Machine learning

14. 概率图模型

汇报人：何思成

概率图模型

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，概率模型的核心在于基于可观测的变量推测出未知变量的条件分布 。生成式模型计算联合分布，比如同时建模天气和活动的关系；判别式模型直接计算条件分布，例如直接根据声音信号推测对应的文字。

- **概率模型**框架：
  - 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*: 目标变量集合
  - *O*: 可观测变量集合
  - *R*: 其他变量集合

概率图模型

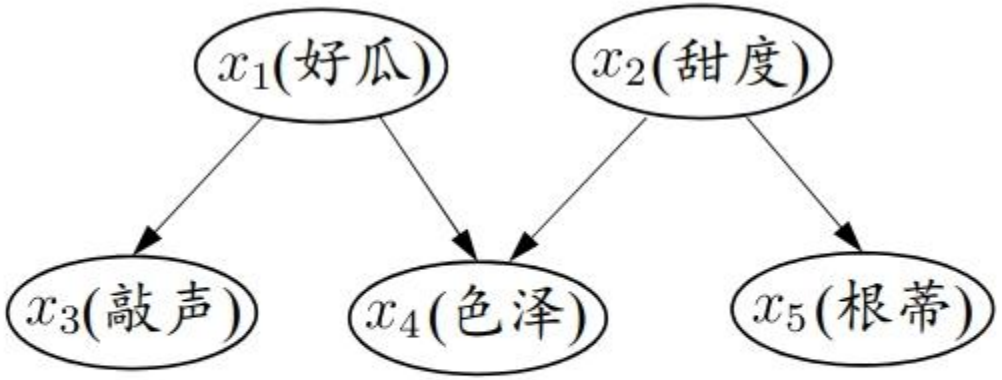
概率图模型

•定义：用图表示变量相关关系的概率模型，用节点表示变量，节点之间的边表示局部变量间的概率依赖关系

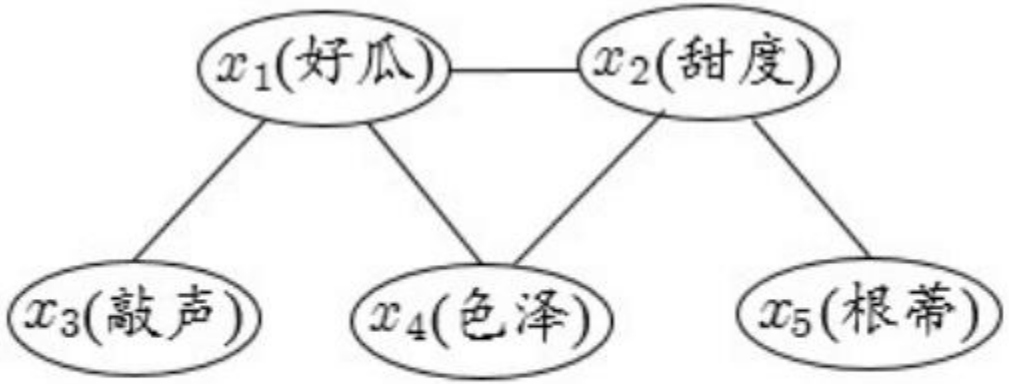
- 结点：随机变量
- 边：变量间依赖关系

•分类：

- **有向图**（贝叶斯网）
- **无向图**（马尔可夫网）



有向图



无向图

隐马尔可夫模型

隐马尔可夫模型（HMM）：是一种结构简单的动态贝叶斯网

- **齐次马尔可夫性**
- **观测独立性**

•组成:

- 状态变量 {y1,y2,...,yn}（隐变量）
- 观测变量 {x1,x2,...,xn}

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

•联合概率:

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

图中的箭头表示变量之间的依赖关系. 在任意时刻，观测变量的取值仅依赖于状态变量。系统下一时刻的状态仅由当前状态决定，不依赖于以往的任何状态

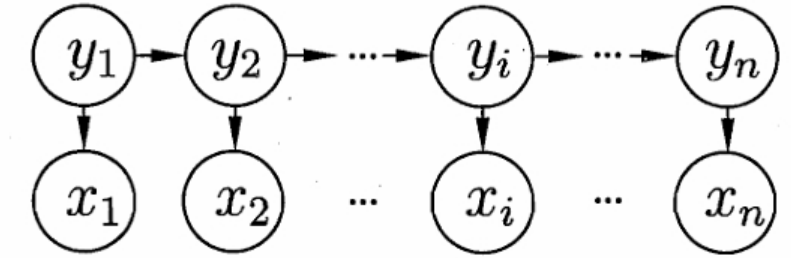


图 14.1 隐马尔可夫模型的图结构

隐马尔可夫模型

确定一个HMM需要三组参数

•**参数**:

- 状态转移概率：模型在各个状态间转换的概率
  表示在任意时刻t，若状态为s_i，下一状态为s_j的概率

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

- 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

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

隐马尔可夫模型

通过指定状态空间y， 观测空间 x和三组参数(状态转移概率、输出观测概率、初始状态概率）， 就能确定一个隐马尔可夫模型。给定λ=[A,B,π]， 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

隐马尔可夫模型

隐马尔可夫模型的三个基本问题:
对于模型 λ=[A,B,π]，给定观测序列{x1,x2,...,xn}

- **评估问题：**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(x| λ)
- **解码问题：**根据观测序列“推测”隐 藏的模型状态
- **参数学习问题：**如何调整模型 参数λ=[A,B,π], 以使得该序列出现的概率 P(x| λ) 最大

具体应用

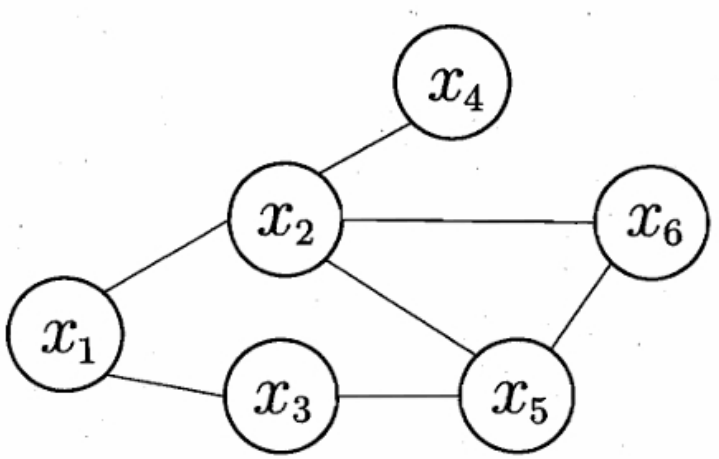
- 根据以往的观测序列 x={x1,x2,...,xn} 预测当前时刻最 有可能的观测值 xn
- 语音识别：根据观测的语音 信号推测最有可能的状态序 列（即：对应的文字）
- 通过数据学习参数（模型训练）

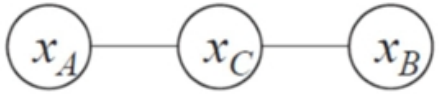
马尔可夫随机场

马尔可夫随机场（MRF）
用节点表示变量，节点之间的边表示局部变量间的概率依赖关系
团和极大团：若其中任意两节点间都有边连接，则称该结点子集为团。如果在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”
•典型的无向图模型，基于极大团定义联合分布（基于极大团分解为多个因子的乘积）：

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

- ψQ: 团Q的势函数（势函数用以描述团内变量的相关性）
- Z: 归一化因子（保证概率和为1）





假定变量均为二值变量，定义势函数：

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

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

马尔科夫随机场

得到图模型的联合概率为：

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

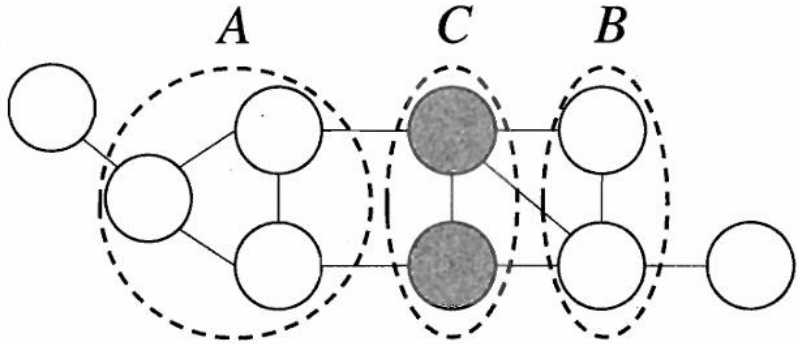


图 14.3 结点集 A 和 B 被结点集 C 分离

- **马尔可夫性**：
  - 全局马尔可夫性：给定分离集，变量子集条件独立
  - 局部马尔可夫性：给定邻接变量，节点独立于其他变量
  - 成对马尔可夫性：给定所有其他变量，两非邻接变量条件独立

条件随机场

条件随机场（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x={x1,x2,...,xn}为观测序列，y={y1,y2,...,yn}为对应的标记序列，CRF的目标是构建条件概率模型P(y|x)

- 条件随机场使用势函数和图结构上的团来定义条件概率P(y | x)

