# 机器学习
# Machine Learning

北京航空航天大学计算机学院
School of Computer Science and Engineering, Beihang University
刘庆杰　陈佳鑫

2025年春季学期
Spring 2025

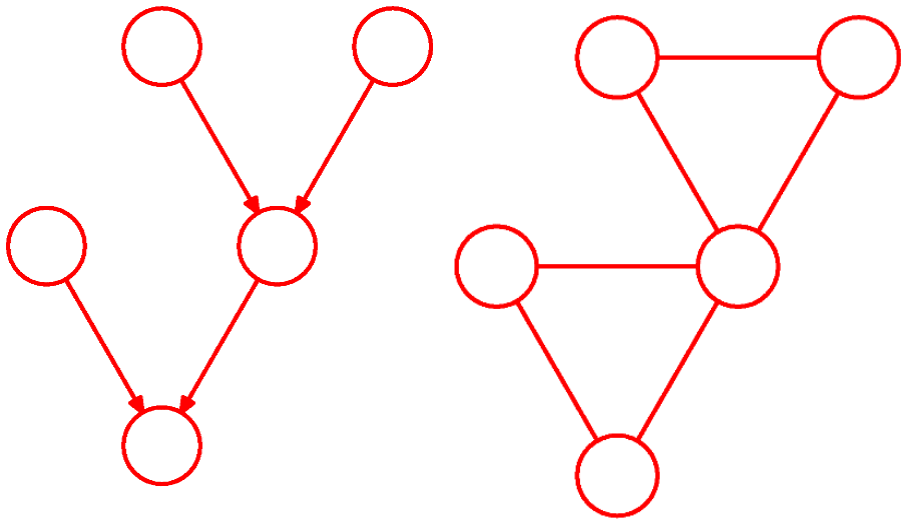
部分内容来源于C. Bishop和A. NG等人的课程以及互联网资源

# 10.1 什么是概率图模型?

- 概率图模型的定义
- 概率图模型的分类

# 什么是概率图模型?

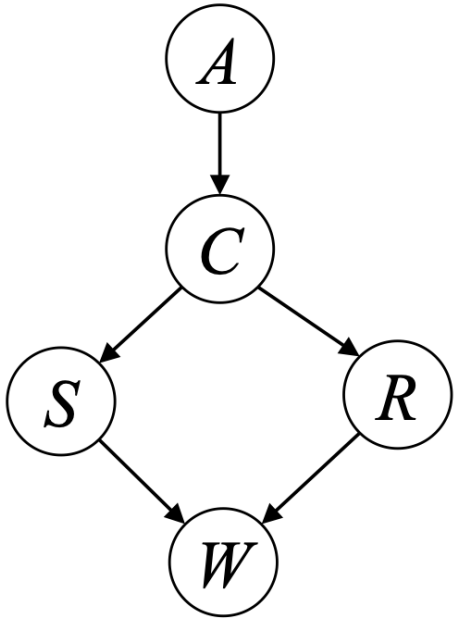
- 概率图模型（Probabilistic Graphical Model, PGM）的定义
  - 概率图模型是一种用图结构来表示和推断多元随机变量之间关系的概率模型。



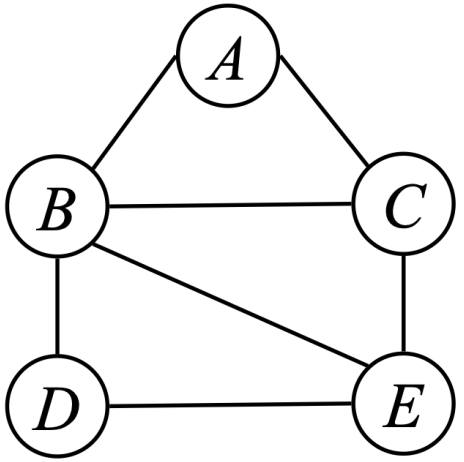
  - 结点：随机变量或一组随机变量
  - 连接弧：随机变量之间的关系

# 概率图模型的分类

- **概率图模型的分类**
  - 根据图模型的结构的不同，概率图模型主要分为**贝叶斯网络**和**马尔可夫场**。贝叶斯网络由**有向无环图**组成，马尔可夫场由**无向图**组成。



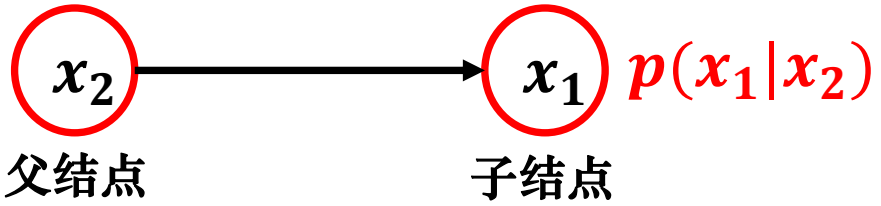

贝叶斯网络




马尔可夫场

# 贝叶斯网络的概率分布

- 贝叶斯网络的概率分布
  - 贝叶斯网络是一种**有向无环**概率图模型，表示了随机变量之间的**直接依赖关系**。每个结点定义了一个关于其父结点的**条件概率分布**。




  - 考虑含有K个随机变量$X = \{x_1, \dots, x_K\}$的贝叶斯网络，其联合概率分布为每个结点上的条件概率分布的乘积：

$$p(X) = \prod_{k=1}^{K} p(x_k|\mathrm{pa}_k)$$

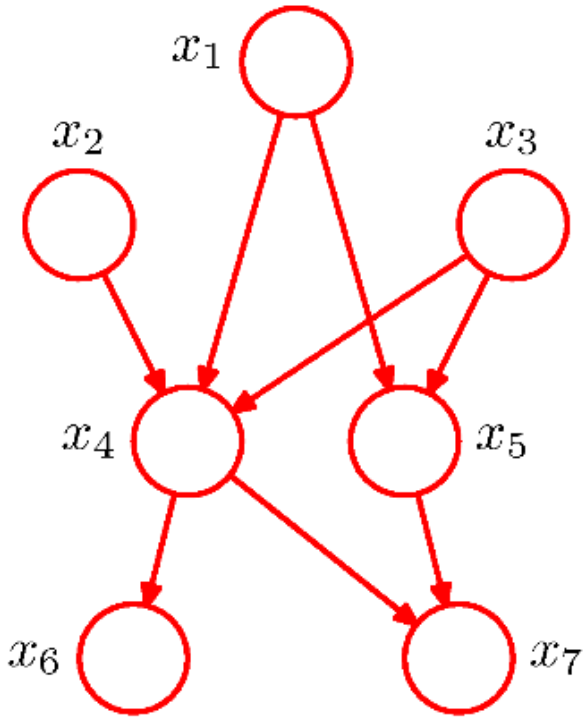
有向图的因子分解

其中，$pa_k$代表子节点$\mathrm{x}_k$的父节点集合

# 贝叶斯网络的概率分布

## 贝叶斯网络的概率分布

- 例子：考虑包含随机变量$X = \{x_1, \ldots, x_7\}$的贝叶斯网络，联合分布为：

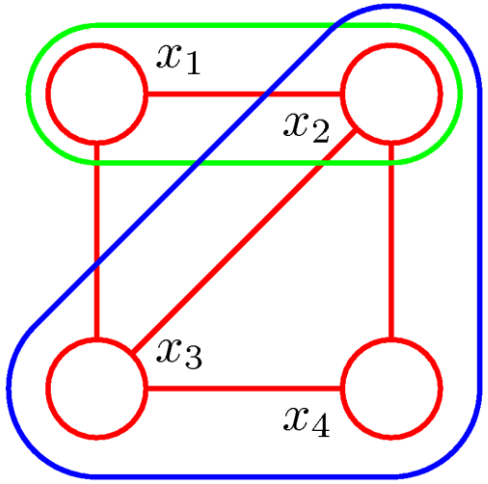



$$p(x_1, \ldots, x_7) = p(x_1)p(x_2)p(x_3)p(x_4|x_1, x_2, x_3)$$
$$p(x_5|x_1, x_3)\,p(x_6|x_4)\,p(x_7|x_4, x_5)$$

- $x_1, x_2, x_3$没有父节点
- $x_4$父节点为$x_1, x_2, x_3$
- $x_5$父节点为$x_1, x_3$
- $x_6$父节点为$x_4$
- $x_7$父节点为$x_4, x_5$

# 马尔可夫场的概率分布

- 马尔可夫场的概率分布
  - 马尔可夫场是一种无向概率图模型，通过对变量间联合概率分布的建模描述变量间关系。马尔可夫场的概率分布通过团上的势函数定义。
  - 无向图的团(clique)：任意两个结点之间都有连接的子集




两个结点：$\{x_1, x_2\}\{x_2, x_3\}\{x_3, x_4\}\{x_2, x_4\}\{x_1, x_3\}$

三个结点：$\{x_1, x_2, x_3\}\{x_2, x_3, x_4\}$

最大团：不能被其他团所包含的团

# 马尔可夫场的概率分布

- 马尔可夫场的概率分布
  - 势函数：对于团$Q$，$Q$中的元素为$X_Q$，定义函数$\psi_C(Q_C)$为$Q_C$的势函数，表示了团内局部变量的偏好，例如：

$$\psi(x_1, x_2) = \begin{cases} 1, if\ x_1 = x_2 \\ 0.1\ otherwise \end{cases}$$

  - 马尔可夫场的联合概率分布基于最大团分解为多个因子的乘积，设所有最大团构成的集合为$C$，则联合概率分布为：

$$P(X) = \frac{1}{Z}\prod_{Q \in C}\psi_Q(X_Q)$$

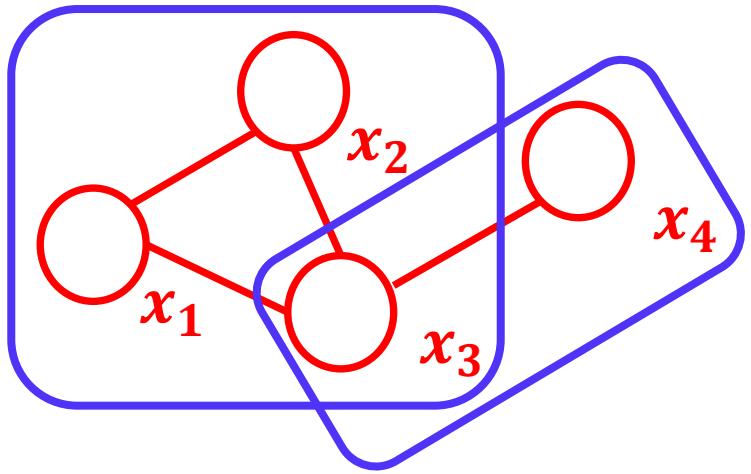
其中，Z为归一化因子

$$Z = \sum_X \prod_{Q \in C}\psi_Q(X_Q)$$

# 马尔可夫场的概率分布

- 马尔可夫场的概率表示

➢ 例子：考虑包含随机变量$X = \{x_1, x_2, x_3, x_4\}$的马尔可夫随机场，其联合概率分布为：



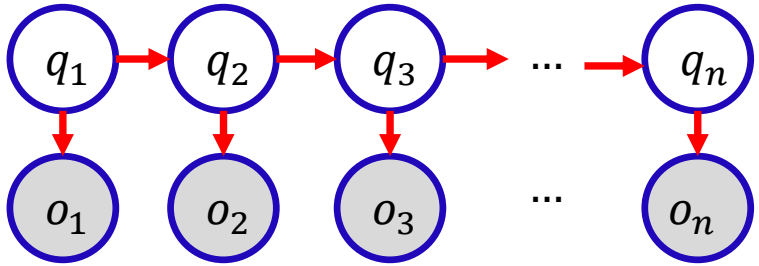

➢ 寻找最大团：$\{x_1, x_2, x_3\}, \{x_3, x_4\}$

➢ 根据最大团的势函数写出概率分布

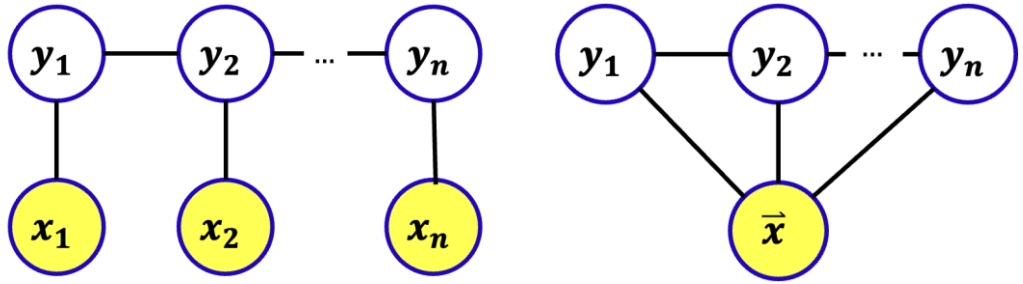
$$P(x_1, x_2, x_3, x_4) = \frac{1}{Z}\psi_{123}(x_1, x_2, x_3)\psi_{34}(x_3, x_4)$$

