第九章 概率模型

第 26 讲 概率模型与图表示

黄定江

DaSE @ ECNU

djhuang@dase.ecnu.edu.cn

1. 26.1 概率模型的有向图表示
2. 26.2 概率模型的无向图表示

概率图模型

- 机器学习中很多模型会涉及多元随机向量的概率分布。
- 如果采用单个函数来描述整个随机变量的联合分布是非常低效的 (无论是计算上还是统计上)，因为这些随机变量中涉及到的直接相互作用通常只介于非常少的变量之间的。
- 利用随机变量之间的条件独立性关系，可以将随机变量的联合分布分解为一些因式的乘积，得到简洁的概率表示。
- 我们可以采用图论中的“图”的概率来表示这种分解，得到概率图模型：图中的节点表示随机变量，边表示随机变量之间的直接作用。
- 有向图和无向图均可以用于表示条件独立性，两者的主要差异是从图中读出独立性的规则不同。

下面我们首先介绍概率模型的有向图表示，其中的一个典型模型就是朴素贝叶斯模型。

26.1.1 有向图与条件独立性：1. 有向图

定义 1

一个**有向图**$\mathcal{G}$ 是由节点集 V 及连接一对有序节点的边集 E 组成的。

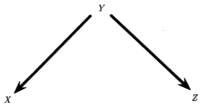

图 1: 节点集为 $V = \{X, Y, Z\}$ 且边集为 $E = \{(Y, X), (Y, Z)\}$

图1 给出了一个有向图的例子。若 $(Y, X) \in E$，则存在一条有向边从 Y 指向 X。通常一个被赋予某种概率分布的有向图常被称为贝叶斯网络，每个节点对应一个随机变量，每条边展现随机变量间的关联关系。图在表示变量间的独立性关系方面是很有用处的，还可以用来代替反事实去表示因果关系。

2. 条件独立性

在进行关于有向非循环图 (DAGs) 的讨论之前，需要先讨论一下条件独立性。

定义 2

令 X, Y 和 Z 为随机变量。在给定 Z 的条件下，如果下式对于所有的 x，y 和 z 均成立，

$$f_{X,Y|Z}(x, y \mid z) = f_{X|Z}(x \mid z) f_{Y|Z}(y \mid z)$$

则 X 和 Y 称为条件独立的，记作 $X \perp\!\!\!\perp Y \mid Z$。

直观地理解，上述定义表明知道了 Z, Y 并没有提供关于 X 的额外信息。

注：一个等价的定义为

$$f(x \mid y, z) = f(x \mid z)$$

条件独立性的基本性质

条件独立性具有一些基本的性质。

定理 1

下列各蕴涵关系成立：

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

$$X \perp\!\!\!\perp Y \mid Z \text{ 且 } U = h(X) \Rightarrow U \perp\!\!\!\perp Y \mid Z$$

$$X \perp\!\!\!\perp Y \mid Z \text{ 且 } U = h(X) \Rightarrow X \perp\!\!\!\perp Y \mid (Z, U)$$

$$X \perp\!\!\!\perp Y \mid Z \text{ 且 } X \perp\!\!\!\perp W \mid (Y, Z) \Rightarrow X \perp\!\!\!\perp (W, Y) \mid Z$$

$$X \perp\!\!\!\perp Y \mid Z \text{ 且 } X \perp\!\!\!\perp Z \mid Y \Rightarrow X \perp\!\!\!\perp (Y, Z).$$

3. 有向非循环图 (DAG)

有向路与无向路

定义 3

- 若一条有向边连接两个随机变量 X 和 Y(取任意一个方向)，就称 X 和 Y 是**邻接的**。
- 若一条有向边从 X 指向 Y，则称 X 是 Y 的**母节点**，而 Y 是 X 的**子节点**。X 的所有母节点的集合记作 π_X 或 $\pi(X)$。
- 两变量间的一条**有向路**是由一系列的同方向的有向边构成的，如下所示：

$$\mathbf{X} \longrightarrow \cdots \longrightarrow \boldsymbol{Y}$$

- 一个从 X 开始至 Y 结束的邻接节点的序列，但是忽略其有向边的方向性，就称该序列为一个**无向路**。
- 若存在一条有向路从 X 指向 Y （或 $X = Y$），则称 X 是 Y 的**祖节点**。也可以说 Y 是 X 的**后裔节点**。

