11)$

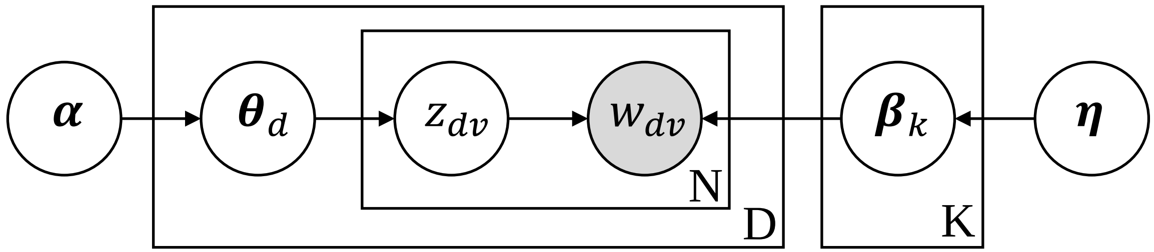


$\boldsymbol{\alpha} = (14, 9, 5)$



$\boldsymbol{\alpha} = (6,2,6)$

隐狄里克雷分配 LDA



$$p(\boldsymbol{w},\boldsymbol{z},\boldsymbol{\beta},\boldsymbol{\Theta}|\boldsymbol{\alpha},\boldsymbol{\eta}) = \prod_{d}^{D} \boxed{p(\boldsymbol{\theta}_d|\alpha)} \prod_{k}^{K} \boxed{p(\boldsymbol{\beta}_k|\eta)} \left(\prod_{v=1}^{N_d} P(w_{dv}|z_{dv},\boldsymbol{\beta}_k) P(z_{dv}|\boldsymbol{\theta}_d) \right)$$

狄里克雷分布

LDA模型的**参数估计**

- 给定训练数据$\boldsymbol{w} = \{\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_D\}$, 参数通过极大似然法估计，寻找$\boldsymbol{\alpha}$和$\boldsymbol{\eta}$以最大化对数似然

$$LL(\boldsymbol{\alpha}, \boldsymbol{\eta}) = \sum_{d=1}^{D} \ln p(\boldsymbol{w}_d | \boldsymbol{\alpha}, \boldsymbol{\eta})$$

- 求解算法
  - 可以通过吉布斯采样求解
  - 可以通过变分法求解

LDA模型的**参数估计**

- 给定训练数据$\boldsymbol{w} = \{\boldsymbol{w}_1, \boldsymbol{w}_2, \dots, \boldsymbol{w}_D\}$, 参数通过极大似然法估计，寻找$\boldsymbol{\alpha}$和$\boldsymbol{\eta}$以最大化对数似然

$$LL(\boldsymbol{\alpha}, \boldsymbol{\eta}) = \sum_{d=1}^{D} \ln p(\boldsymbol{w}_d | \boldsymbol{\alpha}, \boldsymbol{\eta})$$

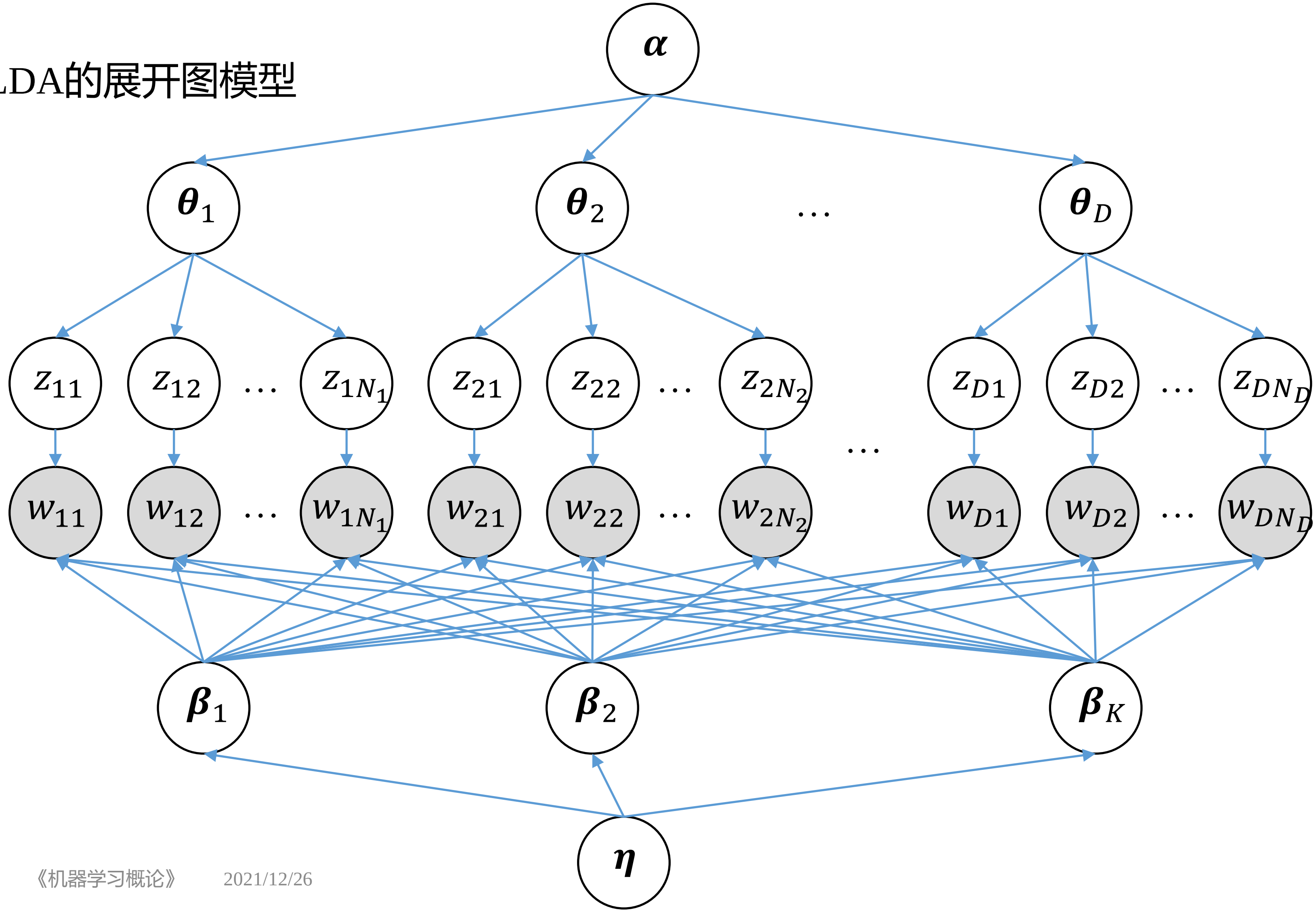
- 能否进行模型推断后用EM算法呢?

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

模型推断的挑战

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

LDA的展开图模型



LDA模型的参数估计

- 给定训练数据$\boldsymbol{w}=\{\boldsymbol{w}_1,\boldsymbol{w}_2,\dots,\boldsymbol{w}_D\}$, 参数通过极大似然法估计，寻找$\boldsymbol{\alpha}$和$\boldsymbol{\eta}$以最大化对数似然

$$LL(\boldsymbol{\alpha},\boldsymbol{\eta})=\sum_{d=1}^{D}\ln p(\boldsymbol{w}_d|\boldsymbol{\alpha},\boldsymbol{\eta})$$

- 能否进行模型推断后用EM算法呢?

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

无法进行

- 求解算法
  - 可以通过吉布斯采样求解：通过使用随机化方法完成近似
  - 可以通过变分法求解：使用确定性近似完成推断

近似推断：采样法

- 核心思想：用 一组样本 近似 分布

  - 设某个计算机程序可以产生正态分布的样本，但是参数未知。那么可以不断调用该程序，产生一组样本，从而通过样本来估计均值和方差

  - 考虑计算$\mathbb{E}_p[f(x)] = \int p(x)f(x)dx$ 或者 $\mathbb{E}_p[f(x)] = \sum_x p(x)f(x)$，可以通过从分布$p(x)$中抽样n个样本$x^{(1)}, \cdots, x^{(n)}$

$$\mathbb{E}_p[f(x)] = \frac{1}{n}\sum_i f\left(x^{(i)}\right)$$

关键难题：那应该如何从$p(x)$从采样呢?