# 常见的概率图模型

- **隐马尔可夫模型**
  **Hidden Markov Model**




- **条件随机场**
  **Conditional Random Field**




- **…**

# 10.2 贝叶斯网络

- 认识贝叶斯网络
- 条件独立性
- D-分离

# 认识贝叶斯网络

- **贝叶斯网络（Bayesian Network）**
  - **一个贝叶斯网络 $B$ 由结构 $G$ 和参数 $\Theta$ 两部分构成，即$B = \langle G, \Theta \rangle$**
    - **其中网络结构 $G$ 是一个有向无环图，图中每个节点对应一个随机变量。若两个随机变量间有直接依赖关系，则它们由一条边连接起来**
    - **参数 $\Theta$ 定量描述了上述依赖关系。假设随机变量 $a$ 在 $G$ 中的父节点集合为 $\boldsymbol{\pi}_a$，则 $\Theta$ 中包含了每个随机变量的条件概率 $\boldsymbol{\theta}_{a|\boldsymbol{\pi}_a} = \boldsymbol{P}(a|\boldsymbol{\pi}_a)$**
  - **贝叶斯网络通过有向无环图结构提供了对联合分布中随机变量条件独立性的紧凑表示。**

# 回顾条件独立性

- **随机变量的独立性**

  - 考虑两个随机变量$\{a,b\}$，若$\{a,b\}$满足:

$$p(a|b)=p(a) \Longleftrightarrow p(a,b)=p(a)p(b)$$ ，称 $a$ 与 $b$ 独立，记为 $a \perp\!\!\!\perp b$

独立符号

事件$b$ 发生与否
对 $a$ 不产生影响

- **随机变量的条件独立性**

  - 考虑三个随机变量$\{a,b,c\}$，若$\{a,b,c\}$满足:

$$p(a|b,c)=p(a|c) \Longleftrightarrow p(a,b|c)=p(a|c)p(b|c)$$

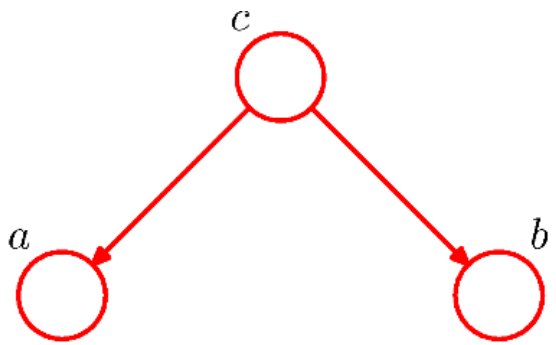
称在给定 $c$ 的条件下，$a$ 与 $b$ 条件独立，记为 $a \perp\!\!\!\perp b \mid c$

独立符号　条件符号

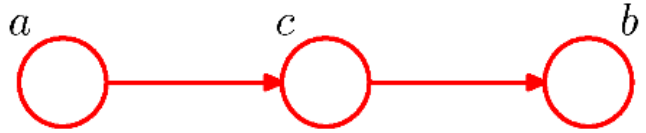
在事件 $c$ 发生的条件下，
事件 $b$ 发生与否
对 $a$ 不产生影响

# 贝叶斯网络的条件独立性

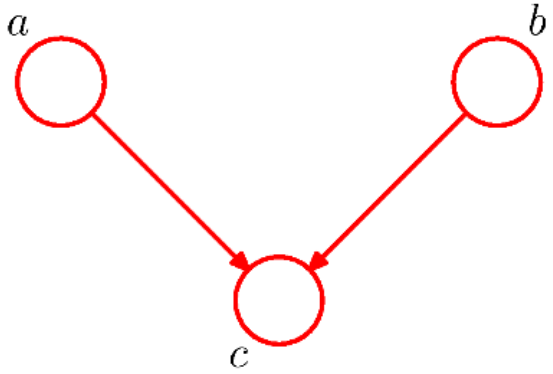
- **贝叶斯网络的条件独立性**
  - **在贝叶斯网络中，随机变量以节点的形式存在**
  - **贝叶斯网络中的三个节点$\{a, b, c\}$，存在如下三种结构：**




尾尾相连（tail-to-tail）




头尾相连（head-to-tail）




头头相连（head-to-head）

# 贝叶斯网络的条件独立性

- **条件独立性（尾尾相连）**
  - $a$ 与 $b$ 的独立性关系

$$p(a,b) \overset{?}{=} p(a)p(b)$$

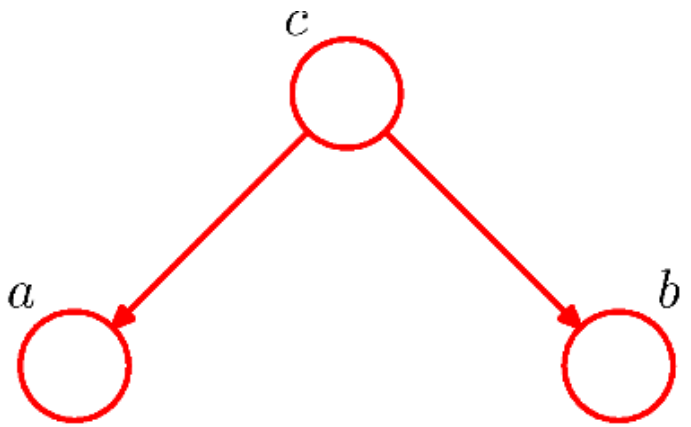
$$p(a,b) = \sum_c p(a,b,c)$$

**因子分解**

$$= \sum_c p(a|c)p(b|c)p(c) \quad \neq p(a)p(b)$$

**链式法则**

$$= \sum_c p(a|b,c)p(b|c)p(c) \quad \neq p(a)p(b)$$




尾尾相连（tail-to-tail）

  - $a$ 与 $b$ 不独立

$$a \not\perp b \mid \emptyset$$

**无条件情况下**

# 贝叶斯网络的条件独立性

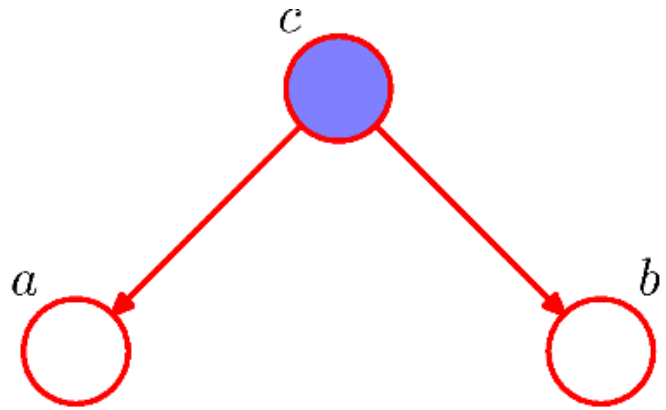
- **条件独立性（尾尾相连）**
  - $a, b, c$ **之间的条件独立性关系。**

$$p(a|c) \overset{?}{=} p(a|b,c)$$

$$p(a,b,c) = p(a|c)p(b|c)p(c)$$ **因子分解**

$$= p(a|b,c)p(b|c)p(c)$$ **链式法则**

$$\Rightarrow \quad p(a|c) = p(a|b,c)$$



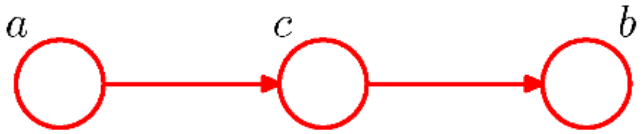

尾尾相连（tail-to-tail）

  - $a$ **与** $b$ **关于** $c$ **条件独立**

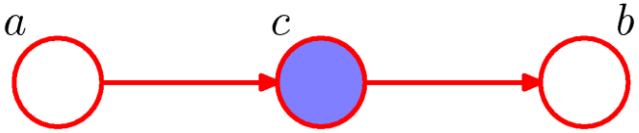
$$a \perp\!\!\!\perp b \mid c$$

# 贝叶斯网络的条件独立性

- **条件独立性（头尾相连）**




头尾相连
(head-to-tail)




➢ $a$ 与 $b$ 的独立性关系。

$$p(a,b) = \sum_c p(a,b,c)$$

$$= \sum_c p(a)p(c|a)p(b|c)$$

$$\neq p(a)p(b)$$

$$a \not\perp\!\!\!\perp b \mid \emptyset$$

➢ $a, b, c$ 之间的条件独立性关系。

$$p(a,b,c) = p(a|c)p(c|b)p(b)$$

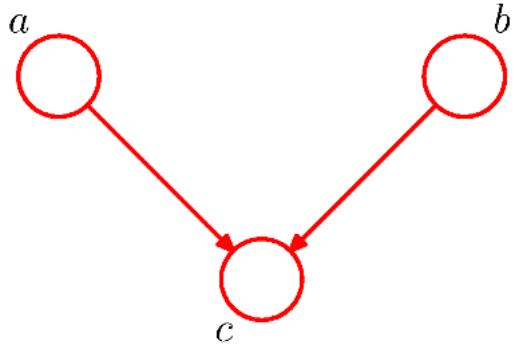
$$= p(a|b,c)p(c|b)p(b)$$

$$\Rightarrow p(a|c) = p(a|b,c)$$

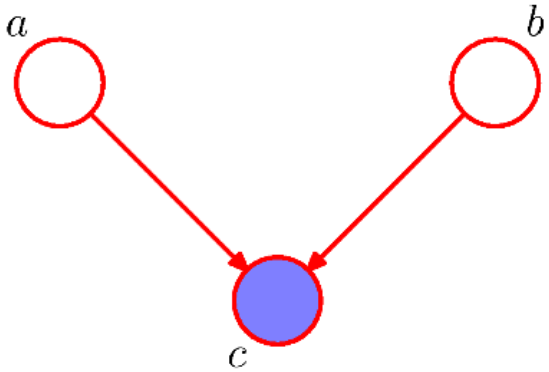
$$a \perp\!\!\!\perp b \mid c$$

# 贝叶斯网络的条件独立性

- **条件独立性（头头相连）**




头头相连
(head-to-head)




