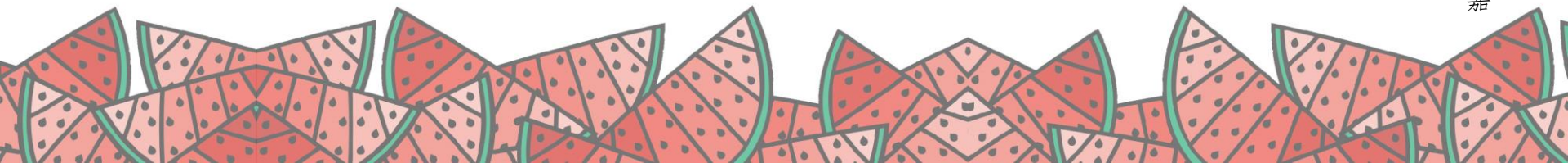
周志华 著

MACHINE LEARNING

机器学习

清华大学出版社

本章课件致谢：
叶翰嘉



第十四章：概率图模型

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

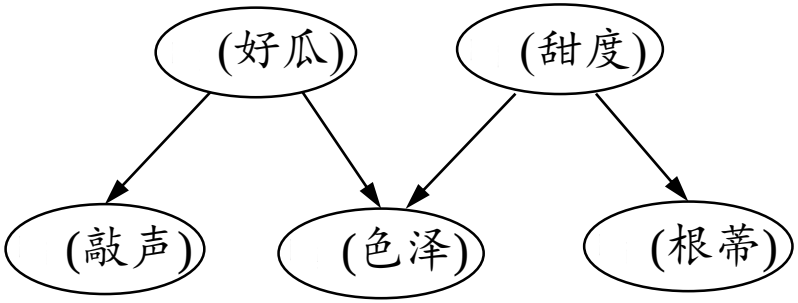
概率图模型

- 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

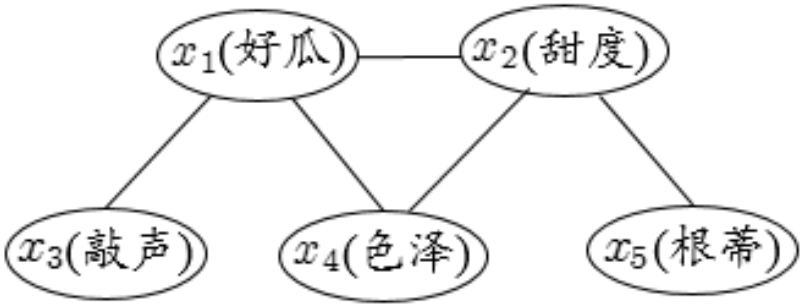
- 分类：
  - 有向图：贝叶斯网
    - 使用有向无环图表示变量之间的依赖关系
  - 无向图：马尔可夫网
    - 使用无向图表示变量间的相关关系

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



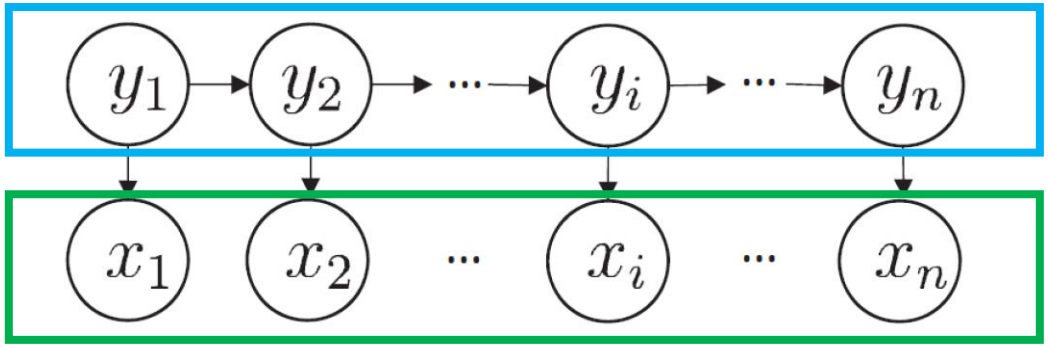
无向图

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
- 学习与推断
  - 精确推断
  - 近似推断
    - 随机化方法
    - 确定性方法
- 实例：话题模型

隐马尔可夫模型

- 隐马尔可夫模型（Hidden Markov Model，HMM）
- 组成
  - 状态变量：$\{y_1, y_2, \ldots, y_n\}$　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：$\{x_1, x_2, \ldots, x_n\}$　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 $\{o_1, o_2, \ldots, o_M\}$



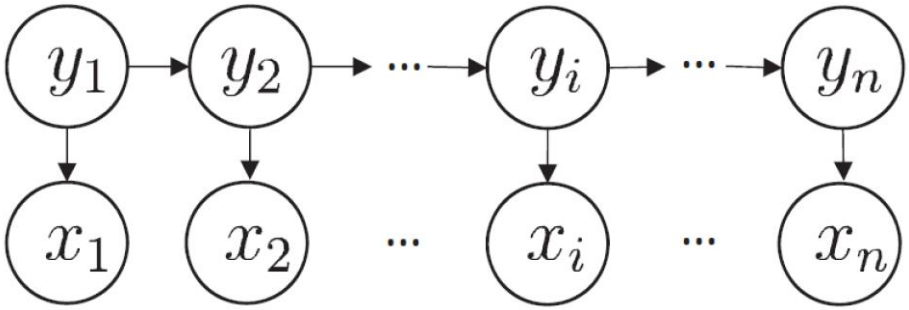
隐马尔可夫模型

- 隐马尔可夫模型（Hidden Markov Model）
- t时刻的状态x_t仅依赖于$t-1$时刻状态x_{t-1}，与其余$\mathrm{n}-2$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

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



联合概率

HMM的基本组成

- 确定一个HMM需要三组参数 $\lambda = [A, B, \pi]$
  - 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

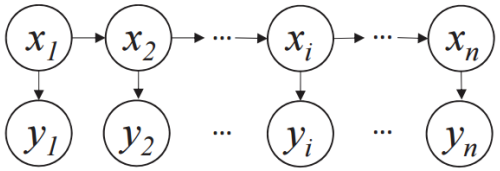
$$A = [a_{ij}]_{N\times N} \qquad a_{ij} = p(y_{t+1} = s_j \mid y_t = s_i)\ , \qquad 1 \le i, j \le N$$

  - 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，则在下一时刻状态为s_j的概率

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

  - **初始**状态概率：模型在初始时刻各个 状态出现的概率

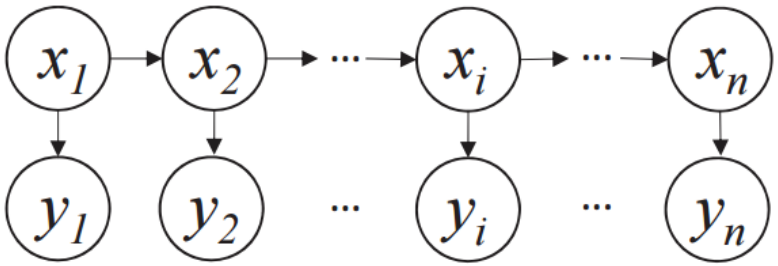
$$\pi = [\pi_1, \ldots, \pi_n] \qquad \pi_i = P(y_1 = s_i), \qquad 1 \le i \le N$$



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

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



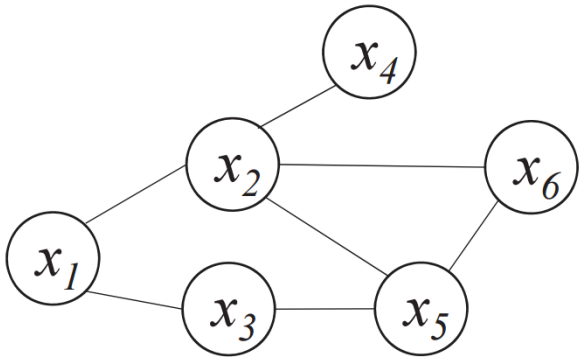
HMM的基本问题

- 对于模型 $\lambda = [A, B, \pi]$，给定观测序列 $\mathrm{x} = \{x_1, x_2, \ldots x_n\}$
  - 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 $P(\mathbf{x} \mid \lambda)$
  - 根据观测序列“推测”隐藏的模型状态 $y = \{y_1, y_2, \ldots y_n\}$
  - 参数学习：如何调整模型参数 $\lambda = [A, B, \pi]$以使得该序列出现的概率 $P(\mathbf{x} \mid \lambda)$ 最大