有向非循环图

定义 4

如下形式的结构：

$$\boldsymbol{X} \longrightarrow \boldsymbol{Y} \longleftarrow \boldsymbol{Z}$$

称作在 Y 处**相遇**。不具有该种形式的结构称作**不相遇**。

例如，

$$\boldsymbol{X} \longrightarrow \boldsymbol{Y} \longrightarrow \boldsymbol{Z}$$

不相遇。相遇的性质是依赖于路的。

定义 5

一条开始和结束都在同一个变量处的有向路是一个圈。若一个有向图没有圈，则它是**非循环的**。在这种情况下，称这种图为一个**有向非循环图或 *DAG***。

4. 概率与 DAGs

令 $\mathcal{G}$ 为一个具有节点集 $V = (X_1, \cdots, X_k)$ 的 DAG。

定义 6

若 P 为 V 的分布，它的概率函数为 f，若下式成立：

$$f(v) = \prod_{i=1}^{k} f(x_i \mid \pi_i)$$

就说 P 是关于 $\mathcal{G}$ 是**马尔可夫的**，或称 $\mathcal{G}$ 表示 P，其中，π_i 为 X_i 的母节点。由 $\mathcal{G}$ 表示的分布集记为 $M(\mathcal{G})$。

马尔可夫举例

例 1

对于图2 中的 *DAG* 来说，$\mathbb{P} \in M(\mathcal{G})$ 当且仅当其概率函数 f 具有以下形式：

$$f(x, y, z, w) = f(x)f(y)f(z \mid x, y)f(w \mid z)$$

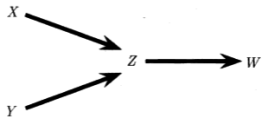

图 2: 另一个 DAG

马尔可夫条件成立定理

下述定理表明 $\mathbb{P} \in M(\mathcal{G})$ 当且仅当马尔可夫条件成立。

定理 2

一个分布 $\mathbb{P} \in M(\mathcal{G})$ 当且仅当下面的马尔可夫条件成立：对于每个变量 W，

$$W \perp\!\!\!\perp \widetilde{W} \,|\, \pi_W$$

其中，$\widetilde{W}$ 表示除了 W 的母节点和后裔节点以外的所有其他变量。

粗略地讲，马尔可夫条件意味着每个变量 W 在给定其母节点的情况下与“过去”是独立的。

独立性成立举例

例 2

考虑图3 中的 DAG。在这种情况下，概率函数分解如下：

$$f(a,b,c,d,e) = f(a)f(b \mid a)f(c \mid a)f(d \mid b,c)f(e \mid d)$$

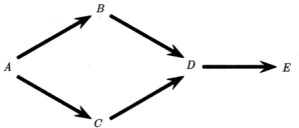

图 3: 另一个 DAG

马尔可夫条件意味着下面的独立性关系：

$$D \perp\!\!\!\perp A \mid \{B,C\}, \quad E \perp\!\!\!\perp \{A,B,C\} \mid D \text{ 且 } B \perp\!\!\!\perp C \mid A.$$

5. DAGs 的估计

在 DAGs 中有两个首先要考虑的估计问题。

- 第一，给定一个 DAG 为 $\mathcal{G}$ 和来自与 $\mathcal{G}$ 相符的分布为 f 的数据 $V_1, \cdots, V_n$，如何去估计 f?
- 第二，给定数据 $V_1, \cdots, V_n$，又如何去估计 $\mathcal{G}$?

第一个问题是一个纯粹的估计问题，而第二个问题则涉及到模型的选择。这些都是非常复杂的问题。这里仅简要介绍其主要思想，我们将在具体的模型中体会这一点。

分布的估计

通常，对于每个条件密度，人们常选择用某个参数模型 $f(x \mid \pi_x; \theta_x)$，则其似然函数为

$$\mathcal{L}(\theta) = \prod_{i=1}^{n} f(V_i; \theta) = \prod_{i=1}^{n} \prod_{j=1}^{m} f(X_{ij} \mid \pi_j; \theta_j)$$