➢ **$a$ 与 $b$ 的独立性关系**

$$p(a,b) = \sum_c p(a,b,c)$$

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

$$= p(a)p(b)$$

$$a \perp\!\!\!\perp b \mid \emptyset$$

➢ **$a, b, c$ 之间的条件独立性关系**

$$p(a,b,c) = p(a)p(b)p(c|a,b)$$

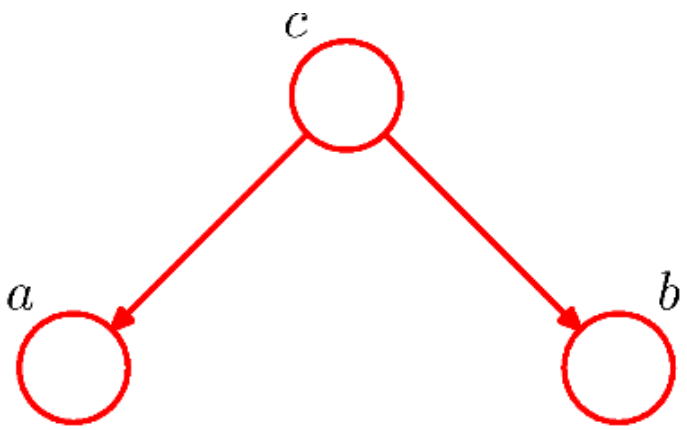
$$= p(a|b,c)p(c|b)p(b)$$

$$\Rightarrow p(a|c) \neq p(a|b,c)$$

$$a \not\perp\!\!\!\perp b \mid c$$

# 贝叶斯网络的条件独立性

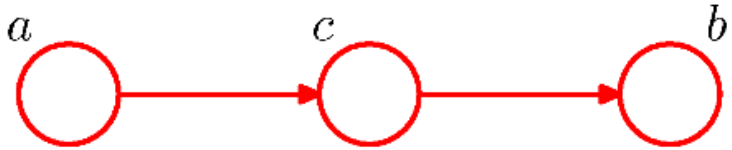
- **总结**




尾尾相连（tail-to-tail）

$$a \not\perp\!\!\!\perp b \mid \emptyset$$

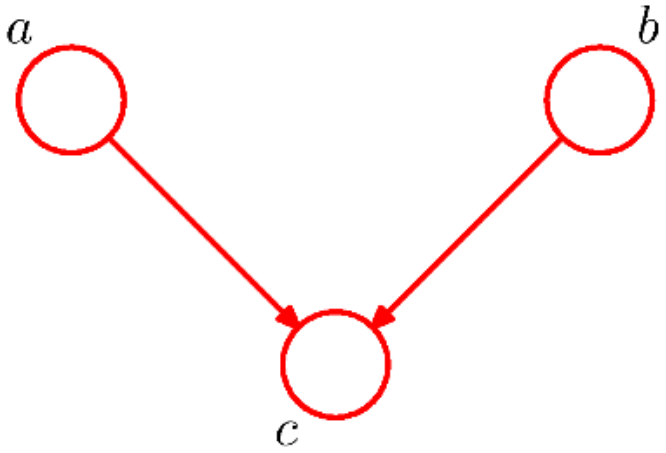
$$a \perp\!\!\!\perp b \mid c$$




头尾相连（head-to-tail）

$$a \not\perp\!\!\!\perp b \mid \emptyset$$

$$a \perp\!\!\!\perp b \mid c$$




头头相连（head-to-head）

$$a \perp\!\!\!\perp b \mid \emptyset$$

$$a \not\perp\!\!\!\perp b \mid c$$

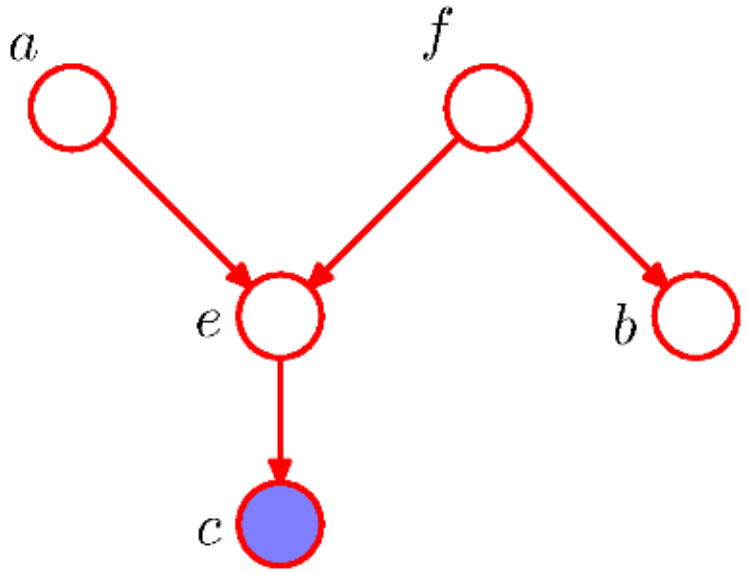
# D-分离

- D-分离（Directed-Separation）
  - 是一种用于判断概率图中节点集合间条件独立性的准则
  - 具体而言，对于概率图中的三个节点集合{A, B, C}，如果满足如下任一条件
    - 条件1：C中的结点满足“头尾相连”或“尾尾相连”；
    - 条件2：“头头相连”的节点和它的任何后裔节点都不在C中

    则A与B关于C条件独立

# D-分离

- 例子1
  - $a$ 与 $b$ 是否关于 $c$ 条件独立?




$$a \not\perp b \mid c$$

**D-分离**

如满足如下任一条件

- 条件1： C中的节点满足“头尾相连”或“尾尾相连”；
- 条件2：“头头相连”的节点和它的任何后裔节点都不在C中

则A与B关于C条件独立。

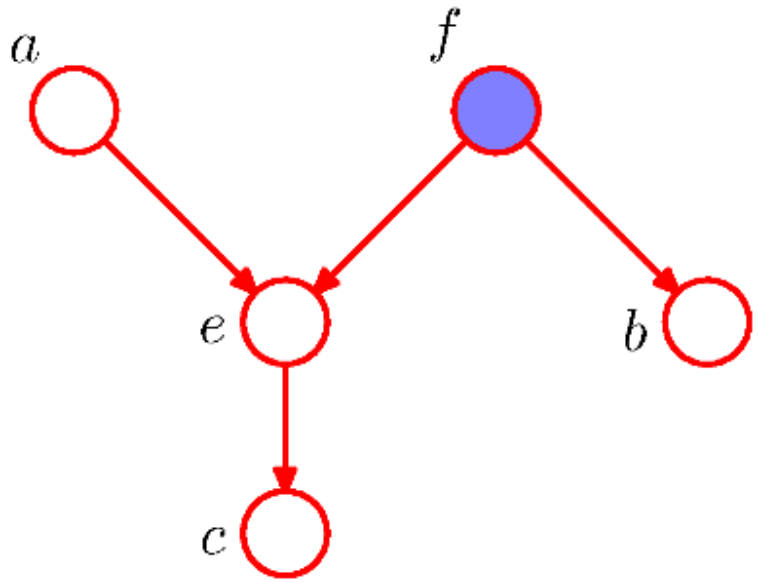
- A= $\{a\}$，B= $\{b\}$，C= $\{c\}$

两个条件均不满足

$a$ 与 $b$ 关于 $c$ 不条件独立

# D-分离

- 例子2
  - $a$ 与 $b$ 是否关于 $f$ 条件独立?




$$a \perp\!\!\!\perp b \mid f$$

**D-分离**

如满足如下任一条件

- 条件1： C中的节点满足“头尾相连”或“尾尾相连”；
- 条件2：“头头相连”的节点和它的任何后裔节点都不在C中

则A与B关于C条件独立。

■ A= $\{a\}$，B= $\{b\}$，C= $\{f\}$

节点 $f$ 满足“尾尾相连”，满足第一个条件

$a$ 与 $b$ 关于 $f$ 条件独立

# 10.3 贝叶斯推理

- 什么是贝叶斯推理
- 变量消除
- 信念传播

# 贝叶斯推理

- 什么是贝叶斯推理?
  - 贝叶斯推理是基于贝叶斯定理来求解某些假设概率的一种方法，通常用于根据已知的数据来推断未知的模型参数或预测结果
  - 假设贝叶斯网络对应的随机变量集合$X = \{x_1, x_2, \ldots, x_N\}$能分为$X_E$和$X_F$两个不相交的变量集，贝叶斯推理的目标就是计算目标变量$X_F$的边缘概率$P(X_F)$或条件概率$P(X_F|X_E)$。根据条件概率定义

$$P(X_F|X_E) = \frac{P(X_F, X_E)}{P(X_E)} = \frac{P(X_F, X_E)}{\sum_{X_F} P(X_F, X_E)}$$

其中，联合概率$P(X_F, X_E)$可基于贝叶斯网络的因子分解获得

# 贝叶斯推理方法分类

- 贝叶斯推理方法可大致分为两类
  - **精确推理**方法：希望计算出目标变量的边缘分布或条件分布的精确值。适用于模型简单或数据维度较小的情况
    - 变量消除（Variable Elimination）
    - 信念传播（Belief Propagation）
  - **近似推理**方法：希望在较低的时间开销下获得概率的近似估计值。适用于模型较为复杂或数据维度较高的情况
    - 马尔可夫链蒙特卡洛法 （Markov Chain Monte Carlo）
    - 变分推理（Variational Inference）

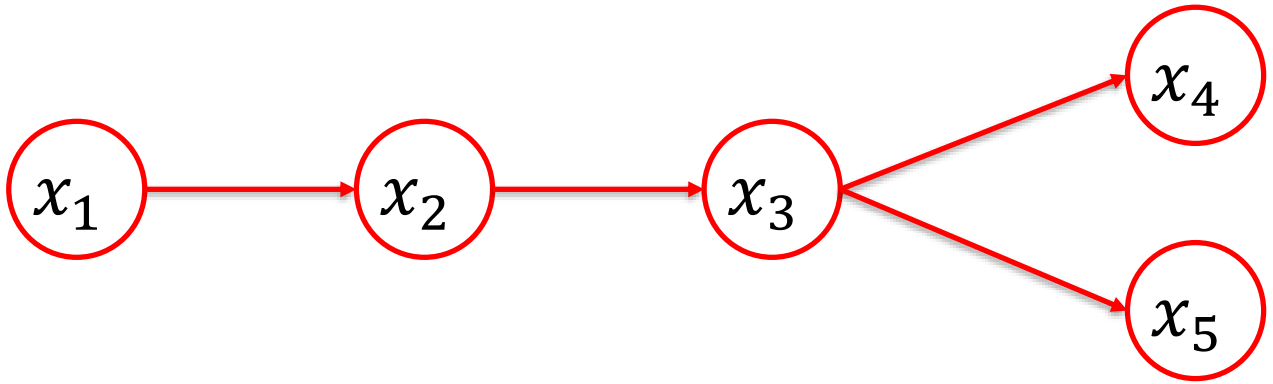
# 变量消除

- **变量消除（Variable Elimination）** 【1994年Nevin L. Zhang提出】
  - 变量消除是一种动态规划算法，它利用贝叶斯网络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
  - 变量消除是最直观的精确推理算法，也是构建其他精确推理算法的基础

- **算法基本流程**

# 变量消除-示例

- 给定如下贝叶斯网络，求目标随机变量 $x_5$ 的边缘概率分布$p(x_5)$



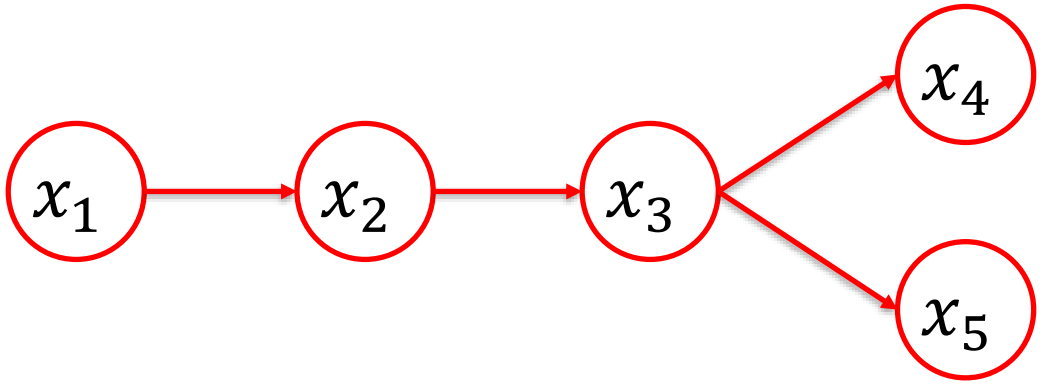

$$p(x_5) = \sum_{x_4}\sum_{x_3}\sum_{x_2}\sum_{x_1} p(x_1, x_2, x_3, x_4, x_5)$$

因子分解

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

# 变量消除-示例

- 给定如下贝叶斯网络，求目标随机变量 $x_5$ 的边缘概率分布$p(x_5)$




  - 选用变量**消除顺序**$x_1 \rightarrow x_2 \rightarrow x_4 \rightarrow x_3$

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

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

$$= \sum_{x_4}\sum_{x_3}\sum_{x_2} p(x_5|x_3)p(x_4|x_3)p(x_3|x_2)m_{12}(x_2) \quad (11.2.1)$$

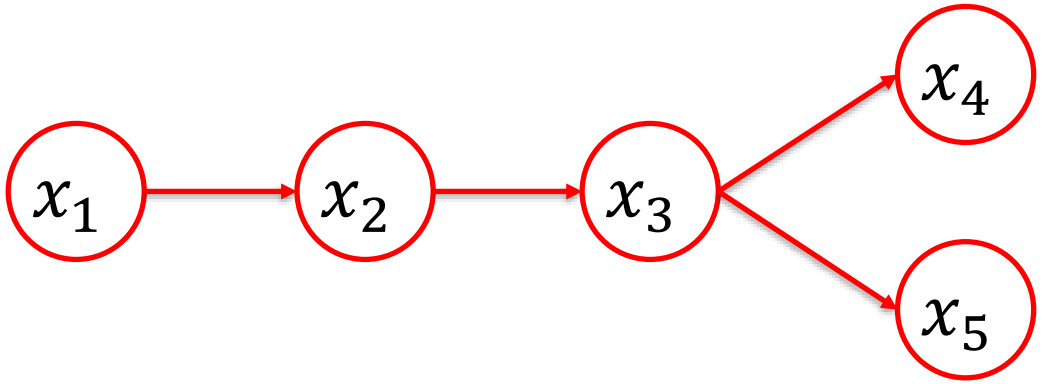
合并变量 $x_1$ 的全部因子

消除变量 $x_1$

代表对$x_1$求和的结果，是一个关于$x_2$的函数

# 变量消除-示例

- 给定如下贝叶斯网络，求目标随机变量 $x_5$ 的边缘概率分布$p(x_5)$




$$p(x_5) = \sum_{x_4}\sum_{x_3}\sum_{x_2} p(x_5|x_3)p(x_4|x_3)p(x_3|x_2)m_{12}(x_2)$$