标准分布采样

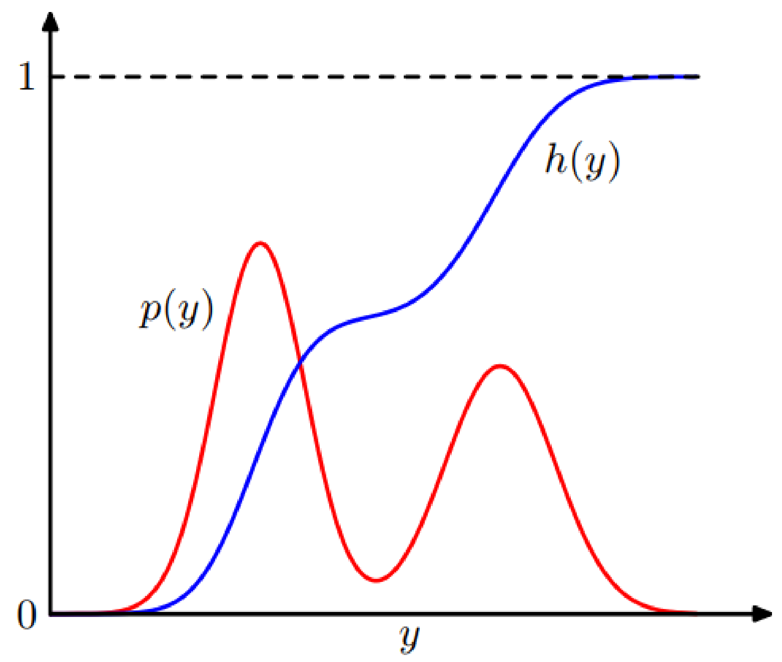
- 伯努利分布采样：$P(x) = p^x(1-p)^{1-x}$

$$z \sim U(0,1), \qquad x = \begin{cases} 1, z \le p \\ 0, othewise \end{cases}$$

$$P(z \le p) = p$$

标准分布采样

- 假设z满足均匀分布，$z \sim U(0,1)$, 通过对z进行变换可以求得相应分布$p(y)$
- $p(y) = p(z)\left|\frac{dz}{dy}\right| = \left|\frac{dz}{dy}\right|$
- 左右两边积分后，可得$z = h(y) = \int_{-\infty}^{y} p(\hat{y})d\hat{y}$
- 若$\mathrm{y} = \mathrm{h}^{-1}(\mathrm{z})$，那么$y \sim p(y)$



标准分布采样

- 指数分布采样：$p(y) = \lambda \exp(-\lambda y)$
  - $h(y) = 1 - \exp(-\lambda y)$
  - $y = h^{-1}(z) = -\lambda^{-1} \ln(1 - z)$
- 标准柯西分布采样：$p(y) = \frac{1}{\pi} \frac{1}{1+y^2}$
  - $h(y) = \frac{1}{\pi} \arctan(y) + \frac{1}{2}$
  - $y = h^{-1}(z) = \tan\left(z\pi - \frac{\pi}{2}\right)$

标准分布采样

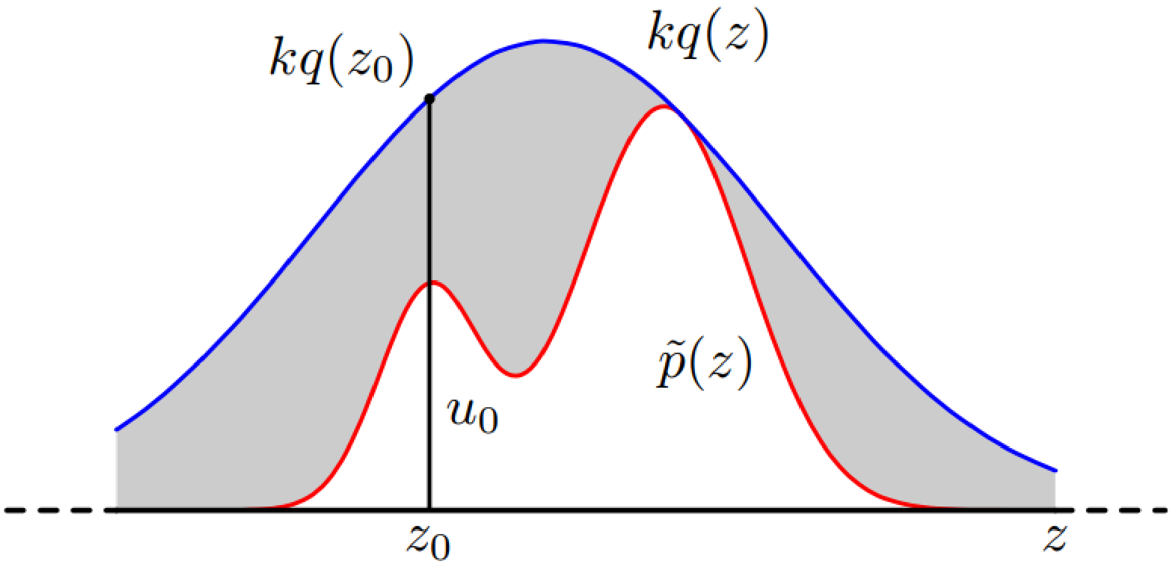
- 标准正态分布
  - $z_1, z_2 \sim U(-1,1)$
  - 如果$z_1^2 + z_2^2 > 1$，则丢掉这对样本，那么 $p(z_1, z_2) = \frac{1}{\pi}$
  - 令$y_1 = z_1 \left(\frac{-2\ln z_1}{z_1^2+z_2^2}\right)^{\frac{1}{2}}, y_2 = z_2 \left(\frac{-2\ln z_2}{z_1^2+z_2^2}\right)^{\frac{1}{2}}$, 则有
  - $p(y_1, y_2) = p(z_1, z_2) \left|\frac{\partial(z_1,z_2)}{\partial(y_1,y_2)}\right| = \left[\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}y_1^2\right)\right]\left[\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}y_2^2\right)\right]$
  - y_1和y_2 独立，且均值为0，方差为1

非标准分布采样

- 难以计算分布的累计概率分布并求逆函数
- 借助于两种重要而且通用的方法
  - 拒绝采样 Rejection sampling
  - 重要性重采样 Importance resampling

拒绝采样

- 从$p(z)$采样很难，但是计算$p(z)$很容易，可以不包括归一化项
  - $p(z) = \frac{1}{Z_p}\tilde{p}(z)$
  - $\tilde{p}(z)$很容易计算，但Z_p未知
- 提议分布 (proposal distribution): $q(z)$，可以容易从中采样
- 寻找常数k，满足$kq(z) \geq \tilde{p}(z)$



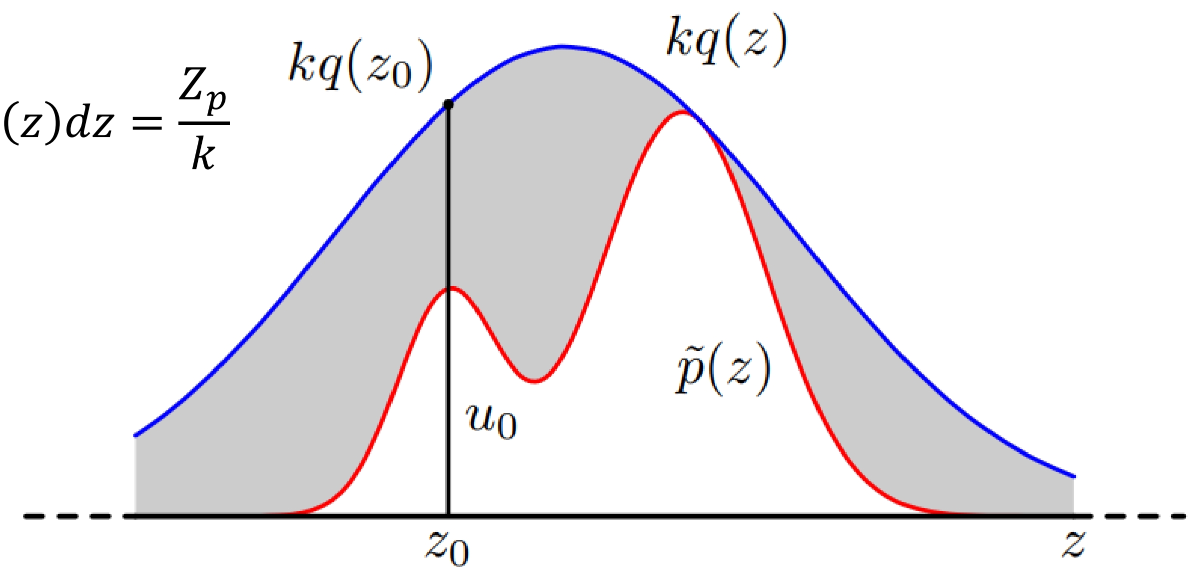
拒绝采样

采样过程

- 从$q(z)$从采样z_0
- 从$U\big(0, kq(z_0)\big)$采样u_0
- 如果$u_0 > \tilde{p}(z_0)$，样本被拒绝，否则被保留