- 具体应用
  - 根据以往的观测序列$\mathrm{x} = \{x_1, x_2, \ldots x_n\}$预测当前时刻最有可能的观测值$x_n$
  - 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）
  - 通过数据学习参数（模型训练）

马尔可夫随机场

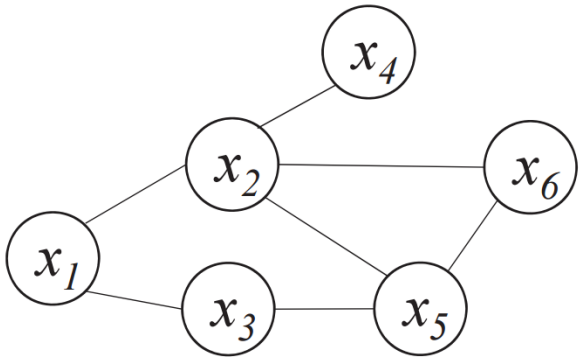
- 马尔可夫随机场（Markov Random Field，MRF）
  - 是典型的马尔可夫网
  - 著名的**无向图**模型



- 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

- 马尔可夫随机场（Markov Random Field，MRF）
- 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



  - 多个变量之间的连续分布可基于团分解为多个因子的**乘积**

马尔可夫随机场

- 马尔可夫随机场（Markov Random Field，MRF）
- 基于**极大团**的势函数（因子）：
  - 多个变量之间的连续分布可基于团分解为多个因子的**乘积**，每个因子只与一个团相关
  - 对于n个变量$\mathbf{x} = \{x_1, x_2, \ldots x_n\}$，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$\mathbf{x}_Q$，则联合概率定义为：

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

  - 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

- 基于**极大团**的势函数：
  - 通过极大团构造势函数。若团Q不是一个极大团，则必然被一个极大团Q^*包含，这意味着变量$\mathbf{x}_Q$的关系不仅体现在势函数ψ_Q中，还体现在ψ_{Q^*}中
  - 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C^*

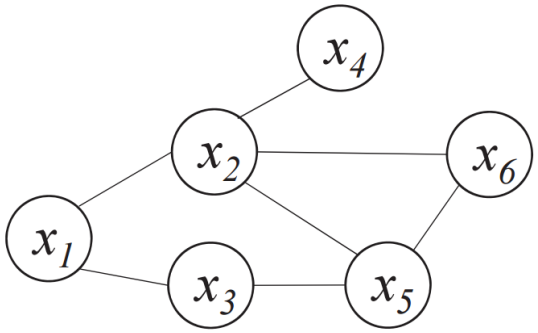
$$P(\mathbf{x}) = \frac{1}{Z^*} \prod_{Q \in C^*} \psi_Q(\mathbf{x}_Q)$$

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

联合概率分布定义举例

- 基于**极大团**的势函数：
  - 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C^*

- 图模型



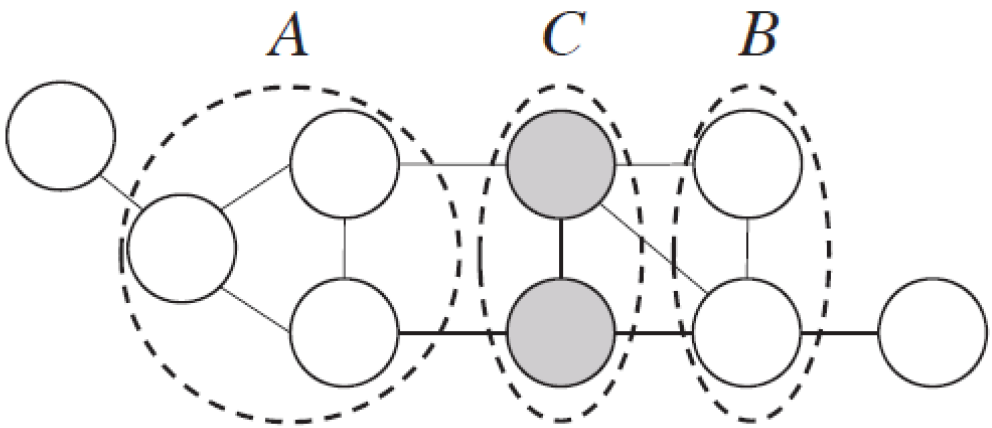
$$P(\mathbf{x}) = \frac{1}{Z^*} \prod_{Q \in C^*} \psi_Q(\mathbf{x}_Q)$$

- 联合概率分布

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

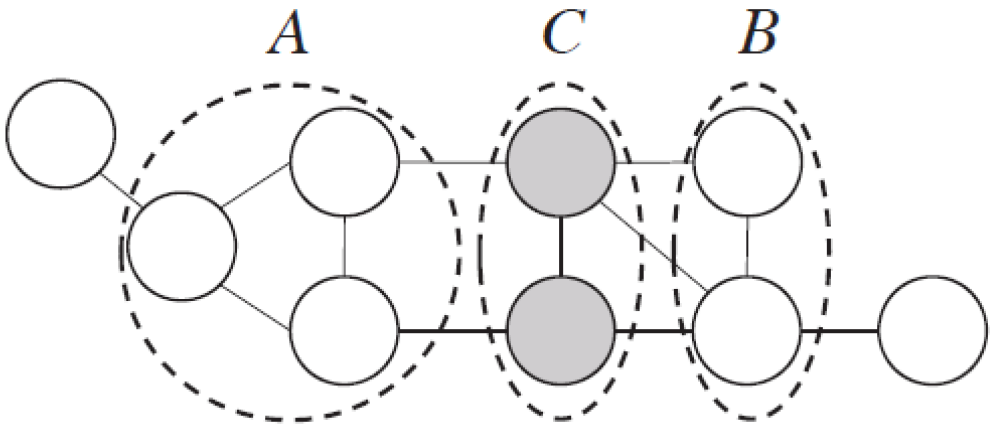
马尔可夫随机场中的分离集

- 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）



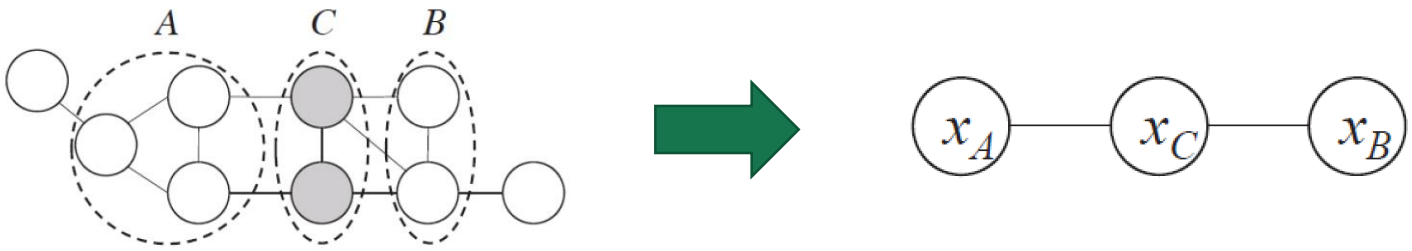
全局马尔可夫性

- 马尔可夫随机场（Markov Random Field，MRF）
- 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$\mathbf{x}_A, \mathbf{x}_B, \mathbf{x}_C$，则$\mathbf{x}_A$和$\mathbf{x}_B$在$\mathbf{x}_C$给定的条件下独立，记为$\mathbf{x}_A \perp \mathbf{x}_B \mid \mathbf{x}_C$



全局马尔可夫性的验证

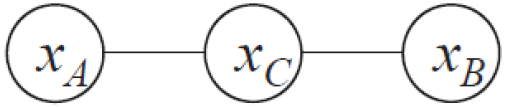
- **全局**马尔可夫性（global Markov property）：在给定**分离集**的条件下， 两个变量子集条件独立
  - 若令A,B,C对应的变量集分别为$\mathrm{x_A}, \mathrm{x_B}, \mathrm{x}_C$，则$\mathrm{x_A}$和$\mathrm{x_B}$在$\mathrm{x}_C$给定的条件下独立，记为$\mathrm{x_A} \perp \mathrm{x_B} \mid \mathrm{x}_C$
- 图模型简化：