合并 $x_2$ 的全部因子

$$= \sum_{x_4}\sum_{x_3} p(x_5|x_3)p(x_4|x_3)\sum_{x_2} p(x_3|x_2)m_{12}(x_2)$$

消除 $x_2$

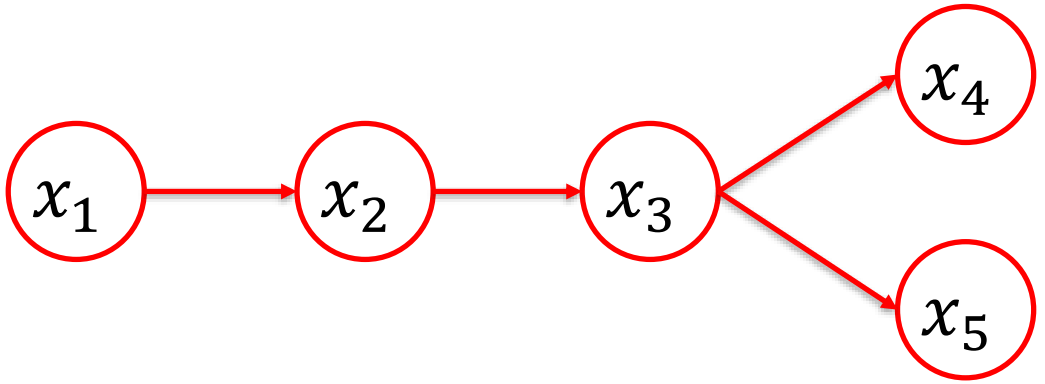
$$= \sum_{x_4}\sum_{x_3} p(x_5|x_3)p(x_4|x_3)m_{23}(x_3) \qquad (11.2.2)$$

合并 $x_4$ 的全部因子

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

# 变量消除-示例

- 给定如下贝叶斯网络，求目标随机变量 $x_5$ 的边缘概率分布$p(x_5)$




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

$$= \sum_{x_3} p(x_5|x_3)m_{23}(x_3)m_{43}(x_3) \quad (11.2.3)$$

$$= m_{35}(x_5) \quad (11.2.4)$$

消除 $x_4$

消除 $x_3$

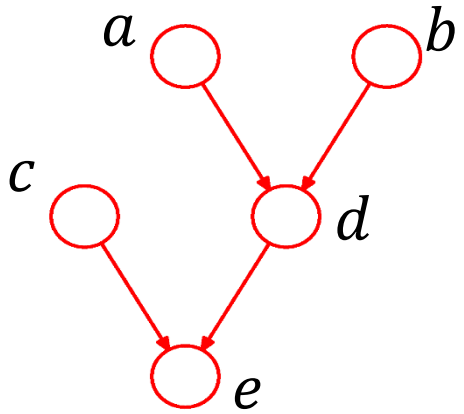
是一个关于$x_5$的函数，仅与$x_5$的取值有关

- 变量消除法把多个变量的积的求和问题，转化为对部分变量交替进行**求积与求和**的问题，使每次运算限制在局部，从而简化了计算

# 变量消除

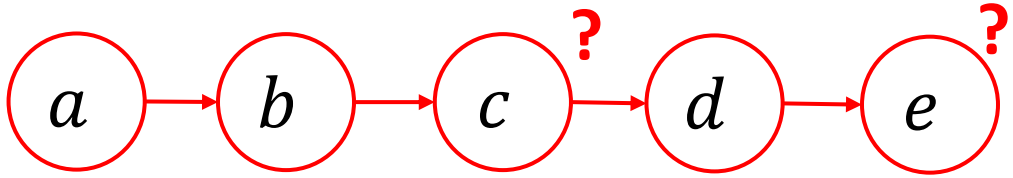
- **变量消除的局限性**
  - **变量消除顺序难以确定**

$$p(e) = \sum_{a,b,c,d} p(a,b,c,d,e)$$




**最优的变量消除顺序是一个NP-Hard问题【1999年Rina Dechter】**

  - **重复计算**




$$p(e) = \sum_{d} p(e|d) \sum_{c} p(d|c) \sum_{b} p(c|b) \sum_{a} p(a)p(b|a)$$

$$p(c) = \sum_{d} p(d|c) \sum_{e} p(e|d) \sum_{b} p(c|b) \sum_{a} p(a)p(b|a)$$

**若计算多个边缘分布，变量消除法存在大量重复计算**

# 信念传播

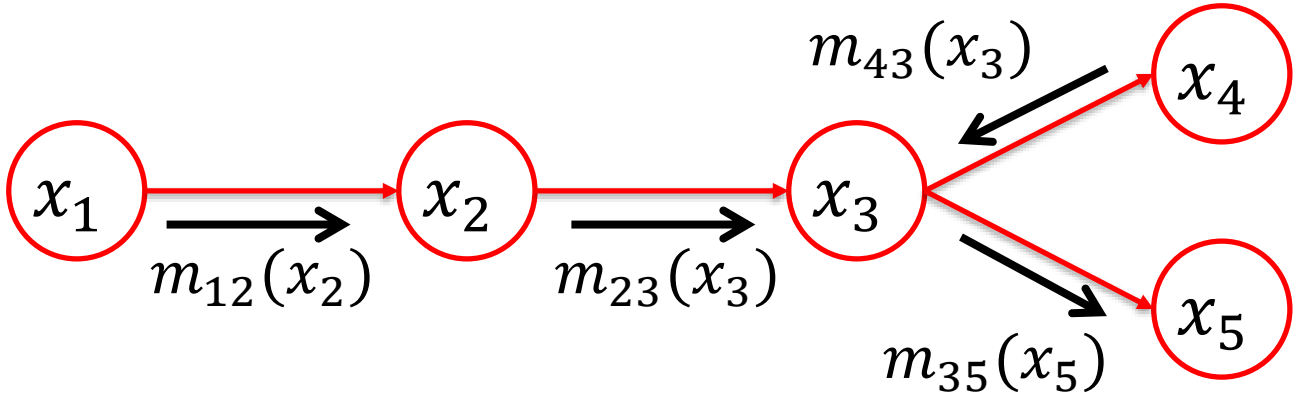
- **信念传播（Belief Propagation）**
  - 信念传播是一种精确推理算法，其将变量消除中的求和操作看作节点间**消息传递**的过程，通过预先存储每个节点的消息，较好的解决了求解多个边缘分布时的重复计算问题
  - 在**变量消除法**中，通过如下的求和操作消去变量$x_i$，其中$n(i)$表示结点$x_i$的邻居结点：

$$m_{ij}(x_j) = \sum_{x_i} p(x_j|x_i) \prod_{k \in n(i) \setminus j} m_{ki}(x_i)$$

在信念传播算法中，上述操作被看作从$x_i$向$x_j$，传递了一个**消息**$m_{ij}(x_j)$

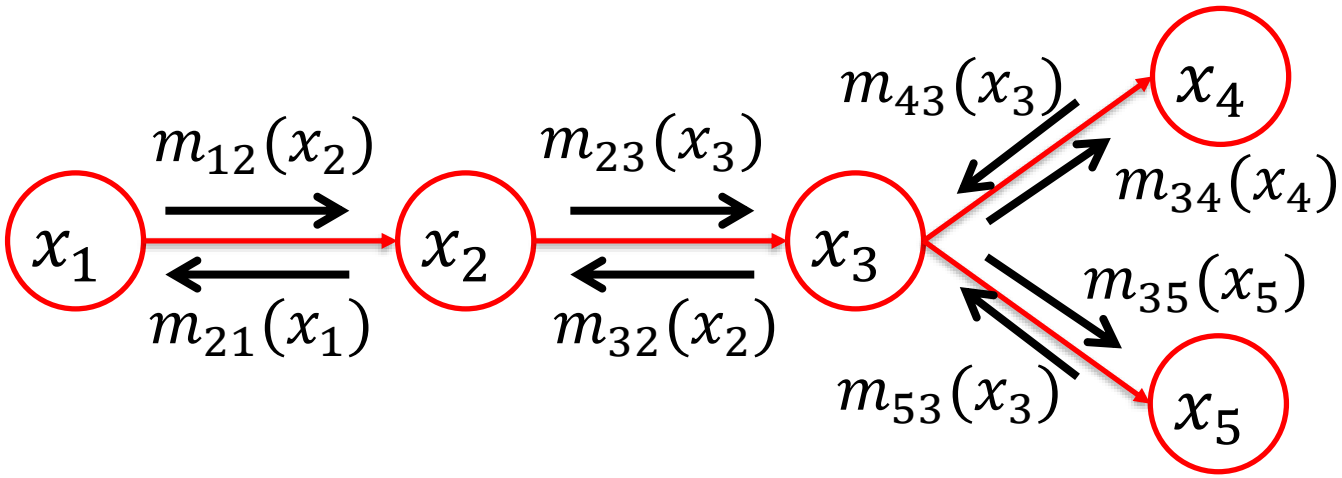
# 信念传播

- **消息传递在信念传播中的表示**
  - **在变量消除的示例中，公式11.2.1～11.2.4所描述的变量消除过程就能描述为如下的消息传递过程**



# 信念传播

- 信念传播的消息传递过程
  - 指定一个根节点（例如$x_1$），所有叶结点向根节点传递消息，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

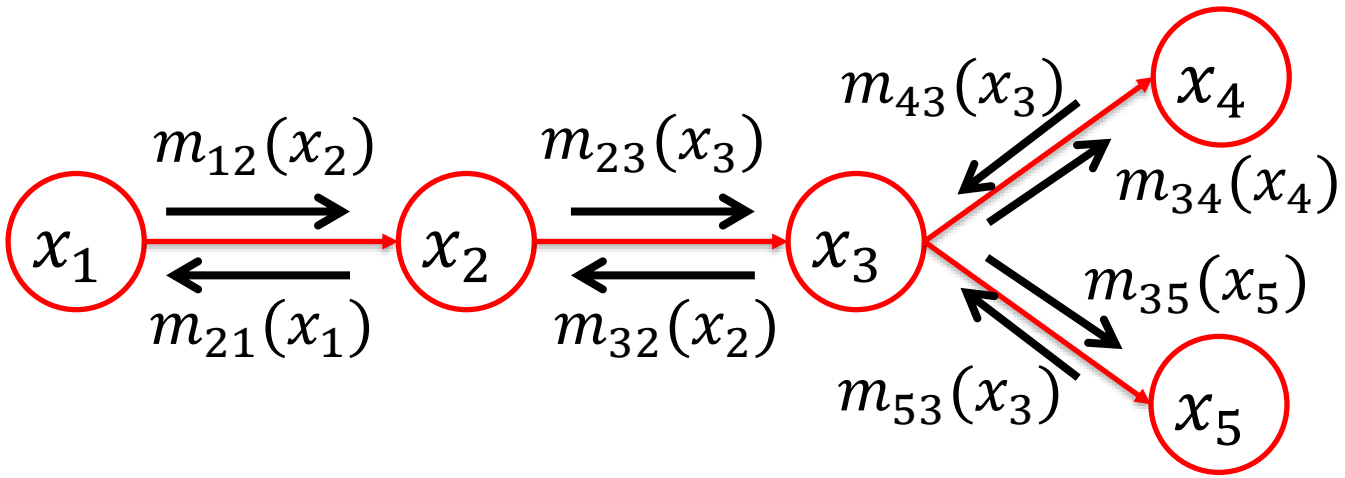


# 信念传播

- **边缘概率计算**
  - 在结束消息传播之后，对于节点$x_i$的边缘概率分布$\mathbf{p}(x_i)$，其正比于所接收消息的乘积。其中，$\alpha$为归一化因子。

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

  - 例如，在信念传播结束后，可得$p(x_3) = \alpha m_{23}(x_3) m_{43}(x_3) m_{53}(x_3)$



# 近似推理

- **近似推理**
  - 当**模型较为复杂**或**数据维度较高**时，精确推理方法（变量消除、信念传播）通常需要很大的计算开销，此时可以使用近似推理方法估计后验概率
  - 近似推理方法可大致分为两类
    - 基于采样的方法：通过使用随机化方法完成近似，如**马尔科夫链蒙特卡洛方法**（Markov Chain Monte Carlo，MCMC）（课件 12.5）
    - 使用已知的简单分布逼近需要推断的复杂分布：典型代表为**变分推断**（Variational Inference）