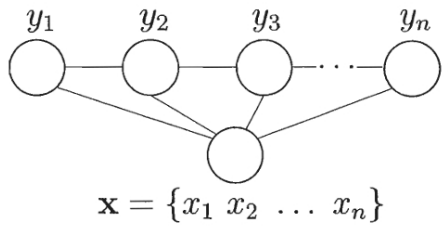
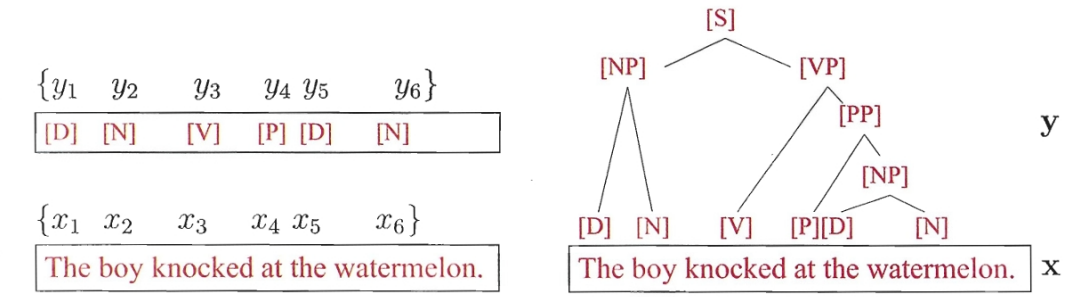


图 14.6 链式条件随机场的图结构

- **判别式模型**，建模 P(Y | X)
- **特征函数**：
  - 转移特征 $t_j(y_{i+1}, y_i, x, i)$ 捕捉相邻标签的关系（如动词后接介词）
  - 状态特征 $s_k(y_i, x, i)$ 捕捉观测与标签的关系（如“knock”对应动词）
- **条件概率公式**：

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

- 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。



(a) 词性标注　　(b) 语法分析

图 14.5 自然语言处理中的词性标注和语法分析任务

推断

- 基于概率图模型定义的分布，可以对目标变量的边际分布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

假设图模型所对应的变量集$x=\{x_1,x_2,x_3,x_4\cdots x_n\}$能分为$X_E$ 和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$p(X_F)$或者条件概率$p(X_F|X_E)$。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$p(X_F, X_E)$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

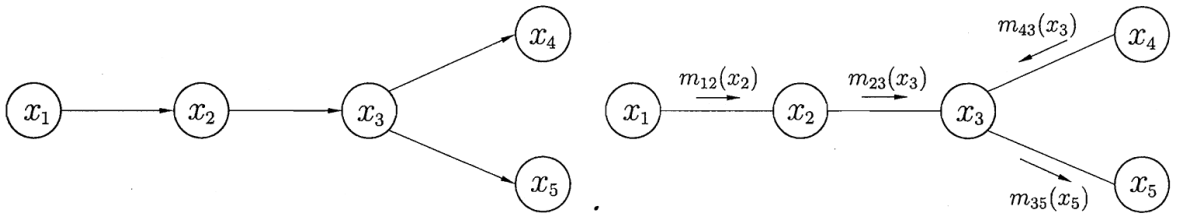
精确推断

概率图模型推断方法主要包括：
1. 精确推断（极大团规模较大时计算复杂度指数级上升）
2. 近似推断（在较低的时间复杂度下获得原问题的近似解）

1.精确推断：

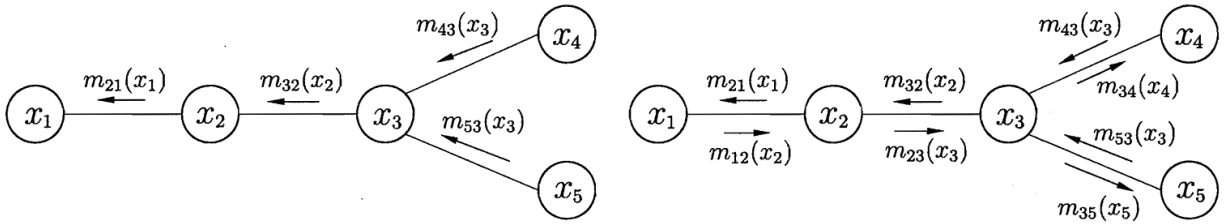
1. **变量消去法**：按顺序边缘化无关变量（动态规划思想），适合小规模网络。
    1. 示例：计算 P(x5) 时，逐步消去 x1,x2,x3,x4*x1,x2,x3,x4*。