- 得到图模型的联合概率为：

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

全局马尔可夫性的验证



- 联合概率：

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

- 条件概率

$$\begin{aligned} P(x_A, x_B \mid x_C) &= \frac{P(x_A, x_B, x_C)}{P(x_C)} = \frac{P(x_A, x_B, x_C)}{\sum_{x'_A}\sum_{x'_B} p(x'_A, x'_B, x_C)} \\ &= \frac{\frac{1}{Z}\psi_{AC}(x_A, x_C)\psi_{BC}(x_B, x_C)}{\sum_{x'_A}\sum_{x'_B}\frac{1}{Z}\psi_{AC}(x'_A, x_C)\psi_{BC}(x'_B, x_C)} \\ &= \frac{\psi_{AC}(x_A, x_C)}{\sum_{x'_A}\psi_{AC}(x'_A, x_C)} \cdot \frac{\psi_{BC}(x_B, x_C)}{\sum_{x'_B}\psi_{BC}(x'_B, x_C)} . \end{aligned}$$

$$\begin{aligned} P(x_A \mid x_C) &= \frac{P(x_A, x_C)}{P(x_C)} = \frac{\sum_{x'_B} P(x_A, x'_B, x_C)}{\sum_{x'_A}\sum_{x'_B} P(x'_A, x'_B, x_C)} \\ &= \frac{\sum_{x'_B}\frac{1}{Z}\psi_{AC}(x_A, x_C)\psi_{BC}(x'_B, x_C)}{\sum_{x'_A}\sum_{x'_B}\frac{1}{Z}\psi_{AC}(x'_A, x_C)\psi_{BC}(x'_B, x_C)} \\ &= \frac{\psi_{AC}(x_A, x_C)}{\sum_{x'_A}\psi_{AC}(x'_A, x_C)}. \end{aligned}$$

- 验证：

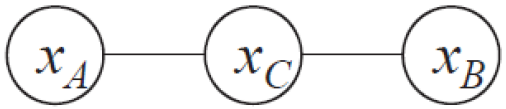
$$P(x_A, x_B \mid x_C) = P(x_A \mid x_C)P(x_B \mid x_C)$$

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
  - 若令A,B,C对应的变量集分别为$\mathrm{x}_A, \mathrm{x}_B, \mathrm{x}_C$，则$\mathrm{x}_A$和$\mathrm{x}_B$在$\mathrm{x}_C$给定的条件下独立，记为$\mathrm{x}_A \perp \mathrm{x}_B \mid \mathrm{x}_C$
- 由全局马尔可夫性可以导出：
  - **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$\psi_Q(\mathbf{x}_Q)$的作用是定量刻画变量集$\mathbf{x}_Q$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：

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

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

- 说明模型偏好$\mathbf{x}_A$与$\mathbf{x}_C$有相同的取值，$\mathbf{x}_B$与$\mathbf{x}_C$有不同的取值，换言之，$\mathbf{x}_A$与$\mathbf{x}_C$正相关，$\mathbf{x}_B$与$\mathbf{x}_C$负相关。所以令$\mathbf{x}_A$与$\mathbf{x}_C$相同且$\mathbf{x}_B$与$\mathbf{x}_C$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$\psi_Q(\mathbf{x}_Q)$的作用是定量刻画变量集$\mathbf{x}_Q$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：

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

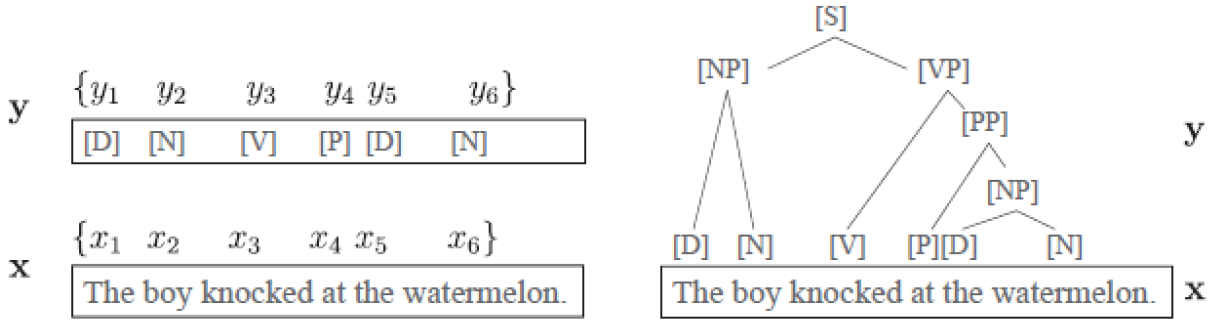
- 其中，$H_Q(\mathbf{x}_Q)$是一个定义在变量$\mathbf{x}_Q$上的实值函数，常见形式为：

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

- 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$\mathrm{x} = \{\mathrm{x}_1, \mathrm{x}_2, \ldots, \mathrm{x}_n\}$为观测序列，$\mathrm{y} = \{y_1, y_2, \ldots, y_n\}$为对应的标记序列， CRF的目标是构建条件概率模型$P(\boldsymbol{y}|\mathrm{x})$
- 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

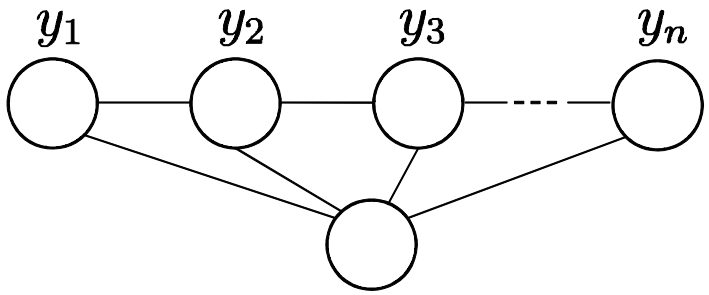
条件随机场

- 令$\mathrm{G} = \langle V, E \rangle$表示结点与标记变量y中元素一一对应的无向图。无向图中，$\boldsymbol{y}_v$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，

$$P(y_v \mid \mathbf{x}, \mathbf{y}_{V\setminus\{v\}}) = P(y_v \mid \mathbf{x}, \mathbf{y}_{n(v)})$$

则$(\boldsymbol{y}, \boldsymbol{x})$构成条件随机场。

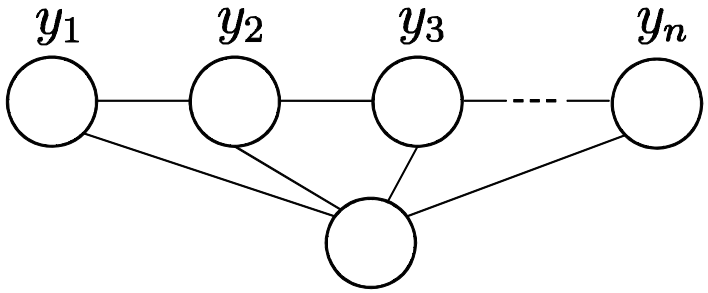
- CRF使用势函数和图结构上的团来定义$P(\boldsymbol{y}|\mathrm{x})$
- 本章仅考虑**链式**条件随机场（chain-structured CRF），如下所示：



$$\boldsymbol{x} = \{x_1, x_2, \ldots, x_n\}$$

条件随机场

- **链式**条件随机场（chain-structured CRF）
- 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：

$$P(\mathbf{y} \mid \mathbf{x}) = \frac{1}{Z} \exp\left(\sum_{j} \sum_{i=1}^{n-1} \lambda_j t_j(y_{i+1}, y_i, \mathbf{x}, i) + \sum_{k} \sum_{i=1}^{n} \mu_k s_k(y_i, \mathbf{x}, i) \right)$$

  - $t_j(y_{i+1}, y_i, \mathrm{x}, i)$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
  - $s_k(y_i, \mathrm{x}, i)$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：

y	$\{y_1$	y_2	y_3	y_4	y_5	$y_6\}$
	[D]	[N]	[V]	[P]	[D]	[N]

x	$\{x_1$	x_2	x_3	x_4	x_5	$x_6\}$
	The	boy	knocked	at	the	watermelon.