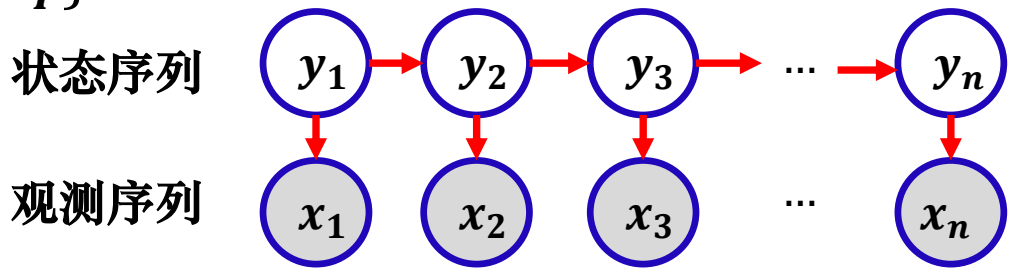
- **此部分内容推荐阅读**
  - ***Pattern Recognition and Machine Learning*** Christopher M. Bishop（第10章）
  - **《机器学习》**周志华 （第14章）

# 10.4 隐马尔可夫模型

- 认识隐马尔可夫模型
- 观测序列概率计算
- 模型参数学习
- 隐状态预测
- 隐马尔可夫模型的典型应用

# 隐马尔可夫模型(Hidden Markov Model)

- 什么是隐马尔可夫模型（HMM）
  - 是一种用于建模时间序列数据的有向概率图模型。由一个隐藏的马尔可夫链随机生成隐藏状态序列$\mathbf{Y} = \{y_1, y_2, \dots, y_T\}$，随后由状态产生观测随机序列$\mathbf{X} = \{x_1, x_2, \dots, x_T\}$




**齐次马尔可夫假设**：任意时刻的隐藏状态只依赖于它前一个隐藏状态
**独立观测假设**：任意时刻的观察状态仅依赖于当前时刻的隐藏状态

  - 隐马尔可夫模型的联合概率分布为：

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

# 隐马尔可夫模型(Hidden Markov Model)

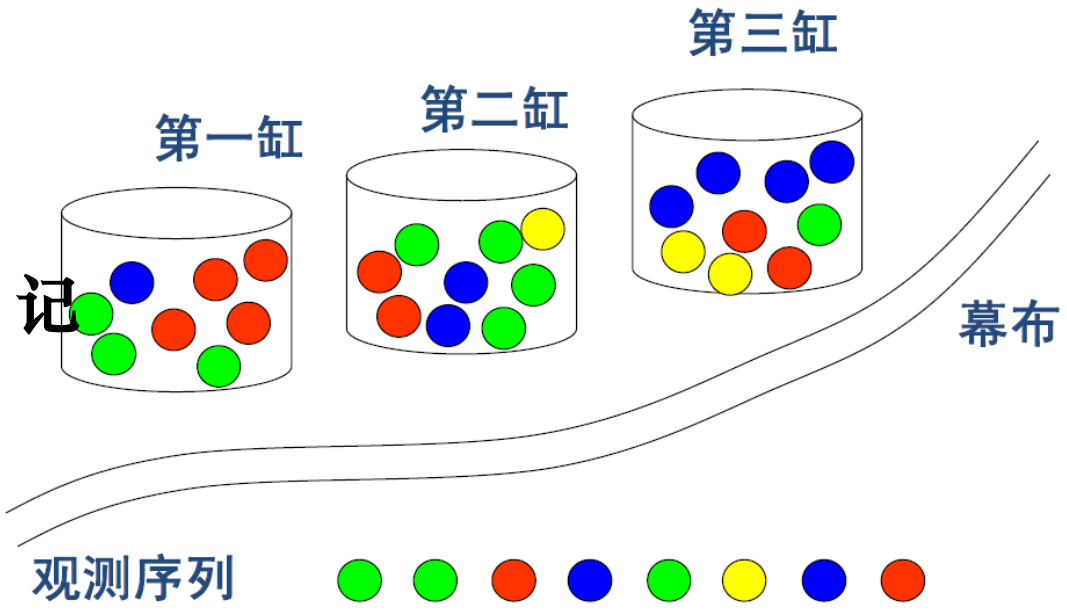
- 例子-采样球模型构建HMM观测序列

➢ 有N个缸，每个缸中装有很多个彩球，彩球共有M种颜色，**球的颜色即为观测值，缸的编号为隐藏状态**。

➢ 以**初始概率Π**从N个缸中进行采样

➢ 重复T次:从缸$y_t$中以采样概率B采样到一个球，记录颜色$x_t$ ,并放回,以转移概率A转移到缸$y_{t+1}$

➢ 得到观测序列$X = \{x_1, x_2, \dots, x_T\}$



# 隐马尔可夫模型(Hidden Markov Model)

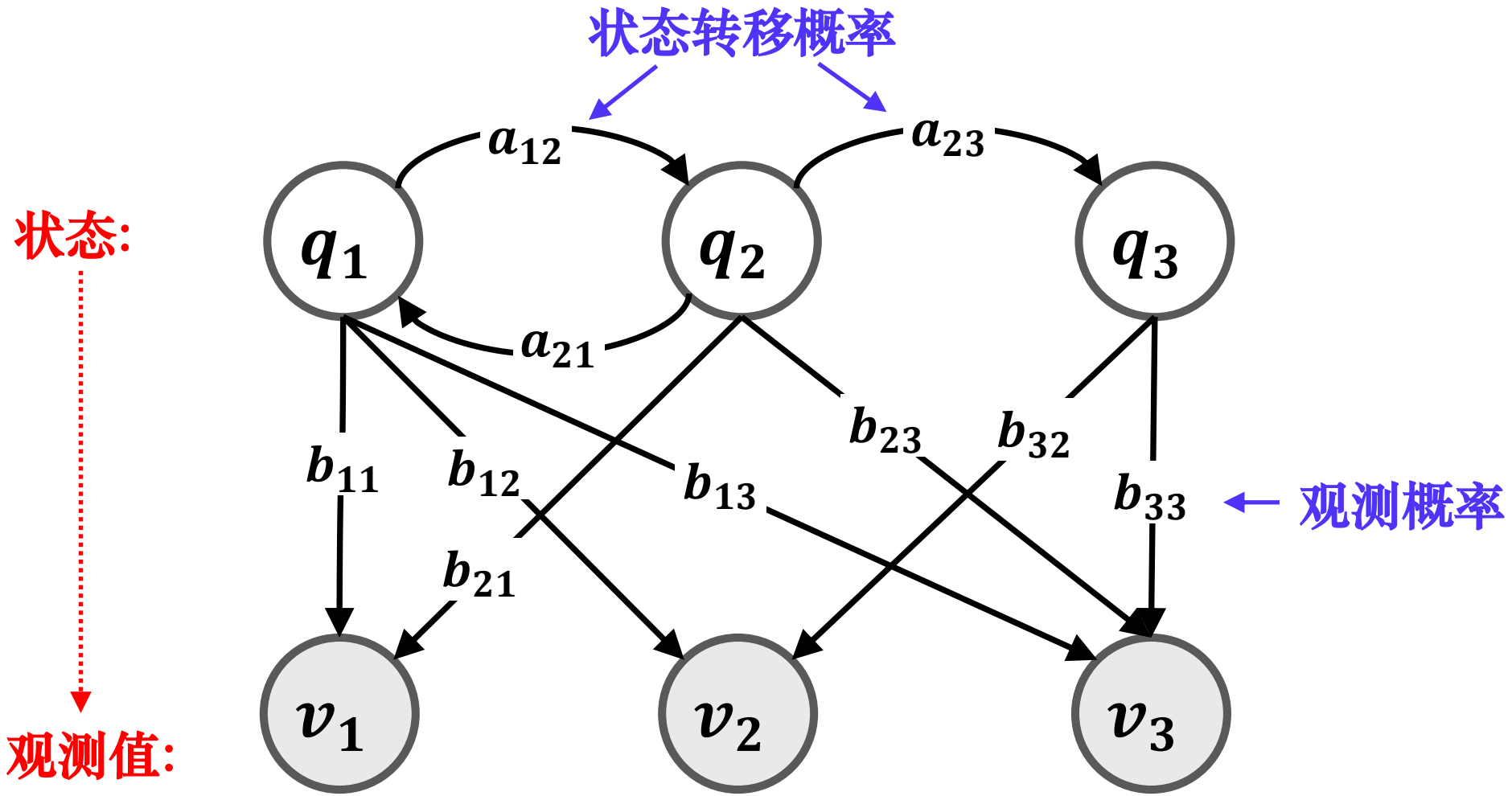
- **隐马尔可夫模型参数的定义：**

设隐藏状态集合$\mathrm{Q}=\{q_1,q_2,\ldots,q_N\}$和可能的观测状态集合$\mathrm{V}=\{v_1,v_2,\ldots,v_M\}$，则定义：

- 状态转移矩阵A=$[a_{ij}]_{N\times N}$，其中$\mathrm{a_{ij}}=\mathrm{P}(y_{t+1}=\mathrm{q_j}|y_t=\mathrm{q_i})$表示任意时刻t，从状态$q_i$转移到$q_j$的概率
- 观测概率矩阵B=$[b_{ij}]_{N\times M}$，其中$\mathrm{b_{ij}}=\mathrm{P}(x_t=v_i|y_t=\mathrm{q}_j)$表示任意时刻t，从状态$q_j$产生观测$v_i$的概率
- 此外定义t=1时隐藏状态概率分布$\Pi=[\pi_i]_{\mathrm{N}}$，其中$\pi_i=P(y_1=q_i)$

一个HMM模型，可以由三元组$\lambda=(\Pi,\mathrm{A},\mathrm{B})$进行表示

# 隐马尔可夫模型（Hidden Markov Models）



# 隐马尔可夫模型(Hidden Markov Model)

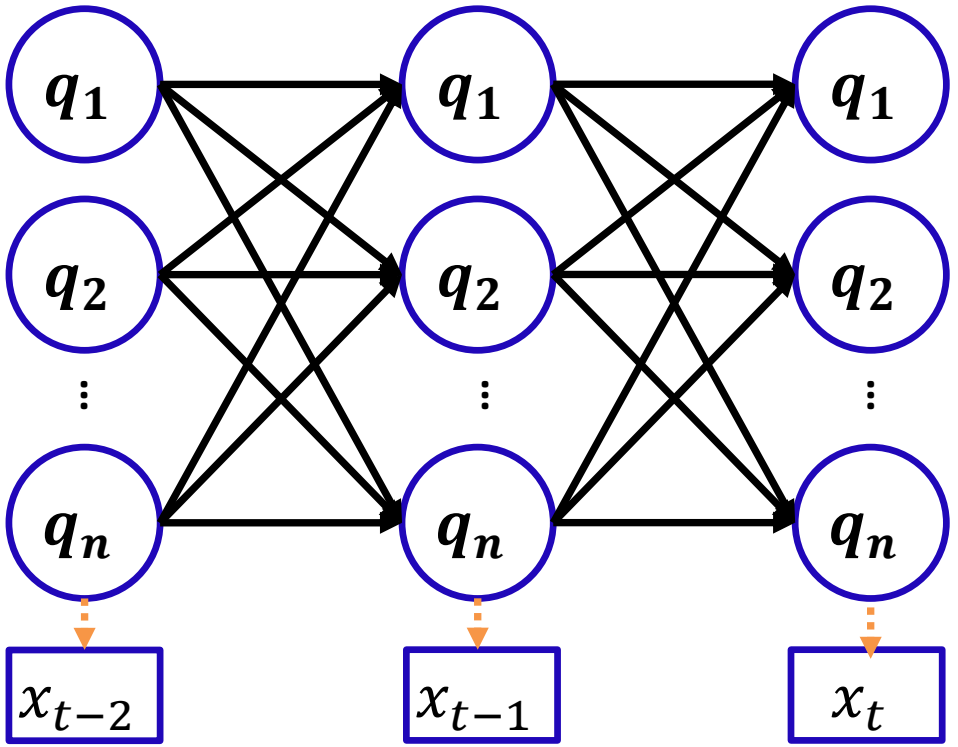
- 隐马尔可夫模型要解决的三个基本问题
  - **概率计算问题**：给定采样球颜色观测序列$\mathbf{X}=\{x_1,x_2,\dots,x_\mathrm{T}\}$，以及模型$\boldsymbol{\lambda}=(\boldsymbol{\Pi},\mathbf{A},\mathbf{B})$，如何计算产生该颜色序列的概率$\boldsymbol{P}(\boldsymbol{O}|\boldsymbol{\lambda})$
  - **学习问题**：给定采样球颜色观测序列$\mathrm{X}=\{x_1,x_2,\dots,x_\mathrm{T}\}$，估计模型参数$\boldsymbol{\lambda}=(\boldsymbol{\Pi},\mathbf{A},\mathbf{B})$，使得该模型下出现观测序列的概率最大
  - **预测问题**：给定采样球颜色观测序列$\mathrm{X}=\{x_1,x_2,\dots,x_\mathrm{T}\}$，以及模型$\boldsymbol{\lambda}=(\boldsymbol{\Pi},\mathbf{A},\mathbf{B})$，如何找到最合理的缸采样序列$\mathrm{Y}=\{y_1,y_2,\dots,y_T\}$

