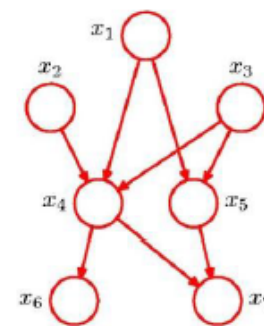
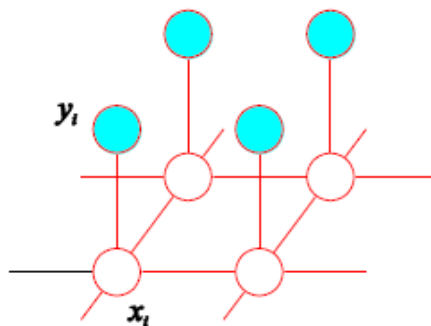
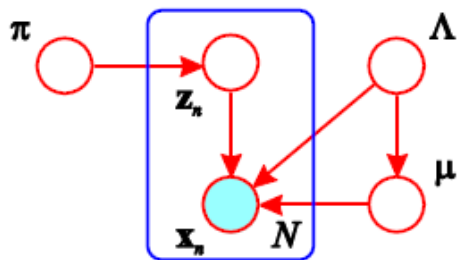


# 概率图模型



郭嘉丰

中国科学院大学，中国科学院计算技术研究所

课程代码: [https://github.com/lixinsu/tutorials2018/blob/master/graph\\_model.ipynb](https://github.com/lixinsu/tutorials2018/blob/master/graph_model.ipynb)

# 大纲

- 概率图模型简介
- 两类概率图模型
  - 有向概率图模型
  - 无向概率图模型
- 学习和推断
- 典型的概率图模型
  - 从HMM到CRF
  - 从PLSA到LDA
- 总结

Markov random field

K-means clustering

logistic regression

random forest

neural networks

support vector machines

ICA

Gaussian process

RVM

HMM

deep networks

linear regression

Radial basis functions

factor analysis

Gaussian mixture

Kalman filter

principal components

kernel PCA

Boltzmann machines

decision trees

传统的机器学习:

- “如何把我的问题映射到标准的方法”?

基于模型的机器学习:

- “什么是适合我的问题的好的模型”?



目标:  
用一个框架创建一系列定制模型

## ■ 目标: 创建能够自适应、学习和推理的智能软件



Player skill



Game Result



Movie Preference



Ratings



Word



Ink

能够用模型去描述这些任务

## ■ 目标: 创建能够自适应、学习和推理的智能软件



Player skill

Game Result



Movie Preference

Ratings



Word

Ink



给定观测数据能够逆向推理



## ➤ 在具有不确定性的情况下进行推理

- 用户适合玩什么关卡？
- 用户将要看哪一个电影？
- 用户要写哪些内容？
- 用户要说什么？
- 用户要找的是哪个网页？
- 用户可能会点击哪个链接？
- 用户可能会买哪类的产品？
- 用户可能会做哪个动作？
- 其他...



60%



40%

使用**概率**处理不确定性

1. 能够用**模型**去描述这些任务

2. 给定观测**数据**能够逆向推理

3. 使用**概率**处理不确定性

■ 如何**表示**世界（**representation**）：

- 把世界表示为一组随机变量 $X_1 \dots X_n$ ，其联合分布为 $P(X_1, \dots, X_n)$
- 如何编码我们的领域知识 / 假设 / 限制？（简约的结构）

■ 如何进行**推断**（**inference**）：

- 如何根据模型和给定的数据回答问题或者查询？
- 利用已知变量推测未知变量的分布 $P(X_i|D)$

■ 如何**学习**模型（**learning**）：

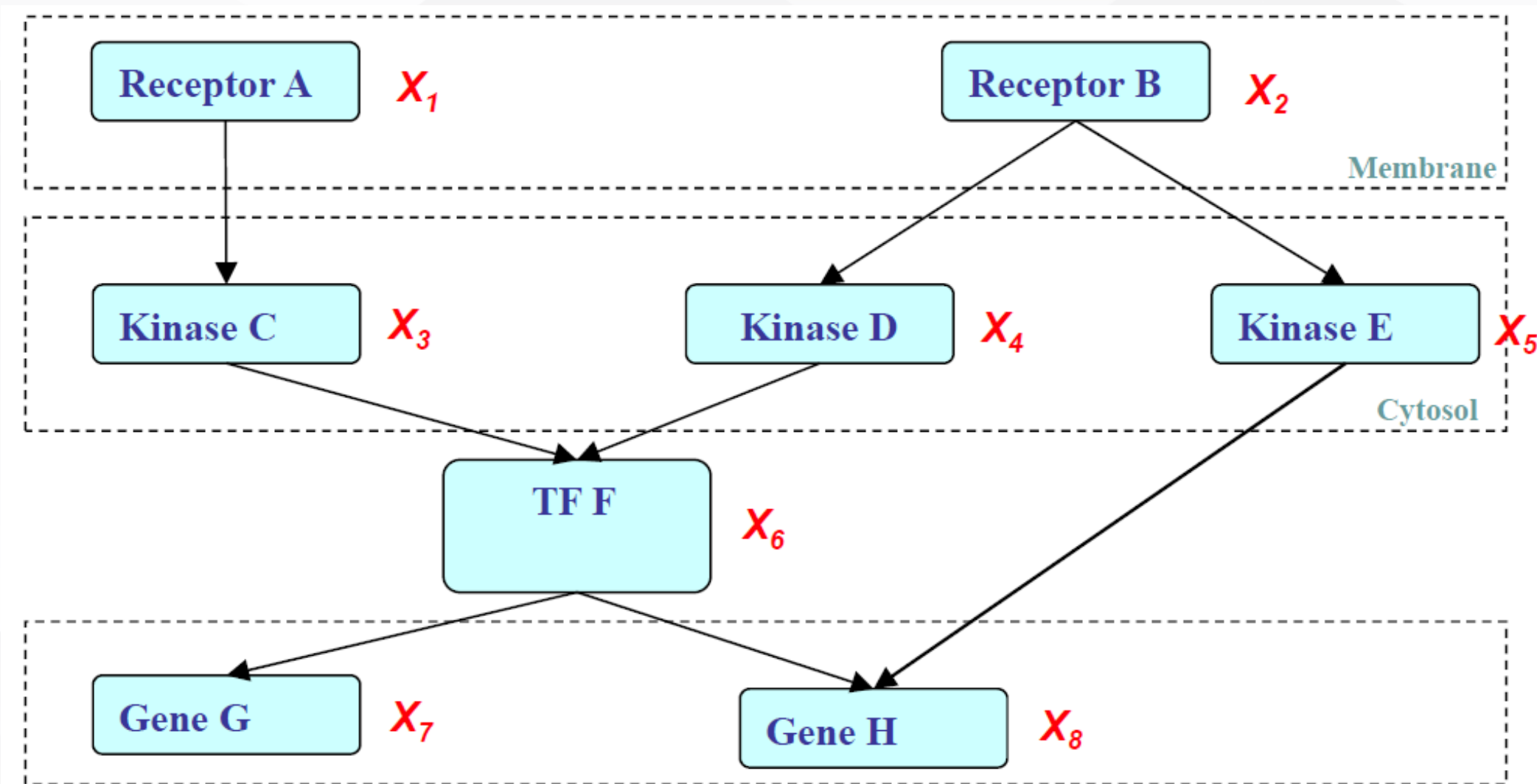
- 哪个模型最适合我的数据？
- 利用数据学习这个联合分布 $\mathcal{M} = \arg \max_{\mathcal{M} \in \mathcal{M}} F(D; \mathcal{M})$



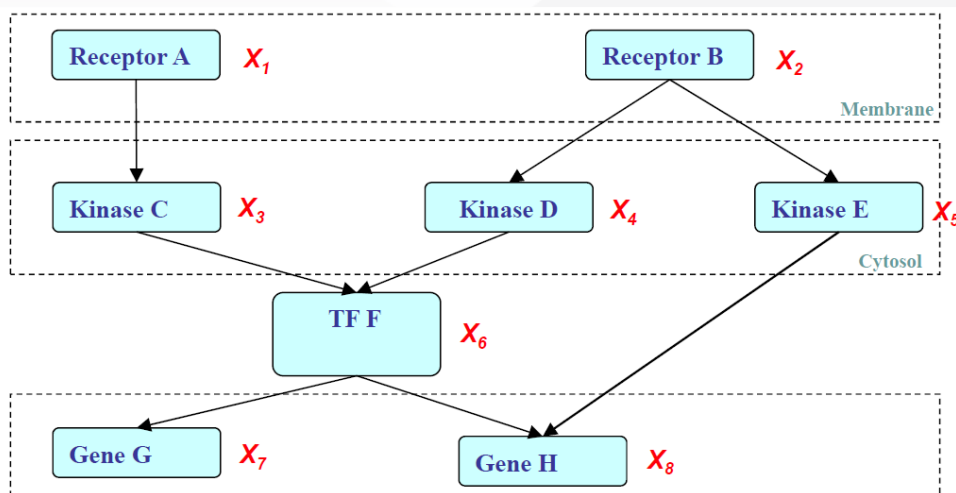
## 如何简洁地表示世界?

把世界表示为一组随机变量，加入我们的领域知识 / 假设 / 限制

变量之间的依赖（简约的结构）



## Graph



+



概率图模型 = 概率 + 结构



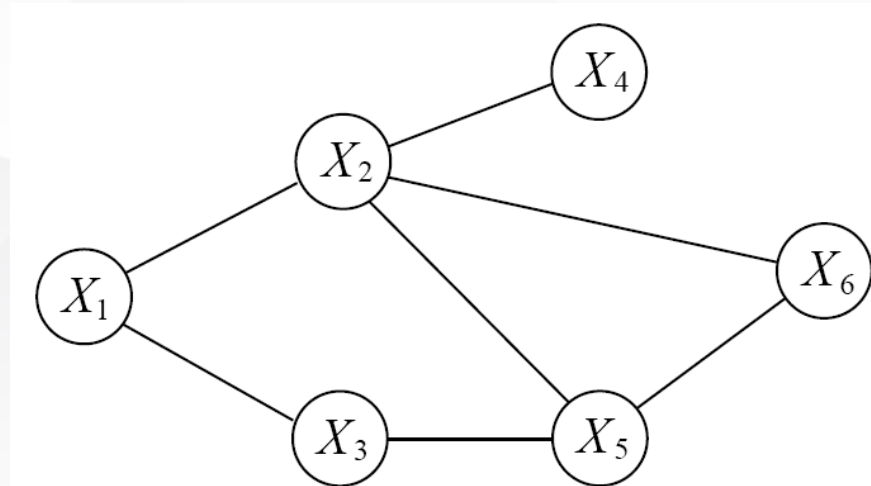
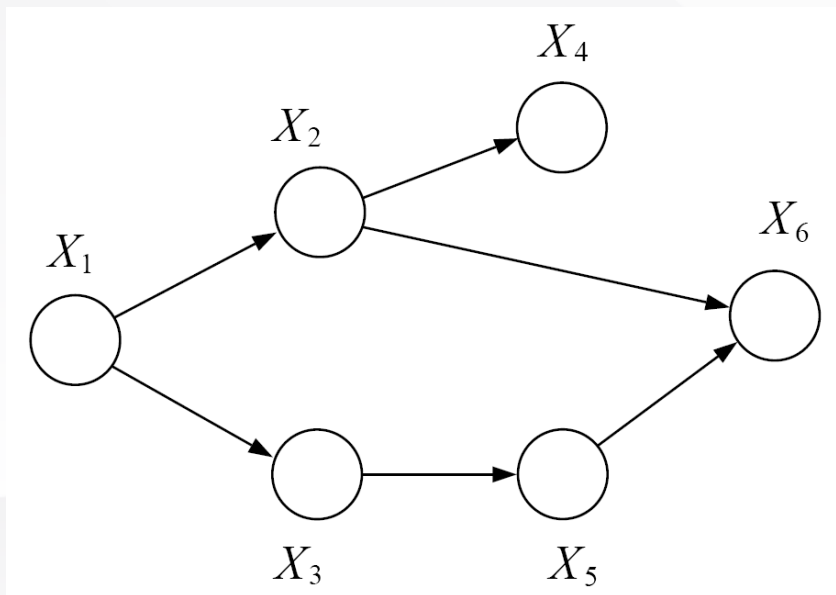
刻画我们的世界  
进行推断与学习

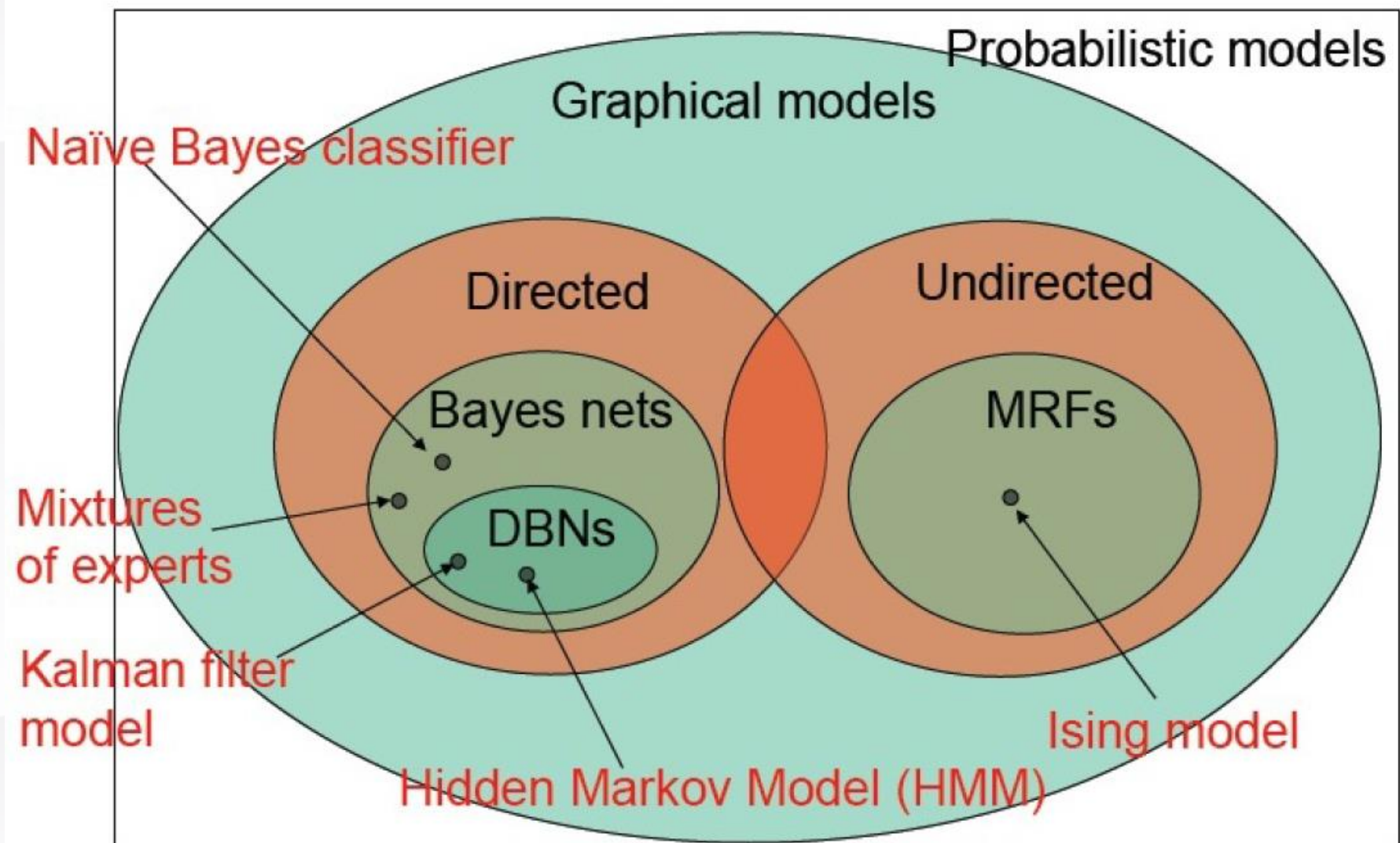
A language for communication  
A language for computation  
A language for development

■ **节点**表示随机变量/状态，**边**表示概率关系

■ **类型**

- 有向概率图模型 或 贝叶斯网络: 因果关系
- 无向图模型 或 马尔科夫随机场: 关联关系



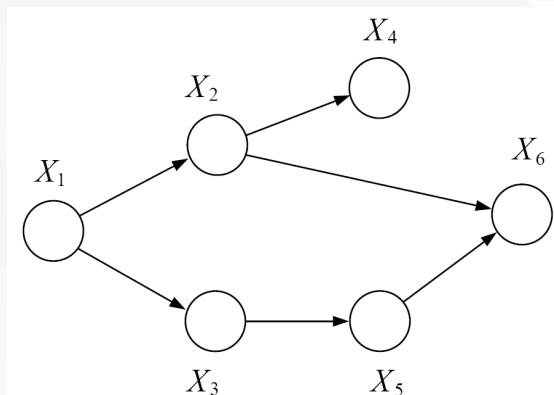


# 大纲

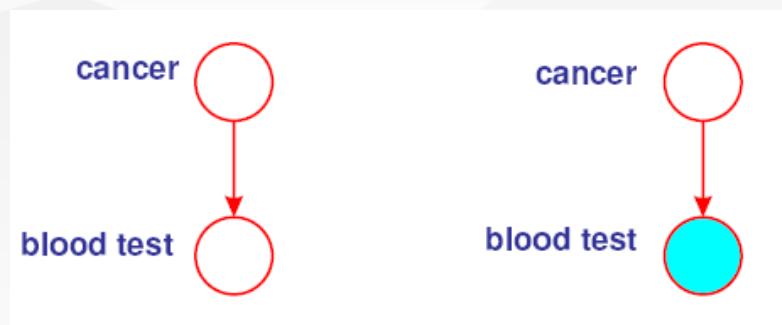
- 概率图模型简介
- 两类概率图模型
  - 有向概率图模型
  - 无向概率图模型
- 学习和推断
- 典型的概率图模型
  - 从HMM到CRF
  - 从PLSA到LDA
- 总结

# 有向概率图模型

- 定义: 有向图  $G = (V, E)$  包含一个点集合  $V$  和一个边的集合  $E$ ，其中每条边为有序点对



- 有向图模型可以表示因果关系
- 我们经常观察子变量并去推断出父变量的分布
- 例子:

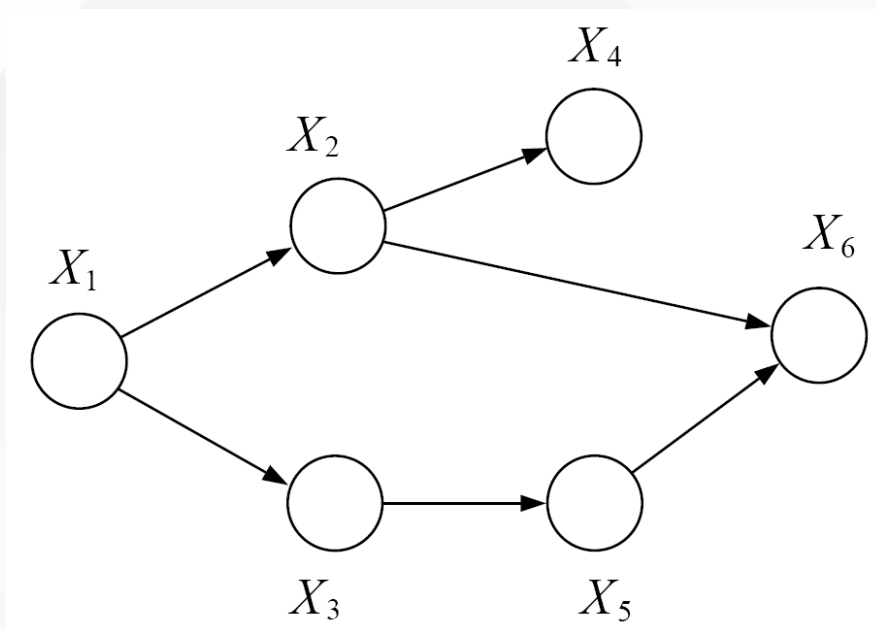




# 有向图的例子

- 隐马尔科夫模型
- 卡尔曼滤波
- 因子分析
- 概率主成分分析
- 独立成分分析
- 混合高斯
- 转换成分分析
- 概率专家系统
- Sigmoid 信念网络
- 层次化混合专家
- 等等...