- 采用特征函数：

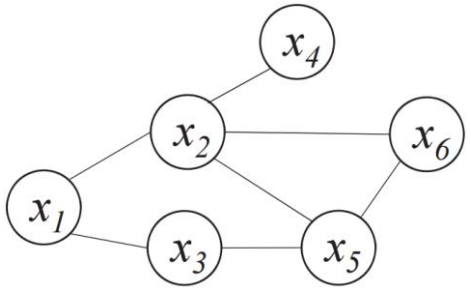
$$t_j(y_{i+1}, y_i, \mathbf{x}, i) = \begin{cases} 1, & \text{if } y_{i+1} = [P]\ y_i = [V], \text{ and } x_i = \text{“knock”}; \\ 0, & \text{otherwise,} \end{cases}$$

  - 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$[V], [P]$.

MRF与CRF的对比

MRF

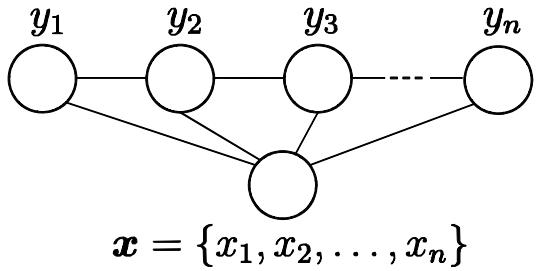
- 使用团上的势函数定义概率
- 对**联合概率**建模



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

CRF

- 使用团上的势函数定义概率
- 有观测变量，对**条件概率**建模



$$P(\mathrm{y} \mid \mathrm{x}) = \frac{1}{Z}\exp\left(\sum_j \sum_{i=1}^{n-1} \lambda_j t_j(y_{i+1}, y_i, \mathrm{x}, i) + \sum_k \sum_{i=1}^{n} \mu_k s_k(y_i, \mathrm{x}, i)\right)$$

模型推断

- 基于概率图模型定义的分布，能对目标变量的边际分布（marginal distribution）或某些可观测变量为条件的条件分布进行推断
- 对概率图模型，还需确定具体分布的参数，称为**参数估计或学习问题**，通常使用极大似然估计或后验概率估计求解。单若将参数视为待推测的变量，则参数估计过程和推断十分相似，可以“吸收”到推断问题中

参数估计（学习）

参数视为待推测的变量

图模型推断

模型推断

- 假设图模型所对应的变量集 $\mathbf{x}=\{x_1,x_2,\cdots,x_n\}$ 能分为 $\mathbf{x}_E$ 和 $\mathbf{x}_F$两个**不相交**的变量集，推断问题的目标就是计算**边际概率** $p(\mathbf{x}_F)$ 或者**条件概率** $p(\mathbf{x}_F|\mathbf{x}_E)$。同时，由条件概率定义容易有

$$p(\mathbf{x}_F|\mathbf{x}_E)=\frac{p(\mathbf{x}_F,\mathbf{x}_E)}{p(\mathbf{x}_E)}=\frac{p(\mathbf{x}_F,\mathbf{x}_E)}{\sum_F p(\mathbf{x}_F,\mathbf{x}_E)}$$

联合分布

边际分布

- 其中，联合概率 $p(\mathbf{x}_F,\mathbf{x}_E)$ 可基于图模型获得，所以推断问题的关键就在于如何高效计算边际分布，即

$$p(\mathbf{x}_E)=\sum_F p(\mathbf{x}_F,\mathbf{x}_E)$$

模型推断分类

- 联合概率 $p(\mathbf{x}_F, \mathbf{x}_E)$可基于图模型获得，所以推断问题的关键就在于如何高效计算边际分布：

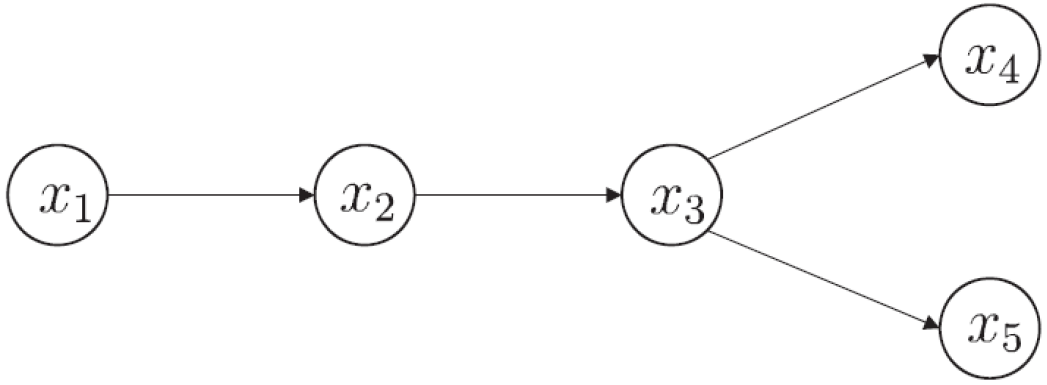
$$p(\mathbf{x}_E) = \sum_F p(\mathbf{x}_F, \mathbf{x}_E)$$

- 概率图模型的推断方法可以分两类
  - **精确**推断方法
    - 计算出目标变量的边际分布或条件分布的精确值
    - 一般情况下，该类方法的计算复杂度随极大团规模增长呈指数增长，适用范围有限
  - **近似**推断方法
    - 在较低的时间复杂度下获得原问题的近似解
    - 在实际问题中更常用

精确推断：变量消去

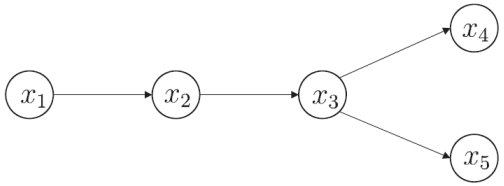
- 精确推断实质是一种**动态规划**算法，利用图模型所描述的条件独立性来削减计算目标概率值所需的计算量
- 变量消去是最直观的精确推断方法，也是构建其它精确推断算法的基础

- 例：计算边缘概率$p(x_5)$



精确推断：变量消去

- 计算边缘概率 $p(x_5)$



$$p(x_5) = \sum_{x_4}\sum_{x_3}\sum_{x_2}\sum_{x_1} p(x_1)p(x_2 \mid x_1)p(x_3 \mid x_2)p(x_4 \mid x_3)p(x_5 \mid x_3)$$

$$= \sum_{x_3} p(x_5 \mid x_3) \sum_{x_4} p(x_4 \mid x_3) \sum_{x_2} p(x_3 \mid x_2) \sum_{x_1} p(x_1)p(x_2 \mid x_1)$$

$$p(x_5) = \sum_{x_3} p(x_5 \mid x_3) m_{23}(x_3) \sum_{x_4} p(x_4 \mid x_3)$$

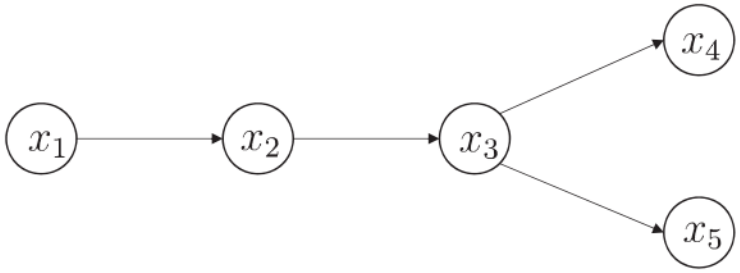
$$= \sum_{x_3} p(x_5 \mid x_3) m_{23}(x_3) m_{43}(x_3) \quad = \quad m_{35}(x_5) \ .$$

