第十四章 概率图模型

14.1 隐马尔科夫模型

14.1.1 概率模型

机器学习最重要的任务, 是**根据一些已经观察到的证据** (例如训练样本) 来**对**感兴趣的**未知变量** (例如类别标记) 进行**估计和推测.**

概率模型将学习任务归结于**计算变量的概率分布**. 在概率模型中, 利用已知变量推测未知变量的分布称为"推断", 其核心是**如何基于可观测变量推测出未知变量的条件分布**.

具体的说, 假定所关心的变量集合为 Y , 可观测变量集合为 O , 其他变量的集合为 R .

- 生成式模型: 考虑联合分布 $P(Y,R,O)$
- 判别式模型: 考虑条件分布 $P(Y,R|O)$

在给定一组观测变量值, 推断就是要由 $P(Y,R,O)$ **或** $P(Y,R|O)$ **得到条件概率分布** $P(Y|O)$

14.1.2 概率图模型

概率图模型 (probabilistic graphical model) 是一类**用图来表达变量相关关系的概率模型**.

- 它以图为表示工具, 最常见的是用**一个结点**表示**一个或一组随机变量**
- 结点之间的**边**表示**变量间的概率相关关系**, 即"变量关系图"

根据边的性质的不同, 概率图模型可大致分类两类:

- **有向图模型**或**贝叶斯网** (Bayesian network): 第一类是**使用有向无环图表示变量间的依赖关系**, 称为有向图模型或贝叶斯网 (Bayesian network);
- **无向图模型**或**马尔可夫网** (Markov network): 第二类是**使用无向图表示变量间的相关关系**, 称为无向图模型或马尔可夫网 (Markov network).

14.1.3 隐马尔科夫模型 (HMM)

隐马尔可夫模型 (Hidden Markov Model，简称 HMM) 是结构最简单的**动态贝叶斯网** (dynamic Bayesian network), 这是**一种著名的有向图模型**, 主要用于时序数据建模, 在语音识别、自然语言处理等领域有广泛应用.

1 隐马尔科夫中变量的概念

如图 14.1 所示

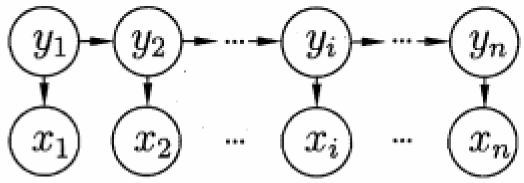

图 14.1 隐马尔可夫模型的图结构

隐马尔可夫模型中的变量可分为**两组**,

- **状态变量**: 第一组是状态变量 $\{y_1, y_2, \ldots, y_n\}$, **其中 $y_i \in \mathcal{Y}$ 表示第 i 时刻的系统状态**. 通常假定**状态变量是隐藏的、不可被观测**的, 因此状态变量亦称**隐变量** (hidden variable).
- **观测变量**: 第二组是现测变量 $\{x_1, x_2, \ldots, x_n\}$, **其中 $x_i \in \mathcal{X}$ 表示第 i 时刻的观测值**

在隐马尔可夫模型中, 系统通常在多个状态 $\{s_1, s_2, \ldots, s_N\}$ 之间转换, 因此**状态变量 y_i 的取值范围** $\mathcal{Y}$ (称为状态空间) 通常是有 N 个可能取值的离散散空间. **观测变量** x_i 可以是**离散型也可以是连续型**. 这里, 仅考虑离散型观测变量, 并假定其取值范围 $\mathcal{X}$ 为 $\{o_1, o_2, \ldots, o_M\}$

2 隐马尔科夫中变量间的依赖关系

图 14.1 中的箭头表示了变量间的依赖关系.

- **观测变量的取值**: 在任一时刻, **观测变量的取值仅依赖于状态变量, 即 x_t 由 y_t 确定**, 与其他状态变量及观测变量的取值无关.
- 状态变量的取值: t **时刻的状态 y_t 仅依赖于 $t-1$ 时刻的状态仙 y_{t-1}** , 与其余 $n-2$ 个状态无关.

这就是所谓的**"马尔可夫链"** (Markov chain) , 即: **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**.