其中，X_{ij} 是对于第 i 个数据点的 X_j 的值，θ_j 是第 j 个条件密度的参数。这样就可以通过极大似然方法来估计参数。

$\mathcal{G}$ 的估计

- 为了估计 DAG 自身的结构，几乎可以通过极大似然方法来估计每个可能的 DAG，且用 AIC（或其他的方法）来选择一个 DAG。
- 然而存在很多可能的 DAGs，所以需要很多数据来确保该方法是可靠的。而且从所有可能的 DAGs 中搜索是一个相当大的计算上的挑战。
- 对于一个 DAG 结构产生一个有效的精确的置信集可能需要天文数字般的样本容量。若知道关于 DAG 的结构的部分先验信息，计算和统计上的问题至少可以部分地改善。

26.1.2 朴素贝叶斯模型

例 3

设输入空间 $\mathcal{X} \subseteq R^n$ 为 n 维向量的集合，输出空间为类标记集合 $\mathcal{Y} = \{c_1, c_2, \cdots, c_K\}$。输入为特征向量 $x \in \mathcal{X}$，输出为类标记 $y \in \mathcal{Y}$。X 是定义在输入空间 $\mathcal{X}$ 上的随机向量，Y 是定义在输出空间 $\mathcal{Y}$ 上的随机变量。考虑训练数据集为

$$T = \{(x_1, y_1), (x_2, y_2), \cdots, (x_N, y_N)\}$$

的分类任务，该训练数据集由 $P(X, Y)$ 独立同分布产生。朴素贝叶斯法通过训练数据集学习联合概率分布 $P(X, Y)$。具体地学习以下先验概率分布和条件概率分布：

$$P(Y = c_k), \quad P(X = x | Y = c_k) = P(X^{(1)} = x^{(1)}, \cdots, X^{(n)} = x^{(n)} | Y = c_k), \quad k = 1, 2, \cdots, K,$$

再根据贝叶斯定理求得后验概率分布 $P(Y \mid X)$，其中 x 为 n 维输入特征向量 $\{x^{(1)}, x^{(2)}, ..., x^{(n)}\}$，$x \in \mathcal{X}$，$c_k$ 为类标记。

因为条件概率分布 $P(X=x|Y=c_k)$ 有指数级数量的参数，其估计实际是不可行的。事实上，假设 $x^{(j)}$ 可取值有 S_j 个，$k=1,2,\cdots,n$，Y 可取值有 K 个，那么参数个数为 $K\prod_{j=1}^{n}S_j$。因此，朴素贝叶斯法需要对类条件概率进行独立性假设。即：

$$P(X|Y=c_k)=\prod_j P(X^{(j)}=x^{(j)}|Y=c_k).$$

上述独立性假设，恰好相当于假设随机变量 X 与 Y 满足如下 DAG：

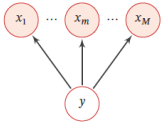

图 4

朴素贝叶斯法实际上学习到生成数据的机制，所以属于生成模型。条件独立假设等于是说用于分类的特征在类确定的条件下都是条件独立的。这一假设使朴素贝叶斯法变得简单，但有时会牺牲一定的分类准确率。

朴素贝叶斯法分类时，对给定的输入 x，通过学习到的模型计算后验概率分布 $P(Y=c_k|X=x)$，将后验概率最大的类作为 x 的类输出。后验概率计算根据贝叶斯定理进行:

$$P(Y=c_k \mid X=x)=\frac{P(X=x \mid Y=c_k)\,P(Y=c_k)}{\sum_k P(X=x \mid Y=c_k)\,P(Y=c_k)}$$

这是朴素贝叶斯法分类的基本公式。于是，朴素贝叶斯分类器可表示为

$$y=f(x)=\arg\max_{c_k}\frac{P(Y=c_k)\prod_j P\left(X^{(j)}=x^{(j)} \mid Y=c_k\right)}{\sum_k P(Y=c_k)\prod_j P\left(X^{(j)}=x^{(j)} \mid Y=c_k\right)}$$