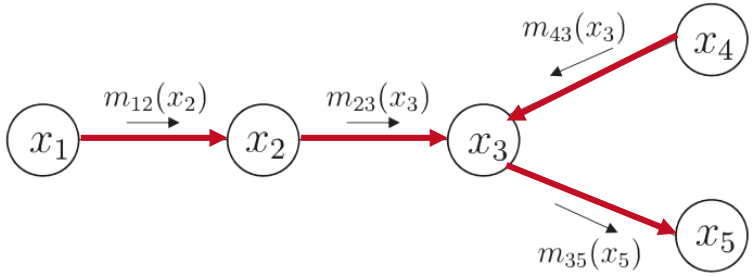
$m_{12}(x_2)$

- $m_{ij}\ (x_j\ )$ 是求加过程中的中间结果，下标i表示此项是对x_i求加的结果，下标j表示此项中还**剩余**的其它变量；显然，$m_{ij}\ (x_j\ )$是关于x_j的函数

精确推断：变量消去

- 计算边缘概率 $p(x_5)$





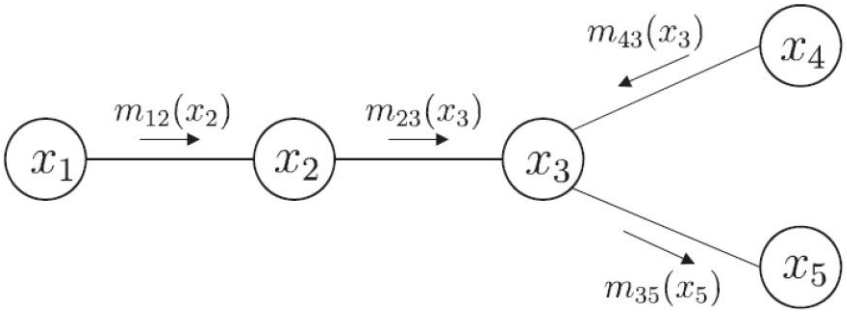
变量消去法实际上是利用了**乘法对加法的分配律**，将**对多个变量的积的求和**问题转化为**对部分变量交替进行求积和求和**的问题。这种转化使得每次的求和和求积运算被**限制在局部**，仅和部分变量有关，从而简化了计算

信念传播

- 计算**多个边际分布**，重复使用变量消去法会造成大量的冗余计算。利用**信念传播**（Belief Propagation）方法避免冗余计算
- 信念传播算法中，一个结点仅在接收到来自其它所有结点的消息后才能向另一个结点发送消息，且结点的边际分布正比于它所接收的消息的乘积：

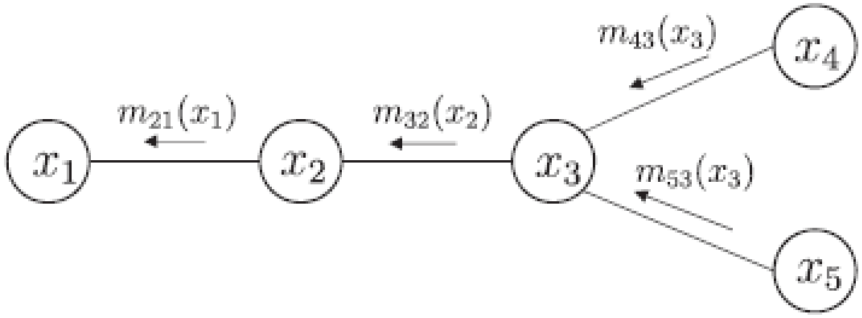
$$p(x_i) \propto \prod_{k \in n(i)} m_{ki}(x_i)$$

- 例：下图中，结点x_3要向x_5发送消息，必须事先收到来自结点x_2和x_4的消息，且传递到x_5的消息$m_{35}(x_5)$ 恰为概率$P(x_5)$:

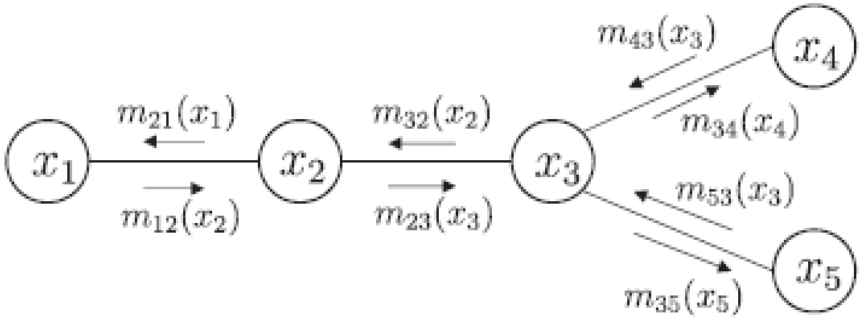


信念传播

- 若图中**没有环**，则信念传播算法经过**两个步骤**即可完成所有消息传递，进而能计算**所有变量**上的边际分布：
  - 指定一个根节点，从所有叶结点开始向根节点传递消息，直到根节点收到所有邻接结点的消息
  - 从根结点开始向叶结点传递消息，直到所有叶结点均收到消息



(a) 消息传向根结点



(b) 消息从根结点传出

近似推断

- 精确推断方法需要很大的计算开销，因此在现实应用中近似推断方法更为常用。
- 近似推断方法大致可以分为**两类**：
  - 采样法（sampling）：通过使用随机化方法完成近似，如MCMC采样
  - 变分推断（variational inference)：使用确定性近似完成推断

近似推断：采样法

- 很多任务中，我们关心的并非概率分布本身，而是基于概率分布的**期望**，并且还能基于期望进一步作出决策。
- 若直接计算或逼近这个期望比推断概率分布更容易，则直接操作无疑将使推断问题更为高效
- 采样法基于这个思路，假定目标是计算函数$f(x)$在概率密度函数$p(x)$下的期望：

$$\mathbb{E}_p[f] = \int f(x)p(x)dx$$

- 则可根据$p(x)$抽取一组样本$\{x_1, x_2, \ldots, x_N\}$，然后计算$f(x)$在这些**样本上的均值**

$$\hat{f} = \frac{1}{N}\sum_{i=1}^{N} f(x_i)$$

  - 用于近似期望$\mathbb{E}[f]$，如果样本$\{x_1, x_2, \cdots, x_N\}$独立，基于**大数定律**，这种通过大量取样的方法就获得较高的近似精度。问题的**关键**就变成了**如何取样**，即如何高效地**从图模型所描述的概率**分布中获取样本。

近似推断：MCMC方法

- 马尔可夫链蒙特卡罗（Markov Chain Monte Carlo， MCMC）是图模型中最常用的采样技术。
- 给定连续变量$x \in X$的概率密度函数$p(x)$，x在区间A中的概率为：

$$P(A) = \int_A p(x)dx$$

- 若有函数$f: X \to \mathbb{R}$, 可计算$f(x)$的期望：

$$p(f) = \mathbb{E}_p\left[f(X)\right] = \int_x f(x)p(x)dx,$$

- 若x为高维多原变量且服从一个复杂分布，积分操作会很困难。
- MCMC先构造出**服从p分布的独立同分布随机变量**$x_1, x_2, \ldots, x_N$，再得到无偏估计：

$$\tilde{p}(f) = \frac{1}{N}\sum_{i=1}^{N} f(x_i),$$

近似推断：MCMC方法

- 如何在概率分布复杂的情况下产生独立同分布的样本?
- **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$\mathbf{x}$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。
- 判定马尔可夫链的平稳态：
  - 假定马尔可夫链T的状态转移概率为$T(\mathbf{x}'|\mathbf{x})$，$t$时刻状态的分布为$p(\mathbf{x}^t)$，则若在某个时刻马尔可夫链满足平稳条件：

$$p(\mathbf{x}^t)T(\mathbf{x}^{t-1} \mid \mathbf{x}^t) = p(\mathbf{x}^{t-1})T(\mathbf{x}^t \mid \mathbf{x}^{t-1})$$

  - 则$p(\mathbf{x})$是该马尔可夫链的平稳分布，且马尔可夫链在满足该条件时已经收敛到平稳状态
- MCMC方法先设法构造一条马尔可夫链，使其收敛至平稳分布恰为待估计参数的后验分布，然后通过该马尔可夫链产生样本，用这些样本进行估计