- 图中阴影部分的样本被拒绝

$$p(Accept) = \int \left\{ \frac{\tilde{p}(z)}{\mathrm{k}q(z)} \right\} q(z)dz = \frac{Z_p}{k}$$



拒绝采样在高维分布的问题

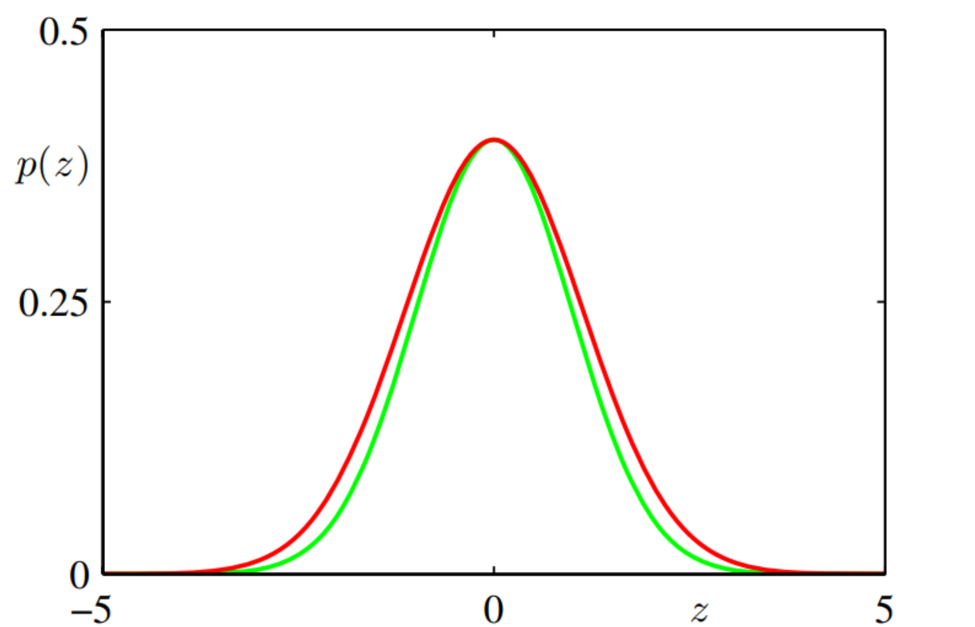
- $p(z)$是多维高斯分布，均值为0，协方差矩阵为$\sigma_p I$
- $q(z)$是多维高斯分布，均值为0，协方差矩阵为$\sigma_q I$

- 满足$\sigma_q \geq \sigma_p, k^{\star} = \left(\frac{\sigma_q}{\sigma_q}\right)^D$

- $P(Accept) = \frac{1}{k}$

维度越高，接收概率指数级减小

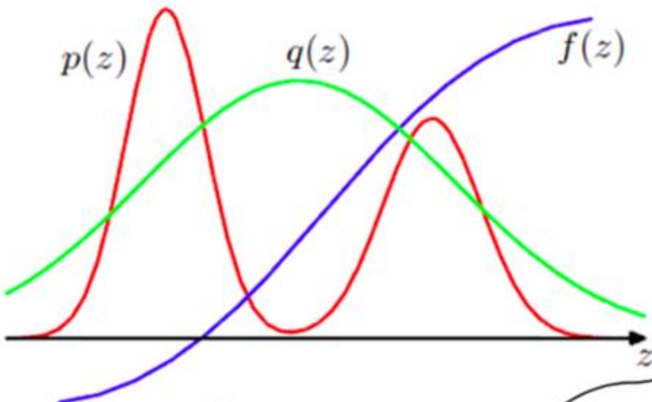
在高维情况下效率极低



重要性重采样 Importance resampling

- 并不直接丢弃样本，因此采样效率提升
- 会根据提议分布和采样分布上概率的差别给每个样本加权
  - 如果采样分布$p(z)$概率小，而提议分布$q(z)$概率大，权重越小
  - 如果采样分布$p(z)$概率大，而提议分布$q(z)$概率小，权重越大
- 有一个前提条件：在$p(z)$显著的地方$q(z)$不能太小
- 同样难以解决从高维概率分布采样的问题

重要性重采样 Importance resampling



重要性权重，用来矫正从"错误"分布采样带来的偏差

$$\mathrm{E}[f] = \int f(z) p(z) dz$$

$$= \int f(z) \frac{p(z)}{q(z)} q(z) dz$$

$$\approx \frac{1}{L} \sum_{l=1}^{L} \boxed{\frac{p\left(z^{(l)}\right)}{q\left(z^{(l)}\right)}} f\left(z^{(l)}\right)$$

$$\boxed{\omega\left(z^{(l)}\right) = p\left(z^{(l)}\right) / q\left(z^{(l)}\right)}$$

Importance Sampling

- 产生样本 $z_1, \ldots, z_L \sim q(z)$
- 计算

$$\hat{I} = \frac{1}{L} \sum_{l=1}^{L} f\left(z^{(l)}\right) \omega\left(z^{(l)}\right)$$

马尔可夫链蒙特卡罗方法（MCMC）

- 拒绝采样和重要性采样在高维情形下效率很低，通过马尔可夫链蒙特卡罗方法可以更好地扩展到高维情形

- **MCMC算法的关键在于通过“构造平稳分布为p的马尔可夫链”来产生样本**：当马尔可夫链运行足够长的时间（收敛到平稳状态），则产出的样本x近似服从p分布；并且通过多次重复运行、遍历马尔可夫链就可以取得多个服从该分布的独立同分布样本。

- 前提
  - 难以从$p(z) = \tilde{p}(z)/Z_p$中采样，但是很容易计算$\tilde{p}(z)$
  - 提议分布$q(z|z^\tau)$, 容易采样：z^τ为当前状态

- $z_1, z_2, \cdots$形成一个马尔可夫链。算法的每一轮，从提议分布中采样一个样本，按照某种规则来决定是否接受这个样本

关于MCMC的一些知识

- 一阶马尔可夫链有初始概率$p^{(0)}(z)$和转移概率$p(z'|z)$确定

- 在t时刻的状态采样概率为 $P^{(t)}(z)$，用向量$\boldsymbol{\pi}^{(t)}$来表示所有状态的概率$P^{(t)}(z)$

- 转移概率用转移矩阵表示： $A_{i,j} = p(z' = i|z = j)$

- 马尔科夫链的动态性用如下方式表示

$$P^{(t)}(z = i) = \sum_{j} P^{(t-1)}(z = j)P(z' = i|z = j)$$

$$\boldsymbol{\pi}^{(t)} = \boldsymbol{A}\boldsymbol{\pi}^{(t-1)} = \boldsymbol{A}^{(t)}\boldsymbol{\pi}^{(0)}$$

关于MCMC的一些知识

- 马尔科夫链的动态性用如下方式表示

$$\boldsymbol{\pi}^{(t)} = A\boldsymbol{\pi}^{(t-1)} = A^{(t)}\boldsymbol{\pi}^{(0)}$$

- A的每一列代表一个概率分布，A称为随机矩阵
- 任意状态之间在可达的情况下，A的最大特征值等于1；
- 当$t \to \infty$时，不等于1的特征值都衰减到0
- 在一些额外宽松条件下，A只有一个特征值为1的特征向量
  - MCMC会收敛到平稳分布 $\boldsymbol{\pi} = \boldsymbol{A}\boldsymbol{\pi}$， 即$\boldsymbol{\pi}$为对应的特征向量

关于MCMC的一些知识

- 遍历马尔可夫链有唯一的平稳分布 $p^{\star}(z)$
  - 平稳分布满足：$p^{\star}(z) = \sum_{z'} p(z|z')p^{\star}(z')$
  - 遍历性 $\lim_{\tau\to\infty} p^{(\tau)}(z) = p^{\star}(z)$，不管初始概率分布的选择
- $p^{\star}(z)$是平稳分布的充分条件：$p^{\star}(z')p(z|z') = p(z'|z)p^{\star}(z)$