## ➤ 有向概率图模型 (贝叶斯网)

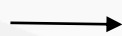


1. 概率分布



用于查询/推断

2. 表示



具体实现

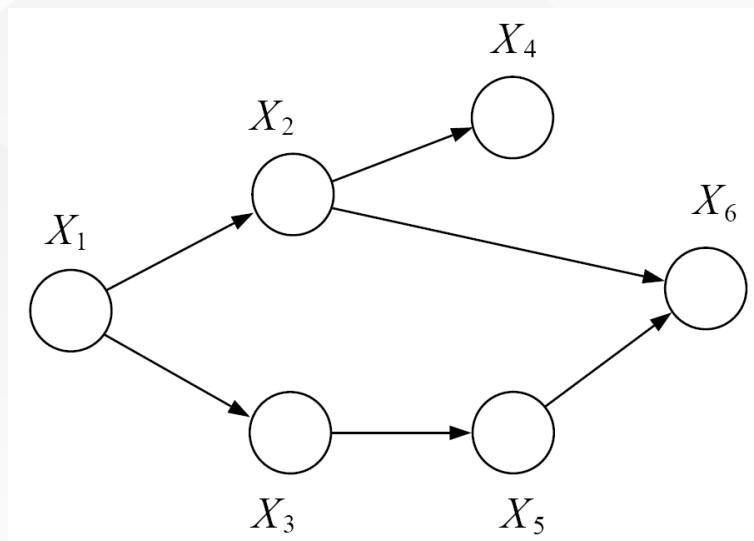
3. 条件独立



模型的解释

# ➤ 1. 概率分布

一个概率图模型对应着一族概率分布 (a family of probability distribution)



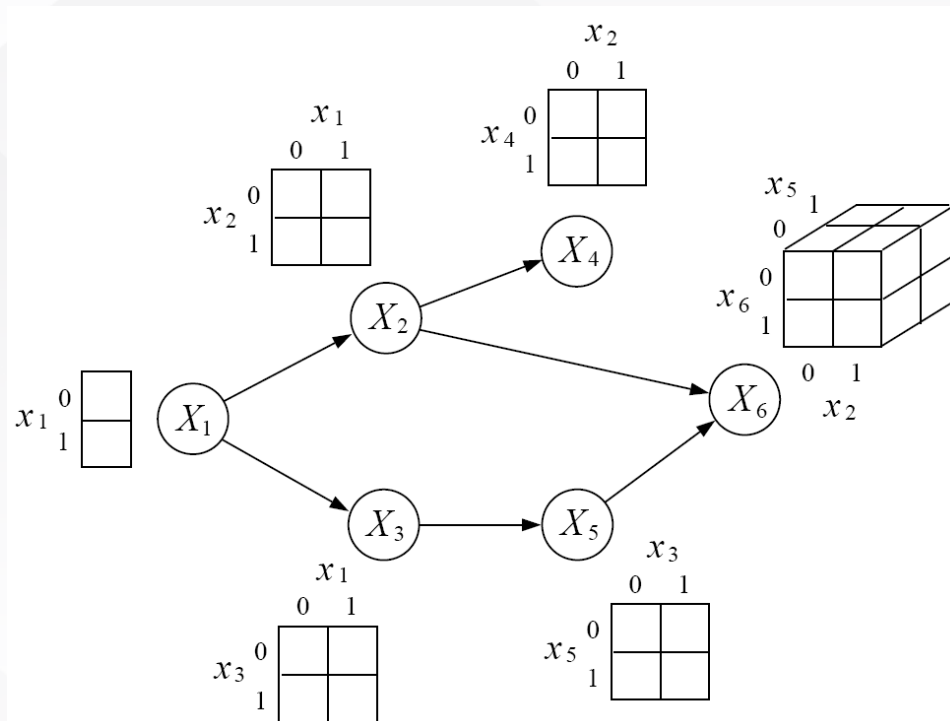
每个结点对应一个条件概率分布  $p(x_i|x_{\pi i})$

联合概率分布可以表示为:

$$P(x_1, x_2, \dots, x_n) = \prod_{i=1}^n p(x_i|x_{\pi i})$$

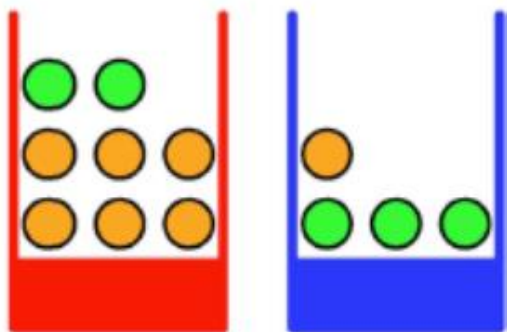
$$P(x_1, x_2, x_3, x_4, x_5, x_6) = P(x_1)P(x_2|x_1)P(x_3|x_1)P(x_4|x_2)P(x_5|x_3)P(x_6|x_2, x_5)$$

## 2. 表示



$$O(2^n) \rightarrow O(n \cdot 2^k)$$

贝叶斯网使用一系列变量间的“局部”关系“紧凑”地表示联合概率分布



$x_1$

$x_2$



$$p(x_1, x_2) = p(x_1)p(x_2|x_1)$$

- $x_1$  = Box: red/blue
- $x_2$  = Fruit: a/o

$$p(x_1)$$

$$p(x_1=\text{red}) = 4/10$$

$$p(x_1=\text{blue}) = 6/10$$

$$p(x_2|x_1)$$

$$p(x_2=\text{a}|x_1=\text{red}) = 1/4$$

$$p(x_2=\text{o}|x_1=\text{red}) = 3/4$$

$$p(x_2=\text{a}|x_1=\text{blue}) = 3/4$$

$$p(x_2=\text{o}|x_1=\text{blue}) = 1/4$$

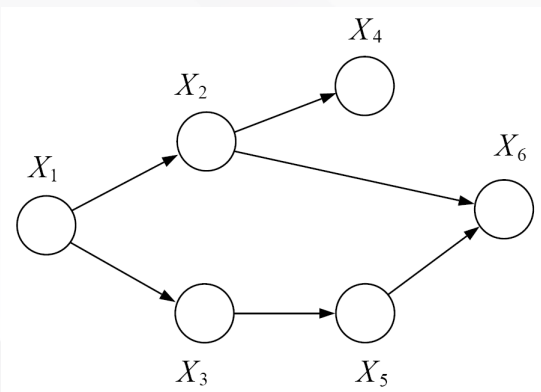
## 3. 条件独立

### 利用条件独立解释缺少的边

定义图 $G$ 中节点的顺序 $I$ ，如果对每个节点 $i \in V$ ，他的父节点都在这个顺序中出现在它之前，那么我们称 $I$ 为拓扑排序. 例如 $I = \{1,2,3,4,5,6\}$ 是图的一种拓扑排序.

对于节点 $i$ ，假设  $v_i$ 表示在 $I$ 中除了 $\pi_i$ 之外所有出现在 $i$ 个节点之前的节点

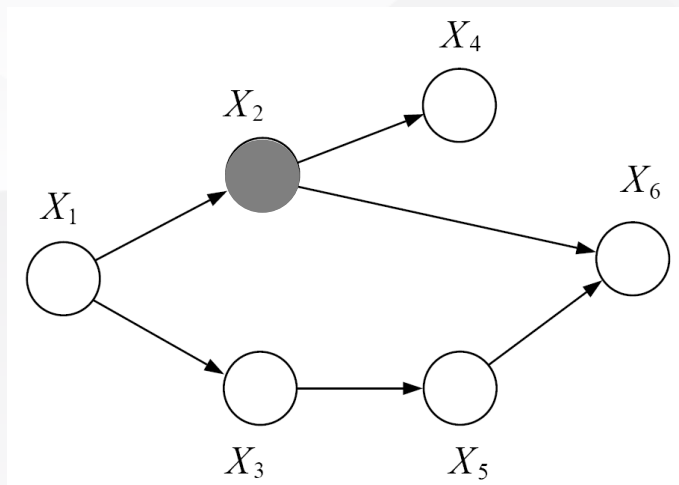
给定图 $G$ 的拓扑排序 $I$  我们将这样一组条件独立性陈述 $\{X_i \perp X_{v_i} | X_{\pi_i}\}$ 和图关联起来



断言： 给定一个节点的父节点，该节点和其祖先条件独立



### 3. 条件独立



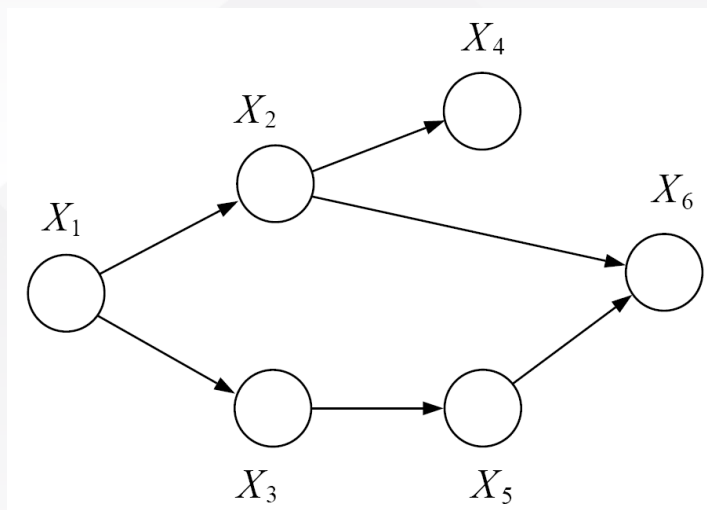
$$X_4 \perp \{X_1, X_3\} | X_2$$

$$\begin{aligned} P(x_1, x_2, x_3, x_4) &= \sum_{x_5} \sum_{x_6} P(x_1, x_2, x_3, x_4, x_5, x_6) = \sum_{x_5} \sum_{x_6} P(x_1) P(x_2 | x_1) P(x_3 | x_1) P(x_4 | x_2) P(x_5 | x_3) P(x_6 | x_2, x_5) \\ &= P(x_1) P(x_2 | x_1) P(x_3 | x_1) P(x_4 | x_2) \sum_{x_5} P(x_5 | x_3) \sum_{x_6} P(x_6 | x_2, x_5) = P(x_1) P(x_2 | x_1) P(x_3 | x_1) P(x_4 | x_2) \end{aligned}$$

$$P(x_1, x_2, x_3) = \sum_{x_4} P(x_1) P(x_2 | x_1) P(x_3 | x_1) P(x_4 | x_2) = P(x_1) P(x_2 | x_1) P(x_3 | x_1)$$

$$P(x_4 | x_1, x_2, x_3) = P(x_4 | x_2)$$

### 3. 条件独立



对于这样一族联合分布，是否存在其他的条件独立陈述？  
这些陈述是否有对应的图解释？

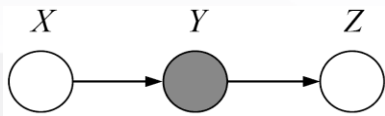
图分离？

给定  $X_2$  和  $X_3$ ,  $X_1$ 和 $X_6$  独立

给定 $X_1$  和 $X_6$ ,  $X_2$  不一定独立于 $X_3$

我们需要准确的刻画图上的“隔开”的概念

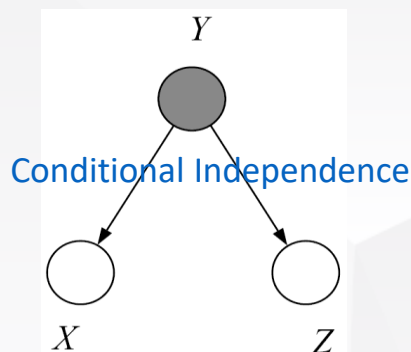
# 三种经典图



$$\begin{aligned} P(x, y, z) &= P(x)P(y|x)P(z|y) \\ P(z|x, y) &= \frac{P(x, y, z)}{P(x, y)} \\ &= \frac{P(x)P(y|x)P(z|y)}{P(x)P(y|x)} \\ &= P(z|y) \end{aligned}$$

$$X \perp\!\!\!\perp Z \mid Y$$

经典的马尔科夫链  
“过去”, “现在”, “未来”

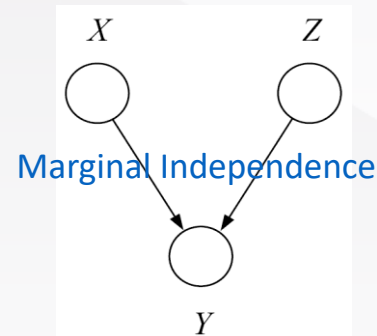


Conditional Independence

$$\begin{aligned} P(x, y, z) &= P(y)P(x|y)P(z|y) \\ P(x, z|y) &= \frac{P(y)P(x|y)P(z|y)}{P(y)} \\ &= P(x|y)P(z|y) \end{aligned}$$

$$X \perp\!\!\!\perp Z \mid Y$$

共同的起因 (Common Cause)  
Y “解释” X 和 Z 之间所有的依赖



Marginal Independence

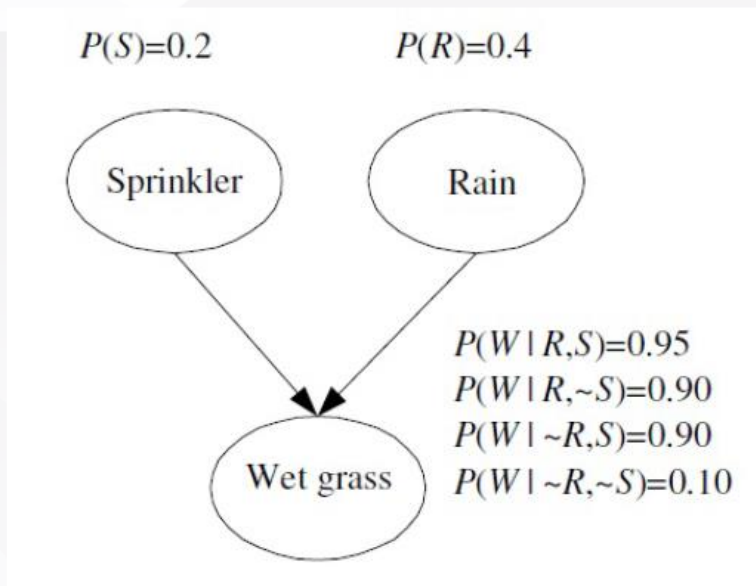
$$\begin{aligned} P(x, y, z) &= P(x)P(z)P(y|x, z) \\ &= P(x)P(z) \frac{P(x, y, z)}{P(x, z)} \\ P(x, z) &= P(x)P(z) \end{aligned}$$

$$X \perp\!\!\!\perp Z$$

共同效应 (Common effect)  
多个相互竞争的解释

# 解释消除 (Explaining Away)

## ■ 阐述:



$$P(S|R, W) = \frac{P(W|R, S)P(S|R)}{P(W|R)} = \frac{P(W|R, S)P(S)}{P(W|R)} = 0.21$$

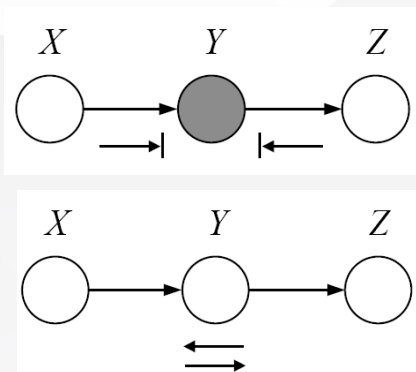
$$P(S|W) = \frac{0.92 \times 0.2}{0.52} = 0.35$$

$$0.21 = P(S|R, W) < P(S|W) = 0.35$$

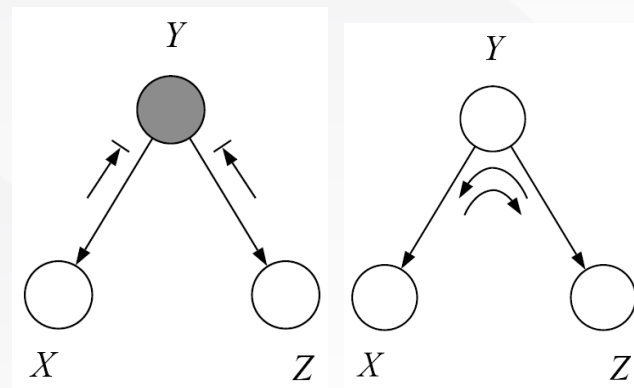
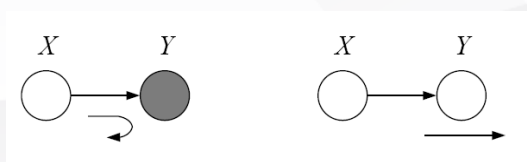
- 已知天下过雨会减少预测洒水器打开的概率
- 知道草地是湿的，那么下过雨和开过洒水器变成互相依赖的

$$P(S|R, W) \neq P(S|W)$$

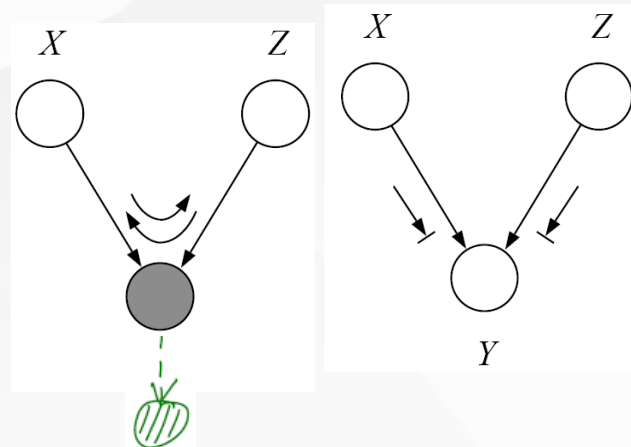
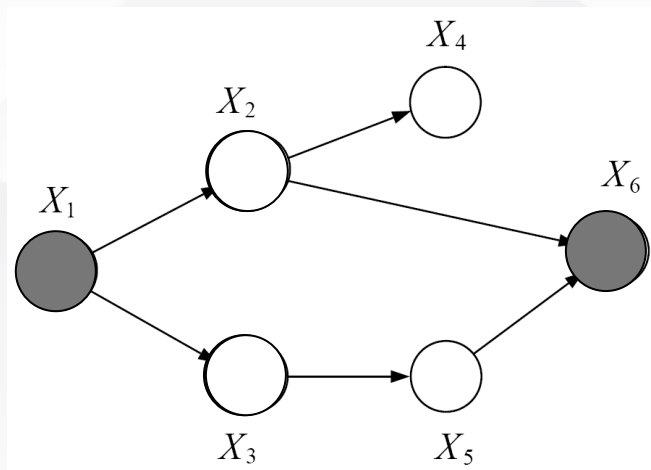
## 贝叶斯球算法(规则)



一个输入箭头和  
一个输出箭头



两个输出箭头



两个输入箭头

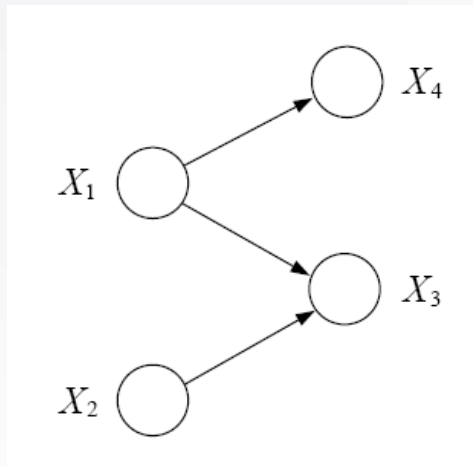
检查通过可达性

$$X_1 \perp\!\!\!\perp X_6 \mid \{X_2, X_3\}$$

$$X_2 \perp\!\!\!\perp X_3 \mid \{X_1, X_6\}$$

# 有向图模型小结

- 图模型和概率分布族关联
- 通过两种等价的方式定义:



$$p(x_1, \dots, x_n) \triangleq \prod_{i=1}^n p(x_i \mid x_{\pi_i}).$$

$$X_1 \perp\!\!\!\perp X_2$$

$$X_2 \perp\!\!\!\perp X_4$$

$$X_2 \perp\!\!\!\perp X_4 \mid X_1$$

$$X_3 \perp\!\!\!\perp X_4 \mid X_1$$


$$X_2 \perp\!\!\!\perp X_4 \mid \{X_1, X_3\}$$

$$\{X_2, X_3\} \perp\!\!\!\perp X_4 \mid X_1$$

参数化的方式或者利用条件独立陈述来对联合概率分布进行描述是等价的！




# 2011图灵奖颁发给贝叶斯网络




acm


MORE ACM AWARDS



A.M. TURING AWARD

A.M. TURING CENTENARY CELEBRATION WEBCAST

 Search




A.M. TURING AWARD WINNERS BY...

ALPHABETICAL LISTING

YEAR OF THE AWARD

RESEARCH SUBJECT



 **Photo-Essay**

**BIRTH:**

September 4, 1936, Tel Aviv.

**EDUCATION:**

B.S., Electrical Engineering (Technion, 1960); M.S., Electronics (Newark College of Engineering, 1961); M.S., Physics (Rutgers University, 1965); Ph.D., Electrical Engineering (Polytechnic Institute of Brooklyn, 1965).

**EXPERIENCE:**


Research Engineer, New York University Medical School (1960–1961); Instructor,


**JUDEA PEARL**


United States – 2011


**CITATION**


For fundamental contributions to artificial intelligence through the development of a calculus for probabilistic and causal reasoning.