# HMM的概率计算问题

问题回顾：给定观测序列$X$和HMM模型$\lambda$，求$P(X|\lambda)$

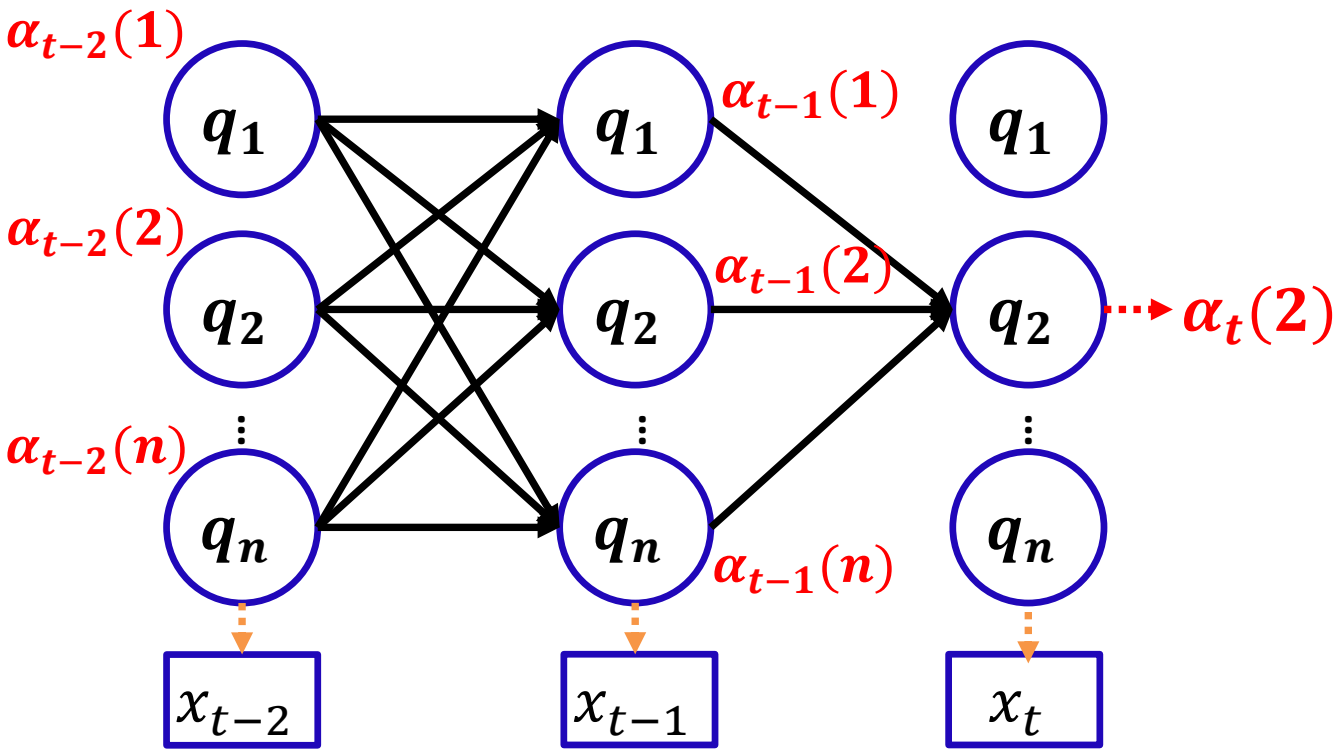
- 暴力求解：遍历所有$T$个时刻的所有$N$个可能状态

由于预测状态有$N^T$种组合，算法复杂度为$O(TN^T)$阶，因此隐藏状态多时计算量会非常大。



# HMM的概率计算问题

- 前向后向算法：利用信念传播法的思想，定义局部概率避免重复计算




**以前向算法为例**

- 定义前向概率$\alpha_t(i)$：当前观测下时刻t时隐藏状态为$q_i$的概率

- 前向概率递推公式为：

$$\alpha_t(i) = \left[\sum_{j=1}^{N} \alpha_{t-1}(j)\right] b_i(x_t)$$

# HMM的概率计算问题

- 前向算法流程

  - 计算$t = 1$时刻下各隐藏状态前向概率

$$\alpha_1(i) = \pi_i b_i(x_1)$$

  - 对t=2….T-1进行递推

$$\alpha_t(i) = \left[\sum_{j=1}^{N} \alpha_{t-1}(j)\right] b_i(x_t)$$

  - $t = T$时算法终止

$$P(O|\lambda) = \sum_{i=1}^{N} \alpha_T(i)$$

后向算法和前向算法基本相同但方向相反

# HMM的学习问题

**问题回顾：给定观测序列数据集$X$，估计HMM参数$\lambda = (\Pi, A, B)$**

- **将观测序列看做观测数据O，状态序列看做隐变量I，HMM可以被建模为含有隐变量的概率模型**

$$P(\boldsymbol{O}|\lambda) = \sum_{t} \boldsymbol{P}(\boldsymbol{O}|\boldsymbol{I}, \boldsymbol{\lambda})\boldsymbol{P}(\boldsymbol{I}|\boldsymbol{\lambda})$$

- **因此，HMM的学习问题可以用EM算法实现，通过迭代优化地方式，最大化观测数据的似然函数。**

# HMM的学习问题

- **EM算法估计HMM参数：鲍姆-韦尔奇（Baum-Welch）算法**

  - **E-step: 基于当前HMM参数估计值$\bar{\lambda}$，计算隐变量期望**

$$\begin{aligned} Q(\lambda, \bar{\lambda}) &= E_I[\log P(O|I,\lambda)|O,\bar{\lambda}] \\ &= \sum_I \log P(O,I|\lambda) P(O,I|\bar{\lambda}) \end{aligned}$$

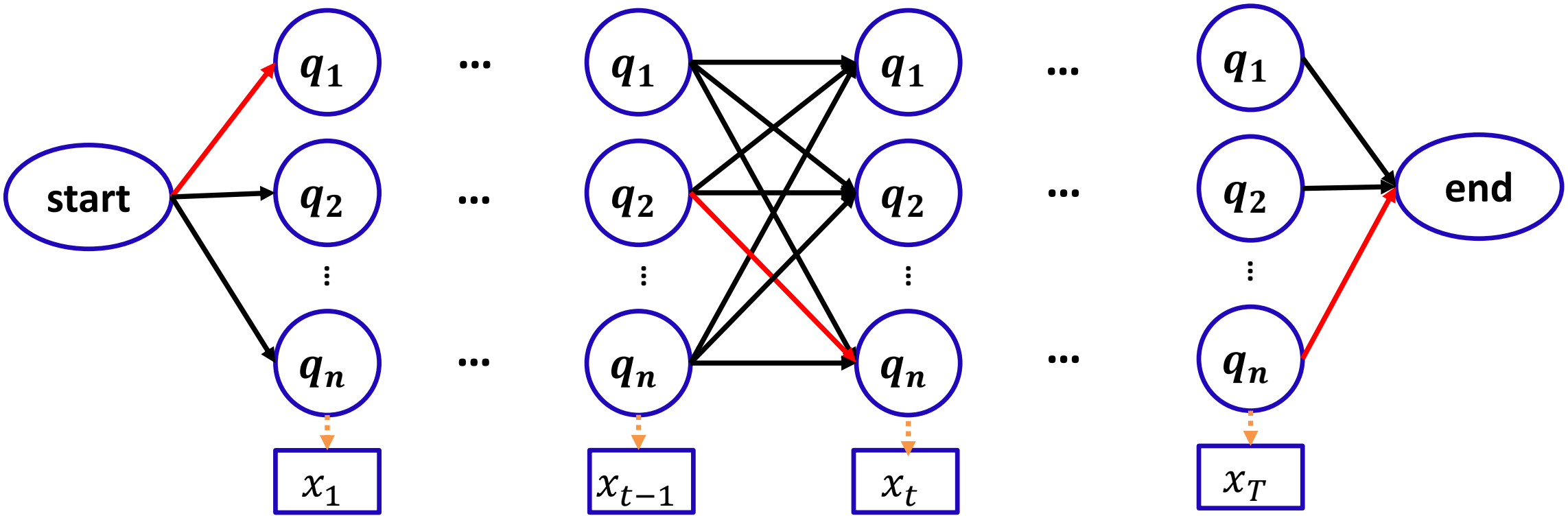
  - **M-step: 最大化似然函数，更新参数Π, A, B**

$$a_{ij} = \frac{\sum_{t=1}^{T-1} P(x_t = q_i|O,\lambda)}{\sum_{t=1}^{T-1} P(x_t = q_i|O,\lambda)} \quad b_j(k) = \frac{\sum_{t=1, y_t=v_k}^{T-1} P(x_t = q_j|O,\lambda)}{\sum_{t=1}^{T-1} P(x_t = q_i|O,\lambda)} \quad \pi_i = P(x_1 = q_i|O,\lambda)$$

详细推导过程：统计学习方法， 李航

# HMM的预测问题

问题回顾：给定观测序列$X$和HMM参数$\lambda$，找到概率最大的隐状态序列$Y^*$




- HMM的预测问题可以转化为在上图中找一条使$P(Y^*|X)$最大的路径

# HMM的预测问题

- 在该图中，如果最优路径在t时刻经过状态$q_t$,则对于t到T的后续所有路径，该路径最优

- **维特比算法（Viterbi）：**利用**动态规划**思想，在t时刻记录每个状态对应的最优路径和概率值，避免后续时刻的重复计算。首先定义两个局部状态用于记录:

- $\delta_t(q_i)$：时刻t状态为$q_i$的所有单个路径概率最大值

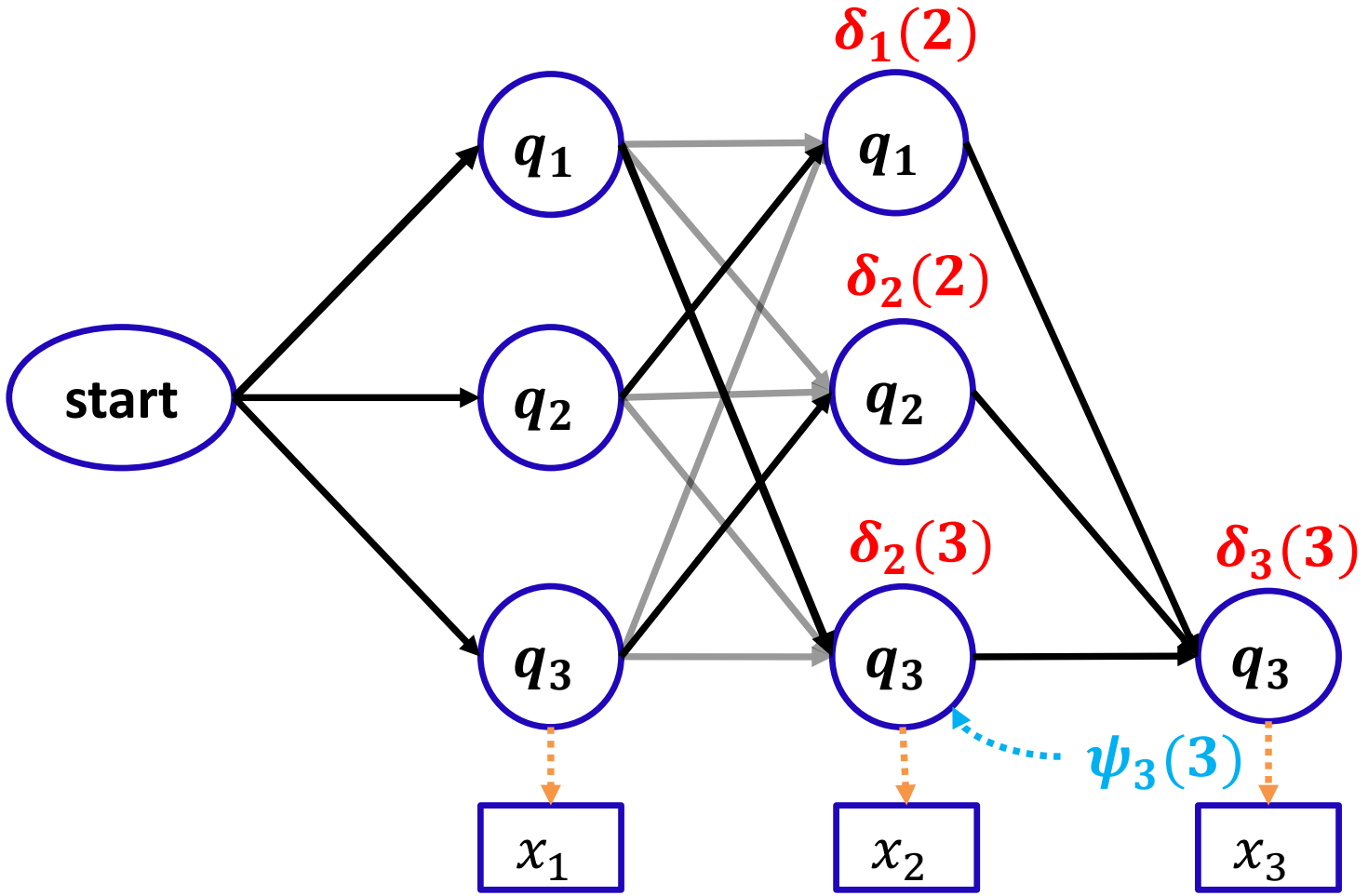
$$\delta_t(i) = \max_{1 \le j \le N} \left[\delta_{t-1}(j) a_{ji}\right] b_i(x_t)$$