注意到，在上式中分母对所有 c_k 都是相同的，所以，

$$y=\arg\max_{c_k} P(Y=c_k)\prod_j P\left(X^{(j)}=x^{(j)} \mid Y=c_k\right)$$

朴素贝叶斯法将实例分到后验概率最大的类中。这等价于期望风险最小化。假设选择 0-1 损失函数:

$$L(Y, f(X)) = \begin{cases} 1, & Y \neq f(X) \\ 0, & Y = f(X) \end{cases}$$

式中 $f(X)$ 是分类决策函数。这时，期望风险函数为

$$R_{\exp}(f) = E[L(Y, f(X))]$$

期望是对联合分布 $P(X, Y)$ 取的。由此取条件期望

$$R_{\exp}(f) = E_X \sum_{k=1}^{K} [L(c_k, f(X))] P(c_k \mid X)$$

为了使期望风险最小化，只需对 $X = x$ 逐个极小化，由此得到

$$
\begin{aligned}
f(x) &= \arg\min_{y \in \mathcal{Y}} \sum_{k=1}^{K} L(c_k, y) P(c_k \mid X = x) \\
&= \arg\min_{y \in \mathcal{Y}} \sum_{k=1}^{K} P(y \neq c_k \mid X = x) \\
&= \arg\min_{y \in \mathcal{Y}} (1 - P(y = c_k \mid X = x)) \\
&= \arg\max_{y \in \mathcal{Y}} P(y = c_k \mid X = x)
\end{aligned}
$$

这样一来，根据期望风险最小化准则就得到了后验概率最大化准则:

$$
f(x) = \arg\max_{c_k} P(c_k \mid X = x)
$$

即朴素贝叶斯法所采用的原理。

上述优化问题可采用极大似然估计或贝叶斯估计进行求解。

26.1.3 隐马尔可夫模型

例 4

隐马尔可夫模型 *(Hidden Markov Model, HMM)* 是用来表示一种含有隐变量的马尔可夫过程，如下图所示。其中 $X_{1:T}$ 为可观测变量，$Y_{1:T}$ 为隐变量。每个可观测标量 X_t 依赖当前时刻的隐变量 Y_t，隐变量构成一个马尔可夫链。

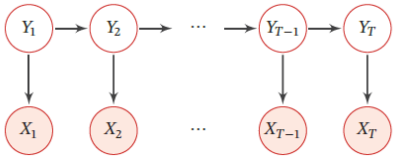

图 5

隐马尔可夫模型

从定义可知，隐马尔可夫模型作了两个基本假设:

- 齐次马尔可夫性假设，即假设隐藏的马尔可夫链在任意时刻 t 的状态只依赖于其前一时刻的状态，与其他时刻的状态及观测无关，也与时刻 t 无关；
- 观测独立性假设，即假设任意时刻的观测只依赖于该时刻的马尔可夫链的状态，与其他观测及状态无关。

隐马尔可夫模型

概率模型：隐马尔可夫模型是生成模型，根据假设知，其联合概率可以分解为

$$p(\boldsymbol{x}, \boldsymbol{y}; \theta) = \prod_{t=1}^{T} p\left(y_t \mid y_{t-1}, \theta_s\right) p\left(x_t \mid y_t, \theta_t\right)$$

除了上述结构信息，要确定一个隐马尔可夫模型还需以下三组参数：状态转移概率、输出观测概率和初始状态概率。在实际应用中，人们常关注隐马尔可夫模型的三个基本问题：概率计算问题、学习问题和预测问题。

设所有观测数据写成 $(\boldsymbol{x}^{(1)}, \boldsymbol{y}^{(1)}), (\boldsymbol{x}^{(2)}, \boldsymbol{y}^{(2)}), \cdots, (\boldsymbol{x}^{(N)}, \boldsymbol{y}^{(N)})$，确定并极大化完全数据的对数似然函数，得**优化问题**：

$$\max_{\theta} \quad \sum_{i=1}^{N} \ln p(\boldsymbol{x}^{(i)}, \boldsymbol{y}^{(i)}; \theta)$$

若未观测到隐变量，则构建可观测变量的联合概率函数。

引言

- 机器学习中概率模型，有些可以用有向图来表示，还有一些可以用无向图来表示，一般称为概率无向图模型。
- 概率无向图模型 (probabilistic undirected graphical models)，又称为马尔可夫随机场，是一个可以由无向图表示的联合概率分布。
- 虽然无向图模型与有向图表达条件独立性规则不同，但是它仍能将联合概率表示分解成一组函数的乘积。它主要是借助于“团”的概念及其势函数，建立随机向量的联合概率。