Metropolis 算法

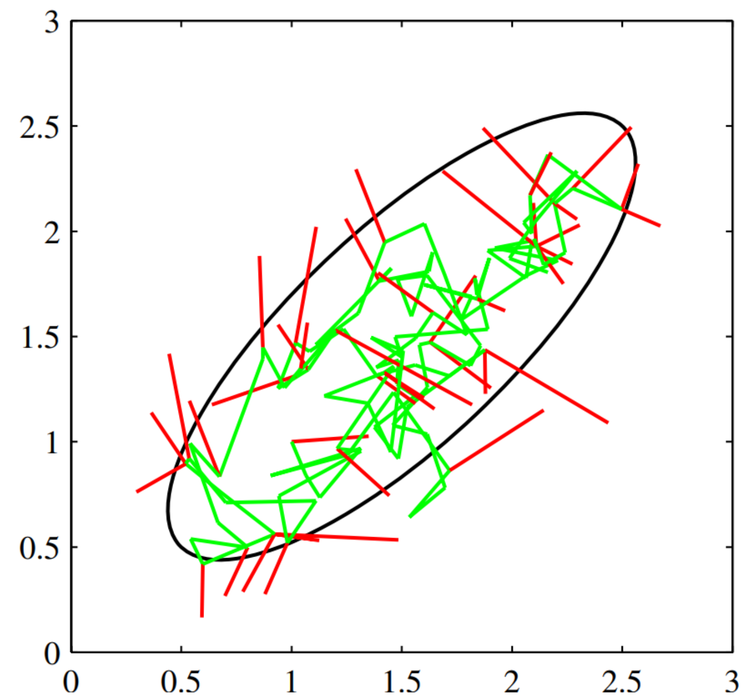
- 提议分布是对称的，满足$q(\boldsymbol{z}_A|\boldsymbol{z}_B) = q(\boldsymbol{z}_B|\boldsymbol{z}_A)$

- 从提议分布中采样的候选样本被接受的概率为

$$A\left(\boldsymbol{z}^\star, \boldsymbol{z}^{(\tau)}\right) = \min\left(1, \frac{\tilde{p}(\boldsymbol{z}^\star)}{\tilde{p}(\boldsymbol{z}^{(\tau)})}\right)$$

  - 如果$\tilde{p}(\boldsymbol{z}^\star) > \tilde{p}\left(\boldsymbol{z}^{(\tau)}\right)$，那么$\boldsymbol{z}^\star$一定被接收

- 如果候选样本被接受，$\boldsymbol{z}^{(\tau+1)} = \boldsymbol{z}^\star$，否则$\boldsymbol{z}^{(\tau+1)} = \boldsymbol{z}^{(\tau)}$
  - 如果$\boldsymbol{z}^\star$被拒绝，$\boldsymbol{z}^{(\tau)}$被拷贝一次



状态$\boldsymbol{z}^{(\tau)}$到状态$\boldsymbol{z}^\star$的转移概率为$q\left(\boldsymbol{z}^\star|\boldsymbol{z}^{(\tau)}\right)A\left(\boldsymbol{z}^\star, \boldsymbol{z}^{(\tau)}\right)$

只要$q(z_A|z_B)$是正的，$z^{(\tau)}$的分布趋近于$p(z)$

Metropolis-Hastings 算法

Algorithm 24.2: Metropolis Hastings algorithm

Initialize x^0 ;
for $s = 0, 1, 2, \ldots$ **do**
  Define $x = x^s$;
  Sample $x' \sim q(x'|x)$;
  Compute acceptance probability

$$\alpha = \frac{\tilde{p}(x')q(x|x')}{\tilde{p}(x)q(x'|x)}$$

  Compute $r = \min(1, \alpha)$;
  Sample $u \sim U(0, 1)$;
  Set new sample to

$$x^{s+1} = \begin{cases} x' & \text{if } u < r \\ x^s & \text{if } u \geq r \end{cases}$$

连续变量提议分布一般选择以当前状态为均值的高斯分布，方差要做权衡

- 推广到非对称的提议分布，只需要修改接受概率

$$A(\boldsymbol{z}^\star, \boldsymbol{z}^{(\tau)}) = \min\left(1, \frac{\tilde{p}(\boldsymbol{z}^\star)q\left(\boldsymbol{z}^{(\tau)}|\boldsymbol{z}^\star\right)}{\tilde{p}(\boldsymbol{z}^{(\tau)})q\left(\boldsymbol{z}^\star|\boldsymbol{z}^{(\tau)}\right)}\right)$$

状态$\boldsymbol{z}^{(\tau)}$到状态$\boldsymbol{z}^\star$的转移概率为$q\left(\boldsymbol{z}^\star|\boldsymbol{z}^{(\tau)}\right)A\left(\boldsymbol{z}^\star, \boldsymbol{z}^{(\tau)}\right)$

Metropolis-Hastings 算法

- $p(z)$是该算法定义的马尔科夫链的平稳分布

$$A\left(z^{\star}, z^{(\tau)}\right)=\min\left(1, \frac{\tilde{p}(z^{\star})q\left(z^{(\tau)}|z^{\star}\right)}{\tilde{p}(z^{(\tau)})q(z^{\star}|z^{(\tau)})}\right)$$

$$\begin{aligned} p\left(z^{(\tau)}\right)q\left(z^{\star}|z^{(\tau)}\right)A\left(z^{\star}, z^{(\tau)}\right) &= \min\left(p\left(z^{(\tau)}\right)q\left(z^{\star}|z^{(\tau)}\right), p(z^{\star})q\left(z^{(\tau)}|z^{\star}\right)\right) \\ &= \min\left(p(z^{\star})q\left(z^{(\tau)}|z^{\star}\right), p\left(z^{(\tau)}\right)q\left(z^{\star}|z^{(\tau)}\right)\right) \\ &= p(z^{\star})q\left(z^{(\tau)}|z^{\star}\right)A\left(z^{(\tau)}, z^{\star}\right) \end{aligned}$$

Gibbs 采样

- Metropolis-Hastings 算法一个特列

$$\text{状态}\boldsymbol{z}^{(\tau)}\text{到状态}\boldsymbol{z}^{\star}\text{的转移概率}q_k\left(\boldsymbol{z}^{\star}|\boldsymbol{z}^{(\tau)}\right)=p(z_k^{\star}|\boldsymbol{z}_{-\mathrm{k}})\mathrm{I}(\boldsymbol{z}_{-\mathrm{k}}^{\star}=\boldsymbol{z}_{-\mathrm{k}})$$

1. Initialize $\{z_i : i = 1, \ldots, M\}$
2. For $\tau = 1, \ldots, T$:
   - Sample $z_1^{(\tau+1)} \sim p(z_1|z_2^{(\tau)}, z_3^{(\tau)}, \ldots, z_M^{(\tau)})$.
   - Sample $z_2^{(\tau+1)} \sim p(z_2|z_1^{(\tau+1)}, z_3^{(\tau)}, \ldots, z_M^{(\tau)})$.

     $\vdots$
   - Sample $z_j^{(\tau+1)} \sim p(z_j|z_1^{(\tau+1)}, \ldots, z_{j-1}^{(\tau+1)}, z_{j+1}^{(\tau)}, \ldots, z_M^{(\tau)})$.

     $\vdots$
   - Sample $z_M^{(\tau+1)} \sim p(z_M|z_1^{(\tau+1)}, z_2^{(\tau+1)}, \ldots, z_{M-1}^{(\tau+1)})$.

可以证明：
Metropolis-Hastings步总是被接受

Gibbs 采样

- 可以证明：Metropolis-Hastings步总是被接受
  - $q_k(\boldsymbol{z}^\star|\boldsymbol{z}) = p(z_k^\star|\boldsymbol{z}_{-\mathrm{k}})\mathrm{I}(\boldsymbol{z}_{-\mathrm{k}}^\star = \boldsymbol{z}_{-\mathrm{k}})$
  - $p(\boldsymbol{z}) = p(z_k|\boldsymbol{z}_{-\mathrm{k}})p(\boldsymbol{z}_{-\mathrm{k}})$
  - $\boldsymbol{z}_{-\mathrm{k}}^\star = \boldsymbol{z}_{-\mathrm{k}}$
  - $A(\boldsymbol{z}^\star, \boldsymbol{z}) = \min\left(1, \frac{p(\boldsymbol{z}^\star)q_k(\boldsymbol{z}|\boldsymbol{z}^\star)}{p(\boldsymbol{z})q_k(\boldsymbol{z}^\star|\boldsymbol{z})}\right)$

$$= \min\left(1, \frac{p(z_k^\star|\boldsymbol{z}_{-\mathrm{k}}^\star)p(\boldsymbol{z}_{-\mathrm{k}}^\star)p(z_k|\boldsymbol{z}_{-\mathrm{k}}^\star)\mathrm{I}(\boldsymbol{z}_{-\mathrm{k}}^\star = \boldsymbol{z}_{-\mathrm{k}})}{p(z_k|\boldsymbol{z}_{-\mathrm{k}})p(\boldsymbol{z}_{-\mathrm{k}})p(z_k^\star|\boldsymbol{z}_{-\mathrm{k}})\mathrm{I}\left(\boldsymbol{z}_{-\mathrm{k}}^\star = \boldsymbol{z}_{-\mathrm{k}}\right)}\right)$$
$$= 1$$

LDA模型的参数估计

- 给定训练数据$\boldsymbol{w}=\{\boldsymbol{w}_1,\boldsymbol{w}_2,\dots,\boldsymbol{w}_D\}$, 参数通过极大似然法估计，寻找$\boldsymbol{\alpha}$和$\boldsymbol{\eta}$以最大化对数似然

$$LL(\boldsymbol{\alpha},\boldsymbol{\eta})=\sum_{d=1}^{D}\ln p(\boldsymbol{w}_d|\boldsymbol{\alpha},\boldsymbol{\eta})$$

- 能否进行模型推断后用EM算法呢?