- $\psi_t(q_i)$：时刻t获得最大路径概率的状态

$$\psi_t(i) = arg\max_{1 \le j \le N} \left[\delta_{t-1}(j) a_{ji}\right]$$

# HMM的预测问题

- 维特比算法




- 根据状态转移概率基于上一时刻的记录求$\delta_3(3)$:

$$\delta_3(3) = \max_{1 \le j \le 3}\left[\delta_2(j)a_{j3}\right]b_3(x_3)$$

- 选择概率最大的对应状态 计为$\psi_3(3)$

$$\psi_3(3) = arg\max_{1 \le j \le 3}\left[\delta_2(j)a_{j3}\right]$$

# HMM的预测问题

- **维特比算法流程**

➢ 初始化

$$\delta_1(q_i) = \pi_i b_i(x_1) \qquad \psi_1(i) = 0$$

➢ 正向递推，对$t = 2, 3, \dots, T$

$$\delta_t(i) = \max_{1 \le j \le N} \left[\delta_{t-1}(j) a_{ji}\right] b_i(x_t) \qquad \psi_t(i) = arg\max_{1 \le j \le N} \left[\delta_{t-1}(j) a_{ji}\right]$$

➢ T时刻递推终止

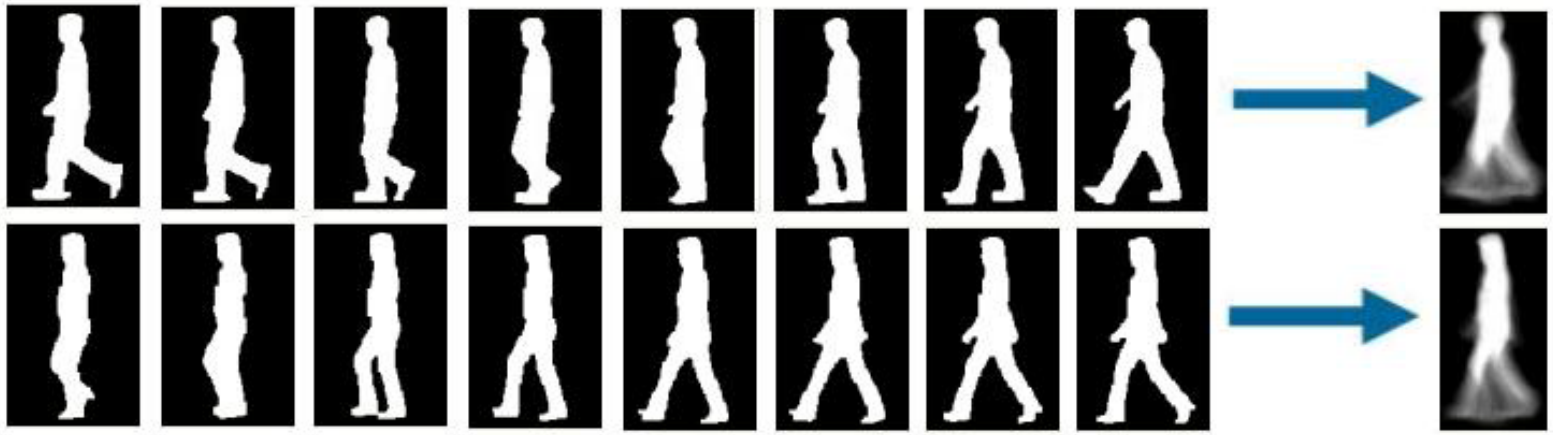
$$P^* = \max_{1 \le i \le N} \delta_T(i) \qquad Y_T^* = \arg\max_{1 \le i \le N} [\delta_T(i)]$$

➢ 路径回溯,对$t = T-1, T-2, \dots, 1$,得到最优路径

$$Y_t^* = \psi_{t+1}(i_{t+1}^*)$$

# 隐马尔可夫模型的应用

- 步态识别



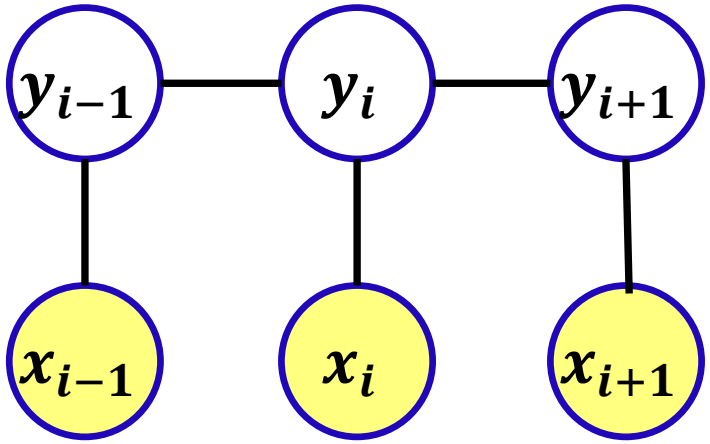
- 步态识别旨在通过连续行走的图像确定人的身份。对于HMM模型，一般将步态图像序列建模为观测序列，将人类行走中潜在的姿态作为隐状态。
- HMM进行步态识别的步骤：
  （1）模型学习：通过学习算法对数据集中每个人的步态用HMM建模
  （2）概率计算：给定测试步态序列，对每个模型使用前向后向算法计算序列概率
  （3）比较所有HMM模型的概率，选择最大概率的模型作为该序列ID

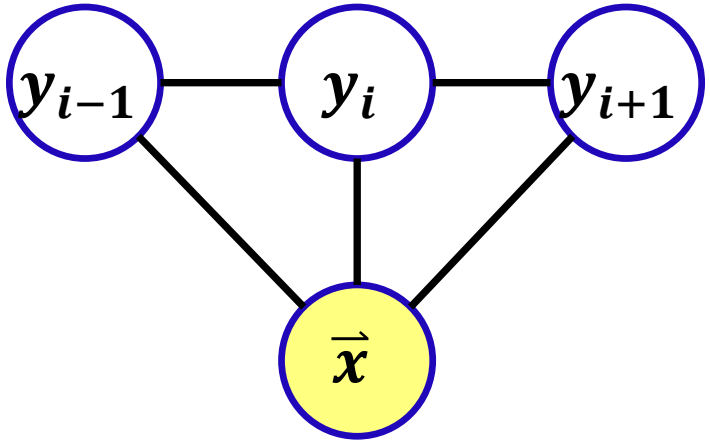
# 10.5 条件随机场

- 认识条件随机场
- 条件随机场的概率计算
- 模型参数学习
- 条件随机场的预测
- 条件随机场的典型应用

# 条件随机场(Conditional Random Field)

- 什么是条件随机场（CRF）？
  - 是一种判别式概率无向图模型，是在给定观测变量X的条件下，随机变量为Y的马尔可夫随机场（节点取值只和相邻位置有关）




  - 相比于隐马尔可夫模型，CRF没有对观测变量做马尔可夫假设，因此可以建模观测间的长距离关系

# 条件随机场(Conditional Random Field)

- **条件随机场的定义**
  - 在给定随机变量X的条件下，若随机变量Y构成一个由无向图$G=(V,E)$表示的马尔可夫场，即：

$$P(Y_v|X,Y_w,w\neq v)=P(Y_v|X,Y_w,w\sim v)$$

对任意节点v成立，则称条件概率分布$P(Y|X)$为条件随机场。式中$w\sim v$表示在图中和节点$v$有边连接的所有节点$w$,$w\neq v$表示节点$v$以外的所有节点。

- **线性链条件随机场**
  - 在实际情况下，通常考虑X和Y具有相同结构的线性链条件随机场，即给定随机变量$X=(X_1,X_2,\dots,X_n)$,$Y=(Y_1,Y_2,\dots,Y_n)$，满足：

$$P(Y_t|X,Y_1,\dots,Y_n)=P(Y_t|X,Y_{i-1},Y_{i+1})$$

# 条件随机场(Conditional Random Field)

- **CRF的参数化形式**

$$P(y|x = \frac{1}{Z(x)}\exp(\sum_{i,k}\lambda_k t_k(y_{i-1}, y_i, x, i) + \sum_{i,l}\mu_l s_l(y_i, x, i))$$

  - $t_k$为转移特征，定义边上的特征函数
  - $s_l$为状态特征，定义节点上的特征函数
  - $Z(x)$为规范化因子，$\lambda_k$和$\mu_l$为不同特征函数对应的权值

类似马尔可夫场的势函数，特征函数取值为1或0，例如：

$$t_1(y_{i-1}, y_i, x, i) = \begin{cases} 1, y_{i-1} = 1, y_i = 2, x, (i = 2, 3) \\ 0, otherwise \end{cases}$$

$$s_1(y_i, x, i) = \begin{cases} 1, y_i = 1, x, (i = 3) \\ 0, otherwise \end{cases}$$

# 条件随机场(Conditional Random Field)

- **CRF的参数化形式简化**

  - 设置 $f_k(y_{i-1}, y_i, x, i) = \begin{cases} t_k(y_{i-1}, y_i, x, i) \\ s_l(y_i, x, i) \end{cases}$ $w_k = \begin{cases} \lambda_k \\ \mu_l \end{cases}$

  - 对所有位置的特征函数求和

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

**CRF简化形式：** $$P(y|x) = \frac{1}{Z(x)} \exp \sum_{k=1}^{K} w_k f_k(y, x)$$

  - 根据CRF简化形式，在给定$x$时，从$y_{i-1}$转移到$y_i$的非规范化概率为：

$$M_i(y_{i-1}, y_i|x) = exp(\sum_{k=1}^{K} w_k f_k(y_{i-1}, y_i, x, i))$$

# 条件随机场(Conditional Random Field)

- 例子-词性标注

| | | | | | | |
|---|---|---|---|---|---|---|
| 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. |

- 给定一段文本作为观测X，我们希望标注对应单词的词性Y，用CRF进行建模
- 使用特征函数：

$$t_k(y_{i+1}, y_i, x, i) = \begin{cases} 1, if\ y_{i+1} = [p]\ y_i = [V], and\ x_i =' knock' \\ 0, otherwise \end{cases}$$

表示第i个观测值$x_i$为单词knock时，对应的$y_i, y_{i+1}$词性很可能为动词[V]和介词[P]

# 条件随机场(Conditional Random Field)

- 条件随机场要解决的三个基本问题：
  - **概率计算问题**：对于一段文本序列$x$和词性序列$y$，计算在CRF模型$P(Y|X)$下某位置的词性概率$P(Y_i = y_i|x)$和相邻位置联合概率$P(Y_i = y_i, Y_{i-1} = y_{i-1}|x)$
  - **学习问题**：给定用于训练的文本数据X和词性数据Y，如何学习CRF的模型参数$w_k$以及条件概率$P_w(y|x)$
  - **预测问题**：给定CRF模型$P(y|x)$，输入文本序列X，求对应的词性序列Y

# CRF的概率计算问题

问题回顾：给定条件序列$x$和状态序列$y = \{y_1, y_2, \dots, y_n\}$,求基于CRF模型$P(Y|X)$的条件概率$P(Y_i = y_i|x)$和$P(Y_i = y_i, Y_{i-1} = y_{i-1}|x)$

- 前向后向算法
  - 计算CRF的条件概率，可以使用和HMM中类似的前向后向算法，通过保留局部概率以简化整个序列的概率计算。
  - 和HMM不同，CRF是一个无向图，所以对于位置$i$的条件概率，需要计算$Y_1 - Y_i$（前向）和$Y_{i+1} - Y_n$（后向）两部分。