接下来我们首先回顾无向图的定义，然后定义无向图表示的随机变量之间存在的成对马尔可夫性和全局马尔可夫性，最后引出团和势函数的概念以及概率无向图模型的因子分解定理。

26.2.1 无向图

1. 无向图的基本概念

定义 7

一个**无向图**$G=(V,E)$ 由一个有限节点集 V 和由每对节点组成的边或 (弧) 集 E 所构成。节点对应着随机变量 $X,Y,Z,\cdots$，而边被记作一些无序对。

例如，$(X,Y)\in E$ 表示 X 和 Y 通过一条边连接起来。图6 给出了一个无向图的例子。

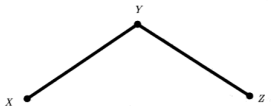

图 6: 节点集为 $V=\{X,Y,Z\}$ 的一个图。其边集为 $E=\{(Y,X),(Y,Z)\}$

完全图

定义 8

- 若两个节点之间存在一条边，则称这两个节点是**邻接的**，记作 $X \sim Y$. 在图6 中，X 和 Y 是邻接的但是 X 和 Z 不是邻接的。若对每个 i 都有 $X_{i-1} \sim X$，则序列 $X_0, \cdots, X_n$ 称为一条路。在图6 中，X, Y, Z 是**一条路**。
- 若一个图中任意两个节点之间都存在一条边，则称这个图是**完全的**（完全图）。一个子节点集 $U \in V$ 连同其边被称作一个子图。

分离

定义 9

设 A, B 和 C 是 V 的不同子集，若从 A 中的一个变量到 B 中的一个变量的路都相交于 C 中的一个变量，就说 C 分离 A 和 B。

例如，在图7 中，Y, W 和 Z 被 Z **分离**。同时，W 和 Z 被 X, Y **分离**。

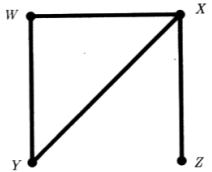

图 7: $\{Y, W\}$ 和 $\{Z\}$ 被 $\{X\}$ 分离。而且，W 和 Z 被 $\{X, Y\}$ 分离

2. 概率与图

定义 10

令 V 为具有分布 $\mathbb{P}$ 的随机变量集。构造一个图，其每个节点对应 V 中的每个变量。略去一对变量之间的边，若它们在给定其余变量的条件下是独立的。即

$$X \text{ 和 } Y \text{ 之间没有边} \Leftrightarrow X \perp\!\!\!\perp Y \mid \text{其余变量},$$

其中，“其余变量”表示除了 X 和 Y 之外的所有其他变量，这样的图称作成对马尔可夫图。

如图8 所示：

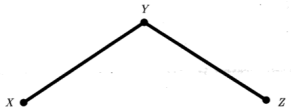

图 8: $X \perp\!\!\!\perp Z \mid Y$

其他条件独立性

图中暗含着一系列的成对条件独立性关系。这些关系可以推出其他的条件独立性关系。如何从图中直接读出其他的条件独立性关系呢？事实上，有如下结论成立：

定理 3

令 $\mathcal{G} = (V, E)$ 是一个分布为 $\mathbb{P}$ 的成对马尔可夫图。令 A, B 和 C 为 V 的不相同的子集使得 C 分离 A 和 B，$A \perp\!\!\!\perp B \mid C$。

注：若 A 和 B 不是连通的（也就是不存在一条从 A 到 B 的路），则可以把 A 和 B 看作被空集分离，则由定理3可知 $A \perp\!\!\!\perp B$。

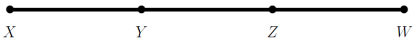

图 9: 若满足成对马尔可夫性，可以得到 $X \perp\!\!\!\perp Z \mid Y$ 吗？

成对马尔可夫性质与全局马尔可夫性质

定理3 中的独立性条件被称作全局马尔可夫性质。将看到成对和全局马尔可夫性质是等价的。更确切的，可以描述为：给定一个图 $\mathcal{G}$，

- 令 $M_{pair}(\mathcal{G})$ 表示满足**成对马尔可夫性质的分布集**，因此 $P \in M_{pair}(\mathcal{G})$，在分布 $\mathbb{P}$ 下，若 $X \perp\!\!\!\perp Y \mid$ 其余变量当且仅当 X 和 Y 之间不存在边。
- 令 $M_{global}(\mathcal{G})$ 为满足**全局马尔可夫性质的分布集**：则 $P \in M_{pair}(\mathcal{G})$，在分布 $\mathbb{P}$ 下，若 $A \perp\!\!\!\perp B \mid C$ 当且仅当 C 分离 A 和 B。