Metropolis-Hastings (MH)算法

- MCMC中如何构造马尔可夫链的转移概率至关重要，不同构造方法产生不同的MCMC算法。
- Metropolis-Hastings (MH)算法是MCMC（Markov Chain Monte Carlo）算法的代表之一。基于“拒绝采样”逼近平稳分布。
  - 算法每次根据上一轮采样结果$\mathbf{x}^{t-1}$来采样获得候选状态样本$\mathbf{x}^*$，但这个样本会以一定概率被拒绝
  - 若从状态$\mathbf{x}^{t-1}$到状态$\mathbf{x}^*$的转移概率为$Q(\mathbf{x}^*|\mathbf{x}^{t-1})A(\mathbf{x}^*|\mathbf{x}^{t-1})$，$\mathbf{x}^*$最终收敛到平稳状态，则：

用户给定的先验概率

$$p(\mathbf{x}^{t-1})Q(\mathbf{x}^* \mid \mathbf{x}^{t-1})A(\mathbf{x}^* \mid \mathbf{x}^{t-1}) = p(\mathbf{x}^*)Q(\mathbf{x}^{t-1} \mid \mathbf{x}^*)A(\mathbf{x}^{t-1} \mid \mathbf{x}^*)$$

$\mathbf{x}^*$被接受的概率

  - 为达到平稳状态，只需将接受率设置为：

$$A(\mathbf{x}^* \mid \mathbf{x}^{t-1}) = \min\left(1, \frac{p(\mathbf{x}^*)Q(\mathbf{x}^{t-1}|\mathbf{x}^*)}{p(\mathbf{x}^{t-1})Q(\mathbf{x}^*|\mathbf{x}^{t-1})}\right)$$

Metropolis-Hastings (MH)算法

输入： 先验概率 $Q(\mathbf{x}^* \mid \mathbf{x}^{t-1})$.

过程：

初始化 $\mathbf{x}^0$;
for $t = 1, 2, \ldots$ **do**
  根据 $Q(\mathbf{x}^* \mid \mathbf{x}^{t-1})$ 采样出候选样本 $\mathbf{x}^*$;
  根据均匀分布从 $(0, 1)$ 范围内采样出阈值 u;
  if $u \leq A(\mathbf{x}^* \mid \mathbf{x}^{t-1})$ **then**
    $\mathbf{x}^t = \mathbf{x}^*$
  else
    $\mathbf{x}^t = \mathbf{x}^{t-1}$
  end if
end for
return $\mathbf{x}^1, \mathbf{x}^2, \ldots$

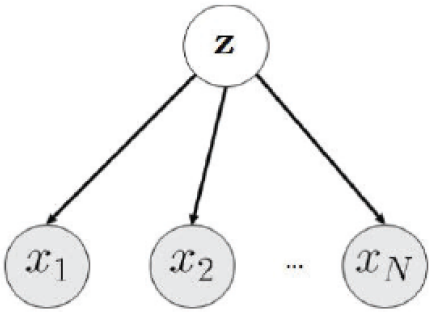
输出： 采样出的一个样本序列 $\mathbf{x}^1, \mathbf{x}^2, \ldots$

吉布斯采样

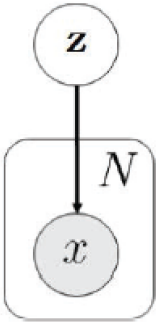
- 吉布斯采样(Gibbs sampling）被视为MH算法的特例，也使用马尔可夫链获取样本，平稳分布也是$p(\mathrm{x})$
- 具体来说，假定文本为$\mathrm{x}_1, \mathrm{x}_2, \ldots, \mathrm{x}_\mathrm{N}$，目标分布为$p(\mathrm{x})$，在初始化x的取值后，通过循环下列步骤来完成采样：
  - 1. 随机或以某个次序选取某变量x_i
  - 2. 根据x中除x_i外的变量的现有取值，计算条件概率$p(x_i \mid \mathrm{x}_i)$，其中$\mathrm{x}_i = \{x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_N\}$
  - 3. 根据$p(x_i \mid \mathrm{x}_i)$对变量x_i采样，用采样值代替原值

近似推断：变分推断

- 变分推断通过使用已知简单分布来逼近需推断的复杂分布，并通过限制近似分布的类型，从而得到一种局部最优、但具有确定解的近似后验分布
- 图模型**盘式记法**(plate notation）：
  - 相互独立的、由相同机制生成的多个变量被放在一个方框（盘）内，并在方框中标出类似变量重复出现的个数N
  - 方框可以嵌套
  - 通常用阴影标注出已知的、能观察到的变量



(a) 普通变量关系图

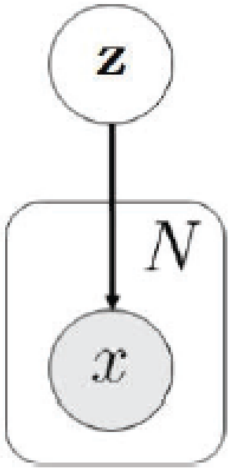


(b) 盘式记法

近似推断：变分推断

- 图模型中，已知变量本 $\mathbf{x} = \{x_1, x_2, \cdots, x_N\}$，$\Theta$是$\mathbf{x}$与$z$服从的分布参数。所有能观察到的变量 $\mathbf{x}$ 的联合分布概率密度函数是：

$$p(\mathbf{x} \mid \Theta) = \prod_{i=1}^{N} \sum_{\mathbf{z}} p(x_i, \mathbf{z} \mid \Theta)$$



- 对应的对数似然：

$$\ln p(\mathbf{x} \mid \Theta) = \sum_{i=1}^{N} \ln \left\{ \sum_{\mathbf{z}} p(x_i, \mathbf{z} \mid \Theta) \right\}$$

- 推断和学习任务主要是由观察变量$\mathbf{x}$来估计隐变量z和分布参数Θ，即求解$p(z \mid \mathbf{x}, \Theta)$和$\Theta$

近似推断：变分推断

- 对数似然：

$$\ln p(\mathbf{x} \mid \Theta) = \sum_{i=1}^{N} \ln \left\{ \sum_{\mathbf{z}} p(x_i, \mathbf{z} \mid \Theta) \right\}$$

- 可使用EM算法最大化对数似然
  - E步：根据t时刻的参数Θ^t对$p(z \mid \mathrm{x}, \Theta^t)$进行推断，并计算联合似然函数$p(\mathrm{x},\ z \mid \Theta^t)$
  - M步：基于E步结果进行最大化寻优，对关于变量Θ的函数$\mathcal{Q}(\Theta; \Theta^t)$进行最大化从而求取：

$$\Theta^{t+1} = \underset{\Theta}{\operatorname{argmax}} \mathcal{Q}(\Theta; \Theta^t)$$

$$= \underset{\Theta}{\operatorname{argmax}} \sum_{\mathbf{z}} p(\mathbf{z} \mid \mathbf{x}, \Theta^t) \ln p(\mathbf{x}, \mathbf{z} \mid \Theta)$$

对数联合似然函数$\ln p(\mathrm{x},\ z \mid \Theta)$在分布$p(z \mid \mathrm{x}, \Theta^t)$下的期望，分布$p(z \mid \mathrm{x}, \Theta^t)$与z的真实后验分布相等，$\mathcal{Q}(\Theta; \Theta^t)$近似于对数似然函数

近似推断：变分方法

- $p(z \mid \mathrm{x}, \Theta^t)$是隐变量z的近似分布，将这个近似分布用$q(z)$表示：

对数似然

$$\ln p(\mathbf{X}) = \mathcal{L}(q) + \mathrm{KL}(q \parallel p)$$

构成下界

$$\mathcal{L}(q) = \int q(\mathbf{z}) \ln \left\{ \frac{p(\mathbf{x}, \mathbf{z})}{q(\mathbf{z})} \right\} \mathrm{d}\mathbf{z}$$

分布q和p的差异度量

$$\mathrm{KL}(q \parallel p) = -\int q(\mathbf{z}) \ln \left\{ \frac{p(\mathbf{z}) | p(\mathbf{x})}{q(\mathbf{z})} \right\} \mathrm{d}\mathbf{z}$$