# CRF的概率计算问题

- 前向后向算法
  - 前向概率$\alpha_i(Y_i|x)$:表示位置$i$的标记为$Y_i$，且到位置$i$的前半部分状态序列和$y$相同的概率。递推公式为:

$$\alpha_{i+1}(Y_{i+1}|x) = \sum_{Y_i \in y} \alpha_i(Y_i|x) M_i(Y_i, Y_{i+1}|x)$$

  - 后向概率$\beta_i(Y_i|x)$:表示位置$i$的标记为$Y_i$，且位置$i$的后半部分状态序列和$y$相同的概率。递推公式为:

$$\beta_i(Y_i|x) = \sum_{Y_{i+1} \in y} \beta_{i+1}(Y_{i+1}|x) M_i(Y_i, Y_{i+1}|x)$$

  - 其中，$M_i(Y_i, Y_{i+1}|x)$ 为从$Y_i$转移到$Y_{i+1}$的概率

# CRF的概率计算问题

- **前向后向算法**
  - 按照前向后向的定义，很容易计算标记序列在位置$i$是标记$y_i$的条件概率和在位置$i-1$与$i$是标记$y_{i-1}$和$y_i$的条件概率：

$$P(Y_i = y_i|x) = \frac{\alpha_i(Y_i = y_i|x)\beta(Y_i = y_i|x)}{Z(x)}$$

$$P(Y_{i=1} = y_{i-1}, Y_i = y_i|x) = \frac{\alpha_{i-1}(Y_{i-1} = y_{i-1}|x)M_i(y_{i-1}, y_i|x)\beta(Y_i = y_i|x)}{Z(x)}$$

其中，归一化因子 $Z(x) = \sum^{n} \alpha_n(Y_n = y_i|x) = \sum^{n} \beta_1(Y_1 = y_i|x)$

# CRF的学习问题

问题回顾：给定训练数据集D $= \{(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)\}$和K个特征函数 $\{f_1, f_2, \dots, f_K\}$，求基于CRF模型参数$w_k$

- CRF的学习问题可以转化为有标签的极大似然估计问题，极大化基于训练数据D的对数似然函数:

$$L(w) = \sum_{x,y} \overline{P}(x,y) log P_w(y|x)$$

$\overline{P}(x,y)$表示基于训练数据D的样本经验分布

- 对参数$w_k$求导可得:

$$\frac{\partial L(w)}{\partial w} = -\sum_{x,y} \overline{P}(x) P_w(y|x) f(x,y) + \sum_{x,y} \overline{P}(x) f(x,y)$$

可用梯度下降法、迭代尺度法或拟牛顿法求解

# CRF的预测问题

**问题回顾：基于CRF模型$P(Y|X)$和条件序列$X$,求概率最大的状态序列$Y^*$**

- 对于线性链CRF的预测问题，使用HMM预测一样的维特比算法。基于动态规划思想，通过记录局部状态简化计算。通过CRF的概率表示定义两个局部状态：

  - $\delta_i(l)$：到达位置$i$且状态$y_i = l$所有可能路径的概率最大值

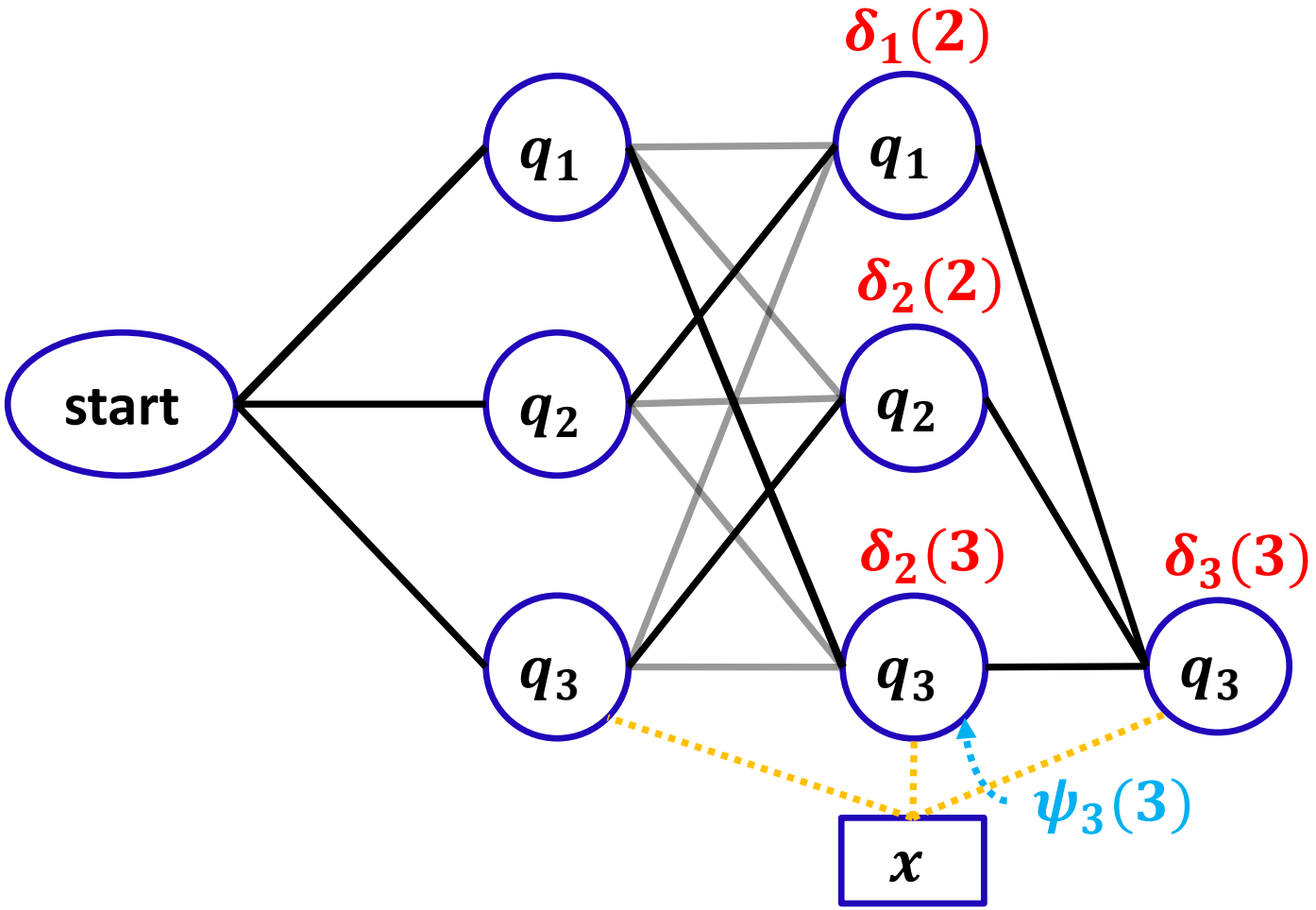
$$\delta_i(l) = \max_{1 \le j \le m} \{\delta_{i-1}(j) + \sum_{k=1}^{K} w_k f_k(y_{i-1} = j, y_i = l, x, i)\}$$

  - $\psi_i(l)$：到达位置$i$且状态$y_i = l$的最优路径中，位置$i-1$的状态取值

$$\Psi_i(l) = \underset{1 \le j \le m}{argmax} \{\delta_{i-1}(j) + \sum_{k=1}^{K} w_k f_k(y_{i-1} = j, y_i = l, x, i)\}$$

# CRF的预测问题

- 维特比算法




- 根据状态转移概率基于上一位置的记录求$\delta_3(3)$:

$$\delta_3(3) = \max_{1 \le j \le 3}\{\delta_2(j) + \sum_{k=1}^{K} w_k f_k\}$$

- 选择概率最大的对应状态 计为$\psi_3(3)$

$$\Psi_3(3) = \operatorname{argmax}_{1 \le j \le 3}\{\delta_2(j) + \sum_{k=1}^{K} w_k f_k\}$$

# CRF的预测问题

- 维特比算法流程

➢ 初始化 $\delta_1(l) = \sum_{k=1}^{K} w_k f_k(y_o = start, y1 = l, x) \quad \Psi_1(l) = start$

➢ 依据局部状态的递推公式正向递推，对i$= 2, 3, \dots, n$

$$\delta_i(l) = \max_{1 \le j \le m} \{\delta_{i-1}(j) + \sum_{k=1}^{K} w_k f_k(y_{i-1} = j, y_i = l, x, i)\}$$

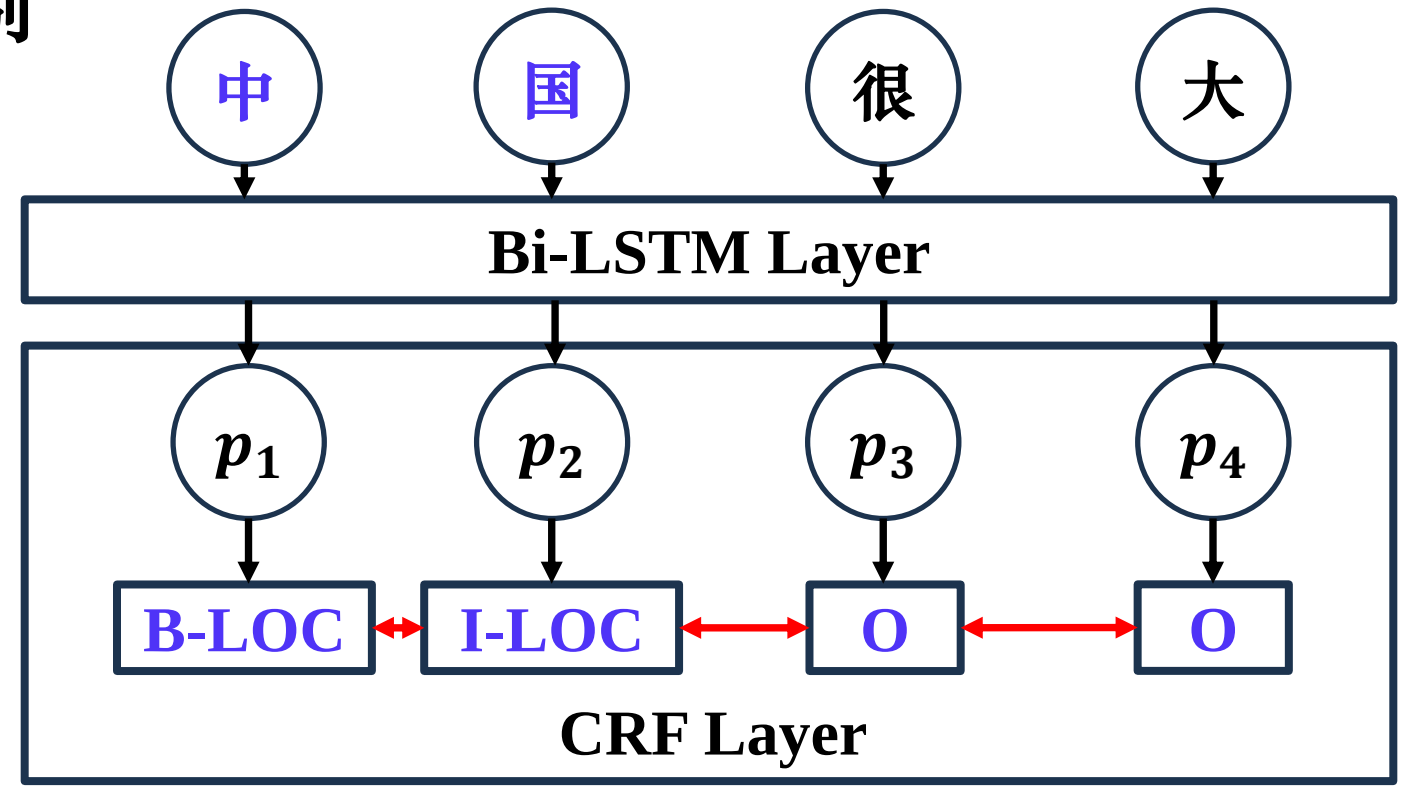
$$\Psi_i(l) = \underset{1 \le j \le m}{argmax} \{\delta_{i-1}(j) + \sum_{k=1}^{K} w_k f_k(y_{i-1} = j, y_i = l, x, i)\}$$

➢ 递推终止 $y_n^* = arg\ max_{1 \le j \le m} \delta_n(j)$

➢ 路径回溯,对$i = n-1, n-2, \dots, 1$,得到最优路径 $y_n^* = \Psi_{i+1}(y_{i+1}^*)$

# 条件随机场的典型应用

- 命名实体识别




- 名实体识别旨在将一串文本中的实体识别出来，比如人名、地名、机构名等等。
- 经过LSTM网络输出的概率分布作为观测X，对应的实体类别作为Y构建CRF模型
- CRF模型通过建模相邻标签间的依赖关系有效提取实体。