成对和全局马尔可夫的等价性

定理 4

令 $\mathcal{G}$ 为一个图，则 $M_{pair}(\mathcal{G}) = M_{global}(\mathcal{G})$。

上述定理保证了可以使用简单的成对性质来构建图，这就使得可以用全局马尔可夫性来推导其他独立关系。

例 5

由图10 可知 $X \perp\!\!\!\perp Y, X \perp\!\!\!\perp Z$ 和 $X \perp\!\!\!\perp (Y, Z)$。

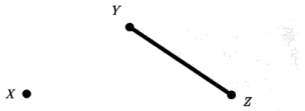

图 10: $X \perp\!\!\!\perp Y$

概率无向图模型

定义 11

设有联合概率分布 $P(Y)$, 由无向图 $G=(V,E)$ 表示, 在图 G 中, 结点表示随机变量, 边表示随机变量之间的依赖关系。如果联合概率分布 $P(Y)$ 满足成对或全局马尔可夫性, 就称此联合概率分布为**概率无向图模型** *(probabilistic undirected graphical model)*, 或**马尔可夫随机场** *(Markov random field)*。

在实际中，我们关心的是如何求其联合概率分布。为便于模型的学习与计算, 对于给定的概率无向图模型，我们希望将整体的联合概率写成若干子联合概率的乘积的形式, 也就是将联合概率进行因子分解。事实上，概率无向图模型的最大特点就是易于因子分解。

3. 团与势

首先我们给出无向图中的团与极大团的定义。

定义 12

若一个图的变量集中的任意两个对应的节点都是邻接的，则称该集为一个**团**。若一个团任意增加一个节点后就不能成为团，则称之为一个**极大团**。

例 6

图 11 中的极大团为 $\{X_1, X_2\}$， $\{X_1, X_3\}$， $\{X_2, X_4\}$， $\{X_3, X_5\}$， $\{X_2, X_5, X_6\}$

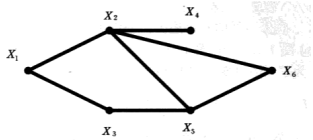

图 11: 极大团示例

势函数

一个势就是任意一个正函数。在特定的条件下，可以证明分布 $\mathbb{P}$ 关于无向图 $\mathcal{G}$ 是马尔可夫的当且仅当其概率函数 f 可以写作图中所有极大团 $\mathcal{C}$ 上的函数 $\psi_C(x_C)$ 的乘积形式，即

$$f(x) = \frac{\prod_{C\in\mathcal{C}} \psi_C(x_C)}{Z}$$

其中，$\mathcal{C}$ 是一个极大团集，ψ_C 是一个势，且

$$Z = \sum_x \prod_{C\in\mathcal{C}} \psi_C(x_C),$$

称为规范化因子。规范化因子保证 f 构成一个概率分布。函数 $\psi_C(x_C)$ 称为势函数，一般要求为严格正函数，通常定义为指数函数：

$$\psi_C(x_C) = \exp(-E(x_C))$$

概率无向图模型的因子分解

将概率无向图模型的联合概率分布表示为其极大团上的随机变量的函数的乘积形式的操作，称为概率无向图模型的因子分解（factorization）。

定理 5

*(Hammersley-Clifford 定理）*概率无向图模型的联合概率分布 $P(Y)$ 可以表示为如下形式:

$$P(Y)=\frac{1}{Z}\prod_{C}\Psi_{C}(Y_{C})$$

$$Z=\sum_{Y}\prod_{C}\Psi_{C}(Y_{C})$$

其中，C 是无向图的最大团，Y_C 是 C 的结点对应的随机变量，$\Psi_C(Y_C)$ 是 C 上定义的严格正函数，乘积是在无向图所有的最大团上进行的。

因子分解举例

例 7

前面已知图*11* 中的极大团是

$$\{X_1, X_2\}, \quad \{X_1, X_3\}, \quad \{X_2, X_4\}, \quad \{X_3, X_5\}, \quad \{X_2, X_5, X_6\}$$