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

- 求解算法
  - 可以通过收缩的吉布斯采样求解：通过使用随机化方法完成近似

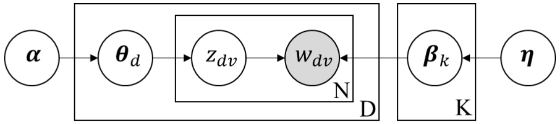
LDA模型的**参数估计**—吉布斯采样

- 基本思想是通过对隐变量θ和β积分，得到边缘概率 $P(\boldsymbol{z}_d|\boldsymbol{w}_d, \boldsymbol{\alpha}, \boldsymbol{\eta})$
- 对后验概率进行吉布斯抽样，得到分布$P(\boldsymbol{z}_d|\boldsymbol{w}_d, \boldsymbol{\alpha}, \boldsymbol{\eta})$的样本集合
- 利用这个样本集合对参数$\boldsymbol{\alpha}$ 和$\boldsymbol{\eta}$ 进行参数估计

LDA模型的**参数估计**—吉布斯采样

- 基本思想是通过对隐变量θ和β积分，得到边缘概率$P(\boldsymbol{z}|\boldsymbol{w},\boldsymbol{\alpha},\boldsymbol{\eta})$

$$P(\boldsymbol{z}|\boldsymbol{w},\boldsymbol{\alpha},\boldsymbol{\eta}) = \frac{P(\boldsymbol{z},\boldsymbol{w}|\boldsymbol{\alpha},\boldsymbol{\eta})}{P(\boldsymbol{w}|\boldsymbol{\alpha},\boldsymbol{\eta})}$$



$$\propto P(\boldsymbol{w}|\boldsymbol{z},\boldsymbol{\alpha},\boldsymbol{\eta})P(\boldsymbol{z}|\boldsymbol{\alpha},\boldsymbol{\eta}) = P(\boldsymbol{w}|\boldsymbol{z},\boldsymbol{\eta})P(\boldsymbol{z}|\boldsymbol{\alpha})$$

$$P(\boldsymbol{w}|\boldsymbol{z},\boldsymbol{\eta}) = \int \boxed{p(\boldsymbol{w}|\boldsymbol{z},\boldsymbol{\beta})}\, p(\boldsymbol{\beta}|\boldsymbol{\eta})d\boldsymbol{\beta}$$

$$p(\boldsymbol{w}|\boldsymbol{z},\boldsymbol{\beta}) = \prod_{d=1}^{D}\prod_{w=1}^{N_d}\beta_{z_{dw}w} = \prod_{d=1}^{D}\prod_{w=1}^{N_d}\prod_{k=1}^{K}\beta_{kw}^{\mathbb{I}(z_{dw}=k)}$$

$$= \prod_{v=1}^{V}\prod_{k=1}^{K}\beta_{kv}^{\sum_d\sum_{w=v}\mathbb{I}(z_{dw}=k)}$$

$$= \prod_{v=1}^{V}\prod_{k=1}^{K}\beta_{kv}^{n_{kv}}$$

数据中第k个话题生成第v个单词的次数

$$= \int \prod_{v=1}^{V}\prod_{k=1}^{K}\beta_{kv}^{n_{kv}}\prod_{k=1}^{K}\left(\frac{1}{B(\boldsymbol{\eta})}\prod_{v=1}^{V}\beta_{kv}^{\eta_v-1}\right)d\boldsymbol{\beta}$$

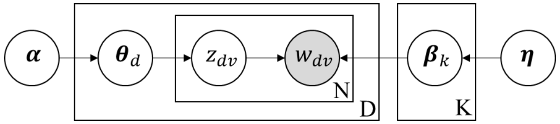
$$= \prod_{k=1}^{K}\frac{1}{B(\boldsymbol{\eta})}\left(\int\prod_{v=1}^{V}\beta_{kv}^{n_{kv}+\eta_v-1}\,d\boldsymbol{\beta}_k\right)$$

$$= \prod_{k=1}^{K}\frac{B(\boldsymbol{\eta}+\boldsymbol{n}_k)}{B(\boldsymbol{\eta})}$$

LDA模型的**参数估计**—吉布斯采样

- 基本思想是通过对隐变量θ和β积分，得到边缘概率$P(\boldsymbol{z}|\boldsymbol{w},\boldsymbol{\alpha},\boldsymbol{\eta})$

$$P(\boldsymbol{z}|\boldsymbol{w},\boldsymbol{\alpha},\boldsymbol{\eta}) = \frac{P(\boldsymbol{z},\boldsymbol{w}|\boldsymbol{\alpha},\boldsymbol{\eta})}{P(\boldsymbol{w}|\boldsymbol{\alpha},\boldsymbol{\eta})}$$



$$\propto P(\boldsymbol{w}|\boldsymbol{z},\boldsymbol{\alpha},\boldsymbol{\eta})P(\boldsymbol{z}|\boldsymbol{\alpha},\boldsymbol{\eta}) = P(\boldsymbol{w}|\boldsymbol{z},\boldsymbol{\eta})P(\boldsymbol{z}|\boldsymbol{\alpha})$$

$$P(\boldsymbol{z}|\boldsymbol{\alpha}) = \int \boxed{p(\boldsymbol{z}|\boldsymbol{\theta})} p(\boldsymbol{\theta}|\boldsymbol{\alpha})d\boldsymbol{\theta}$$

$$p(\boldsymbol{z}|\boldsymbol{\theta}) = \prod_{d=1}^{D} p(\boldsymbol{z}_d|\boldsymbol{\theta}_d) = \prod_{d=1}^{D}\prod_{w=1}^{V}\theta_{d\, z_{dw}}$$

$$= \prod_{d=1}^{D}\prod_{k=1}^{K}\theta_{dk}^{\sum_w \mathbb{I}(z_{dw}=k)}$$

$$= \prod_{d=1}^{D}\prod_{k=1}^{K}\theta_{dk}^{n_{dk}}$$

数据中第d个文档生成第k个单词的次数

$$= \int \prod_{d=1}^{D}\prod_{k=1}^{K}\theta_{dk}^{n_{dk}} \prod_{d=1}^{D}\left(\frac{1}{B(\boldsymbol{\alpha})}\prod_{k=1}^{K}\theta_{dk}^{\alpha_k-1}\right)d\boldsymbol{\theta}$$

$$= \prod_{d=1}^{D}\frac{1}{B(\boldsymbol{\alpha})}\left(\int\prod_{k=1}^{K}\theta_{dk}^{n_{dk}+\alpha_k-1}\, d\boldsymbol{\theta}_d\right)$$

$$= \prod_{d=1}^{D}\frac{B(\boldsymbol{\alpha}+\boldsymbol{n}_d)}{B(\boldsymbol{\alpha})}$$

LDA模型的**参数估计**—吉布斯采样

- 基本思想是通过对隐变量θ和β积分，得到边缘概率$P(\boldsymbol{z}|\boldsymbol{w},\boldsymbol{\alpha},\boldsymbol{\eta})$

$$P(\boldsymbol{z}|\boldsymbol{w},\boldsymbol{\alpha},\boldsymbol{\eta}) = \frac{P(\boldsymbol{z},\boldsymbol{w}|\boldsymbol{\alpha},\boldsymbol{\eta})}{P(\boldsymbol{w}|\boldsymbol{\alpha},\boldsymbol{\eta})}$$

$$\propto P(\boldsymbol{w}|\boldsymbol{z},\boldsymbol{\alpha},\boldsymbol{\eta})P(\boldsymbol{z}|\boldsymbol{\alpha},\boldsymbol{\eta}) = P(\boldsymbol{w}|\boldsymbol{z},\boldsymbol{\eta})P(\boldsymbol{z}|\boldsymbol{\alpha})$$

$$P(\boldsymbol{w}|\boldsymbol{z},\boldsymbol{\eta}) = \prod_{k=1}^{K}\frac{B(\boldsymbol{\eta}+\boldsymbol{n}_k)}{B(\boldsymbol{\eta})} \qquad P(\boldsymbol{z}|\boldsymbol{\alpha}) = \prod_{d=1}^{D}\frac{B(\boldsymbol{\alpha}+\boldsymbol{n}_d)}{B(\boldsymbol{\alpha})}$$

$$\Rightarrow \quad P(\boldsymbol{z}|\boldsymbol{w},\boldsymbol{\alpha},\boldsymbol{\eta}) \propto \prod_{k=1}^{K}\frac{B(\boldsymbol{\eta}+\boldsymbol{n}_k)}{B(\boldsymbol{\eta})}\prod_{d=1}^{D}\frac{B(\boldsymbol{\alpha}+\boldsymbol{n}_d)}{B(\boldsymbol{\alpha})}$$