- 现实任务中，E部的推断可能很复杂
  - 借助变分推断，简化q的形式（增加假设）

$$q(\mathbf{z}) = \prod_{i=1}^{M} q_i(\mathbf{z}_i)$$

近似推断：变分方法

- 假设z的分布：$$q(\mathbf{z}) = \prod_{i=1}^{M} q_i(\mathbf{z}_i)$$

- 假设复杂的多变量z可以拆解为一系列相互独立的多变量z_i，同时令q_i分布相对简单或有很好的结构（如为指数族分布）：

$$\begin{aligned}
\mathcal{L}(q) &= \int \prod_i q_i \left\{ \ln p(\mathbf{x}, \mathbf{z}) - \sum_i \ln q_i \right\} d\mathbf{z} \\
&= \int q_j \left\{ \int \ln p(\mathbf{x}, \mathbf{z}) \prod_{i \neq j} q_i \mathrm{d}\mathbf{z}_i \right\} \mathrm{d}\mathbf{z}_j - \int q_j \ln q_j \mathrm{d}\mathbf{z}_j + \text{const.} \\
&= \int q_j \ln \tilde{p}(\mathbf{x}, \mathbf{z}_j) \mathrm{d}\mathbf{z}_j - \int q_j \ln q_j \mathrm{d}\mathbf{z}_j + \text{const}
\end{aligned}$$

$$\ln \tilde{p}(\mathbf{x}, \mathbf{z}_j) = \mathbb{E}_{i \neq j} \left[\ln p(\mathbf{x}, \mathbf{z}) \right] + \text{const}$$

$$\mathbb{E}_{i \neq j} \left[\ln p(\mathbf{X}, \mathbf{z}) \right] = \int \ln p(\mathbf{x}, \mathbf{z}) \prod_{i \neq j} q_i \mathrm{d}\mathbf{z}_i .$$

- 要得到q_j，可固定$q_{i \neq j}$再对$\mathcal{L}(q)$最大化，可发现上式等于$-KL\left(q_j \,||\, \tilde{p}(\mathrm{x}, z_j)\right)$，当$q_j = \tilde{p}(\mathrm{x}, \mathrm{z_j})$时$\mathcal{L}(q)$最大，可得**最优近似**：

$$\ln q_j^*(\mathbf{z}_j) = \mathbb{E}_{i \neq j} \left[\ln p(\mathbf{x}, \mathbf{z}) \right] + \text{const} \quad \Rightarrow \quad q_j^*(\mathbf{z}_j) = \frac{\exp(\mathbb{E}_{i \neq j}[\ln p(\mathbf{x}, \mathbf{z})])}{\int \exp(\mathbb{E}_{i \neq j}[\ln p(\mathbf{x}, \mathbf{z})]) d\mathbf{z}_j}$$

近似推断：变分方法

- 变量子集z_j所服从的最优分布q_j^*应满足：

$$q_j^*(\mathbf{z}_j) = \frac{\exp(\mathbb{E}_{i\neq j}[\ln p(\mathbf{x},\mathbf{z})])}{\int \exp(\mathbb{E}_{i\neq j}[\ln p(\mathbf{x},\mathbf{z})])d\mathbf{z}_j}$$

- 因此，通过恰当分割变量子集z_j并选择q_i服从的分布，$\mathbb{E}_{i\neq j}[\ln p(\mathrm{x},\mathrm{z})]$往往有闭式解，使得上式能对隐变量高效推断
- 由于在对z_j所服从的分布q_j^*估计时融合了z_j之外的其它$z_{i\neq j}$的信息，这是通过联合似然函数$\ln p(\mathrm{x},\mathrm{z})$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
- 在实际应用中，最重要的是**考虑如何对隐变量进行拆解**，以及**假设各变量子集服从何种分布**，在此基础之上结合EM算法对概率图模型进行推断和参数估计

话题模型

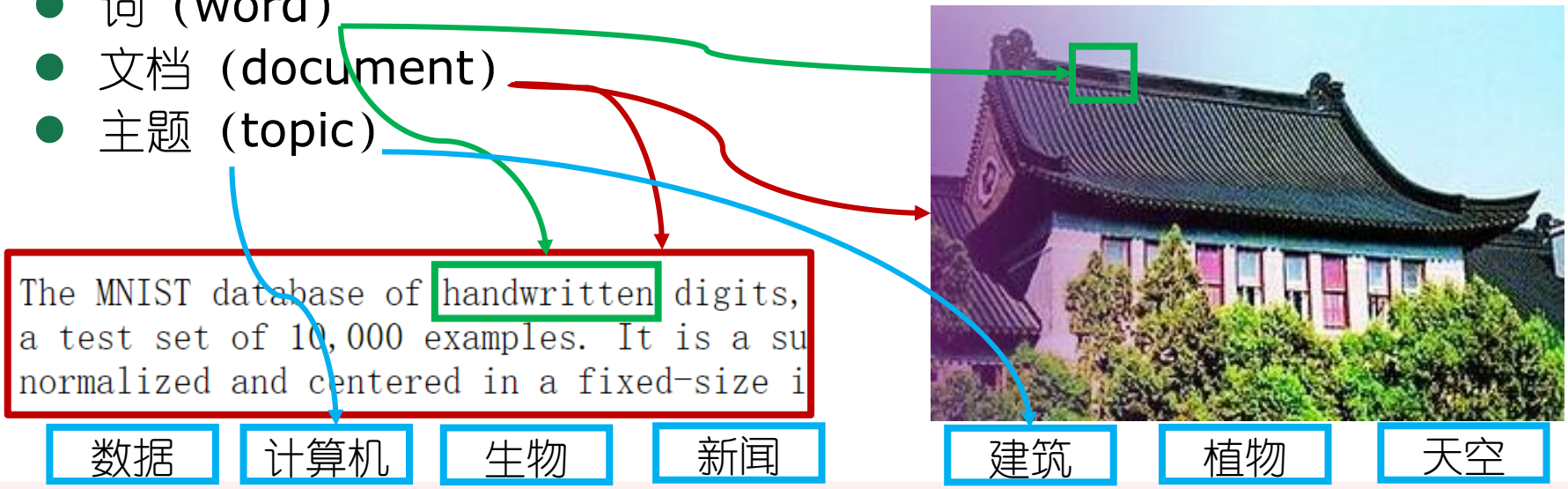
- 话题模型（topic model）是一类生成式**有向图模型**，主要用来处理离散型的数据集合（如文本集合）。作为一种**非监督**产生式模型，话题模型能够有效利用海量数据发现文档集合中**隐含的语义**。隐狄里克雷分配模型（Latent Dirichlet Allocation, **LDA**）是话题模型的典型代表。

- 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

话题模型的基本单元

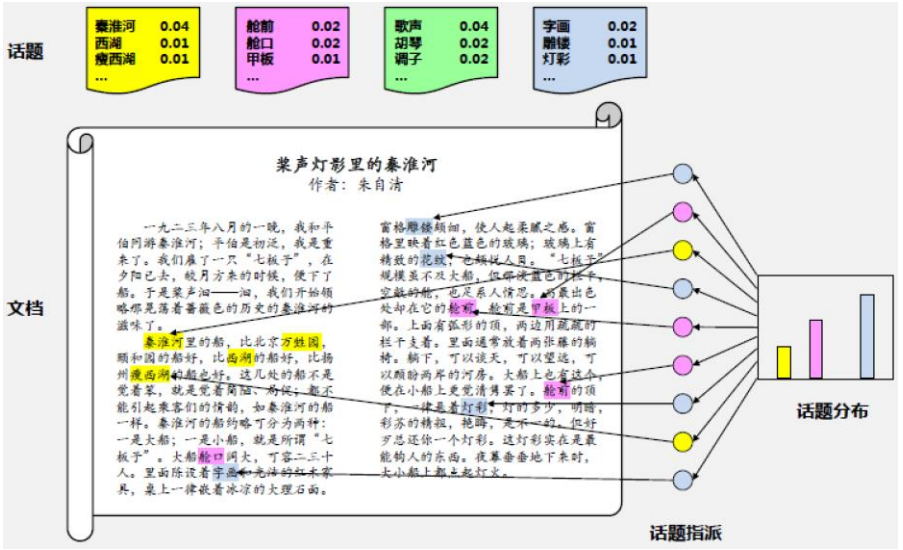
- 话题模型（topic model）是一类生成式**有向图模型**，主要用来处理离散型的数据集合（如文本集合）。作为一种**非监督**产生式模型，话题模型能够有效利用海量数据发现文档集合中**隐含的语义**。隐狄里克雷分配模型（Latent Dirichlet Allocation, **LDA**）是话题模型的典型代表。