因此，可以把概率函数写为

$$f(x_1, x_2, x_3, x_4, x_5, x_6) \propto \psi_{12}(x_1, x_2)\, \psi_{13}(x_1, x_3)\, \psi_{24}(x_2, x_4) \times \psi_{35}(x_3, x_5)\, \psi_{256}(x_2, x_5, x_6)$$

因子分解举例

例 8

图中对应的概率分布可以分解为

$$p(a,b,c,d,e) = \frac{1}{Z}\phi^1(a,b,c)\phi^2(b,d)\phi^3(c,e)$$

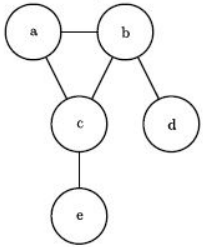

图 12

26.2.2 条件随机场

在实际问题中，我们经常需要建立条件概率模型 $P(Y \mid X)$。条件随机场是给定一组输入随机变量条件下另一组输出随机变量的条件概率分布模型，其特点是假设输出随机变量构成马尔可夫随机场。也即条件随机场是给定随机变量 X 条件下，随机变量 Y 的马尔可夫随机场。条件随机场可以用于不同的预测问题。

定义 13

设 X 与 Y 是随机变量，$P(Y \mid X)$ 是在给定 X 的条件下 Y 的条件概率分布。若随机变旦 Y 构成一个由无向图 $G=(V,E)$ 表示的马尔可夫随机场，即

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

对任意结点 v 成立，则称条件概率分布 $P(Y \mid X)$ 为**条件随机场**。式中 $w \sim v$ 表示在图 $G=(V,E)$ 中与结点 v 有边连接的所有结点 $w, w \neq v$ 表示结点 v 以外的所有结点，Y_v, Y_u 与 Y_w 为结点 v, u 与 w 对应的随机变量。

线性链条件随机场模型

在条件随机场的定义中，并未对无向图的结构进行任何假定，也没有要求 X 和 Y 具有相同的结构。现实中，一般假定 X 和 Y 有相同的图结构。这里，我们考虑无向图具有下图所示的线性链结构，即

$$\mathcal{G} = (V = \{1, 2, \cdots, n\}, E = \{(i, i+1)\}), \quad i = 1, 2, \cdots, n-1$$

在此情况下，$X = (X_1, X_2, \cdots, X_n)$，$Y = (Y_1, Y_2, \cdots, Y_n)$，最大团是相邻两个节点的集合。

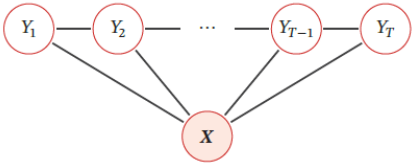

图 13

线性链条件随机场模型

定义 14

设 $X=(X_1, X_2, \cdots, X_n)$ 与 $Y=(Y_1, Y_2, \cdots, Y_n)$ 均为线性链表示的随机变量序列，若在给定随机变量序列 X 的条件，随机变量序列 Y 的条件概率分布 $P(Y \mid X)$ 构成条件随机场，即满足马尔可夫性

$$P(Y_i \mid X, Y_1, \cdots, Y_{i-1}, Y_{i+1}, \cdots, Y_n) = P(Y_i \mid X, Y_{i-1}, Y_{i+1})$$

$$i=1,2,\cdots,n(\text{ 在}i=1\text{ 和}n\text{ 时只考虑单边})$$

则称 $P(Y \mid X)$ 为线性链条件随机场。在标注问题中，X 表示输入观测序列，Y 表示对应的输出标记序列或状态序列。

线性链条件随机场模型

根据概率无向图模型的因式分解定理，可以给出线性链条件随机场 $P(Y|X)$ 的因子分解式，各因子是定义在相邻两个节点（最大团）上的势函数。

定理 6

(线性链条件随机场的参数化形式) 设 $P(Y \mid X)$ 为线性链条件随机场，则在随机变量 X 取值为 x 的条件下，随机变量 Y 取值为 y 的条件概率具有如下形式：