基于这种依赖关系, 所有变量的联合概率分布为:

$$P\left(x_1, y_1, \ldots, x_n, y_n\right)=P\left(y_1\right) P\left(x_1 | y_1\right) \prod_{i=2}^n P\left(y_i | y_{i-1}\right) P\left(x_i | y_i\right) \tag{14.1}$$

3 隐马尔科夫中的另外三组参数

- **状态转移概率**: **模型在各个状态间转换的概率**, 通常记为矩阵 $\mathbf{A}=\left[a_{ij}\right]_{N \times N}$, 其中:
  $$a_{ij}=P\left(y_{t+1}=s_j | y_t=s_i\right), \quad 1 \leqslant i, j \leqslant N$$
  表示在任意时刻 t , 若状态为 s_i , 则在下一时刻状态为 s_j 的概率.
- **输出观测概率**: **模型根据当前状态获得各个观测值的概率**, 通常记为矩阵 $\mathbf{B}=\left[b_{ij}\right]_{N \times M}$, 其中:
  $$b_{ij}=P\left(x_t=o_j | y_t=s_i\right), \quad 1 \leqslant i \leqslant N, 1 \leqslant j \leqslant M$$
  表示在任意时刻 t , 若状态为 s_i , 则观测值 o_j 被获取的概率.
- **初始状态概率**: **模型在初始时刻各状态出现的概率**, 通常记为 $\boldsymbol{\pi}=\left(\pi_1, \pi_2, \ldots, \pi_N\right)$, 其中:
  $$\pi_i=P\left(y_1=s_i\right), \quad 1 \leqslant i \leqslant N$$
  表示模型的初始状态为 s_i 的概率.

通过指定**状态空间** $\mathcal{Y}$ 、**观测空间** $\mathcal{X}$ 和**上述三组参数**, 就能**确定一个隐马尔科夫模型**. 通常用其参数 $\lambda=[\mathbf{A}, \mathbf{B}, \boldsymbol{\pi}]$ 来指代.

给定隐马尔可夫模型 λ , 它按如下过程产生观测序列 $\{x_1, x_2, \ldots, x_n\}$:

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

其中, $y_t \in \{s_1, s_2, \ldots, s_N\}$ 和 $x_t \in \{o_1, o_2, \ldots, o_M\}$ 分别为第 t 时刻的状态和观测值

14.2 马尔可夫随机场

14.2.1 马尔可夫随机场的基本概念

马尔可夫随机场 (Markov Random Field, 简称 MORA) 是**典型的马尔可夫网**, 是一种著名的**无向图模型**.

圈中**每个结点**表示**一个或一组变量**, **结点之间的边**表示**两个变量之间的依赖关系**.

马尔可夫随机场有**一组势函数** (potential functions) , 亦称**"因子"** (factor) , 这是**定义在变量子集上的非负实函数**, 主要用于定义概率分布函数.

- **团 (clique):** 对于图中结点的一个子集, 若其中**任意两结点间都有边连接**, 则称**该结点子集**为一个"团" (clique).

- **极大团 (maximal clique)**: 若在一个团中**加入另外任何一个结点都不再形成团**, 则称该团为"极大团"(maximal clique); 换句话说, 极大团就是不能被其他团所包含的团.

图 14.2 显示出一个简单的马尔可夫随机场

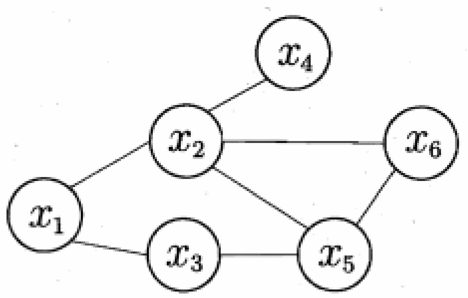

图 14.2　一个简单的马尔可夫随机场

例如, 在图 14.2 中关于团和极大团的结点子集有:

团: $\{x_1, x_2\}$, $\{x_1, x_3\}$, $\{x_2, x_4\}$, $\{x_2, x_5\}$, $\{x_2, x_6\}$, $\{x_3, x_5\}$, $\{x_5, x_6\}$ 和 $\{x_2, x_5, x_6\}$

极大团: 除了 $\{x_2, x_5\}$, $\{x_2, x_6\}$ 和 $\{x_5, x_6\}$ 这三个团以外, 都是极大团.

14.2.2 马尔可夫随机场的数学表示

在马尔可夫随机场中, **多个变量之间的联合概率分布**能**基于团分解为多个因子的乘积**, **每个因子仅与一个团相关**.