LDA模型的**参数估计**—吉布斯采样

即第d'文档中的第w个词

记所有文本的单词序列的第i个单词为$w_i = v'$，对应话题为$z_i = \mathrm{k}'$的条件概率为

$$P(z_i = k' | \boldsymbol{z}_{-i}, \boldsymbol{\alpha}, \boldsymbol{\eta}) \propto \frac{P(\boldsymbol{z}|\boldsymbol{w}, \boldsymbol{\alpha}, \boldsymbol{\eta})}{P(\boldsymbol{z}_{-i}|\boldsymbol{w}, \boldsymbol{\alpha}, \boldsymbol{\eta})} = \frac{\prod_{k=1}^{K} \frac{B(\boldsymbol{\eta}+\boldsymbol{n}_k)}{B(\boldsymbol{\eta})} \prod_{d=1}^{D} \frac{B(\boldsymbol{\alpha}+\boldsymbol{n}_d)}{B(\boldsymbol{\alpha})}}{\prod_{k=1}^{K} \frac{B\left(\boldsymbol{\eta}+\boldsymbol{n}_k^{(-i)}\right)}{B(\boldsymbol{\eta})} \prod_{d=1}^{D} \frac{B\left(\boldsymbol{\alpha}+\boldsymbol{n}_d^{(-i)}\right)}{B(\boldsymbol{\alpha})}}$$

$$= \frac{\prod_{k=1}^{K} B(\boldsymbol{\eta}+\boldsymbol{n}_k)}{\prod_{k=1}^{K} B\left(\boldsymbol{\eta}+\boldsymbol{n}_k^{(-i)}\right)} \cdot \frac{\prod_{d=1}^{D} B(\boldsymbol{\alpha}+\boldsymbol{n}_d)}{\prod_{d=1}^{D} B\left(\boldsymbol{\alpha}+\boldsymbol{n}_d^{(-i)}\right)}$$

$$\frac{1}{B(\boldsymbol{\alpha})} = \frac{\Gamma\left(\sum_{k=1}^{K} \alpha_k\right)}{\prod_{k=1}^{K} \Gamma(\alpha_k)}$$

$$= \frac{\dfrac{\prod_{k,v} \Gamma(\eta_v+n_{kv})}{\prod_k \Gamma\left(\sum_v \eta_v+n_{kv}\right)}}{\dfrac{\Gamma\left(\eta_{v'}+n_{k'v'}-1\right) \prod_{k\neq k', v\neq v'} \Gamma(\eta_v+n_{kv})}{\Gamma\left(-1+\sum_v \eta_v+n_{k'v}\right) \prod_{k\neq k'} \Gamma\left(\sum_v \eta_v+n_{kv}\right)}} \cdot \frac{\dfrac{\prod_{d,k} \Gamma(\alpha_k+n_{dk})}{\prod_d \Gamma\left(\sum_k \alpha_k+n_{dk}\right)}}{\dfrac{\Gamma\left(\alpha_{k'}+n_{d'k'}-1\right) \prod_{d=d', k\neq k'} \Gamma(\alpha_k+n_{dk})}{\Gamma\left(-1+\sum_k \alpha_k+n_{d'k}\right) \prod_{d'\neq d} \Gamma\left(\sum_k \alpha_k+n_{dk}\right)}}$$

$$= \frac{\dfrac{\Gamma\left(\eta_{v'}+n_{k'v'}\right)}{\Gamma\left(\sum_v \eta_v+n_{k'v}\right)}}{\dfrac{\Gamma\left(\eta_{v'}+n_{k'v'}-1\right)}{\Gamma\left(-1+\sum_v \eta_v+n_{k'v}\right)}} \cdot \frac{\dfrac{\Gamma\left(\alpha_{k'}+n_{d'k'}\right)}{\Gamma\left(\sum_k \alpha_k+n_{d'k}\right)}}{\dfrac{\Gamma\left(\alpha_{k'}+n_{d'k'}-1\right)}{\Gamma\left(-1+\sum_k \alpha_k+n_{d'k}\right)}} = \frac{\eta_{v'} + n_{k'v'}}{\sum_v \eta_v + n_{k'v}} \cdot \frac{\alpha_{k'} + n_{d'k'}}{\sum_k \alpha_k + n_{d'k}}$$

LDA模型的**参数估计**—吉布斯采样

- 估计参数$\boldsymbol{\theta}_d$

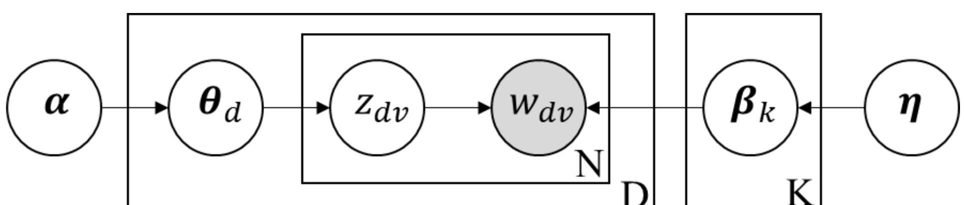
$$p(\boldsymbol{\theta}_d|\boldsymbol{z}_d,\boldsymbol{\alpha}) \propto p(\boldsymbol{\theta}_d|\boldsymbol{\alpha})P(\boldsymbol{z}_d|\boldsymbol{\theta}_d) \propto \prod_{k=1}^{K} \theta_{dk}^{\alpha_k-1+n_{dk}}$$

$$p(\boldsymbol{\theta}_d|\boldsymbol{z}_d,\boldsymbol{\alpha}) = Dir(\boldsymbol{\theta}_d|\boldsymbol{n}_d+\boldsymbol{\alpha}) \Rightarrow \theta_{dk} = \frac{n_{dk}+\alpha_k}{\sum_k(n_{dk}+\alpha_k)}$$

- 估计参数$\boldsymbol{\beta}_k$

$$p(\boldsymbol{\beta}_k|\boldsymbol{w},\boldsymbol{\eta}) \propto P(\boldsymbol{w}|\boldsymbol{\beta}_k)p(\boldsymbol{\beta}_k|\boldsymbol{\eta}) \propto \prod_{v=1}^{V} \beta_{kv}^{n_{kv}+\eta_v-1}$$

$$p(\boldsymbol{\beta}_k|\boldsymbol{w},\boldsymbol{\eta}) = Dir(\boldsymbol{\beta}_k|\boldsymbol{n}_k+\boldsymbol{\eta}) \Rightarrow \beta_{kv} = \frac{n_{kv}+\eta_v}{\sum_v(n_{kv}+\eta_v)}$$



LDA模型的参数估计

- 给定训练数据$\boldsymbol{w}=\{\boldsymbol{w}_1,\boldsymbol{w}_2,\ldots,\boldsymbol{w}_D\}$, 参数通过极大似然法估计，寻找$\boldsymbol{\alpha}$和$\boldsymbol{\eta}$以最大化对数似然

$$LL(\boldsymbol{\alpha},\boldsymbol{\eta})=\sum_{d=1}^{D}\ln p(\boldsymbol{w}_d|\boldsymbol{\alpha},\boldsymbol{\eta})$$

- 能否进行模型推断后用EM算法呢?