 SHORT ANNOTATED BIBLIOGRAPHY

 ACM DL AUTHOR PROFILE

 ACM TURING AWARD LECTURE VIDEO

 RESEARCH SUBJECTS

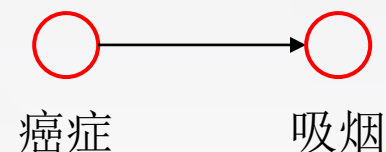
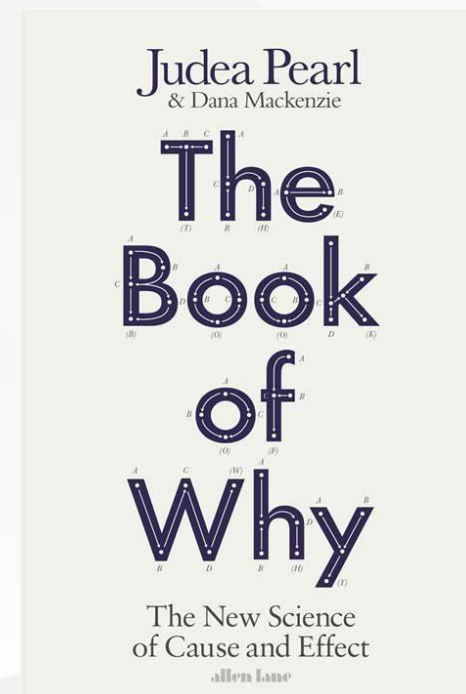
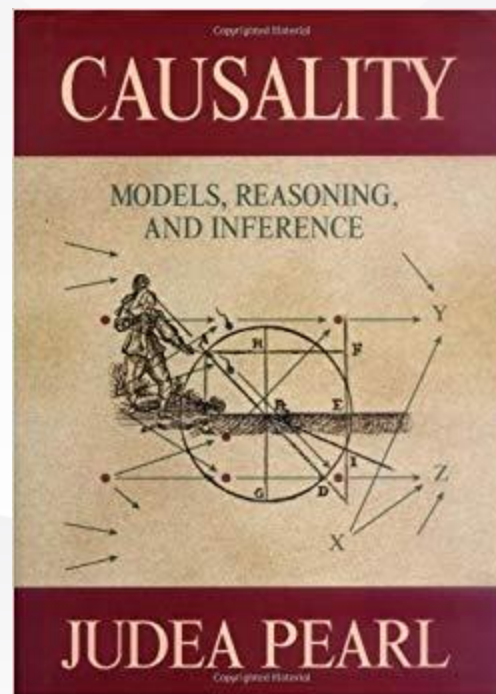
 ADDITIONAL MATERIALS

Judea Pearl created the representational and computational foundation for the processing of information under uncertainty.

He is credited with the invention of *Bayesian networks*, a mathematical formalism for defining complex probability models, as well as the principal algorithms used for inference in these models. This work not only revolutionized the field of artificial intelligence but also became an important tool for many other branches of engineering and the natural sciences. He later created a mathematical framework for *causal inference* that has had significant impact in the social sciences.

Judea Pearl was born on September 4, 1936, in Tel Aviv, which was at that time administered under the British Mandate for Palestine. He grew up in *Bnei Brak*, a Biblical town his grandfather went to reestablish in 1924. In 1956, after serving in the Israeli army and joining a Kibbutz, Judea decided to study engineering. He attended the Technion, where he met his wife, Ruth, and received a B.S. degree in Electrical Engineering in 1960. Recalling the Technion faculty members in a 2012 interview in the *Technion Magazine*, he emphasized the thrill of discovery:

# 从相关到因果

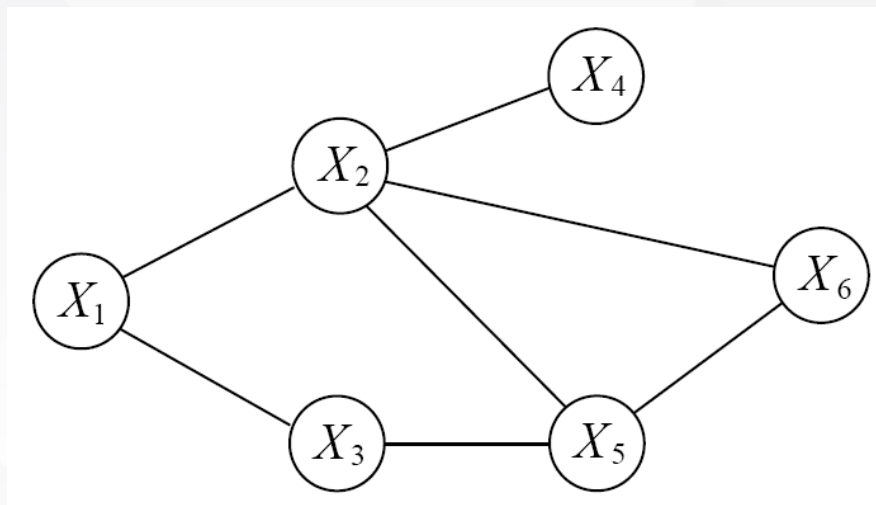


# 大纲

- 概率图模型简介
- 两类概率图模型
  - 有向概率图模型
  - 无向概率图模型
- 学习和推断
- 典型的概率图模型
  - 从HMM到CRF
  - 从PLSA到LDA
- 总结

## ➤ 无向图模型 (马尔科夫随机场)

定义 一个无向图  $G = (V, E)$  包含节点集合  $V$  和边的集合  $E$ ，边由点对组成



1. 概率分布



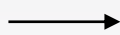
用于查询/推断

2. 表示



具体实现

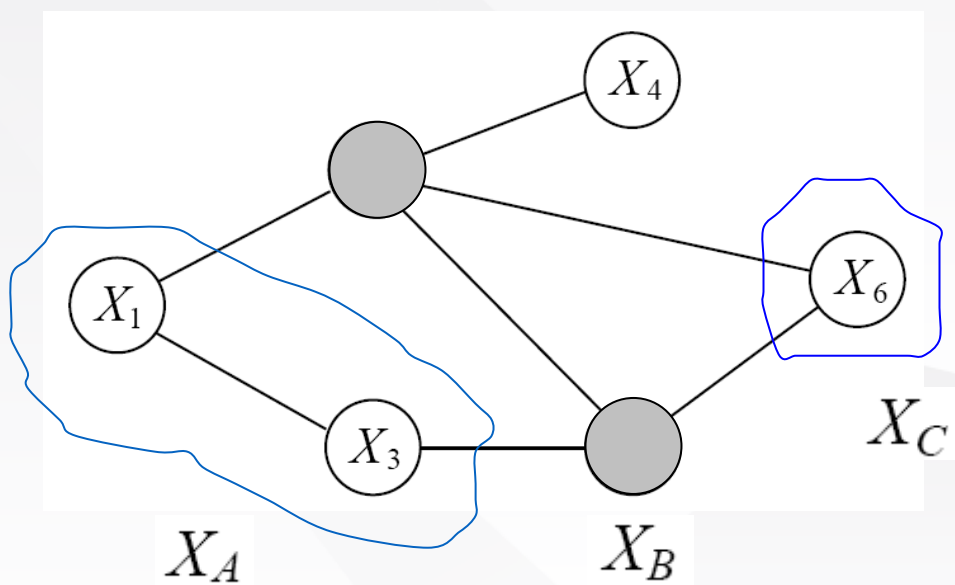
3. 条件独立



模型的解释

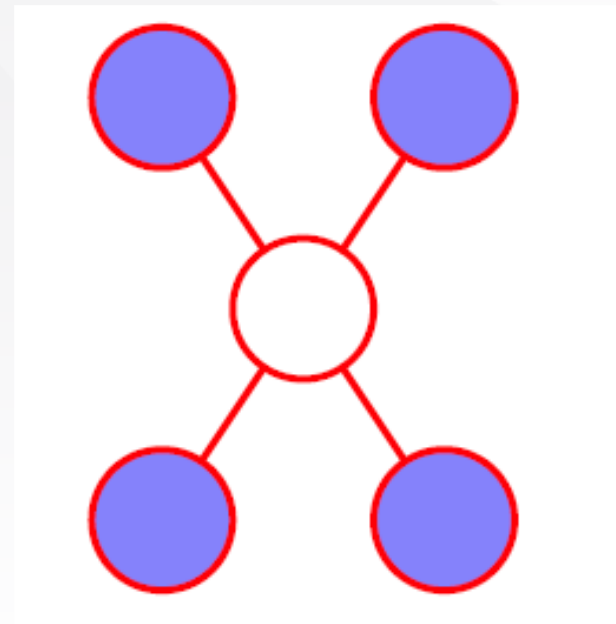
## 3. 条件独立

### 朴素图分割理论



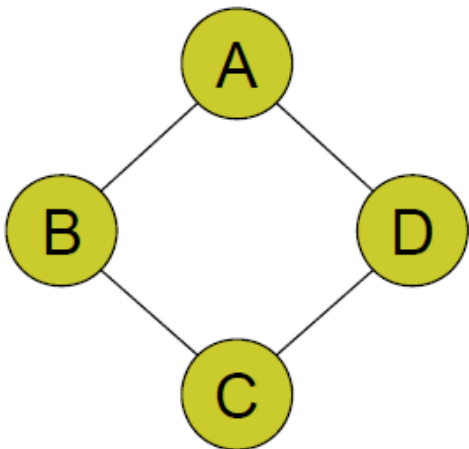
$$X_A \perp X_C | X_B$$

图论中的“可达性”问题.



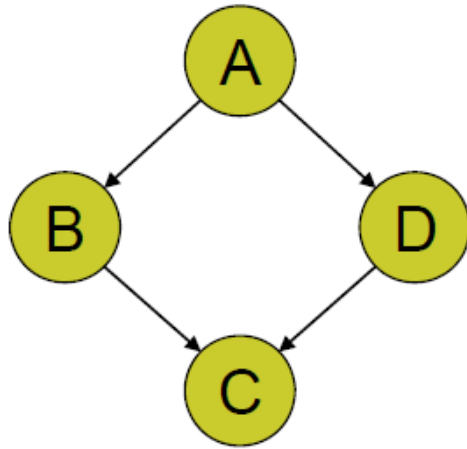
对于一个无向图, 一个节点所有邻居节点构成该节点的马尔科夫包裹(blanket).

是否可以把无向图转化为有向图？



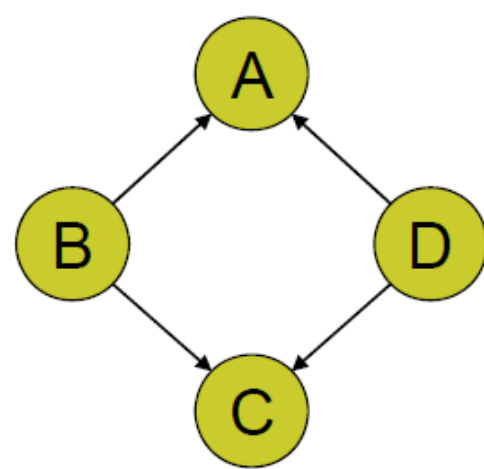
$A \perp C \mid \{B, D\}$

$B \perp D \mid \{A, C\}$



$A \perp C \mid \{B, D\}$

$B \perp D \mid A$  ❌

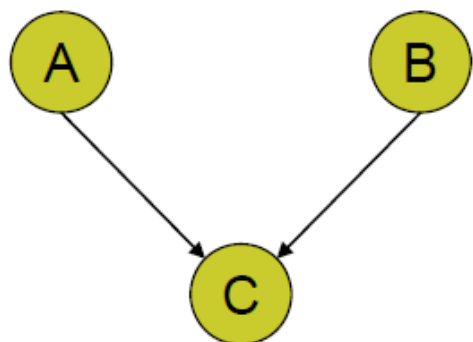


$A \perp C \mid \{B, D\}$

$B \perp D$  ❌

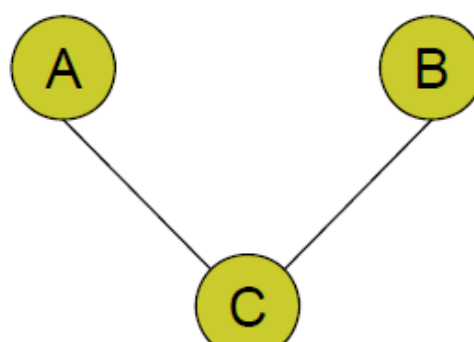


是否可以把有向图转化为无向图？



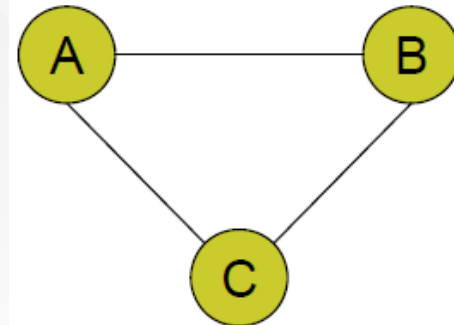
$A \perp B$

$\neg (A \perp B | C)$



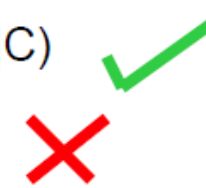
$A \perp B | C$

$\neg (A \perp B)$



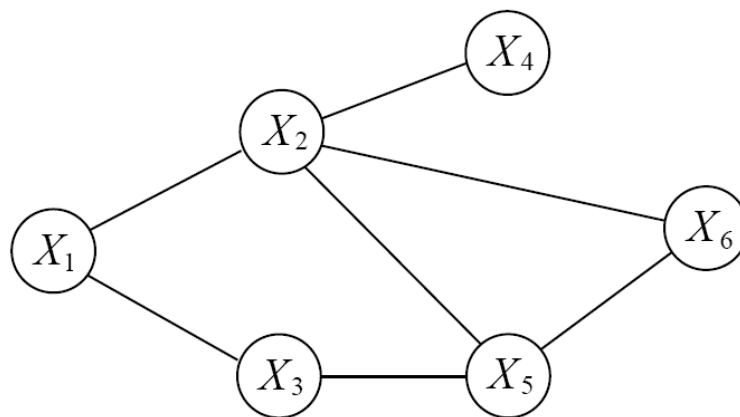
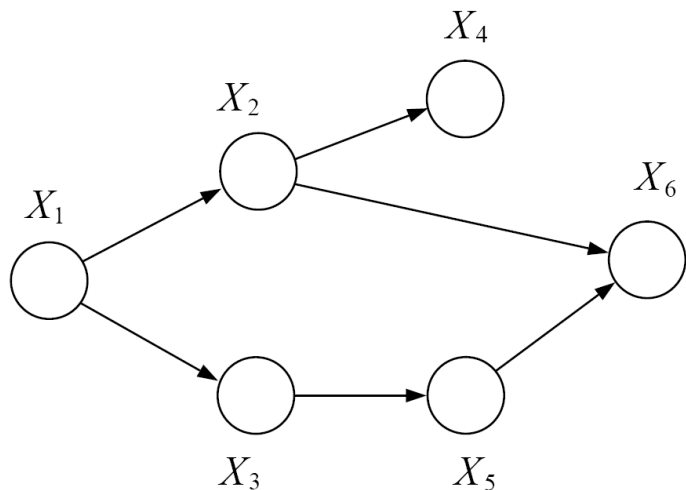
$\neg (A \perp B | C)$

$\neg (A \perp B)$



# ➤ 1. 概率分布

- 有向图: 利用“局部”参数（条件概率）去表示联合概率
- 无向图: 是否也可以用条件概率去表示联合?
  - 一致性问题
- 放弃条件概率
  - 失去局部概率表示
  - 保持独立地任意地选择这些函数的能力
  - 保证所有重要的联合表示可以表示为局部函数的积



# ➤ 1. 概率分布

## ■ 关键问题: 决定局部函数的定义域

- 条件独立: 图分隔

## ■ 团 (Clique)

- 图上的团是一个完全连接的节点子集
- 局部函数 $s$ 不应该被定义在超出团的域上

## ■ 极大团

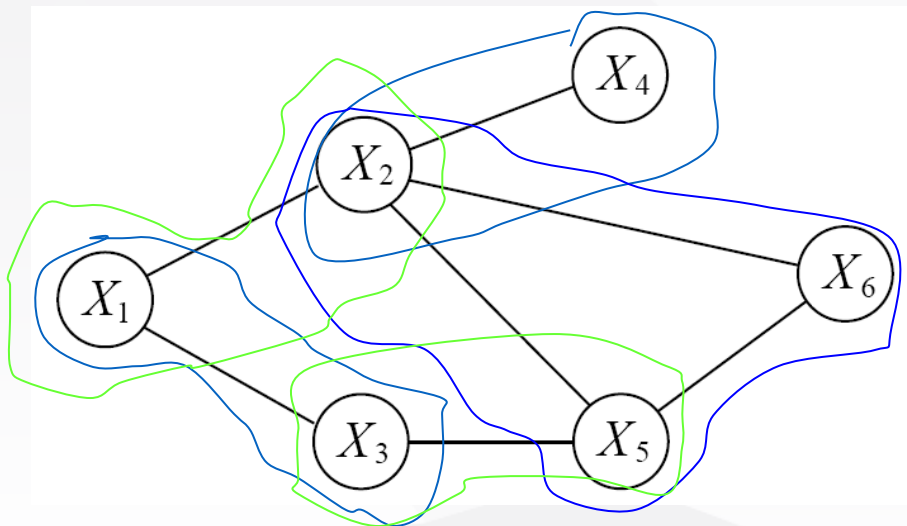
- 图的极大团是指那些没法再增加额外点的团, 否则就会不满足完全连接的性质
- 不失一般性, 我们可以把局部函数定义到极大团上, 因为它包含所有可能的依赖

## ■ 势函数 (局部参数化)

- $\varphi_{x_c}(x_c)$  : 定义在极大团 $x_c$ 上的势函数
- 非负实值函数

# ➤ 1. 概率分布

极大团



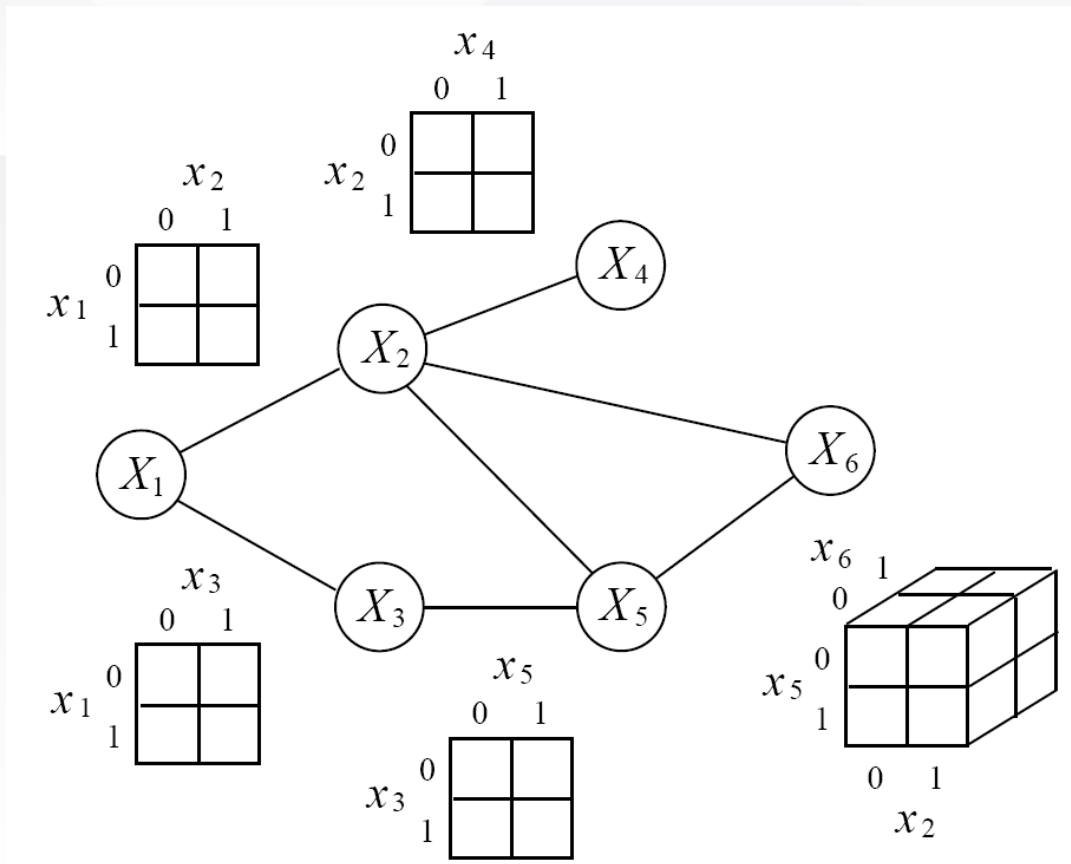
- 联合概率分布

$$P(x) = \frac{1}{Z} \prod_{C \in \mathcal{C}} \varphi_{X_C}(x_C)$$

- 正则化因子

$$Z = \sum_x \prod_{C \in \mathcal{C}} \varphi_{X_C}(x_C)$$

## 2. 表示

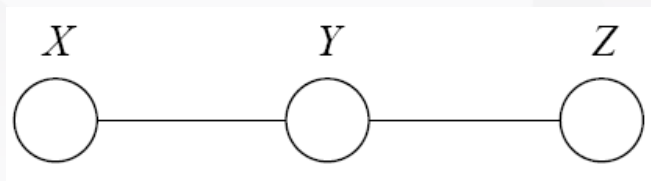


$$O(2^n) \rightarrow O(r \cdot 2^k)$$

## ➤ 势函数的解释 (1)

■ 能否用边际概率 $p(x_c)$ 作为势函数?

$$P(x) = \frac{1}{Z} \prod_{c \in \mathcal{C}} \varphi_{x_c}(x_c)$$



$$P(x, y, z) = P(x, y)P(z)?$$

$$X \perp Z | Y \longrightarrow P(x, y, z) = p(y)p(x|y)P(z|y) \longrightarrow P(x, y, z) \neq P(x, y)P(z)$$

$$P(x, y, z) = P(x, y)P(y, z)? \longrightarrow P(y) = 0 \text{ or } P(y) = 1$$

## 势函数的解释(2)

■ 一般来说, 势函数既不是条件概率也不是边际概率

■ 自然的解释

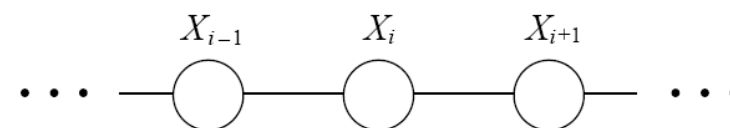
● 一致性, 约束, 或能量

■ 将函数表示为一种无约束形式

$$P(x) = \frac{1}{Z} \prod_c \varphi_{X_c}(X_c) = \frac{1}{Z} \prod_c \exp\{-H_c(X_c)\} = \frac{1}{Z} \exp\left\{-\sum_c H_c(X_c)\right\} = \frac{1}{Z} \exp\{-H(x)\}$$