具体的, 对于 n 个变量 $\mathbf{x}=\{x_1, x_2, \ldots, x_n\}$, 所有团构成的集合为 $\mathcal{C}$, 与团 $Q \in \mathcal{C}$ 对应的变量结合记为 $\mathbf{x}_Q$, 则联合概率 $P(\mathbf{x})$ 定义为:

$$P(\mathbf{x})=\frac{1}{Z} \prod_{Q \in \mathcal{C}} \psi_Q\left(\mathbf{x}_Q\right) \tag{14.2}$$

其中 ψ_Q 为与团 Q 对应的势函数, 用于对团 Q 中的变量关系进行建模, $Z=\sum_{\mathbf{x}} \prod_{Q \in \mathcal{C}} \psi_Q\left(\mathbf{x}_Q\right)$ 为规范化因子, 以确保 $P(\mathbf{x})$ 是被正确定义的概率.

注意到, 若上述团 Q 不是极大团, 那么它必被一个极大团 Q^* 所包含, 即 $\mathbf{x}_Q \subseteq \mathbf{x}_{Q^*}$, 易知, 变量 $\mathbf{x}_Q$ 之间的关系不仅体现在势函数 ψ_Q 中, 还体现在 ψ_{Q^*} 中. 于是, 联合概率 $P(\mathbf{x})$ 可基于极大团来定义. 假定所有极大团构成的集合为 $\mathcal{C}^*$, 则有:

$$P(\mathbf{x})=\frac{1}{Z^*} \prod_{Q \in \mathcal{C}^*} \psi_Q\left(\mathbf{x}_Q\right) \tag{14.3}$$

其中 $Z^*=\sum_{\mathbf{x}} \prod_{Q \in \mathcal{C}^*} \psi_Q\left(\mathbf{x}_Q\right)$ 为规范化因子.

例子:

如图 14.2 中 $\mathbf{x}=\{x_1, x_2, \ldots, x_6\}$, 联合概率分布 $P(\mathbf{x})$ 定义为:

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

14.2.3 马尔可夫随机场的条件独立性

1 分离的概念

如图 14.3 所示

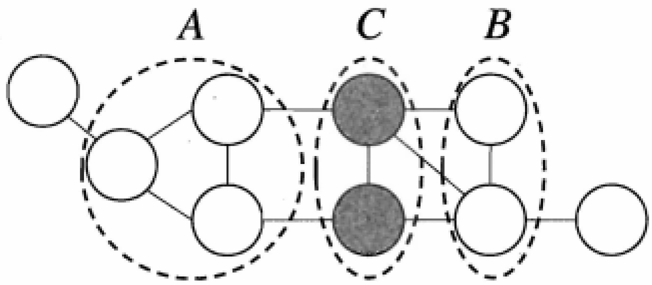

图 14.3 结点集 A 和 B 被结点集 C 分离

若**从结点集 A 中的结点到 B 中的结点都必须经过结点集 C 中的结点**, 则称结点集 A 和 B 被结点集 C 分离, C 称为**"分离集"** (separating set).

2 全局马尔可夫性

对马尔可夫随机场, 有:

"全局马尔可夫性" (global Markov property): **给定两个变量子集的分离集, 则这两个变量子集条件独立.**

例如, 图 14.3 中, 若令 A , B 和 C 对应的变量集分别为 $\mathbf{x}_A, \mathbf{x}_B$ 和 $\mathbf{x}_C$, 则 $\mathbf{x}_A$ 和 $\mathbf{x}_B$ 在给定 $\mathbf{x}_C$ 的条件下独立, 记为: $\mathbf{x}_A \perp \mathbf{x}_B | \mathbf{x}_C$

3 全局马尔可夫性的两个推论

- **局部马尔可夫性** (local Markov property): 给定某变量的邻接变量, 则该变量条件独立于其他变量
- **成对马尔可夫性** (pairwise Markov property): 给定所有其他变量, 两个非邻接变量条件独立.

14.2.4 马尔可夫随机场中的势函数

势函数 $\psi_Q\left(\mathbf{x}_Q\right)$ 的作用是定量刻画变量集 $\mathbf{x}_Q$ 中**变量之间的相关关系**, 它应该是**非负函数**, 且在**所偏好的变量取值上有较大函数值**.

为了满足非负性, 指数函数常被用于定义势函数, 即

$$\psi_Q\left(\mathbf{x}_Q\right)=e^{-H_Q\left(\mathbf{x}_Q\right)} \tag{14.8}$$

$H_Q\left(\mathbf{x}_Q\right)$ 是一个定义在变量 $\mathbf{x}_Q$ 上的实值函数, 常见的形式为:

$$H_Q\left(\mathbf{x}_Q\right)=\sum_{u, v \in Q, u \neq v} \alpha_{u v} x_u x_v+\sum_{v \in Q} \beta_v x_v \tag{11.9}$$

其中, $\alpha_{u v}$ 和 β_v 是参数.