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

- 求解算法
  - 可以通过收缩的吉布斯采样求解：通过使用随机化方法完成近似
  - 可以通过变分法求解：使用确定性近似完成推断

变分推断—EM算法回顾

$$\mathcal{L}(q, \mathbf{\Theta}) = \sum_{\mathbf{z}} q(\mathbf{z}) \ln \frac{p(\mathbf{x}, \mathbf{z} | \mathbf{\Theta})}{q(\mathbf{z})}$$

$$\max_{\mathbf{\Theta}} \ln p(\mathbf{x}|\mathbf{\Theta}) = \max_{\mathbf{\Theta}} \max_{q} \mathcal{L}(q, \mathbf{\Theta})$$

E

基于$\mathbf{\Theta}^t$推断隐变量$\mathbf{z}$的分布$p(\mathbf{z} \mid \mathbf{x}, \boldsymbol{\Theta}^t)$，并计算对数似然$LL(\mathbf{\Theta}|\mathbf{x}, \mathbf{z})$关于$\mathbf{z}$的期望;

$$\mathcal{L}(p(\mathbf{z}|\mathbf{x}, \mathbf{\Theta}), \mathbf{\Theta}) = \mathbb{E}_{\mathbf{z} \sim p(\mathbf{z} \mid \mathbf{x}, \boldsymbol{\Theta}^t)} LL(\mathbf{\Theta}|\mathbf{x}, \mathbf{z}) \quad = Q(\mathbf{\Theta}|\mathbf{\Theta}^t)$$

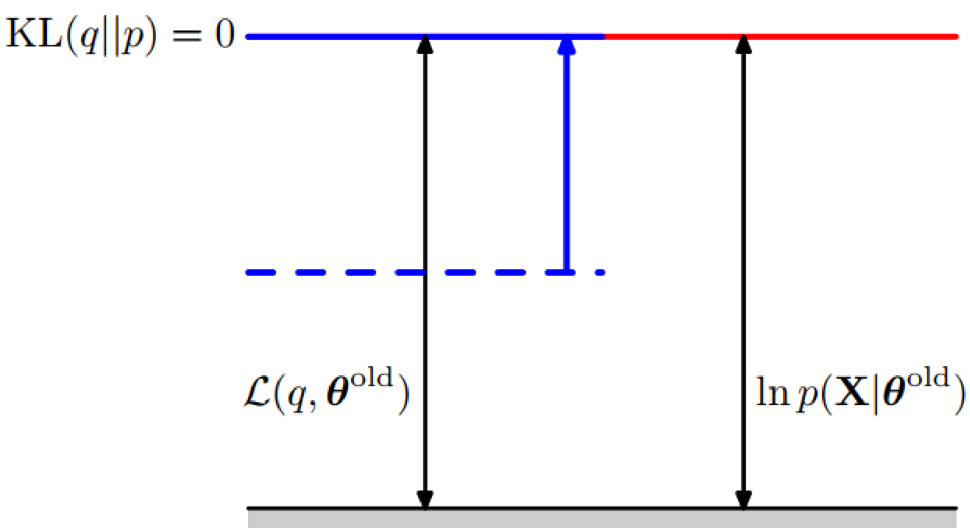
M

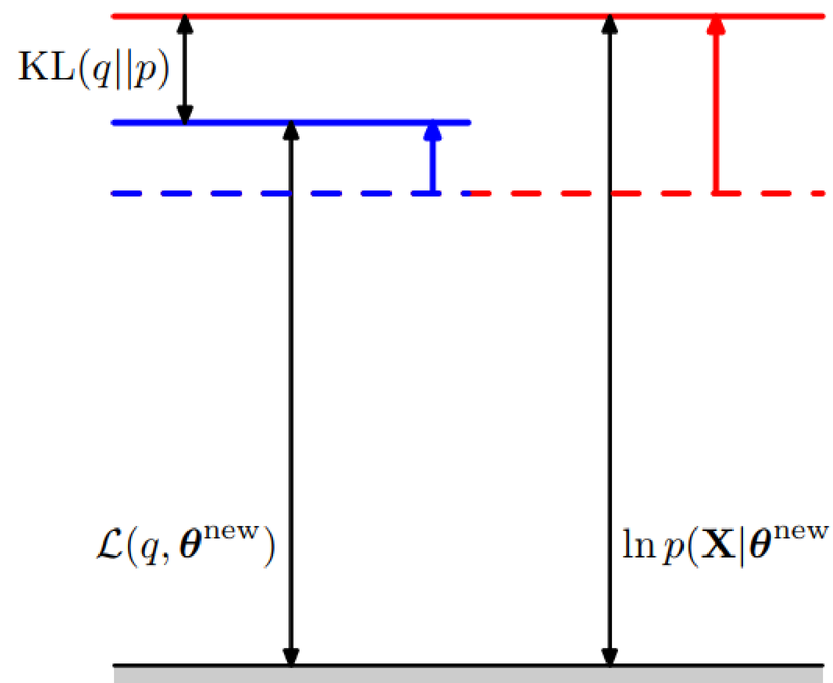
寻找参数最大化期望似然;

$$\mathbf{\Theta}^{t+1} = \arg\max_{\mathbf{\Theta}} Q(\mathbf{\Theta}|\mathbf{\Theta}^t)$$

变分推断—EM算法回顾

$$\max_{\boldsymbol{\Theta}} \ln p(\boldsymbol{x}|\boldsymbol{\Theta}) = \max_{\boldsymbol{\Theta}} \max_{q} \mathcal{L}(q, \boldsymbol{\Theta})$$





EM算法的E步要求能求出最优的后验分布，若无法求解，如何是好?

答：近似后验分布

变分推断

- 平均场近似假设： $q(\boldsymbol{z}) = \prod_{i=1}^{M} q_i(\boldsymbol{z}_i)$

复杂的多变量$\boldsymbol{z}$可以拆解为一系列相互独立的多变量$\boldsymbol{z}_i$

- 此时最优解第i组随机变量的概率分布最优值为

$$q_i^*(\boldsymbol{z}_i) \propto \exp \mathbb{E}_{\boldsymbol{z}_{-i}}[\ln p(\boldsymbol{x}, \boldsymbol{z})]$$

证明

$$\mathcal{L}(q, \boldsymbol{\Theta}) = \sum_{\boldsymbol{z}} q(\boldsymbol{z}) \ln p(\boldsymbol{x}, \boldsymbol{z}|\boldsymbol{\Theta}) - H(q)$$

$$= \sum_{\boldsymbol{z}_i} q_i(\boldsymbol{z}_i) \sum_{\boldsymbol{z}_{-i}} q_{-i}(\boldsymbol{z}_{-i}) \ln p(\boldsymbol{x}, \boldsymbol{z}|\boldsymbol{\Theta}) - \sum_{\boldsymbol{z}_i} q_i(\boldsymbol{z}_i) \ln q_i(\boldsymbol{z}_i) + \text{const}$$

$$= \sum_{\boldsymbol{z}_i} q_i(\boldsymbol{z}_i) \mathbb{E}_{\boldsymbol{z}_{-i}}[\ln p(\boldsymbol{x}, \boldsymbol{z})] - \sum_{\boldsymbol{z}_i} q_i(\boldsymbol{z}_i) \ln q_i(\boldsymbol{z}_i) + \text{const}$$

对$q_i(\boldsymbol{z}_i)$ 求导数，并令其等于0，便可得到上述最优解

变分推断

- 平均场近似假设： $q(\boldsymbol{z})=\prod_{i=1}^{M} q_i(\boldsymbol{z}_i)$

复杂的多变量$\boldsymbol{z}$可以拆解为一系列相互独立的多变量$\boldsymbol{z}_i$

- 此时最优解第i组随机变量的概率分布最优值为

$$q_i^*(\boldsymbol{z}_i) \propto \exp \mathbb{E}_{\boldsymbol{z}_{-i}}[\ln p(\boldsymbol{x},\boldsymbol{z})]$$

由于在对$\boldsymbol{z}_i$所服从的分布q_i^*估计时融合了$\boldsymbol{z}_i$之外的其它$\boldsymbol{z}_{-i}$的信息，即通过联合似然函数$\ln p(\mathrm{x},\mathrm{z})$在$z_j$之外的隐变量分布上求期望得到的，因此亦称为“平均场”（mean field）方法

LDA模型的**参数估计**—变分推断

- 为简单起见，只考虑一个文档，记为$\boldsymbol{w} = (w_1, \dots, w_N)$；对应话题为$\boldsymbol{z} = (z_1, \dots, z_N)$

- 联合概率分布 $p(\boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{w} | \boldsymbol{\alpha}, \boldsymbol{\beta}) = p(\boldsymbol{\theta} | \boldsymbol{\alpha}) \prod_{n=1}^{N} p(z_n | \boldsymbol{\theta}) p(w_n | z_n, \boldsymbol{\beta})$

- 平均场假设 $q(\boldsymbol{\theta}, \boldsymbol{z} | \boldsymbol{\gamma}, \boldsymbol{\phi}) = q(\boldsymbol{\theta} | \boldsymbol{\gamma}) \prod_{n=1}^{N} q(z_n | \boldsymbol{\phi}_n)$

- 证据下界

$$\mathcal{L}(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\phi}) = \mathbb{E}_q[\log p(\boldsymbol{\theta}|\boldsymbol{\alpha})] + \sum_{n=1}^{N} \left(\mathbb{E}_q[p(z_n|\boldsymbol{\theta})] + \mathbb{E}_q[\log p(w_n|z_n, \boldsymbol{\beta})] \right) - \mathbb{E}_q[q(\boldsymbol{\theta}|\boldsymbol{\gamma})] - \sum_{n=1}^{N} \mathbb{E}_q[\log q(z_n|\boldsymbol{\phi}_n)]$$