玻尔兹曼分布

$$Z = \sum_x \prod_c \varphi_{X_c}(X_c) = \sum_x \exp\{-H(x)\}$$



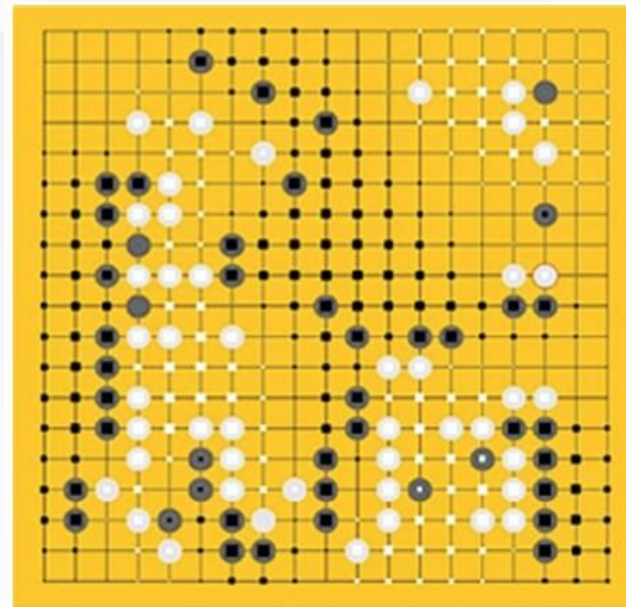
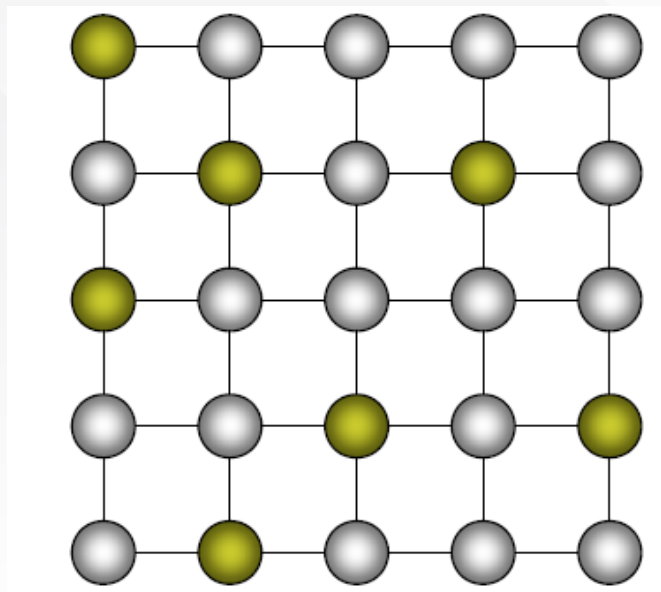
(a)

	$x_i$			$x_{i+1}$	
	-1	1		-1	1
$x_{i-1}$	-1	1.5	0.2	-1	1.5
	1	0.2	1.5	1	0.2

(b)

磁性晶体行为

## ■ 网格模型



■ 在图像处理和晶体物理等领域常用

■ 每个结点表示单个像素,或一个原子

- 由于模式连续性或电磁力等因素, 相邻节点状态耦合.
- 最可能的一组联合配置, 通常对应一个低能量状态



## ➤ 无向图模型小结

### ■ 定义一族概率分布的两种方式:

- U1, 通过枚举所有图上极大团的势函数的可能选择
- U2, 通过声明图 $G$ 上的所有条件独立断言

### ■ Hammersley-Clifford 定理

- U1 和U2 是相同的

对应于图 $G=(V,E)$ 的一个分布具有局部马尔科夫性, 是指如果给定任意一节点的邻居, 该点和其余节点条件独立

**Hammersley-Clifford定理:** 如果分布是严格正并且满足局部马尔科夫性质, 那么它就会像对应的图 $G$ 那样分解

# 大纲

- 概率图模型简介
- 两类概率图模型
  - 有向概率图模型
  - 无向概率图模型
- 学习和推断
- 典型的概率图模型
  - 从HMM到CRF
  - 从PLSA到LDA
- 总结

■ 我们现在有了紧凑的概率分布表示: 概率图模型

■ 概率图 $M$  描述了唯一的概率分布 $P$

■ 典型任务:

- 任务1: 我们如何回答关于  $P_M$  的查询, 例如,  $P_M(X|Y)$  ?

- 我们使用 **推断** 表示计算上述问题答案的过程

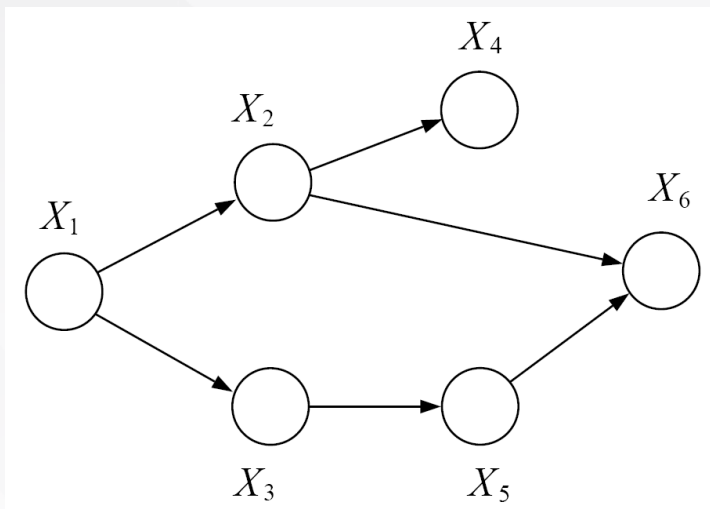
- 任务 2: 我们如何基于数据 $D$ 估计合理的模型  $M$ ?

1. 我们使用**学习**来指代获得  $M$  的点估计过程.

2. 对于**贝叶斯学派**, 他们寻找 $p(M | D)$ , 实际上是一个推断问题.

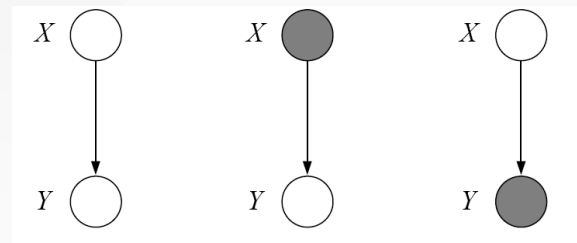
3. 当不是所有的变量都是可观察时, 即使是计算 $M$  的点估计, 也需要进行**推断**处理隐含变量

## ■ 可能的查询:



边际概率 例如.  $P(X_6)$

后验概率 例如.  $P(X_2|X_6 = 1)$



边际概率

$$P(y) = \sum_x P(y|x)P(x)$$

后验概率

$$P(x|y) = \frac{P(y|x)P(x)}{P(y)}$$

## ■ 精确推断:

- 变量消去
- 信念传播
- 较高的计算代价

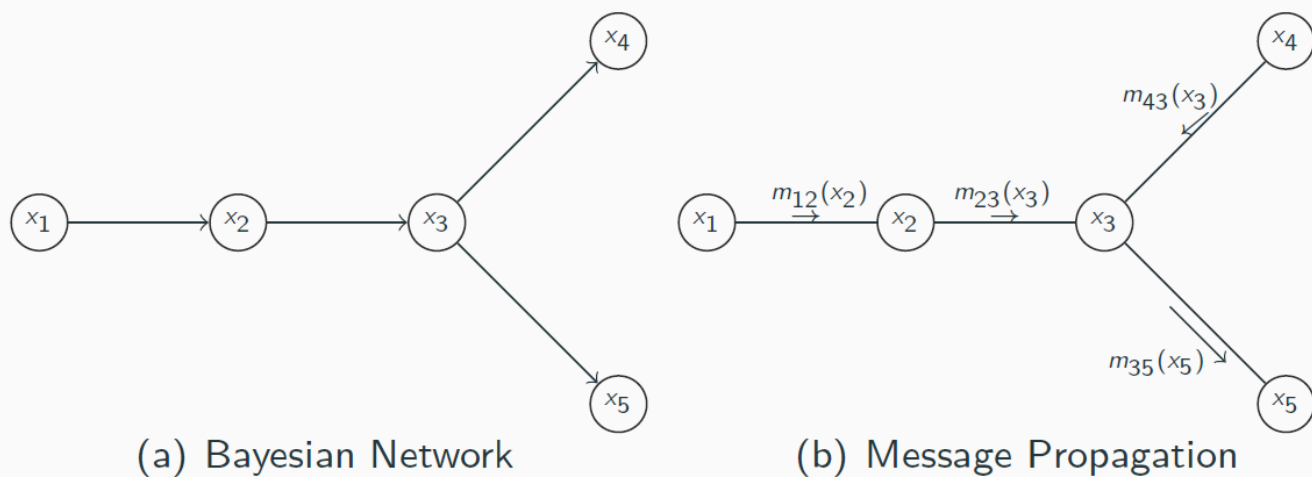
## ■ 近似推断

- 采样
- 变分推断
- 较低的计算复杂度

- 给定  $P(x_1, \dots, x_5)$ , 目标是计算  $P(x_5)$ :

$$P(x_5) = \sum_{x_1, x_2, x_3, x_4} P(x_1)P(x_2|x_1)P(x_3|x_2)P(x_4|x_3)P(x_5|x_3)$$

- 消去其他变量  $\{x_1, \dots, x_4\}$



**Figure 9:** Variables Elimination and Message Propagation

- 利用顺序 $\{x_1, x_2, x_4, x_3\}$  消去其他变量 $\{x_1, \dots, x_4\}$

$$\begin{aligned} P(x_5) &= \sum_{x_1, x_2, x_3, x_4} P(x_1)P(x_2|x_1)P(x_3|x_2)P(x_4|x_3)P(x_5|x_3) \\ &= \sum_{x_3} P(x_5|x_3) \sum_{x_4} P(x_4|x_3) \sum_{x_2} P(x_3|x_2) \sum_{x_1} P(x_1)P(x_2|x_1) \\ &= \sum_{x_3} P(x_5|x_3) \sum_{x_4} P(x_4|x_3) \sum_{x_2} P(x_3|x_2) m_{12}(x_2) \\ &= \sum_{x_3} P(x_5|x_3) \sum_{x_4} P(x_4|x_3) m_{23}(x_3) \\ &= \sum_{x_3} P(x_5|x_3) m_{23}(x_3) m_{43}(x_3) \\ &= m_{35}(x_5) \end{aligned}$$

- Sum-Product算法: 适用于贝叶斯网络和马尔科夫网络

$$m_{i \rightarrow j}(x_j) = \sum_{x_i} \psi(x_i, x_j) \prod_{k \in n(i) \setminus j} m_{ki}(x_i)$$

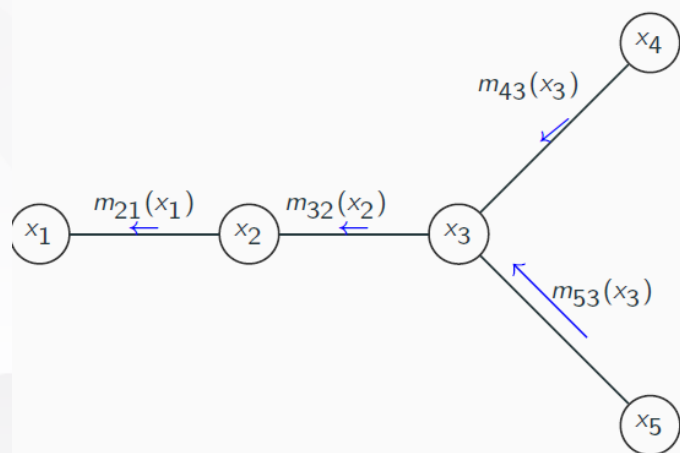
- 计算多个边际概率有很多重复计算

- 信念传播法: 把 $m_{i \rightarrow j}(x_j)$  作为 $x_i$  传递到 $x_j$ 的消息
- 边际分布:

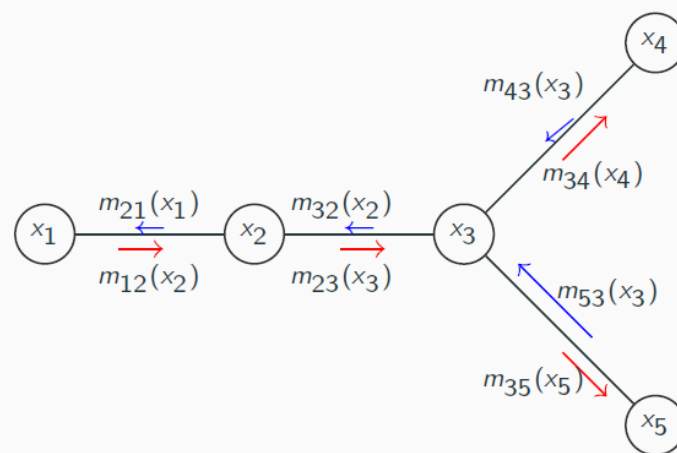
$$P(x_i) \propto \prod_{k \in n(i)} m_{k \rightarrow i}(x_i)$$



1. **叶子到根**: 指派一个根节点, 从叶子节点开始传递信息, 直到根节点接收到所有邻居节点传来的信息
2. **根到叶子**: 从根节点开始传播信息, 直到所有叶子节点接受到信息



(a) message to root node



(b) message from root

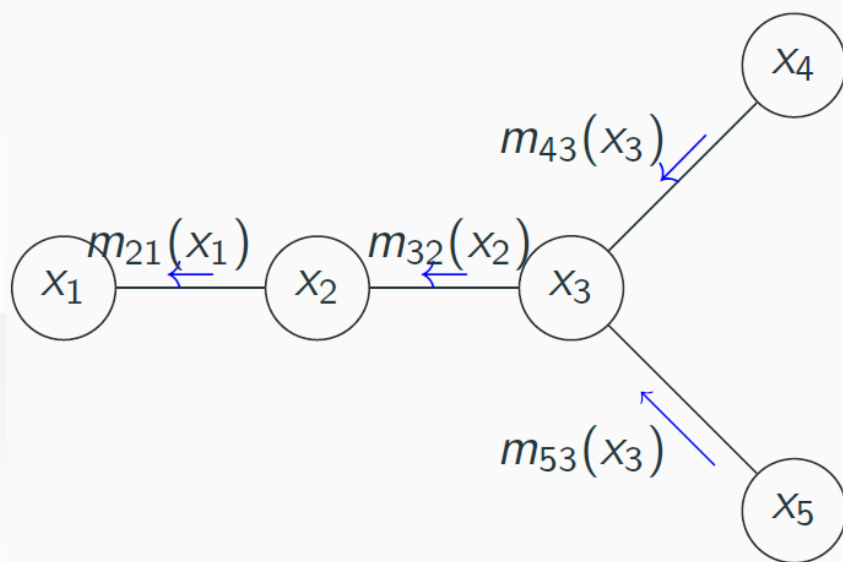
Figure 10: Blief Propagation

## 联合概率

$$P(x_1, x_2, \dots, x_5) = \frac{1}{Z} \psi_{12}(x_1, x_2) \psi_{23}(x_2, x_3) \psi_{34}(x_3, x_4) \psi_{35}(x_3, x_5)$$

$$\text{其中, } Z = \sum_x \psi_{12}(x_1, x_2) \psi_{23}(x_2, x_3) \psi_{34}(x_3, x_4) \psi_{35}(x_3, x_5)$$

## 叶子 → 根:



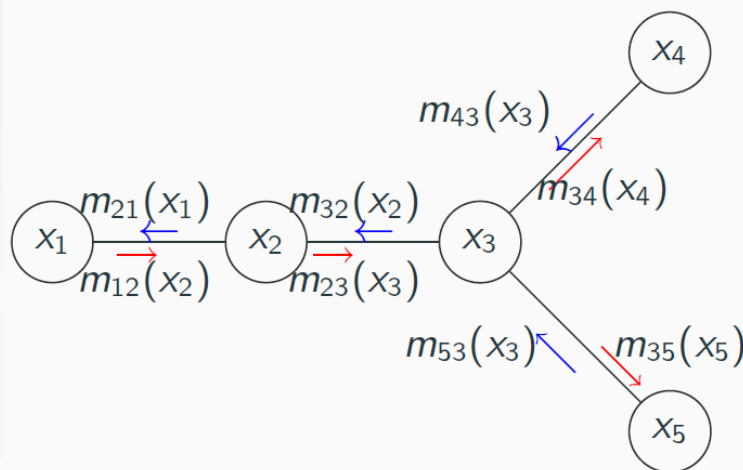
$$m_{43}(x_3) = \sum_{x_4} \psi(x_4, x_3)$$

$$m_{53}(x_3) = \sum_{x_5} \psi(x_5, x_3)$$

$$m_{32}(x_2) = \sum_{x_3} \psi(x_3, x_2) m_{43} m_{53}$$

$$m_{21}(x_1) = \sum_{x_2} \psi(x_2, x_1) m_{32}$$

■ 根 → 叶子：



$$m_{12}(x_2) = \sum_{x_1} \psi(x_1, x_2)$$

$$m_{23}(x_3) = \sum_{x_2} \psi(x_2, x_3) m_{12}(x_2)$$

$$m_{34}(x_4) = \sum_{x_3} \psi(x_3, x_4) m_{23}(x_3) m_{53}(x_3)$$

$$m_{35}(x_5) = \sum_{x_3} \psi(x_3, x_4) m_{23}(x_3) m_{43}(x_3)$$

■ 边际分布

$$P(x_1) \propto m_{21}(x_1) \quad P(x_2) \propto m_{12}(x_2) m_{32}(x_2)$$

$$P(x_3) \propto m_{23}(x_3) m_{43}(x_3) m_{53}(x_3)$$

$$P(x_4) \propto m_{34} \quad P(x_5) \propto m_{35}$$

- 给定结构，并且变量都可以观察到，学习贝叶斯网参数

$$\ell(\theta; D) = \log p(D|\theta) = \log \prod_n \left( \prod_i p(x_{n,i} | x_{n,\pi_i}, \theta_i) \right) = \sum_i \left( \sum_n \log p(x_{n,i} | x_{n,\pi_i}, \theta_i) \right)$$

- MLE: 计数
- MAP: 加上伪计数
- 部分观察: 期望最大化 (EM)

- 对于有向图模型, 对数似然可以分解为多项的和, 每项为一个局部的概率分布族 (节点加上父节点).
- 对于无向图模型, 对数似然无法分解, 因为正则化常数 $Z$ 是所有参数的函数

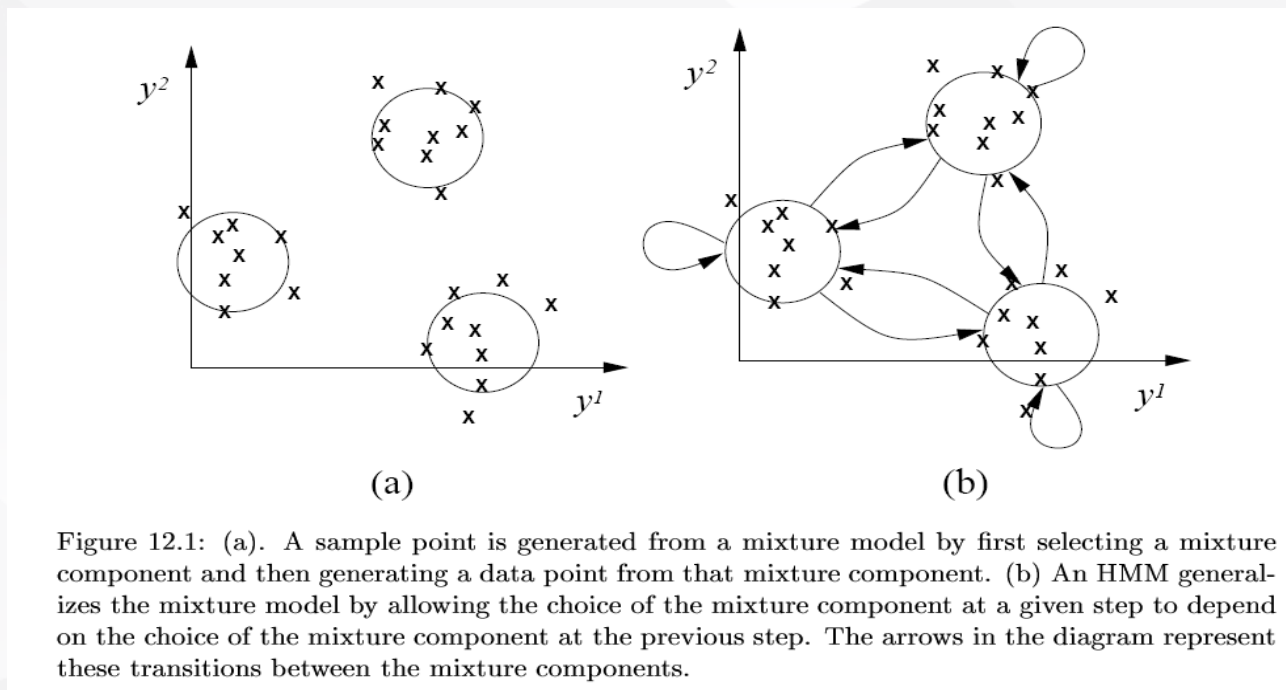
$$P(x_1, \dots, x_n) = \frac{1}{Z} \prod_{c \in C} \psi_c(x_c) \quad Z = \sum_{x_1, \dots, x_n} \prod_{c \in C} \psi_c(x_c)$$