- LDA的基本单元
  - 词（word）
  - 文档（document）
  - 主题（topic）



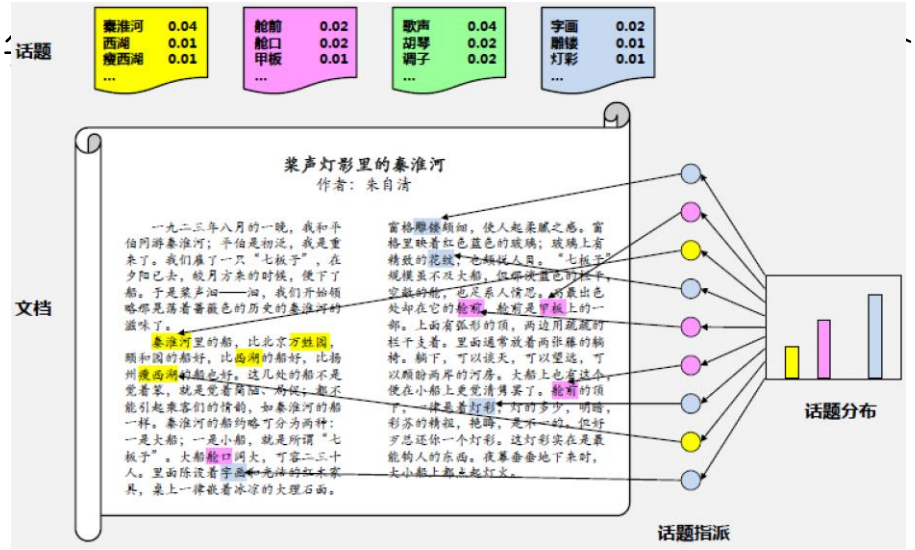
话题模型的构成

- 一个话题就像一个箱子，里面装着这个概念下出现概率较高的词
- 假定数据集中一共包含K个话题和T篇文档，文档中的词来自一个包含N个词的字典
- 用T个N维向量$W = \{w_1, w_2, \ldots, w_T\}$表示文档集合
- 用K个N维向量β_K表示话题
- $w_t \in \mathbb{R}^N$的第n个分量$w_{t,n}$表示文档t中词n的词频，$\beta_K \in \mathbb{R}^N$的第n个分量$\beta_{k,n}$表示话题k中词n的词频



文档的生成过程

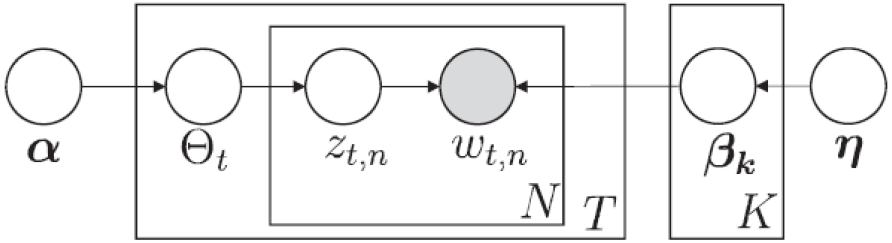
- 现实任务中通过统计文档中出现的词来获得词频向量w_i
- LDA从生成式模型的角度看待文档和话题，认为每篇文档包含多个话题
  - 用$\Theta_t \in \mathbb{R}^K$表示文档$t$中所包含的各个话题$k$的比例
- 生成文档t
  - 从以α为参数的狄利克雷分布中随机采样一个话题分布Θ_t；
  - 按如下步骤产生文档中的N个词
    - 根据Θ_t进行话题指派，得到文档t中词n的话题$z_{t,n}$；
    - 根据指派的话题所对应的的词分布β_k随机采样生成词



LDA的图模型表示

- 生成文档t
  - 从以α为参数的狄利克雷分布中随机采样一个话题分布Θ_t；
  - 按如下步骤产生文档中的N个词
    - 根据Θ_t进行话题指派，得到文档t中词n的话题$z_{t,n}$；
    - 根据指派的话题所对应的的词分布β_k随机采样生成词

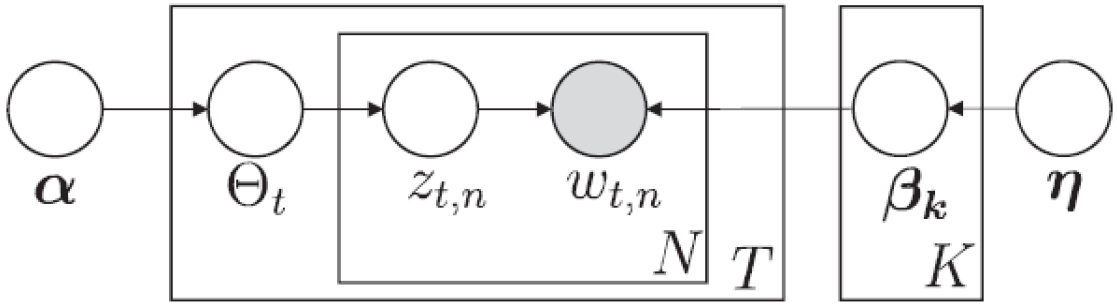
- 生成的文档以不同比例包含多个话题，文档中的每个词来自于一个话题，这个话题是根据话题比例产生的



- 词频$w_{t,n}$是唯一已观测变量，依赖于话题指派$z_{t,n}$及话题对应词频β_k， 话题指派$z_{t,n}$依赖于话题分布Θ_t，Θ_t依赖于α，话题词频依赖于参数η

LDA的图模型表示

- 词频$w_{t,n}$是唯一已观测变量，依赖于话题指派$z_{t,n}$及话题对应词频β_k，话题指派$z_{t,n}$依赖于话题分布Θ_t，Θ_t依赖于α，话题词频依赖于参数η



狄利克雷分布

$$p(\mathbf{W},\mathbf{z},\boldsymbol{\beta},\Theta \mid \boldsymbol{\alpha},\boldsymbol{\eta}) = \prod_{t=1}^{T} p(\Theta_t \mid \boldsymbol{\alpha}) \prod_{i=1}^{K} p(\boldsymbol{\beta}_k \mid \boldsymbol{\eta}) \left(\prod_{n=1}^{N} P(w_{t,n} \mid z_{t,n}, \boldsymbol{\beta}_k) P(z_{t,n} \mid \Theta_t) \right)$$

LDA的基本问题

- 模型**参数估计**
  - 给定训练数据$W = \{w_1, w_2, \ldots, w_T\}$，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(通常采用变分法求取近似解)

$$LL(\boldsymbol{\alpha}, \boldsymbol{\eta}) = \sum_{t=1}^{T} \ln p(\boldsymbol{w}_t \mid \boldsymbol{\alpha}, \boldsymbol{\eta})$$

- 模型**推断**
  - 模型已知，参数$\boldsymbol{\alpha}$和$\boldsymbol{\eta}$已确定，根据词频来推断文档话题结构（推断$\Theta_t, \beta_k, z_{t,n}$）可通过求解（分母常采用吉布斯采样或变分法进行推断）

$$p(\mathbf{z}, \boldsymbol{\beta}, \Theta \mid \mathbf{W}, \boldsymbol{\alpha}, \boldsymbol{\eta}) = \frac{p(\mathbf{W}, \mathbf{z}, \boldsymbol{\beta}, \Theta \mid \boldsymbol{\alpha}, \boldsymbol{\eta})}{p(\mathbf{W} \mid \boldsymbol{\alpha}, \boldsymbol{\eta})}$$

总结

- 概率图模型
  - 表示随机变量之间的（条件）独立性关系
- 图模型的两种表示
  - 有向图
    - 隐马尔可夫模型 HMM（动态贝叶斯网）
  - 无向图
    - 马尔可夫随机场 MRF
    - 条件随机场 CRF
- 图模型推断
  - 联合分布，条件分布，边缘分布
    - 精确推断， 近似推断
  - 参数估计（转化为推断问题）
- 话题模型