LDA模型的**参数估计**—变分推断

$$\mathcal{L}(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\phi}) = \mathbb{E}_q[\log p(\boldsymbol{\theta}|\boldsymbol{\alpha})] + \sum_{n=1}^{N}\left(\mathbb{E}_q[p(z_n|\boldsymbol{\theta})] + \mathbb{E}_q[\log p(w_n|z_n,\boldsymbol{\eta})]\right)$$

$$-\mathbb{E}_q[q(\boldsymbol{\theta}|\boldsymbol{\gamma})] - \sum_{n=1}^{N}\mathbb{E}_q[\log q(z_n|\boldsymbol{\phi}_n)]$$

$$\mathbb{E}_q[\log p(\boldsymbol{\theta}|\boldsymbol{\alpha})] = \mathbb{E}_{\boldsymbol{q}}\left[\log\frac{1}{B(\boldsymbol{\alpha})}\prod_{k=1}^{K}\theta_k^{\alpha_k-1}\right] = \mathbb{E}_{\boldsymbol{q}}\left[\log\frac{1}{B(\boldsymbol{\alpha})}\prod_{k=1}^{K}\theta_k^{\alpha_k-1}\right]$$

$$= \sum_{k}^{K}(\alpha_k - 1)\boxed{\mathbb{E}_{q(\boldsymbol{\theta}|\boldsymbol{\gamma})}[\log\theta_k]} + \log\Gamma\left(\sum_{k=1}^{K}\alpha_k\right) - \sum_{k=1}^{K}\log\Gamma(\alpha_k)$$

$$= \sum_{k}^{K}(\alpha_k - 1)\boxed{\left(\Psi(\gamma_k) - \Psi\left(\sum_{k}^{K}\gamma_k\right)\right)} + \log\Gamma\left(\sum_{k=1}^{K}\alpha_k\right) - \sum_{k=1}^{K}\log\Gamma(\alpha_k)$$

$$\Psi(\gamma_k) = \frac{d}{d\gamma_k}\Gamma(\gamma_k)$$

LDA模型的**参数估计**—变分推断

$$\mathcal{L}(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\phi}) = \mathbb{E}_q[\log p(\boldsymbol{\theta}|\boldsymbol{\alpha})] + \sum_{n=1}^{N}\big(\mathbb{E}_q[p(z_n|\boldsymbol{\theta})] + \mathbb{E}_q[\log p(w_n|z_n,\boldsymbol{\beta})]\big)$$

$$-\mathbb{E}_q[q(\boldsymbol{\theta}|\boldsymbol{\gamma})] - \sum_{n=1}^{N}\mathbb{E}_q[\log q(z_n|\boldsymbol{\phi}_n)]$$

$$\mathbb{E}_q[p(z_n|\boldsymbol{\theta})] = \sum_{k=1}^{K}\phi_{nk}\left(\Psi(\gamma_k) - \Psi\left(\sum_{k=1}^{K}\gamma_k\right)\right)$$

$$\mathbb{E}_q[\log p(w_n|z_n,\boldsymbol{\beta})] = \sum_{k=1}^{K}\sum_{v=1}^{V}\phi_{nk}\mathbb{I}(w_n = v)\log\beta_{kv}$$

$$\mathbb{E}_q[q(\boldsymbol{\theta}|\boldsymbol{\gamma})] = \sum_{k=1}^{K}(\gamma_k - 1)\left(\Psi(\gamma_k) - \Psi\left(\sum_{k}^{K}\gamma_k\right)\right) + \log\Gamma\left(\sum_{k=1}^{K}\gamma_k\right) - \sum_{k=1}^{K}\log\Gamma(\gamma_k)$$

$$\mathbb{E}_q[\log q(z_n|\boldsymbol{\phi})] = \sum_{k=1}^{K}\phi_{nk}\log\phi_{nk}$$

LDA模型的**参数估计**—变分推断

$$\mathcal{L}(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\phi}) = \mathbb{E}_q[\log p(\boldsymbol{\theta}|\boldsymbol{\alpha})] + \sum_{n=1}^{N}\left(\mathbb{E}_q[p(z_n|\boldsymbol{\theta})] + \mathbb{E}_q[\log p(w_n|z_n,\boldsymbol{\beta})]\right)$$

$$-\mathbb{E}_q[q(\boldsymbol{\theta}|\boldsymbol{\gamma})] - \sum_{n=1}^{N}\mathbb{E}_q[\log q(z_n|\boldsymbol{\phi}_n)]$$

对ϕ_{nk}求偏导数，令其等于0，

$$\Psi(\gamma_k) - \Psi\left(\sum_{k=1}^{K}\gamma_k\right) + \sum_{v=1}^{V}\mathbb{I}(w_n = v)\log\beta_{kv} - 1 - \log\phi_{nk} = 0$$

$$\Rightarrow \quad \phi_{nk} \propto \beta_{k\,w_n}\exp\left(\Psi(\gamma_k) - \Psi\left(\sum_{k=1}^{K}\gamma_k\right)\right)$$

LDA模型的**参数估计**—变分推断

$$\mathcal{L}(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\phi}) = \mathbb{E}_q[\log p(\boldsymbol{\theta}|\boldsymbol{\alpha})] + \sum_{n=1}^{N}\left(\mathbb{E}_q[p(z_n|\boldsymbol{\theta})] + \mathbb{E}_q[\log p(w_n|z_n,\boldsymbol{\beta})]\right)$$
$$-\mathbb{E}_q[q(\boldsymbol{\theta}|\boldsymbol{\gamma})] - \sum_{n=1}^{N}\mathbb{E}_q[\log q(z_n|\boldsymbol{\phi}_n)]$$

$$L[\gamma_k] = \textstyle\sum_{k=1}^{K}(\alpha_k - \gamma_k + \sum_n \phi_{nk})\left(\Psi(\gamma_k) - \Psi\left(\sum_{k=1}^{K}\gamma_k\right)\right) - \log\Gamma\left(\sum_{k=1}^{K}\gamma_k\right) + \log\Gamma(\gamma_k)$$

对γ_k求偏导数，令其等于0，

$$\frac{\partial L}{\partial \gamma_k} = (\alpha_k - \gamma_k + \textstyle\sum_n \phi_{nk})\left(\Psi'(\gamma_k) - \Psi'\left(\sum_{l=1}^{K}\gamma_l\right)\right) = 0$$

$$\Rightarrow \quad \gamma_k = \alpha_k + \sum_n \phi_{nk}$$

LDA模型的**参数估计**—变分推断

- 估计参数β
  - 写出全数据集上的对数似然，在等式约束下最大化似然

$$L[\boldsymbol{\beta}] = \sum_{d=1}^{D}\sum_{n=1}^{N_d}\sum_{k=1}^{K}\sum_{v=1}^{V}\phi_{dnk}\mathbb{I}(w_{dn}=v)\log\beta_{kv} \qquad \sum_{v}\beta_{kv}=1$$

$$\Rightarrow \qquad \beta_{kv} = \sum_{d=1}^{D}\sum_{n=1}^{N_d}\eta_{dnk}\mathbb{I}(w_{dn}=v)$$

- 估计参数$\boldsymbol{\alpha}$

$$L(\boldsymbol{\alpha}) = \sum_{d=1}^{D}\left(\sum_{k=1}^{K}(\alpha_k-1)\left(\Psi(\gamma_{dk})-\Psi\left(\sum_{k}^{K}\gamma_{dk}\right)\right)+\log\Gamma\left(\sum_{k=1}^{K}\alpha_k\right)-\sum_{k=1}^{K}\log\Gamma(\alpha_k)\right)$$

用梯度上升法进行优化求解

作业

- PPT 36页 详述除第一项外其余四项的化简过程

假设数据集$D=\{x_1,x_2,\ldots,x_m\}$，任意x_i是从均值为μ、方差为λ^{-1}的正态分布$\mathcal{N}(\mu,\lambda^{-1})$中独立采样而得到。假设$\mu$和$\lambda$的先验分布为$p(\mu,\lambda)=\mathcal{N}(\mu|\mu_0,(\kappa_0\lambda)^{-1})\mathrm{Gam}(\lambda|a_0,b_0)$ 其中 $\mathrm{Gam}(\lambda|a_0,b_0)=\frac{1}{\Gamma(a_0)}b_0^{a_0}\lambda^{a_0-1}\exp(-b_0\lambda)$

(1) 请写出联合概率分布$p(D,\mu,\lambda)$

(2) 请写出证据下界（即变分推断的优化目标），并证明其为观测数据边际似然$\sum_{i=1}^{m}\log p(x_i)$的下界

(3) (3) 请用变分推断法近似推断后验概率$p(\mu,\lambda|D)$