- 一般情况下, 即使变量是完全可观测的, 我们都需要先做推断 (边际化), 以便学习无向图的参数

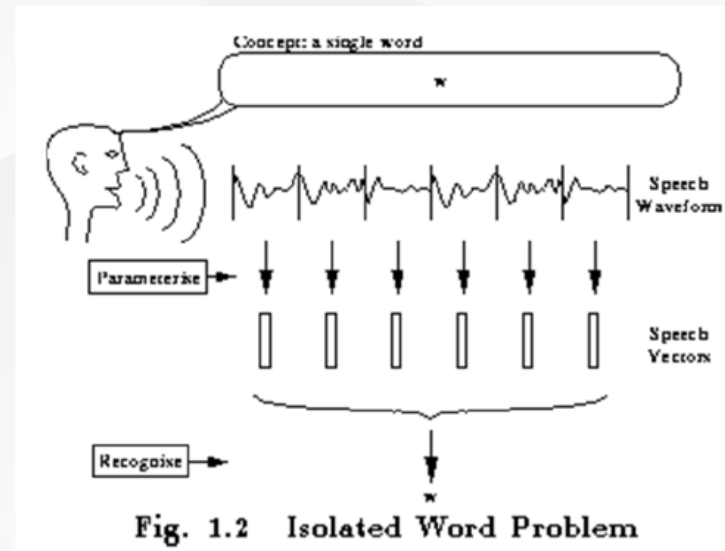
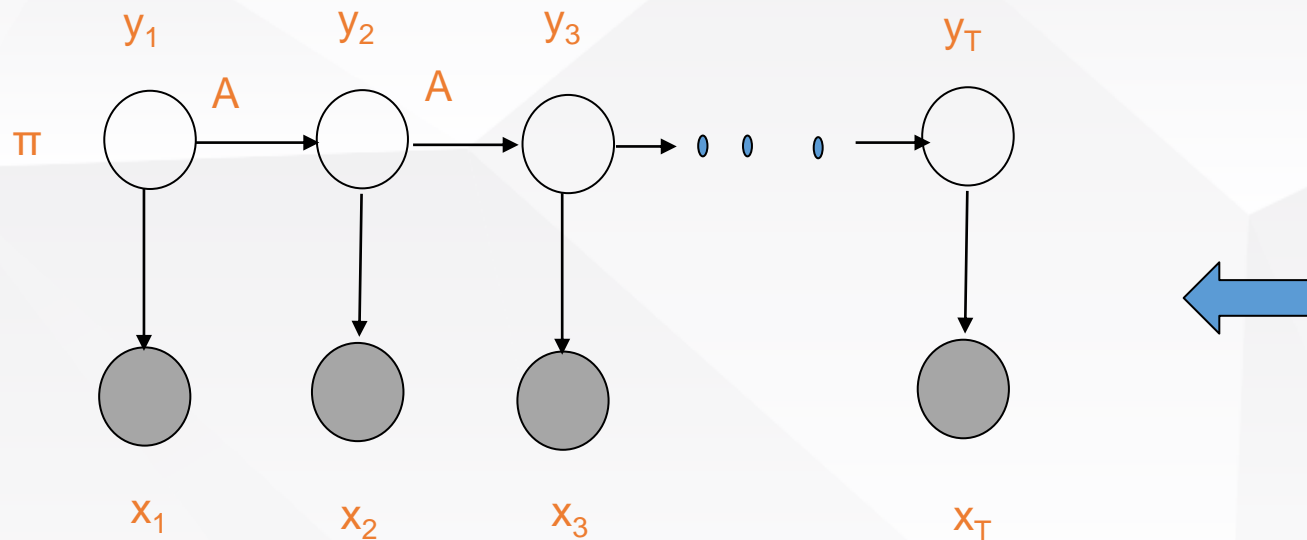
# 大纲

- 概率图模型简介
- 两类概率图模型
  - 有向概率图模型
  - 无向概率图模型
- 学习和推断
- 典型的概率图模型
  - 从HMM到CRF
  - 从PLSA到LDA
- 总结

## ■ 隐马尔科夫模型(Hidden Markov Model ,简称HMM) 是建模序列数据的图模型



## ■ 是混合模型的一种推广

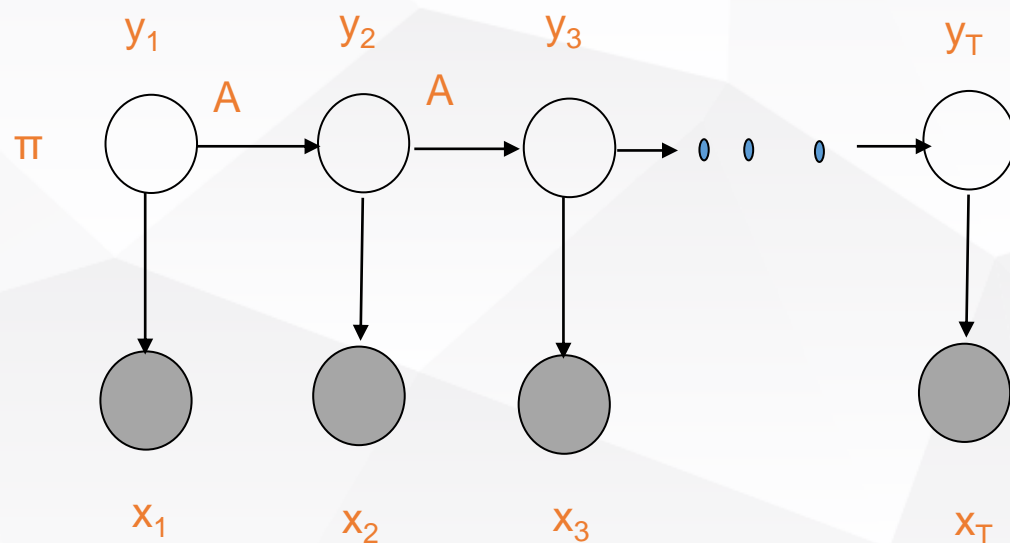


顶层节点表示多项式变量 $y_t$ ，底层节点表示观测变量 $x_t$



## 条件独立

- 给定状态  $y_t$  则  $y_{t-1}$  和  $y_{t+1}$  是独立的.
- 一般来说, 给定状态  $y_t$ , 对于  $s < t$  和  $u > t$ ,  $y_s$  独立于  $y_u$ .
- 当给定状态节点  $y_t$ , 输出节点  $x_s$  和  $x_u$  也相互独立.
- 给定输出节点, 不带来任何条件独立



## ➤ 表示 (参数化)

- 第一个状态节点对应一个非条件分布 $\pi$

$$\pi_i = p(y_1^i = 1)$$

- 状态转移矩阵  $A$  其中  $a_{ij}$  为转移概率

$$p(y_{t+1}^j | y_t^i = 1)$$

- 每个输出节点有一个状态节点作为父节点,因此有发射概率(emission probability )  
 $p(x_t | y_t)$

- 对于特定的参数配置,  $(\mathbf{x}, \mathbf{y}) = (x_1, x_2, \dots, x_T, y_1, y_2, \dots, y_T)$  联合概率可以表示为

$$p(\mathbf{x}, \mathbf{y}) = p(y_1) \prod_{t=1}^{T-1} p(y_{t+1} | y_t) \prod_{t=1}^T p(x_t | y_t)$$

## ➤ 表示 (参数化)

- 为了在联合概率公式引入  $\mathbf{A}$  和  $\boldsymbol{\pi}$ , 我们把转移矩阵索引和初始节点的无条件分布统一写为

$$a_{y_t, y_{t+1}} = \prod_{i,j=1}^M [a_{i,j}]^{y_t^i y_{t+1}^j} \quad \pi_{y_1} \equiv \prod_{i=1}^M [\pi_i]^{y_1^i}$$

- 我们获得联合概率

$$p(\mathbf{x}, \mathbf{y}) = \pi_{y_1} \prod_{t=1}^T a_{y_t y_{t+1}} \prod_{t=1}^T p(x_t | y_t)$$

# 三个基础问题

## 1. 状态序列解码（推断）问题:

- 给定
  - 观察序列  $\mathbf{x}$
  - 模型参数向量  $\theta$
- 寻找
  - 最优的状态序列  $\mathbf{y}$

## 2. 似然评估问题:

- 给定
  - 观察序列  $\mathbf{x}$
  - 模型参数向量  $\theta$
- 寻找
  - 似然函数  $P(\mathbf{x} | \theta)$

## 3. 参数估计问题:

- 给定
  - 观察序列  $\mathbf{x}$
- 寻找
  - $\theta$  的ML 估计使得:  $\theta_{\text{ML}} = \arg \max P(\mathbf{x} | \theta)$

## ➤ 1. 推断

- 一般的推断问题是给定观测序列 $\mathbf{x}$ 计算隐状态序列 $\mathbf{y}$ 的概率
- 给定输出序列 $\mathbf{x}$ ，计算特定隐含状态 $y_t$ 的边际分布
- 给定部分输出，计算条件概率
  - 过滤  $P(y_t|x_1, \dots, x_t)$ : 最后状态
  - 预测  $P(y_t|x_1, \dots, x_s)$  其中  $t > s$ : 未来状态
  - 平滑, 基于已有和未来的数据计算后验概率  $P(y_t|x_1, \dots, x_u)$ , 其中  $t < u$

## ➤ 1. 推断

■ 我们计算 $P(\mathbf{y}|\mathbf{x})$  其中 $\mathbf{x} = (x_1 \cdots x_T)$ 是整个可观测的输出.

■ 
$$P(\mathbf{y}|\mathbf{x}) = \frac{P(\mathbf{x}, \mathbf{y})}{P(\mathbf{x})}$$

■ 但是为了计算 $P(\mathbf{x})$ , 我们需要在所有可能的隐状态值上求和

$$P(\mathbf{x}) = \sum_{y_1} \sum_{y_2} \cdots \sum_{y_T} \pi(y_1) \prod_{t=1}^T a_{y_t, y_{t+1}} \prod_{t=1}^T P(x_t | y_t)$$

● T个隐状态节点, 每个有 M个可能值, 意味着我们需要做 $M^T$ 次求和

# ➤ 1. 推断

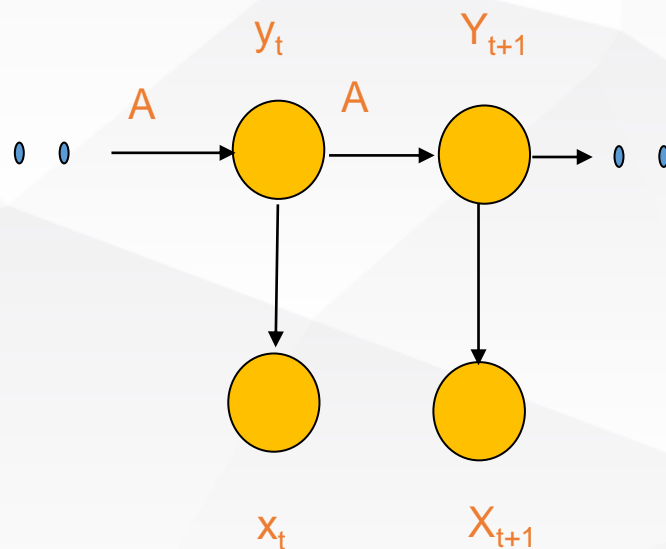
- 其实每个因子只有1、2个状态变量
- 可以把这些求和放到乘法里面
- 把求和移到乘法里并形成递归形式，显著减少计算

$$P(\mathbf{x}) = \sum_{y_1} \sum_{y_2} \cdots \sum_{y_T} \pi(y_1) \prod_{t=1}^T a_{y_t, y_{t+1}} \prod_{t=1}^T P(x_t | y_t)$$

# >> 1. 推断

- 我们先关注一个特定的状态节点  $y_t$  并计算  $P(y_t|\mathbf{x})$
- 我们利用独立性和贝叶斯规则

$$P(y_t|\mathbf{x}) = \frac{P(\mathbf{x}|y_t)P(y_t)}{P(\mathbf{x})} = \frac{P(x_1 \dots x_t|y_t)P(x_{t+1} \dots x_T|y_t)P(y_t)}{P(\mathbf{x})}$$





## ➤ 1. 推断

$$P(y_t|\mathbf{x}) = \frac{P(x_1 \dots x_t, y_t)P(x_{t+1} \dots x_T|y_t)}{P(\mathbf{x})}$$

$$P(y_t|\mathbf{x}) = \gamma(y_t) = \frac{\alpha(y_t)\beta(y_t)}{P(\mathbf{x})}$$

$$P(\mathbf{x}) = \sum_{y_t} \alpha(y_t)\beta(y_t)$$

- 其中 $\alpha(y_t)$ 是产生部分输出序列  $x_1, \dots, x_t$  并结束于  $y_t$  的概率
- 其中 $\beta(y_t)$ 是从  $y_t$  状态开始产生输出序列  $x_{t+1}, \dots, x_T$  的概率

# >> 1. 推断

- 约减到计算 $\alpha, \beta$
- 我们希望能去获得 $\alpha(y_t)$ 和 $\alpha(y_{t+1})$ 的递归关系

$$\alpha(y_{t+1}) = \dots = \sum_{y_t} \alpha(y_t) a_{y_t, y_{t+1}} P(x_{t+1} | y_{t+1})$$

- 算法前向处理，计算复杂度是  $O(M^2T)$
- 初始化: 通过定义 $\alpha$ 在第一个时间步

$$\begin{aligned}\alpha(y_1) &= P(x_1, y_1) \\ &= P(x_1 | y_1) P(y_1) \\ &= P(x_1 | y_1) \pi_{y_1}\end{aligned}$$

## ➤ 1. 推断

- 相似地我们获得后向的递归关系,  $\beta(y_t)$  和  $\beta(y_{t+1})$

$$\beta(y_t) = \sum_{y_{t+1}} \beta(y_{t+1}) a_{y_t, y_{t+1}} P(x_{t+1} | y_{t+1})$$

- 初始化:  $\beta(y_T)$  无需定义

- 如果我们定义  $\beta(y_T)$  为单位向量, 我们可以正确地计算  $\beta(y_{T-1})$

$$\begin{aligned} P(\mathbf{x}) &= \sum_i \alpha(y_T^i) \beta(y_T^i) = \sum_i \alpha(y_T^i) \\ &= \sum_i P(x_1, \dots, x_T, y_T^i) \\ &= P(\mathbf{x}) \end{aligned}$$

- 为了计算所有状态  $y_t$  的后验概率, 我们需要为每一步计算  $\alpha(y_t)$  和  $\beta(y_t)$

## 转移概率的推断

- $\alpha$ - $\beta$  算法可以计算状态的后验概率
  - 可用于估计发射概率的期望充分统计量 (稍后用到)
- 为估计状态转移的后验概率  $P(y_t, y_{t+1} | \mathbf{x})$

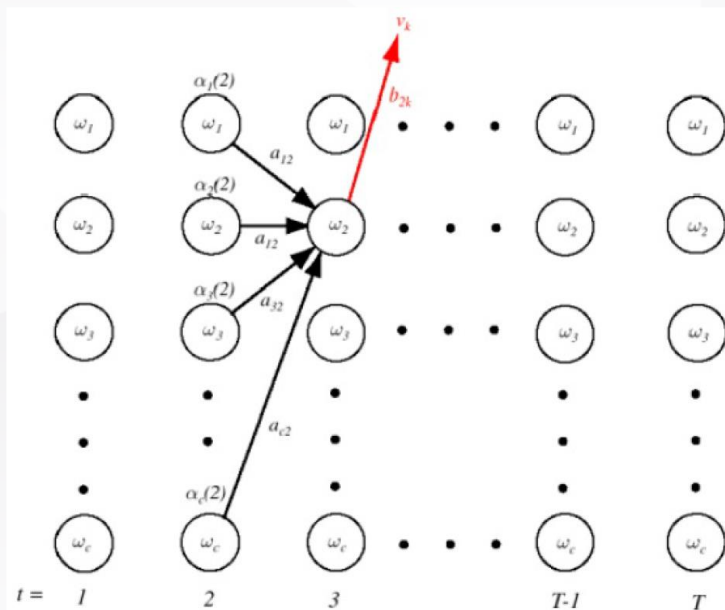
$$\xi(y_t, y_{t+1}) \equiv P(y_t, y_{t+1} | \mathbf{x})$$

- 我们基于  $\alpha$  和  $\beta$  计算  $\xi(y_t, y_{t+1})$

$$\begin{aligned}\xi(y_t, y_{t+1}) &\equiv P(y_t, y_{t+1} | \mathbf{x}) \\ &= \frac{P(\mathbf{x} | y_t, y_{t+1}) P(y_{t+1} | y_t) P(y_t)}{P(\mathbf{x})} \\ &= \frac{P(x_1 \dots x_t | y_t) P(x_{t+1} | y_{t+1}) P(x_{t+2} \dots x_T | y_{t+1}) P(y_{t+1} | y_t) P(y_t)}{P(\mathbf{x})} \\ &= \frac{\alpha(y_t) P(x_{t+1} | y_{t+1}) \beta(y_{t+1}) a_{y_t, y_{t+1}}}{P(\mathbf{x})}\end{aligned}$$

# >> 1. 推断

- 我们可以递归地计算HMM所有的后验概率
- 给定一个观测序列 $\mathbf{x}$ , 我们前向计算 $\alpha$ -递归
- 如果我们需要似然函数, 我们只需简单地求和最终步的 $\alpha$
- 如果我们需要状态的后验概率, 我们在使用  $\beta$ -递归



## ➤ 状态序列解码: 维特比解码

■ 我们现在可以计算

$$P(y_t^k = 1 | \mathbf{x}) = \frac{\alpha(y_t)\beta(y_t)}{P(\mathbf{x})}$$

■ 然后, 我们可以问

- 序列 $\mathbf{x}$ 在位置 $t$ 的最可能的状态 是:

$$k_t^* = \operatorname{argmax} P(y_t^k = 1 | \mathbf{x})$$

- 这是单个隐状态的MAP, 如果我们想要整个序列的最大后验?

- 后验解码:

$$\{y_t^{k_t^*} = 1 : t = 1 \dots T\}$$

- 这是不是整个隐状态序列的MAP?