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

其中，$Z(x)=\sum_y \exp \left(\sum_{i, k} \lambda_k t_k\left(y_{i-1}, y_i, x, i\right)+\sum_{i, l} \mu_l s_l\left(y_i, x, i\right)\right)$ 式中，t_k 和 s_l 是特征函数，λ_k 和 μ_l 是对应的权值。$Z(x)$ 是规范化因子，求和是在所有可能的输出序列上进行的。

线性链条件随机场模型

- 定理中的条件概率表达式是线性链条件随机场模型的基本形式, 表示给定输入序列 x, 对输出序列 y 预测的条件概率。式中, t_k 是定义在边上的特征函数, 称为**转移特征**, 依赖于当前和前一个位置; s_l 是定义在结点上的特征函数, 称为**状态特征**, 依赖于当前位置。t_k 和 s_l 都依赖于位置, 是局部特征函数。
- 通常, 特征函数 t_k 和 s_l 取值为 1 或 0 ; 当满足特征条件时取值为 1, 否则为 0。条件随机场完全由特征函数 t_k, s_l 和对应的权值 λ_k, μ_l 确定。

线性链条件随机场模型

- 很显然线性链条件随机场模型是建立在条件概率分布 $P(Y \mid X)$ 之下，因此是一个判别模型。
- 有了线性链条件随机场的定义和参数表示，一般接下来来会考虑 3 个基本的问题：概率计算问题、学习问题和预测问题。
- 线性链条件随机场可以用于机器学习中的标注等问题。这时，在条件概率模型 $P(Y|X)$ 中，Y 是输出变量，表示标记序列，X 是输入变量，表示需要标注的观测序列。也把标记序列称为状态序列。学习时，利用训练数据集通过极大似然估计或正则化的极大似然估计得到条件概率模型 $\hat{P}(Y|X)$。预测时，对于给定的输入序列 X，求出条件概率 $\hat{P}(y|x)$ 最大的输出序列 $\hat{y}$。

线性链条件随机场模型

已知训练数据集, 由此可知经验概率分布 $\tilde{P}(X, Y)$ 。则训练数据的对数似然函数为

$$L(\boldsymbol{\lambda}, \boldsymbol{\mu}) = L_{\tilde{P}}\left(P_{\boldsymbol{\lambda},\boldsymbol{\mu}}\right) = \log \prod_{x,y} P_{\boldsymbol{\lambda},\boldsymbol{\mu}}(y \mid x)^{\tilde{P}(x,y)} = \sum_{x,y} \tilde{P}(x, y) \log P_{\boldsymbol{\lambda},\boldsymbol{\mu}}(y \mid x)$$

将线性链条件随机场的参数化形式代入上式, 可得对数似然函数为

$$\begin{aligned} L(\boldsymbol{\lambda}, \boldsymbol{\mu}) &= \sum_{x,y} \tilde{P}(x, y) \log P_{\boldsymbol{\lambda},\boldsymbol{\mu}}(y \mid x) \\ &= \sum_{j=1}^{N} \left(\sum_{i,k} \lambda_k t_k\left(y_{i-1}, y_i, x, i\right) + \sum_{i,l} \mu_l s_l\left(y_i, x, i\right) \right) - \sum_{j=1}^{N} \log Z_{\boldsymbol{\lambda},\boldsymbol{\mu}}\left(x_j\right) \end{aligned}$$

因此，可得**优化问题**：

$$\max_{\boldsymbol{\lambda},\boldsymbol{\mu}} \quad \sum_{j=1}^{N} \left(\sum_{i,k} \lambda_k t_k\left(y_{i-1}, y_i, x, i\right) + \sum_{i,l} \mu_l s_l\left(y_i, x, i\right) \right) - \sum_{j=1}^{N} \log Z_{\boldsymbol{\lambda},\boldsymbol{\mu}}\left(x_j\right).$$

本讲小结

概率模型的图表示：

有向图表示

- 条件独立性
- DAGs
- 概率与 DAGs
- DAGS 的估计

无向图表示

- 概率与图
- 图与势
- 概率无向图模型
- 因式分解

三个典型的概率图模型：朴素贝叶斯模型、隐马尔可夫模型、条件随机场。概率无向图模型还包括随机游走模型等，与 pagerank 密切相关，我们这里就不介绍了。