2. **信念传播**：通过消息传递计算所有节点的边际分布，适用于树状结构（如链式CRF）。
    1. 示例：图中，结点x3 要向x5发送消息，必须事先收到来自结点x2和 x4的消息,且传递到x5 的消息m35(x5)恰为概率p(x5)



(a) 贝叶斯网络结构　　(b) 消息传递过程

图 14.7　变量消去法及其对应的消息传递过程



(a) 消息传向根结点　　(b) 消息从根结点传出

图 14.8　信念传播算法图示

近似推断

2. 近似推断:

1. **采样法**基于若直接计算或逼近概率分布的期望比推断概率分布更容易这个思路，假定目标是计算函数f（x）在概率密度函数p（x） 下的期望，则可根据p（x）抽取一组样本{x1,x2,x3⋯xn} ,然后计算f（x）在这些样本上的均值。
    1. **MCMC采样：**先设法构造一条马尔可夫链（吉布斯采样每次更新一个变量），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 。优点是精确但收敛慢。

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

Metropolis-Hastings（MH）算法

•步骤:

- **初始化**：随机选择初始状态 x0。
- **生成候选**：从提议分布 Q(x∗|xt−1) 采样候选样本 x∗。
- **计算接受率**：

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

- **接受/拒绝**：以概率 *A* 接受 x∗，否则保留 xt−1
- **迭代至收敛**

近似推断

2. 变分推断：用简单分布（如高斯分布）来逼近需推断的复杂分布，优化证据下界（ELBO），从而得到一种局部最优、但具有确定解的近似后验分布。速度快但有近似误差。

假设N个变量{x1,x2,x3….xn}均依赖于其他变量z，所有能够观察到的变量x的联合分布的概率密度函数对应的对数似然函数是：

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

- 推断任务是由观察到的变量x来估计隐变量z和分布参数变量Θ，即求解p(z|x, Θ)和Θ。
- 对对数似然函数使用EM算法求解，用EM算法求解的分布q(z)是一个近似分布

EM算法：

1. E步：计算隐变量的后验分布
2. M步：基于期望值最大化似然函数

•**目标**：用简单分布族 $q(z)$ 近似复杂后验 $p(z \mid x)$，最小化KL散度 $\mathrm{KL}(q\|p) = \mathbb{E}_q\left[\log \frac{q(z)}{p(z|x)}\right]$

•**证据下界（ELBO）**：$\log p(x) \geq \mathrm{ELBO} = \mathbb{E}_q[\log p(x,z)] - \mathbb{E}_q[\log q(z)]$

实现步骤

1.选择变分族：

1. 例如平均场假设：$q(z)=\prod_i q_i(z_i)$ 。

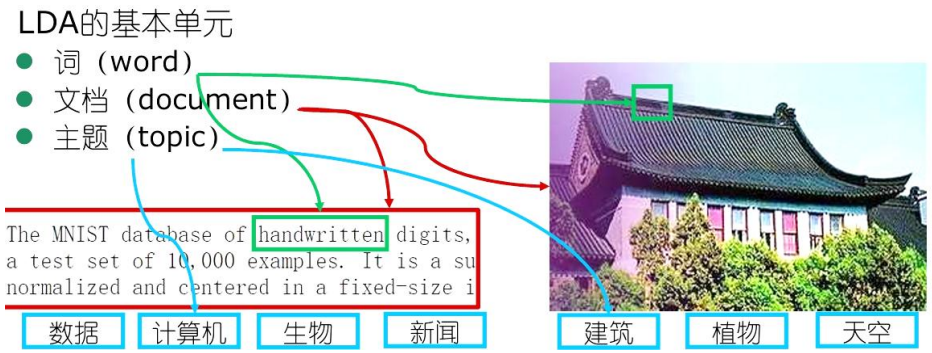
2.初始化参数：设定 q 的初始参数（如高斯分布的均值和方差）。

3.优化迭代：

1. 梯度下降或坐标上升法调整 q 的参数以最大化ELBO。

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



核心思想

•文档是**多个主题的混合**，每个词通过以下步骤生成：

- 表示采样文档包含的的主题分布：$\vartheta \sim \mathrm{Dir}(\alpha)$
- 对每个词：
  - 根据ϑ_t进行话题指派，得到文档t中词n的采样话题：$z_n \sim \mathrm{Mult}(\vartheta)$
  - 根据指派的话题所对应的词频分布随机采样生成采样词：$w_n \sim \mathrm{Mult}(\beta z_n)$

联合概率分布： $p(w,z,\theta,\beta|\alpha,\eta) = \prod_k p(\beta_k|\eta) \prod_d p(\theta_d|\alpha) \prod_n p(z_{d,n}|\theta_d) p(w_{d,n}|z_{d,n},\beta)$

感谢聆听！