- 给定  $\mathbf{x} = x_1, \dots, x_T$ , 我们要找  $\mathbf{y} = y_1, \dots, y_T$ , 使得  $P(\mathbf{y}|\mathbf{x})$  最大

$$\mathbf{y}^* = \underset{\mathbf{y}}{\operatorname{argmax}} P(\mathbf{y}|\mathbf{x}) = \underset{\mathbf{y}}{\operatorname{argmax}} P(\mathbf{x}, \mathbf{y})$$

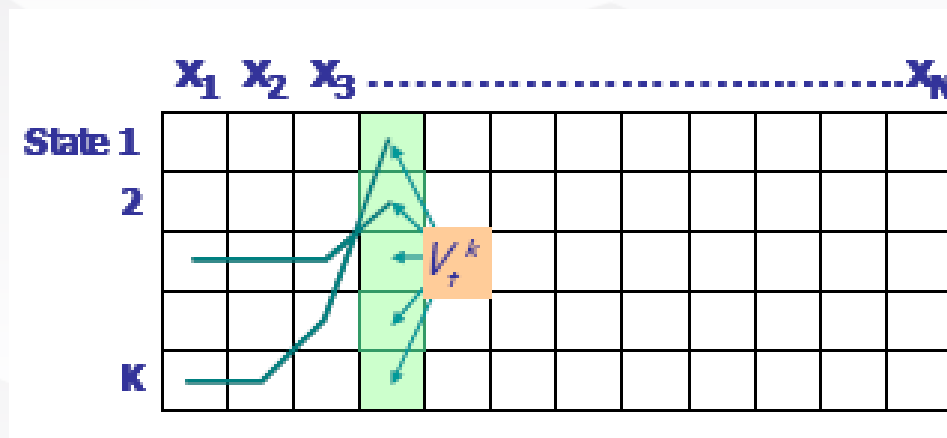
- 令

$$V_t^k = \max_{\{y_1, \dots, y_{t-1}\}} P(x_1, \dots, x_{t-1}, y_1, \dots, y_{t-1}, x_t, y_t^k = 1)$$

= 结尾状态为  $y_t=k$  时, 最可能状态序列的概率

- 递归:

$$V_t^k = P(x_t|y_t^k = 1) \max_i a_{i,k} V_{t-1}^i$$



## ➤ 2. 学习: 参数估计

- HMM的参数是转移矩阵 $A$ , 初始概率分布 $\pi$  和输出分布的参数

- **监督学习**: 当“正确的答案”已知情况下的估计

- 例子:

- 给定:基因组区域 $x = x_1 \dots x_{1,000,000}$ , 我们有针对 CpG islands的实验标记
- 给定:赌场玩家允许我们一晚上观察他, 包括他如何改变骰子投出10000次结果

- **无监督学习**: 当“正确答案”未知情况下的估计

- 例子:

- 给定:基因组区域; 我们不知道CpG islands的频率, 也不知道他们的组合
- 给定: 玩家掷了10,000次骰子, 但我们没看到他什么时候换骰子

- 更新模型的参数 $\theta$  以最大化  $P(x | \theta)$  ---极大似然(ML)估计



## ➤ 2. 学习: 参数估计

■ 令  $\theta = (\pi, A, \eta)$  代表HMM模型所有的参数, 其中  $p(x_t|y_t, \eta)$  是输出分布

■ 对数似然:

$$p(x|\theta) = \log \sum_{y_1} \sum_{y_2} \dots \sum_{y_T} \pi(y_1) \prod_{t=1}^{T-1} a_{y_t, y_{t+1}} \prod_{t=1}^T p(x_t|y_t, \eta)$$

■ 我们的目标是针对 $\theta$  最大化表达

■ EM 算法用于估计HMM参数

## 2. 学习: EM 算法

■ 具体推导, 假设  $x_t$  是多项式变量

$$p(x_t|y_t, \eta) = \prod_{i=1}^M \prod_{j=1}^L [\eta_{ij}]^{y_t^i x_t^j}$$

■ **M**步完全数据的对数似然函数 (以及**E**步需要的充分统计量)

$$\begin{aligned} \log p(x, y) &= \log \left\{ \pi_{y_1} \prod_{t=1}^{T-1} a_{y, y_{t+1}} \prod_{t=1}^T p(x_t|y_t, \eta) \right\} = \log \left\{ \prod_{i=1}^M [\pi_i]^{y_1^i} \prod_{t=1}^{T-1} \prod_{i,j=1}^M [a_{ij}]^{y_t^i y_{t+1}^i} \prod_{t=1}^T \prod_{i=1}^M \prod_{j=1}^L [\eta_{ij}]^{y_t^i x_t^j} \right\} \\ &= \sum_{i=1}^M y_1^i \log \pi_i + \sum_{t=1}^{T-1} \sum_{i,j=1}^M y_t^i y_{t+1}^i \log a_{ij} + \sum_{t=1}^T \sum_{i=1}^M \sum_{j=1}^L y_t^i x_t^j \log \eta_{ij} \end{aligned}$$

■ 从表达式中, 我们看到  $m_{ij} \triangleq \sum_{t=1}^{T-1} y_t^i y_{t+1}^j$  是  $a_{ij}$  的充分统计量

$n_{ij} \triangleq \sum_{t=1}^{T-1} y_t^i x_t^j$  是  $\eta_{ij}$  的充分统计,  $y_1^i$  是  $\pi_i$  的充分统计量

## ➤ 2. 学习: EM 算法

■ 最大似然估计对于完整数据如下:

$$\hat{\alpha}_{ij} = \frac{m_{ij}}{\sum_{k=1}^M m_{ik}}$$

$$\hat{\eta}_{ij} = \frac{n_{ij}}{\sum_{k=1}^N n_{ik}}$$

$$\hat{\pi}_i = y_1^i$$

■ 这些估计都有自然的解释

## ➤ 2. 学习: EM 算法

■ 我们回到EM算法的 **E 步**

■ 考虑第一个充分统计量的期望  $n_{ij} = \sum_{t=1}^T y_t^i x_t^j$

$$\begin{aligned} E(n_{ij} | \mathbf{x}, \theta^{(p)}) &= \sum_{t=1}^T E(y_t^i | \mathbf{x}, \theta^{(p)}) x_t^j \\ &= \sum_{t=1}^T p(y_t^i = 1 | \mathbf{x}, \theta^{(p)}) x_t^j \\ &\triangleq \sum_{t=1}^T \gamma_t^i x_t^j \end{aligned}$$

■ 相似地, 充分统计量  $m_{ij} \triangleq \sum_{t=1}^{T-1} y_t^i y_{t+1}^j$

$$E(m_{ij} | \mathbf{x}, \theta^{(p)}) = \sum_{t=1}^{T-1} E(y_t^i y_{t+1}^j | \mathbf{x}, \theta^{(p)}) = \sum_{t=1}^{T-1} p(y_t^i y_{t+1}^j | \mathbf{x}, \theta^{(p)}) \triangleq \sum_{t=1}^{T-1} \xi_{t,t+1}^{ij}$$

■ 总之, 充分统计量可以通过前向后向的递归过程计算

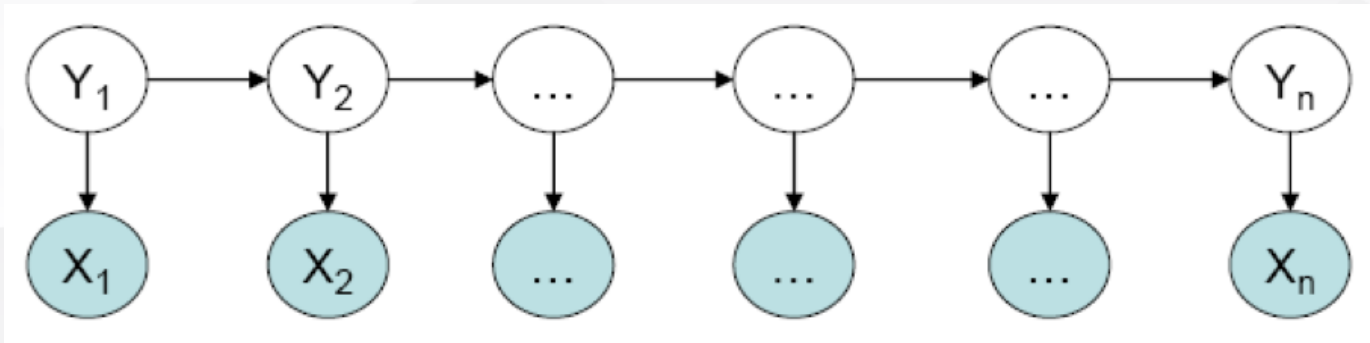
## ➤ 2. 学习: EM 算法

- 已经有了估计的充分统计量, 我们代入到极大似然公式中, 获得M步 (also known, HMM中也称为 “Baum-Welch updates”)

$$\begin{aligned}\hat{\eta}_{ij}^{(p+1)} &= \frac{\sum_{t=1}^T \gamma_t^i x_t^j}{\sum_{k=1}^N \sum_{t=1}^T \gamma_t^i x_t^k} = \frac{\sum_{t=1}^T \gamma_t^i x_t^j}{\sum_{t=1}^T \gamma_t^i} \\ \hat{a}_{ij}^{(p+1)} &= \frac{\sum_{t=1}^{T-1} \xi_{t,t+1}^{i,j}}{\sum_{k=1}^M \sum_{t=1}^{T-1} \xi_{t,t+1}^{i,k}} = \frac{\sum_{t=1}^{T-1} \xi_{t,t+1}^{i,j}}{\sum_{t=1}^{T-1} r_t^i} \\ \hat{\pi}_i^{(p+1)} &= \gamma_1^i\end{aligned}$$

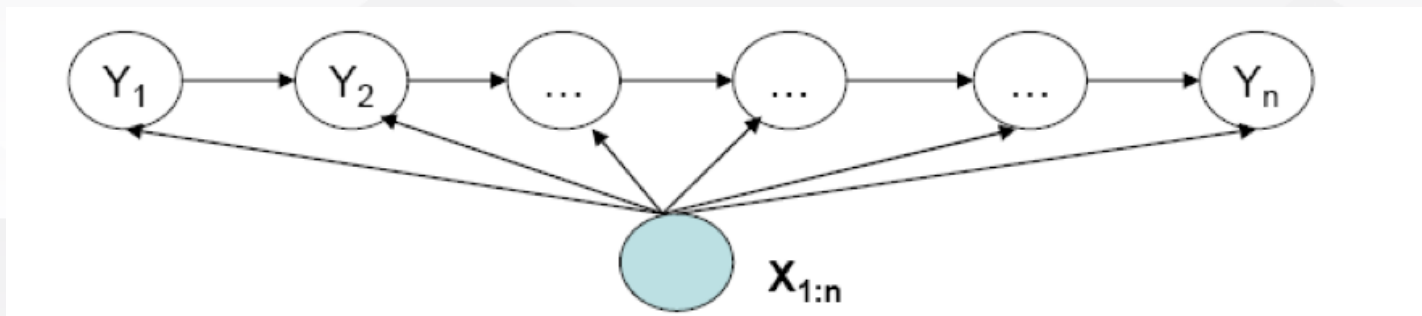
- EM 算法迭代地进行更新 (M 步) 和利用更新值计算前向-后向过程 (E步).

## >> HMM的缺点



- HMM模型仅捕捉了状态之间和状态及其对应输出之间的关系
  - NLP 例子：在一个句子分割的任务中, 每一个分割状态可能不仅依赖单个词 (和临近的分割状态), 还依赖于(非局部)特征, 如整个长度和缩进, 空格的数量等
- 学习目标和预测目标不匹配
  - HMM学习一个状态和观察的联合概率分布  $P(X,Y)$ , 但是在预测任务中, 我们仅需要条件概率  $P(Y|X)$

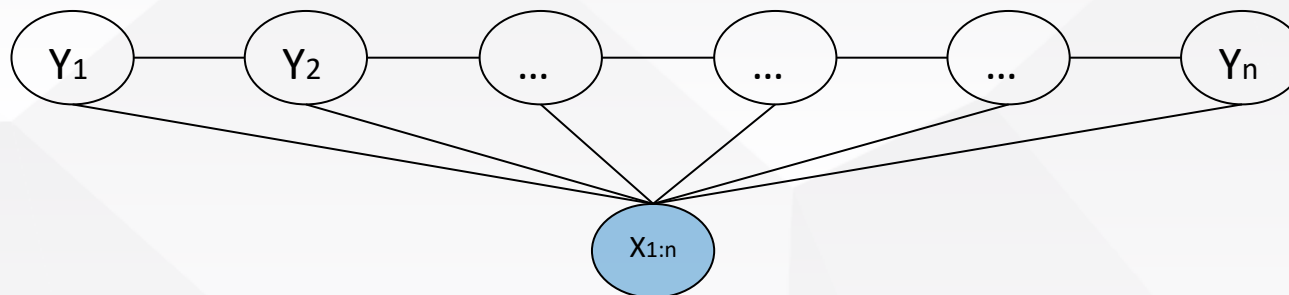
## ➤ 一种方案: 最大熵Markov模型 (MEMM)



$$P(y_{1:n}|x_{1:n}) = \prod_{i=1}^n P(y_i|y_{i-1}, X_{1:n}) = \prod_{i=1}^n \frac{\exp(w^T f(y_i, y_{i-1}, X_{1:n}))}{Z(y_{i-1}, X_{1:n})}$$

- 建模每个隐状态和整个观测序列的依赖关系
  - 比HMM有更强的表达力
- 判别式模型 (Discriminative model)
  - 完全忽略对 $P(X)$ 的建模: 节约了建模的成本
  - 学习目标与预测目标一致:  $P(Y|X)$
- 不足之处: 标记偏差问题 (偏好于转移状态少的状态)

## ➤ 从HMM 到条件随机场 (CRF)



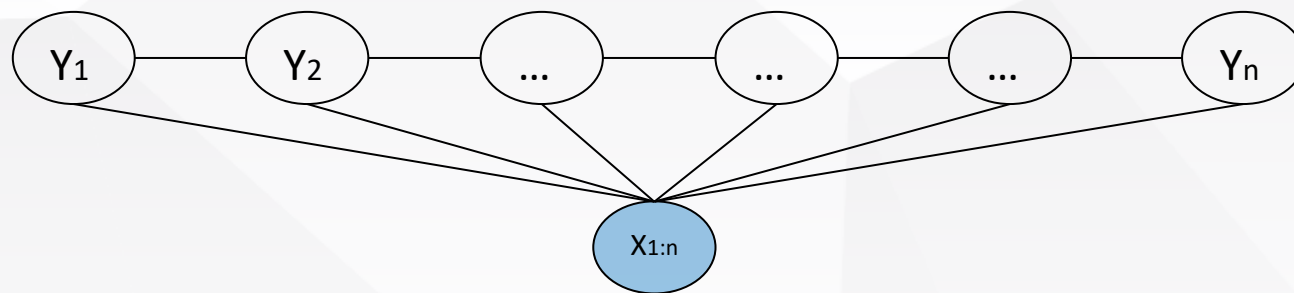
$$P(y_{1:n}|x_{1:n}) = \frac{1}{Z(x_{1:n})} \prod_{i=1}^n \phi(y_i, y_{i-1}, x_{1:n}) = \frac{1}{Z(x_{1:n})} \prod_{i=1}^n \exp(w^T f(y_i, y_{i-1}, x_{1:n}))$$

### ■ CRF 是无向图模型

- 它也是一个判别式模型
- 建模了每个状态和整个观测序列的依赖
- 和MEMM的最大差别，是建模状态序列的全局依赖
- CRF全局归一，MEMM局部归一



## ■ 一般的参数化形式:



$$P(y|x) = \frac{1}{Z(x, \lambda, \mu)} \exp(\sum_{i=1}^n (\sum_k \lambda_k f_k(y_i, y_{i-1}, x) + \sum_l \mu_l g_l(y_i, x)))$$

$$= \frac{1}{Z(x, \lambda, \mu)} \exp(\sum_{i=1}^n (\lambda^T \mathbf{f}(y_i, y_{i-1}, x) + \mu^T \mathbf{g}(y_i, x)))$$

$$\text{where } Z(x, \lambda, \mu) = \sum_y \exp(\sum_{i=1}^n (\lambda^T \mathbf{f}(y_i, y_{i-1}, x) + \mu^T \mathbf{g}(y_i, x)))$$

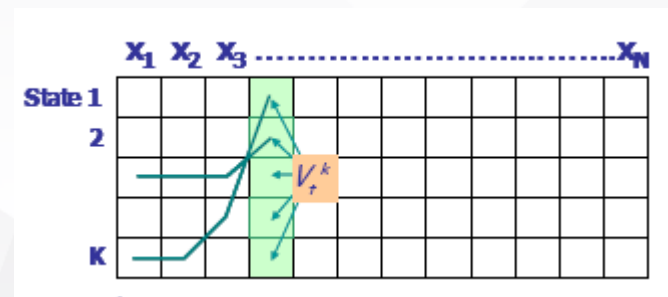
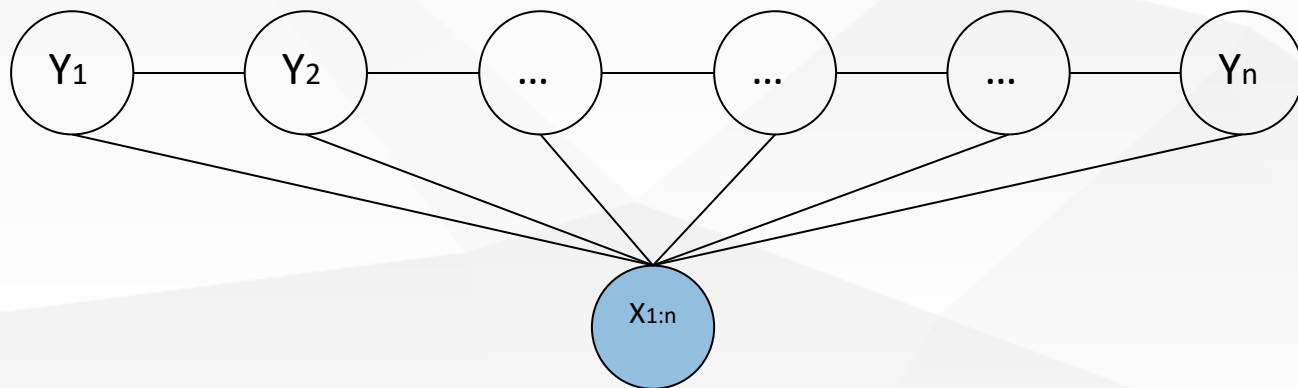
# >> 1. 推断

■ 给定 CRF 参数  $\lambda$  和  $\mu$ , 计算  $\mathbf{y}^*$  最大化  $P(\mathbf{y}|\mathbf{x})$

$$\mathbf{y}^* = \arg \max_{\mathbf{y}} \exp \left( \sum_{i=1}^n (\lambda^T \mathbf{f}(y_i, y_{i-1}, x) + \mu^T \mathbf{g}(y_i, x)) \right)$$

● 可以忽略  $Z(\mathbf{x})$  它不是  $\mathbf{y}$  的函数

■ 在 CRF 上使用 max-product 算法:



和HMM中维特比解码相同

## 2. 学习

给定 $\{(x_d, y_d)\}_{d=1}^N$ , 寻找 $\lambda^*, \mu^*$  使得

$$\begin{aligned}\lambda^*, \mu^* &= \arg \max_{\lambda, \mu} L(\lambda, \mu) = \arg \max_{\lambda, \mu} \log \prod_{d=1}^N P(y_d | x_d, \lambda, \mu) \\ &= \arg \max_{\lambda, \mu} \log \prod_{d=1}^N \frac{1}{Z(x_d, \lambda, \mu)} \exp \left( \sum_{i=1}^n \left( \lambda^T \mathbf{f}(y_{d,i}, y_{d,i-1}, x_d) + \mu^T \mathbf{g}(y_{d,i}, x_d) \right) \right) \\ &= \arg \max_{\lambda, \mu} \sum_{d=1}^N \left( \sum_{i=1}^n \left( \lambda^T \mathbf{f}(y_{d,i}, y_{d,i-1}, x_d) + \mu^T \mathbf{g}(y_{d,i}, x_d) \right) - \log Z(x_d, \lambda, \mu) \right)\end{aligned}$$

针对  $\lambda$  计算梯度:

$$\nabla_{\lambda} L(\lambda, \mu) = \sum_{d=1}^N \left( \sum_{i=1}^n f(y_{d,i}, y_{d,i-1}, x_d) - \sum_y \left( P(y | x_d) \sum_{i=1}^n f(y_{d,i}, y_{d,i-1}, x_d) \right) \right)$$

在指数分布族中, 对数分割函数(log-partition function)的梯度是充分统计量的期望

$$\nabla_{\lambda} L(\lambda, \mu) = \sum_{d=1}^N \left( \sum_{i=1}^n f(y_{d,i}, y_{d,i-1}, x_d) - \sum_y (P(y|x_d) \sum_{i=1}^n f(y_i, y_{i-1}, x_d)) \right)$$

■ 计算模型期望:

● 需要指数量级求和: 可行吗?

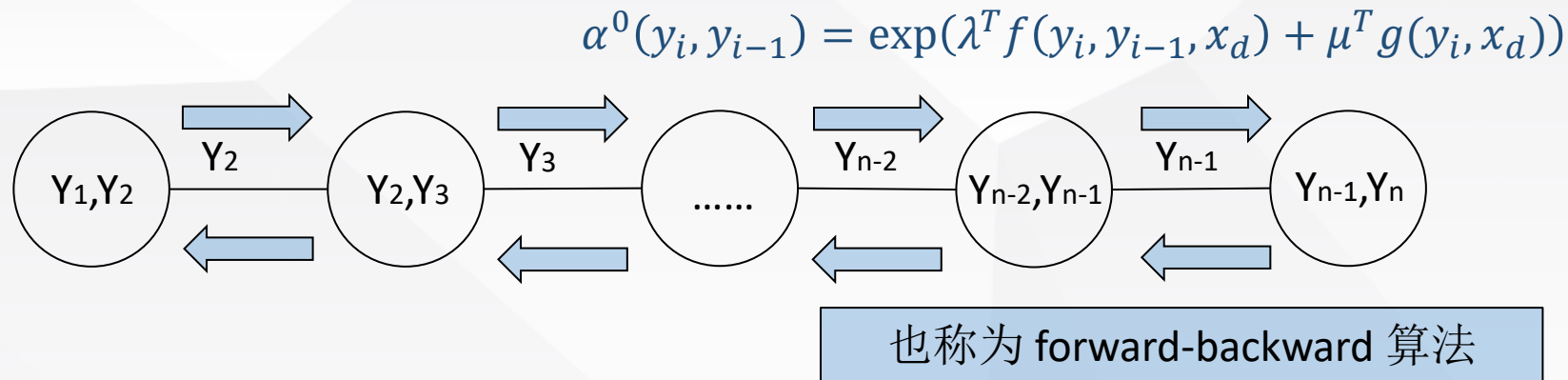
$$\begin{aligned} \sum_y \left( P(y|x_d) \sum_{i=1}^n f(y_i, y_{i-1}, x_d) \right) &= \sum_{i=1}^n \left( \sum_y f(y_i, y_{i-1}, x_d) P(y|x_d) \right) \\ &= \sum_{i=1}^n \sum_{y_i, y_{i-1}} f(y_i, y_{i-1}, x_d) P(y_i, y_{i-1}|x_d) \end{aligned}$$

基于相邻节点边际概率分布的 **f** 函数的期望!!

■ 可行!

● 可以用sum-product算法计算边际分布

### ■ 初始化:



### ■ 校准后的后验概率:

$$P(y_i, y_{i-1} | x_d) \propto \alpha(y_i, y_{i-1})$$
$$\Rightarrow P(y_i, y_{i-1} | x_d) = \frac{\alpha(y_i, y_{i-1})}{\sum_{y_i, y_{i-1}} \alpha(y_i, y_{i-1})} = \alpha'(y_i, y_{i-1})$$

## 2. 学习

### ■ 计算特征期望

$$\sum_{y_i, y_{i-1}} f(y_i, y_{i-1}, x_d) P(y_i, y_{i-1} | x_d) = \sum_{y_i, y_{i-1}} f(y_i, y_{i-1}, x_d) \alpha'(y_i, y_{i-1})$$

### ■ 计算 $\nabla_{\lambda} L(\lambda, \mu)$ :

$$\begin{aligned} \nabla_{\lambda} L(\lambda, \mu) &= \sum_{d=1}^N \left( \sum_{i=1}^n f(y_{d,i}, y_{d,i-1}, x_d) - \sum_y (P(y|x_d) \sum_{i=1}^n f(y_{d,i}, y_{d,i-1}, x_d)) \right) = \\ &\sum_{d=1}^n \left( \sum_{i=1}^n (f(y_{d,i}, y_{d,i-1}, x_d) - \sum_{y_i, y_{i-1}} \alpha'(y_i, y_{i-1}) f(y_{d,i}, y_{d,i-1}, x_d)) \right) \end{aligned}$$

### ■ 学习可以通过梯度上升进行:

$$\begin{aligned} \lambda^{t+1} &= \lambda^t + \eta \nabla_{\lambda} L(\lambda^t, \mu^t) \\ \mu^{t+1} &= \mu^t + \eta \nabla_{\mu} L(\lambda^t, \mu^t) \end{aligned}$$

■ 实际上, 梯度上升收敛非常慢

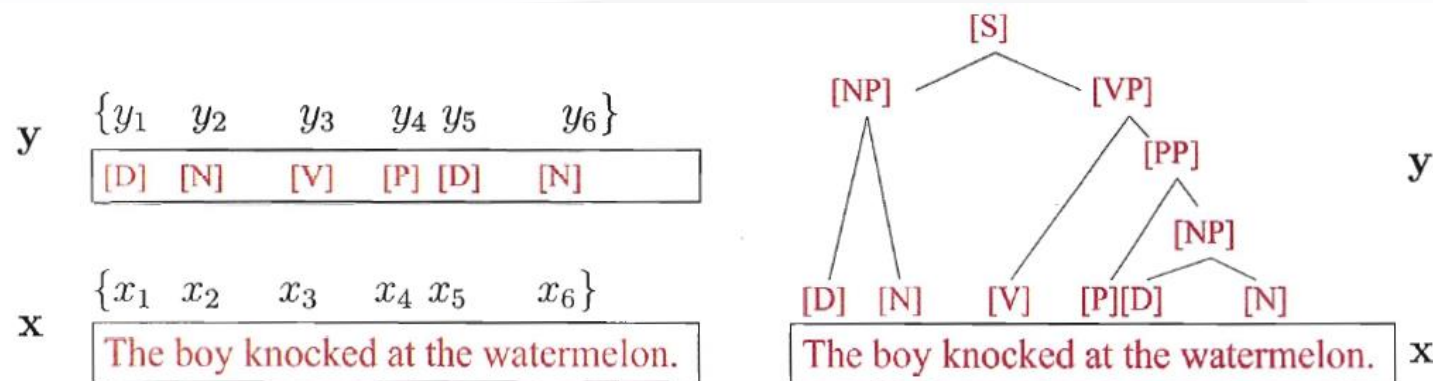
- 替代选择:

- ◆ 共轭梯度方法

- ◆ 内存受限拟牛顿法

# 条件随机场: 一些经验结果

## 词性标注



(a) 词性标注

(b) 语法分析

自然语言处理中的词性标注和语法分析任务

<i>model</i>	<i>error</i>	<i>oov error</i>
HMM	5.69%	45.99%
MEMM	6.37%	54.61%
CRF	5.55%	48.05%
MEMM <sup>+</sup>	4.81%	26.99%
CRF <sup>+</sup>	4.27%	23.76%

<sup>+</sup>Using spelling features

■ 使用相同的特征集合: HMM  $\geq$  CRF  $>$  MEMM

■ 使用额外的重叠特征 CRF<sup>+</sup>  $>$  MEMM<sup>+</sup>  $>$  HMM



---

## Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data

---

John Lafferty<sup>†\*</sup>  
Andrew McCallum<sup>\*†</sup>  
Fernando Pereira<sup>\*‡</sup>

LAFFERTY@CS.CMU.EDU  
MCCALLUM@WHIZBANG.COM  
FPEREIRA@WHIZBANG.COM

\*WhizBang! Labs—Research, 4616 Henry Street, Pittsburgh, PA 15213 USA

<sup>†</sup>School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213 USA

<sup>‡</sup>Department of Computer and Information Science, University of Pennsylvania, Philadelphia, PA 19104 USA

### Abstract

We present *conditional random fields*, a framework for building probabilistic models to segment and label sequence data. Conditional random fields offer several advantages over hidden Markov models and stochastic grammars for such tasks, including the ability to relax strong independence assumptions made in those models. Conditional random fields also avoid a fundamental limitation of maximum entropy Markov models (MEMMs) and other discriminative Markov models based on directed graphical models, which can be biased towards states with few successor states. We present iterative

mize the joint likelihood of training examples. To define a joint probability over observation and label sequences, a generative model needs to enumerate all possible observation sequences, typically requiring a representation in which observations are task-appropriate atomic entities, such as words or nucleotides. In particular, it is not practical to represent multiple interacting features or long-range dependencies of the observations, since the inference problem for such models is intractable.

This difficulty is one of the main motivations for looking at conditional models as an alternative. A conditional model specifies the probabilities of possible label sequences given an observation sequence. Therefore, it does not expend modeling effort on the observations, which at test time

# 大纲

- 概率图模型简介
- 两类概率图模型
  - 有向概率图模型
  - 无向概率图模型
- 学习和推断
- 典型的概率图模型
  - 从HMM到CRF
  - 从PLSA到LDA
- 总结

## ■ 隐语义索引LSI (latent semantic indexing) 发现词语背后的含义；发现文档的主题

### Titles:

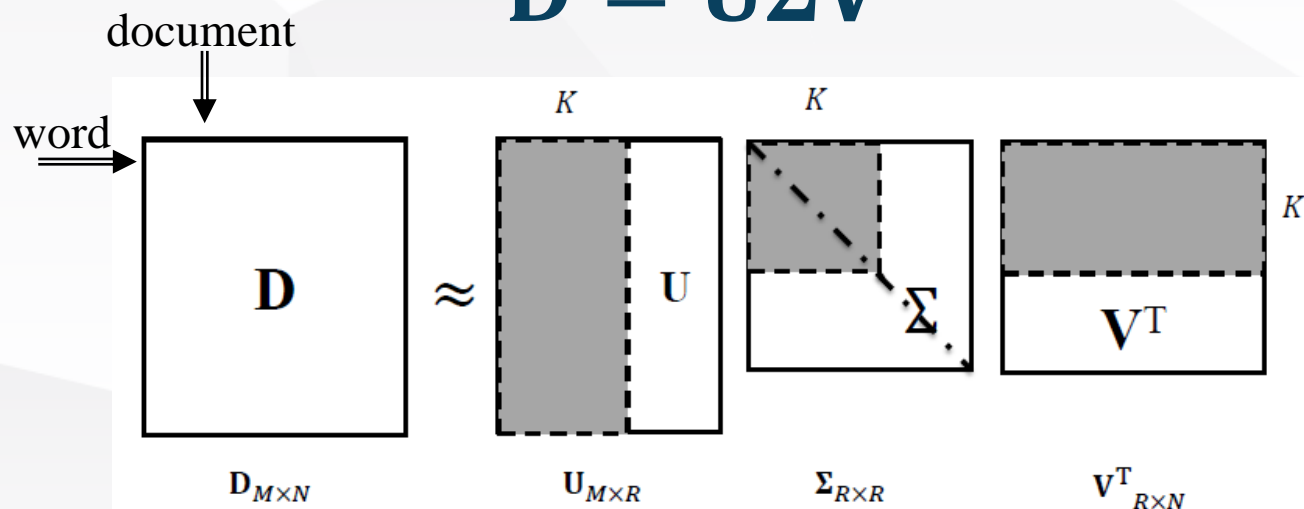
c1: *Human machine interface for Lab ABC computer applications*  
 c2: *A survey of user opinion of computer system response time*  
 c3: *The EPS user interface management system*  
 c4: *System and human system engineering testing of EPS*  
 c5: *Relation of user-perceived response time to error measurement*

m1: *The generation of random, binary, unordered trees*  
 m2: *The intersection graph of paths in trees*  
 m3: *Graph minors IV: Widths of trees and well-quasi-ordering*  
 m4: *Graph minors: A survey*

### Terms

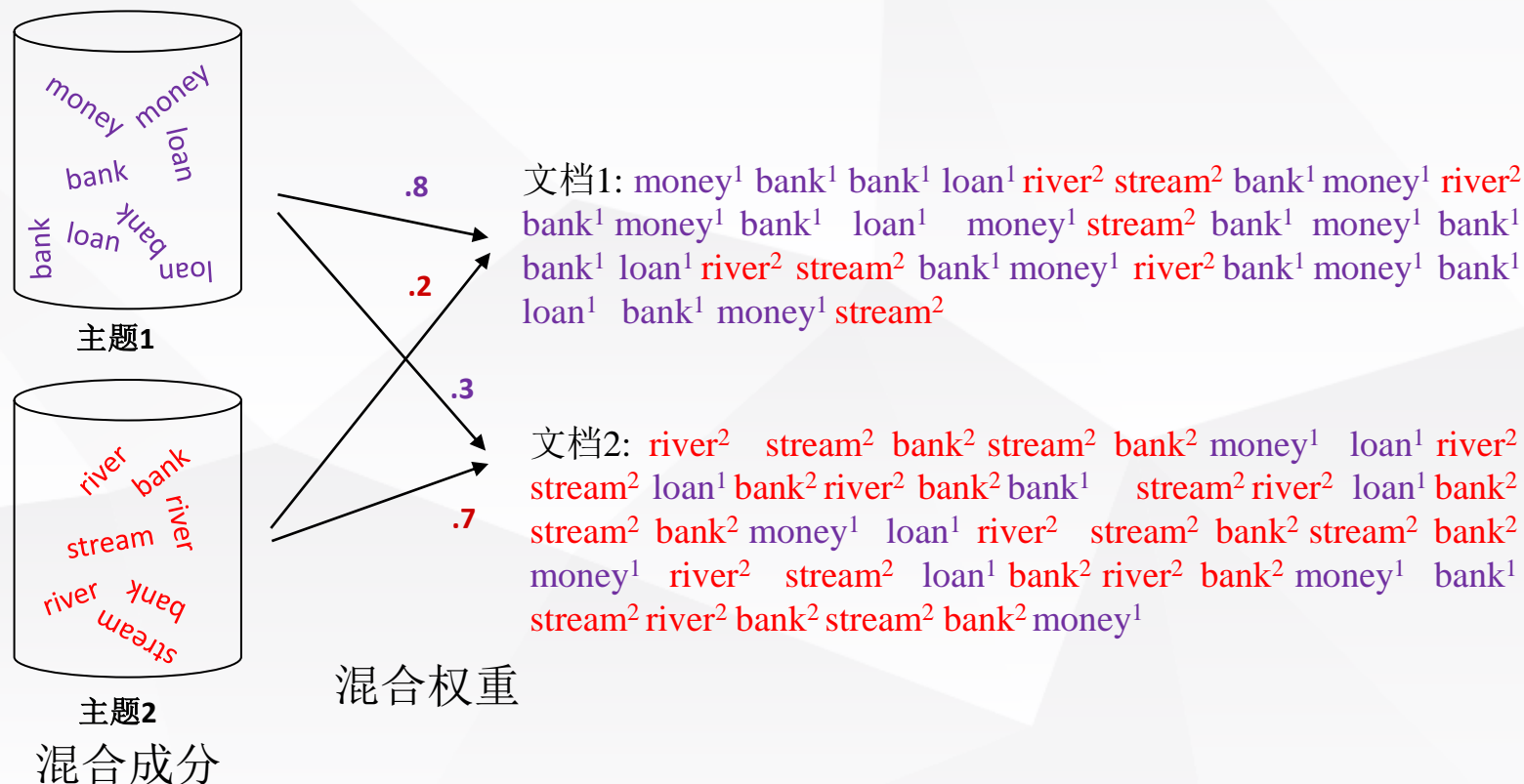
	Documents									
	c1	c2	c3	c4	c5	m1	m2	m3	m4	
human	1	0	0	1	0	0	0	0	0	
interface	1	0	1	0	0	0	0	0	0	
computer	1	1	0	0	0	0	0	0	0	
user	0	1	1	0	1	0	0	0	0	
system	0	1	1	2	0	0	0	0	0	
response	0	1	0	0	1	0	0	0	0	
time	0	1	0	0	1	0	0	0	0	
EPS	0	0	1	1	0	0	0	0	0	
survey	0	1	0	0	0	0	0	0	1	
trees	0	0	0	0	0	1	1	1	0	
graph	0	0	0	0	0	0	1	1	1	
minors	0	0	0	0	0	0	0	1	1	

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



Topic1	Topic2	Topic3	Topic4	Topic5	Topic6	Topic7	Topic8	Topic9	Topic10
OPEC	Africa	contra	school	Noriega	firefight	plane	Saturday	Iran	senate
oil	South	Sandinista	student	Panama	ACR	crash	coastal	Iranian	Reagan
cent	African	rebel	teacher	Panamanian	forest	flight	estimate	Iraq	billion
barrel	Angola	Nicaragua	education	Delval	park	air	western	hostage	budget
price	apartheid	Nicaraguan	college	canal	blaze	airline	Minsch	Iraqi	Trade

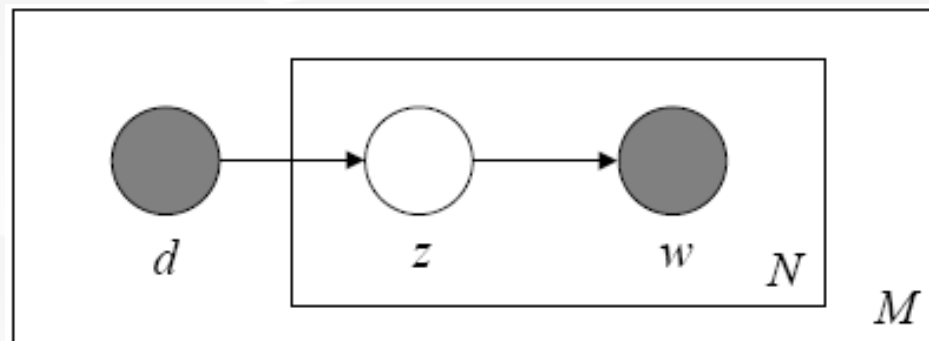
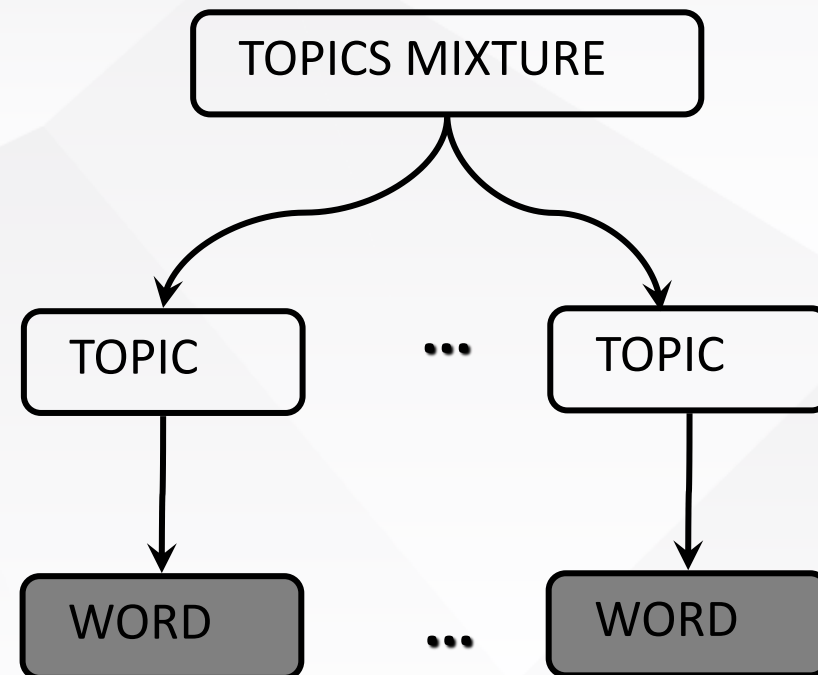
- 将文档中的词看做来自混合模型的采样。
- 每个词来自一个主题，同一个文档中不同词可能来自不同的主题。
- 每个文档被表示为可以被表示为主题（混合成分）的混合



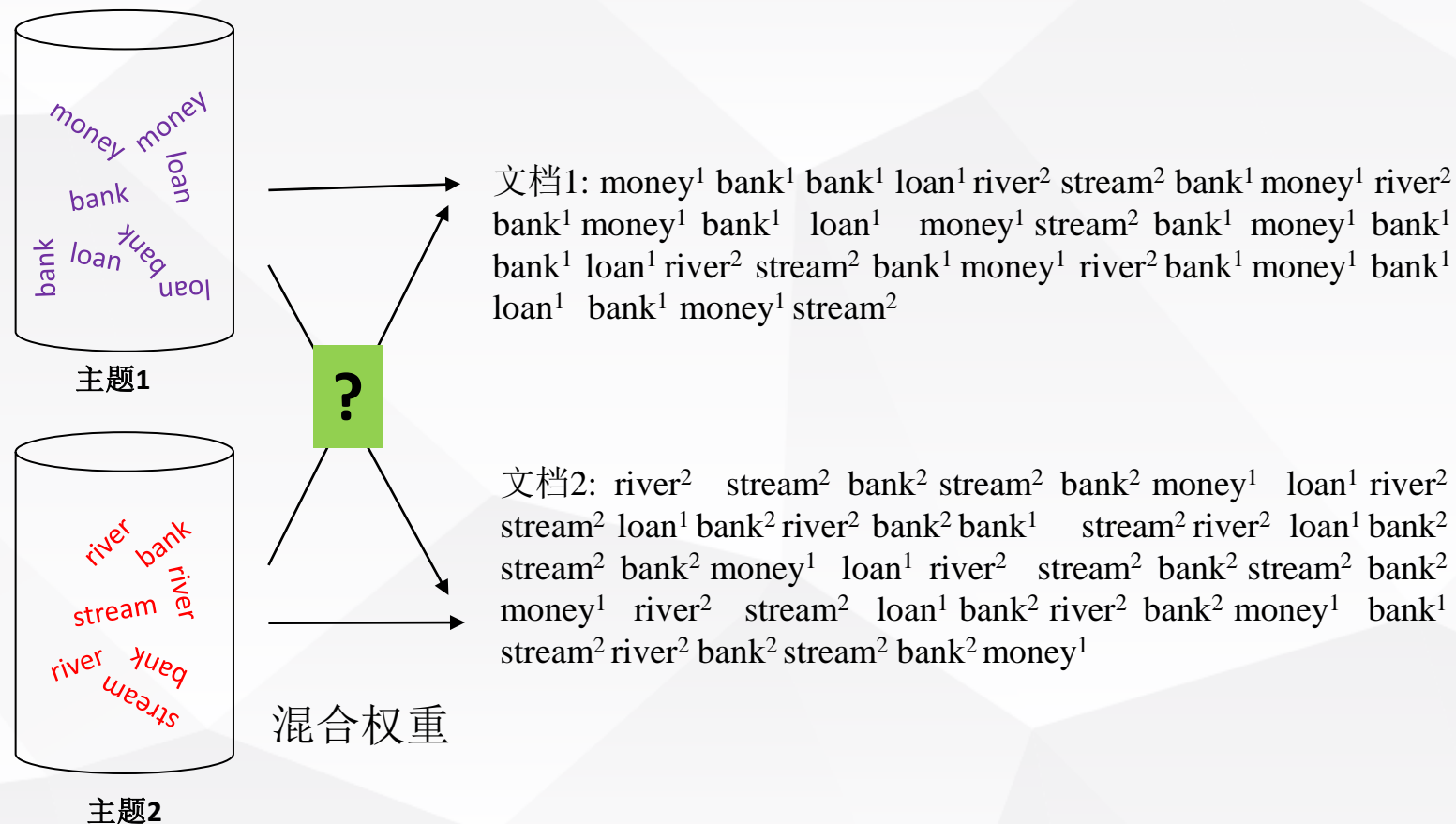
## 文档生成看做概率过程

- 对于每一个文档, 选择一个混合的主题
- 对于每个词, 从主题列表中采样一个词

$$\begin{aligned} p(d, w) &= p(d)p(w|d) \\ &= p(d) \sum_z p(w|z)p(z|d) \\ &= \sum_{z \in Z} P(z)P(d|z)P(w|z) \end{aligned}$$



# >> 推断的问题



$$P(z|d, w) = \frac{P(z)P(d|z)P(w|z)}{\sum_{z' \in Z} P(z')P(d|z')P(w|z')}$$

### ■ E步：计算隐变量的后验概率

$$P(z|d, w) = \frac{P(z)P(d|z)P(w|z)}{\sum_{z' \in Z} P(z')P(d|z')P(w|z')}$$

### ■ M步：更新参数

$$P(w|z) \propto \sum_{d \in D} n(d, w)P(z|d, w)$$

$$P(d|z) \propto \sum_{w \in W} n(d, w)P(z|d, w)$$

$$P(z) \propto \sum_{d \in D} \sum_{w \in W} n(d, w)P(z|d, w)$$

## 例子：从科学杂志论文集合发现主题

universe	0.0439	drug	0.0672	cells	0.0675	sequence	0.0818	years	0.156
galaxies	0.0375	patients	0.0493	stem	0.0478	sequences	0.0493	million	0.0556
clusters	0.0279	drugs	0.0444	human	0.0421	genome	0.033	ago	0.045
matter	0.0233	clinical	0.0346	cell	0.0309	dna	0.0257	time	0.0317
galaxy	0.0232	treatment	0.028	gene	0.025	sequencing	0.0172	age	0.0243
cluster	0.0214	trials	0.0277	tissue	0.0185	map	0.0123	year	0.024
cosmic	0.0137	therapy	0.0213	cloning	0.0169	genes	0.0122	record	0.0238
dark	0.0131	trial	0.0164	transfer	0.0155	chromosome	0.0119	early	0.0233
light	0.0109	disease	0.0157	blood	0.0113	regions	0.0119	billion	0.0177
density	0.01	medical	0.00997	embryos	0.0111	human	0.0111	history	0.0148
bacteria	0.0983	male	0.0558	theory	0.0811	immune	0.0909	stars	0.0524
bacterial	0.0561	females	0.0541	physics	0.0782	response	0.0375	star	0.0458
resistance	0.0431	female	0.0529	physicists	0.0146	system	0.0358	astrophys	0.0237
coli	0.0381	males	0.0477	einstein	0.0142	responses	0.0322	mass	0.021
strains	0.025	sex	0.0339	university	0.013	antigen	0.0263	disk	0.0173
microbiol	0.0214	reproductive	0.0172	gravity	0.013	antigens	0.0184	black	0.0161
microbial	0.0196	offspring	0.0168	black	0.0127	immunity	0.0176	gas	0.0149
strain	0.0165	sexual	0.0166	theories	0.01	immunology	0.0145	stellar	0.0127
salmonella	0.0163	reproduction	0.0143	aps	0.00987	antibody	0.014	astron	0.0125
resistant	0.0145	eggs	0.0138	matter	0.00954	autoimmune	0.0128	hole	0.00824



PRINTING  
PAPER  
PRINT  
PRINTED  
TYPE  
PROCESS  
INK  
PRESS  
IMAGE  
PRINTER  
PRINTS  
PRINTERS  
COPY  
COPIES  
FORM  
OFFSET  
GRAPHIC  
SURFACE  
PRODUCED  
**CHARACTERS**

**PLAY**  
PLAYS  
STAGE  
AUDIENCE  
THEATER  
ACTORS  
DRAMA  
SHAKESPEARE  
ACTOR  
THEATRE  
PLAYWRIGHT  
PERFORMANCE  
DRAMATIC  
COSTUMES  
COMEDY  
TRAGEDY  
**CHARACTERS**  
SCENES  
OPERA  
PERFORMED

TEAM  
GAME  
BASKETBALL  
PLAYERS  
PLAYER  
**PLAY**  
PLAYING  
SOCCER  
PLAYED  
BALL  
TEAMS  
BASKET  
FOOTBALL  
SCORE  
**COURT**  
GAMES  
TRY  
COACH  
GYM  
SHOT

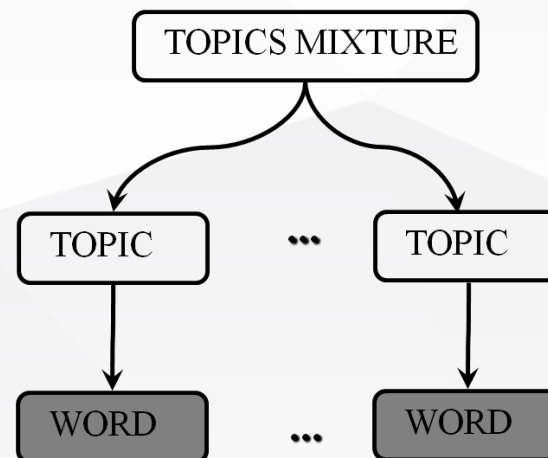
JUDGE  
TRIAL  
**COURT**  
CASE  
JURY  
ACCUSED  
GUILTY  
DEFENDANT  
JUSTICE  
**EVIDENCE**  
WITNESSES  
CRIME  
LAWYER  
WITNESS  
ATTORNEY  
HEARING  
INNOCENT  
DEFENSE  
CHARGE  
CRIMINAL

HYPOTHESIS  
EXPERIMENT  
SCIENTIFIC  
OBSERVATIONS  
SCIENTISTS  
EXPERIMENTS  
SCIENTIST  
EXPERIMENTAL  
**TEST**  
METHOD  
HYPOTHESES  
TESTED  
**EVIDENCE**  
BASED  
OBSERVATION  
SCIENCE  
FACTS  
DATA  
RESULTS  
EXPLANATION

STUDY  
**TEST**  
STUDYING  
HOMEWORK  
NEED  
CLASS  
MATH  
TRY  
TEACHER  
WRITE  
PLAN  
ARITHMETIC  
ASSIGNMENT  
PLACE  
STUDIED  
CAREFULLY  
DECIDE  
IMPORTANT  
NOTEBOOK  
REVIEW

## ■ PLSA的问题:

- 不完整: 没有提供文档层面的概率建模
- 模型的参数数量随着语料规模线性增长
- 没有明确训练数据外的文档如何计算概率



## ➤ 隐含狄利克雷分布(Latent Dirichlet allocation)

- LDA是对整个语料的生成式建模
- 符号约定:一个文档由 $N$ 个词表示 $\mathbf{w} = (w_1, w_2, \dots, w_N)$ ,语料 $D$ 由 $M$ 个文档表示 $D = \{\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_M\}$ ,  $V$  表示词表大小.
- 语料 $D$  中每个文档 $\mathbf{w}$ 的生成过程:

1. 采样 $N \sim \text{Poisson}(\xi)$

2. 从狄利克雷分布随机采样话题分布 $\theta \sim \text{Dir}(\alpha)$  
$$p(\theta|\alpha) = \frac{\Gamma(\sum_{i=1}^k \alpha_i)}{\prod_{i=1}^k \Gamma(\alpha_i)} \theta_1^{\alpha_1-1} \dots \theta_k^{\alpha_k-1}$$

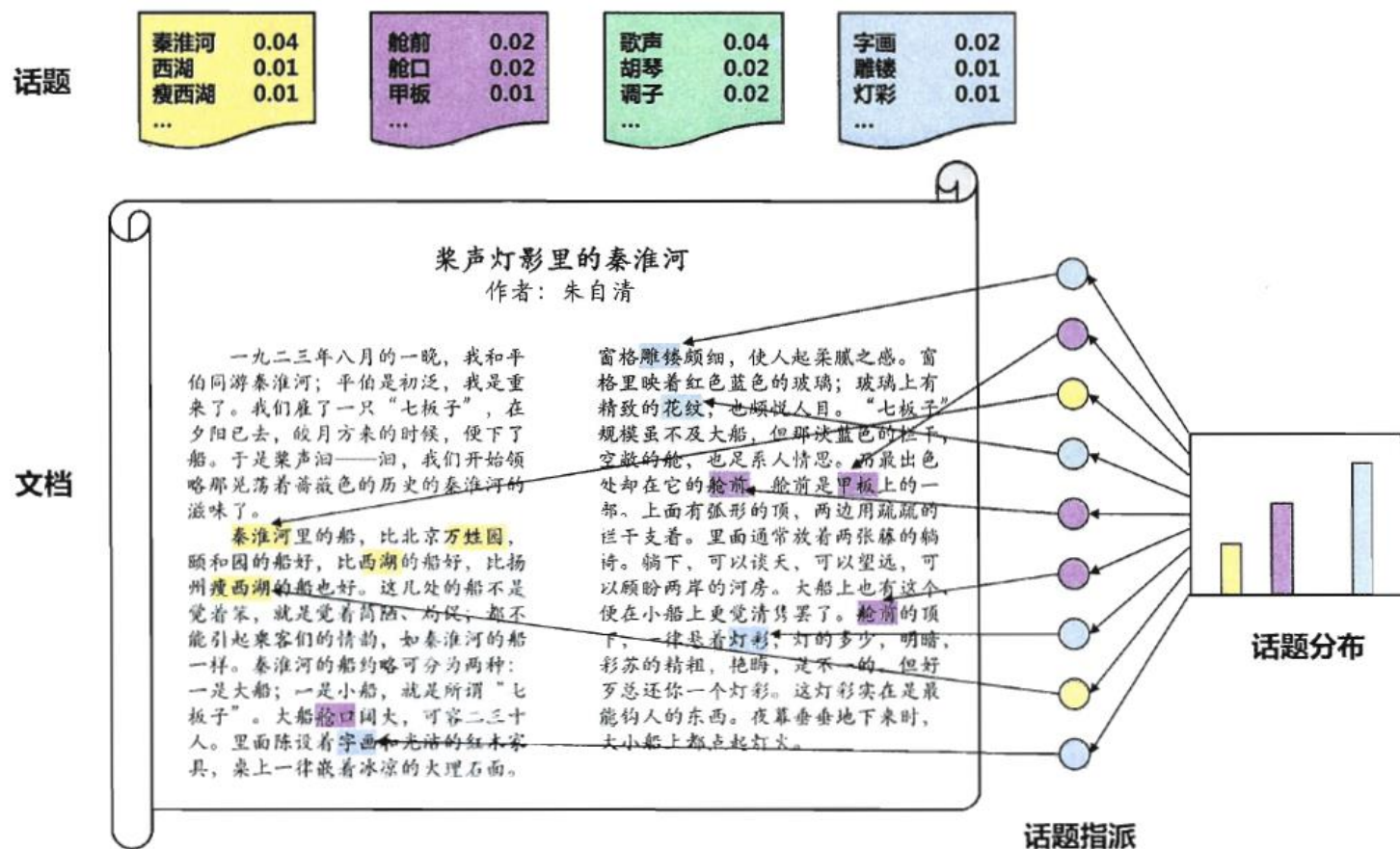
3. 对于 $N$  个词中的每一个词 $w_n$  :

(a) 选择一个话题 $z_n \sim \text{Multinomial}(\theta)$

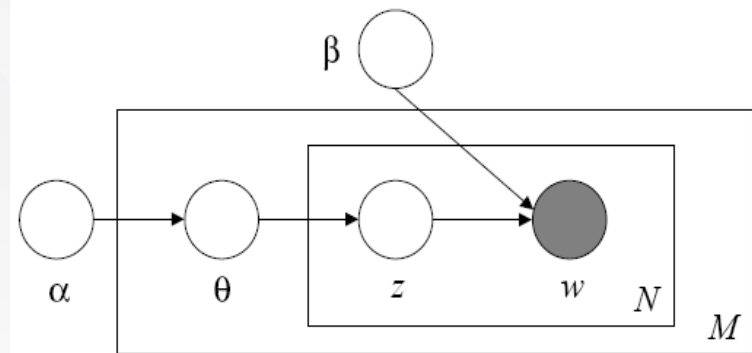
(b) 按 $p(w_n|z_n, \beta)$ 选择一个词 $w_n$ ,  $p(w_n|z_n, \beta)$ 是给定主题 $z_n$ 的条件概率分布

其中 $\beta$ 是 $k \times V$ 的矩阵, 且 $\beta_{ij} = p(w^j = 1|z^i = 1)$ .

# 隐含狄利克雷分布(Latent Dirichlet allocation)



LDA 的文档生成过程示意图



- LDA表示有三种层次
  - $\alpha, \beta$  是语料层次的参数
  - $\theta_d$  是文档层次的参数
  - $z_{dn}, w_{dn}$  是词层次的参数

- 话题分布 $\theta$ , 话题 $\mathbf{z}$ ,  $N$  个词  $\mathbf{w}$ 的联合概率分布:

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

对 $\theta$ 积分, 对 $\mathbf{z}$ 求和, 可以得到文档的边际分布:

$$p(\mathbf{w} | \alpha, \beta) = \int p(\theta | \alpha) \left( \prod_{n=1}^N \sum_{z_n} p(z_n | \theta) p(w_n | z_n, \beta) \right) d\theta$$

- 对每个文档的边际概率求积得到语料的概率:

$$p(D | \alpha, \beta) = \prod_{d=1}^D \int p(\theta_d | \alpha) \left( \prod_{n=1}^{N_d} \sum_{z_{dn}} p(z_{dn} | \theta_d) p(w_{dn} | z_{dn}, \beta) \right) d\theta_d$$

- 关键的推断问题是给定文档计算隐变量的后验分布

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

$$p(\mathbf{w} | \alpha, \beta) = \frac{\Gamma(\sum_i \alpha_i)}{\prod_i \Gamma(\alpha_i)} \int \left( \prod_{i=1}^k \theta_i^{\alpha_i - 1} \right) \left( \prod_{n=1}^N \sum_{i=1}^k \prod_{j=1}^V (\theta_i \beta_{ij})^{w_n^j} \right) d\theta$$

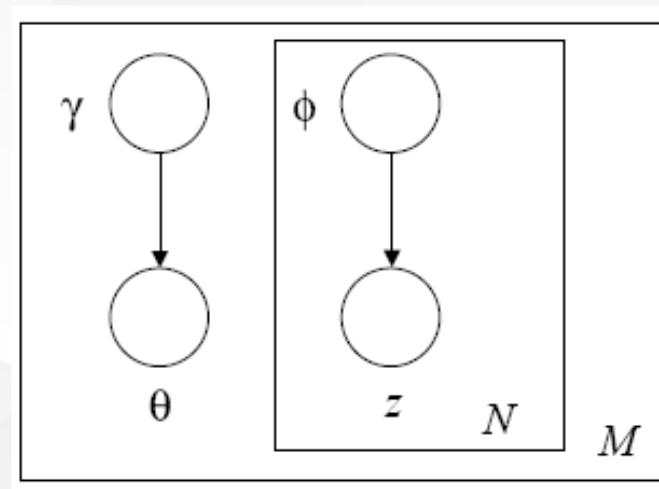
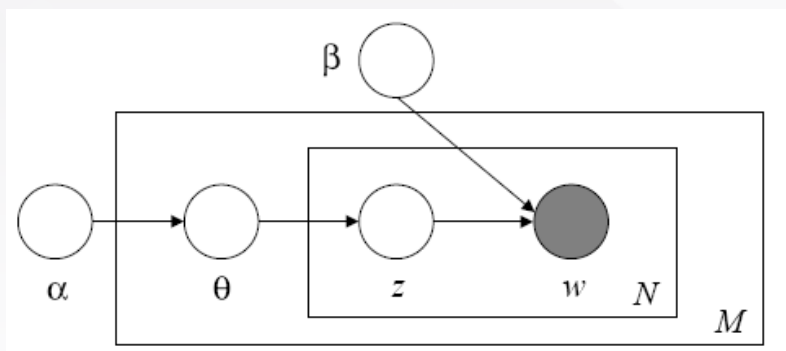
不幸的是该分布难以计算

函数不可计算是因为在对隐含主题求和过程将 $\theta$  和  $\beta$ 在耦合在一起

- 变分推断：E步对 $p(\mathbf{Z} | \mathbf{X}, \theta^{old})$ 的推断因为 $\mathbf{z}$ 模型的复杂难以进行时

$$q(\mathbf{Z}) = \prod_{i=1}^m q_i(\mathbf{Z})$$

- 修改原模型，丢弃一些边( $\theta \rightarrow \mathbf{z}, \mathbf{z} \rightarrow \mathbf{w}$ )和节点 $\mathbf{w}$ 以解耦合 $\theta$  和  $\beta$



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

其中 $\gamma$ 为狄利克雷参数，多项式参数 $(\phi_1, \dots, \phi_n)$ 为自由变分参数

## 对数似然函数的下界

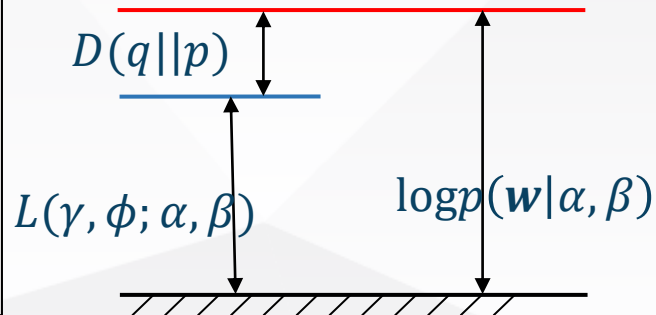
$$\begin{aligned}\log p(\mathbf{w}|\alpha, \beta) &= \log \int \sum_{\mathbf{z}} p(\theta, \mathbf{z}, \mathbf{w}|\alpha, \beta) d\theta = \log \int \sum_{\mathbf{z}} q(\theta, \mathbf{z}|\gamma, \phi) \frac{p(\theta, \mathbf{z}, \mathbf{w}|\alpha, \beta)}{q(\theta, \mathbf{z}|\gamma, \phi)} d\theta \\ &\geq \int \sum_{\mathbf{z}} q(\theta, \mathbf{z}|\gamma, \phi) \log \frac{p(\theta, \mathbf{z}, \mathbf{w}|\alpha, \beta)}{q(\theta, \mathbf{z}|\gamma, \phi)} d\theta = E_q[\log p(\theta, \mathbf{z}, \mathbf{w}|\alpha, \beta)] - E_q[\log q(\theta, \mathbf{z}|\gamma, \phi)]\end{aligned}$$

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

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

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

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



对数似然和其下界的差距在于变分后验和真实后验的KL距离

$$\log p(\mathbf{w}|\alpha, \beta) = \underbrace{E_q[\log p(\theta, \mathbf{z}, \mathbf{w}|\alpha, \beta)] - E_q[\log q(\theta, \mathbf{z}|\gamma, \phi)]}_{L(\gamma, \phi; \alpha, \beta)} + D(q(\theta, \mathbf{z}|\gamma, \phi) || p(\theta, \mathbf{z}|\mathbf{w}, \alpha, \beta))$$



# 变分推断和参数估计

■ **E步**: 针对变分参数  $\gamma, \phi$  寻找对数似然函数紧的下界

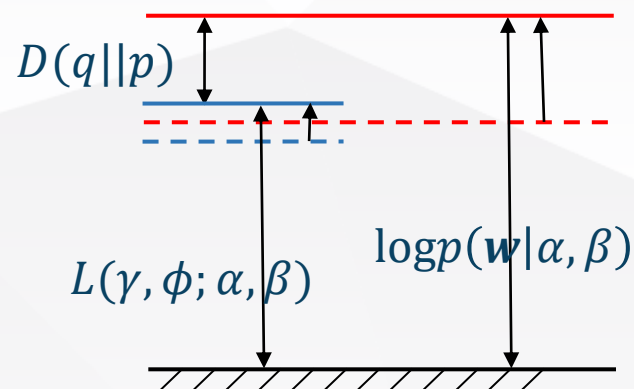
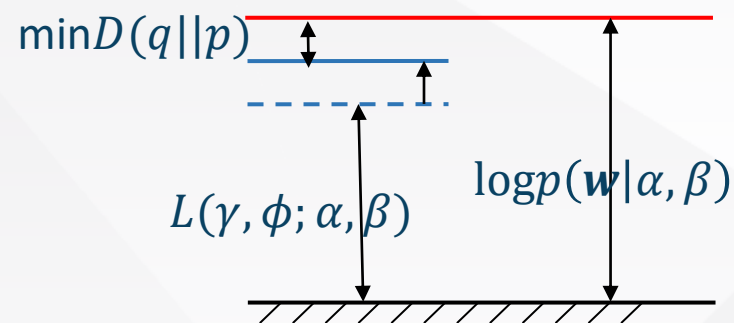
$$(\gamma^*, \phi^*) = \arg \min_{(\gamma, \phi)} D(q(\theta, \mathbf{z} | \gamma, \phi) || p(\theta, \mathbf{z} | \mathbf{w}, \alpha, \beta))$$

$$\phi_{ni} \propto \beta_{iv} \exp \left( \psi(\gamma_i) - \psi \left( \sum_{j=1}^k \gamma_j \right) \right)$$

$$\gamma_i \propto \alpha_i + \sum_{n=1}^N \phi_{ni}$$

■ **M步**: 针参数对  $\alpha$  和  $\beta$  最大化  $L(\gamma, \phi; \alpha, \beta)$ , 即在近似后验下计算最大似然估计

- 对  $L(\gamma, \phi; \alpha, \beta)$  应用坐标下降方法
- 参数  $\beta$  的解析解  $\beta_{ij} \propto \sum_{d=1}^M \sum_{n=1}^{N_d} \phi_{dni} w_{dn}^j$
- 参数  $\alpha$  采用线性时间牛顿法迭代求解



- 概率图模型是一套优美的工具：
  - 表示(数据结构)
  - 推断(算法).
- 两种类型的概率图模型
  - 贝叶斯网络
  - 马尔科夫随机场
- 典型的概率图模型
  - 隐马尔科夫模型是有向的生成模型
    - 三个经典的问题
    - 前后向算法
  - 条件随机场是无向的判别模型
  - PLSA和LDA模型是生成式话题模型
    - LDA比PLSA更加完整

- 课程代码: [https://github.com/lixinsu/tutorials2018/blob/master/graph\\_model.ipynb](https://github.com/lixinsu/tutorials2018/blob/master/graph_model.ipynb)
- 课后作业: 掌握贝叶斯球法则, 能够判断图模型上的条件独立性
- 参考资料:
  - 周志华, 《机器学习》
  - 常虹 《Graphical Models》
  - Michael I. Jordan 《An Introduction to Graphical Models》
  - Christopher Bishop 《Pattern Recognition and Machine Learning》
  - J. Lafferty, A. McCallum and F. Pereira. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. ICML 2001
  - Carlos Guestrin, Ben Taskar and Daphne Koller. Max-margin Markov Networks. NIPS 2003.
  - Eric P. Xing Graphical models Course

**The End**