第十一章　概率图模型

概率图模型（probabilistic graphical model，PGM），简称图模型（graphical model，GM），是指一种用图结构来描述多元随机变量之间条件独立关系的概率模型，从而给研究高维空间中的概率模型带来了很大的便捷性。

对于一个 K 维随机向量 $\mathbf{X} = [X_1, X_2, \cdots, X_K]^{\mathrm{T}}$，其联合概率为高维空间中的分布，一般难以直接建模。假设每个变量为离散变量并有 m 个取值，在不作任何独立假设条件下，则需要 $m^K - 1$ 个参数才能表示其概率分布。当 $m = 2, K = 100$ 时，参数量约为 10^{30}，远远超出了目前计算机的存储能力。

一个有效的减少参数量的方法是独立性假设。我们将 K 维随机向量的联合概率分解为 K 个条件概率的乘积，

$$
\begin{aligned}
p(\mathbf{x}) &\triangleq P(\mathbf{X} = \mathbf{x}) && (11.1)\\
&= p(x_1)p(x_2|x_1)\cdots p(x_K|x_1, \cdots, x_{K-1}), && (11.2)\\
&= \prod_{k=1}^{K} p(x_k|x_1, \cdots, x_{k-1}), && (11.3)
\end{aligned}
$$

其中 x_k 表示变量 X_k 的取值。如果某些变量之间存在条件独立，其参数量就可以大幅减少。

假设有四个二值变量 X_1, X_2, X_3, X_4，在不知道这几个变量依赖关系的情况下，可以用一个联合概率表来记录每一种取值的概率 $P(\mathbf{X}_{1:4})$，共需要 $2^4 - 1 = 15$ 个参数。假设在已知 X_1 时，X_2 和 X_3 独立，即有

$$
\begin{aligned}
p(x_2|x_1, x_3) &= p(x_2|x_1), && (11.4)\\
p(x_3|x_1, x_2) &= p(x_3|x_1). && (11.5)
\end{aligned}
$$

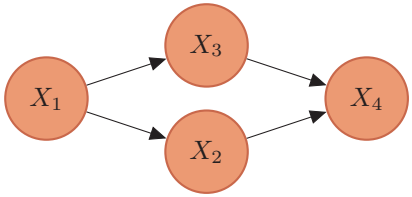


图 11.1: 变量 X_1, X_2, X_3, X_4 之间条件独立性的图形化表示。

在已知 X_2 和 X_3 时，X_4 也和 X_1 独立，即有

$$p(x_4|x_1, x_2, x_3) = p(x_4|x_2, x_3), \tag{11.6}$$

那么其联合概率 $p(\mathbf{x})$ 可以分解为

$$p(\mathbf{x}) = p(x_1)p(x_2|x_1)p(x_3|x_1, x_2)p(x_4|x_1, x_2, x_3), \tag{11.7}$$

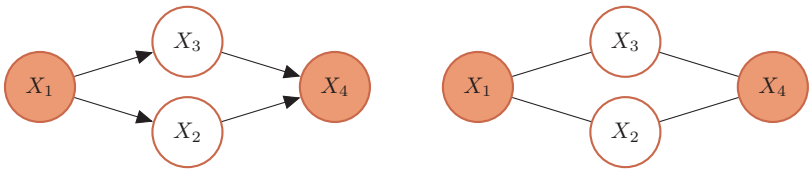
$$= p(x_1)p(x_2|x_1)p(x_3|x_1)p(x_4|x_2, x_3), \tag{11.8}$$

是 4 个局部条件概率的乘积。如果分别用 4 个表格来记录这 4 个条件概率的话，只需要 $1 + 2 + 2 + 4 = 9$ 个独立参数。

当概率模型中的变量数量比较多时，其条件依赖关系也比较复杂。我们可以使用图结构的方式将概率模型可视化，以一种直观、简单的方式描述随机变量之间的条件独立性的性质，并可以将一个复杂的联合概率模型分解为一些简单条件概率模型的组合。图11.1给出了上述例子中 4 个变量之间的条件独立性的图形化描述。图中每个节点表示一个变量，每条连边变量之间的依赖关系。对于一个非全连接的图，都存在一个或多个条件独立性假设，可以根据条件独立性将联合概率分布进行分解，表示为一组局部条件概率分布的乘积。

图模型的基本问题 图模型有三个基本问题：

1. 表示问题：对于一个概率模型，如何通过图结构来描述变量之间的依赖关系。
2. 推断问题：在已知部分变量时，计算其它变量的后验概率分布。
3. 学习问题：图模型的学习包括图结构的学习和参数的学习。在本章我们只关注在给定图结构时的参数学习，即参数估计问题。



(a) 有向图：贝叶斯网络 (b) 无向图：马尔可夫随机场

图 11.2: 有向图和无向图示例。带阴影的节点表示观测变量，不带阴影的节点表示隐变量，连边表示两变量间的条件依赖关系。

图模型与机器学习 很多机器学习模型都可以归结为概率模型（probabilistic model），即建模输入和输出之间的条件概率分布。因此，图模型提供了一种新的角度来解释机器学习模型，并且这种角度有很多优点，比如了解不同机器学习模型之间的联系，方便设计新模型等。在机器学习中，图模型越来越多地用来设计和分析各种学习算法。

11.1 模型表示

图由一组节点和节点之间的边组成。在概率图模型中，每个节点都表示一个随机变量（或一组随机变量），边表示这些随机变量之间的概率依赖关系。

常见的概率图模型可以分为两类：有向图模型和无向图模型。有向图模型的图结构为有向非循环图，如果两个节点之间有连边，表示对于的两个变量为因果关系。无向图模型使用无向图来描述变量之间的关系。每条边代表两个变量之间有概率依赖关系，但是并不一定是因果关系。

图11.2给出了两个代表性图模型（有向图和无向图）的示例，分别表示了四个变量$\{X_1, X_2, X_3, X_4\}$之间的依赖关系

11.1.1 有向图模型

有向图模型（directed graphical model），也称为贝叶斯网络（Bayesian network），或信念网络（belief network，BN），是指用有向图来表示概率分布的图模型。假设一个有向图$G(\mathcal{V}, \mathcal{E})$，节点集合$\mathcal{V} = \{X_1, X_2, \cdots, X_K\}$表示$K$个随机变量，节点$k$对应随机变量$X_k$。$\mathcal{E}$为边的集合，每条边表示两个变量之间的因果关系。

定义 11.1 – 贝叶斯网络： 对于随机向量 $\mathbf{X} = [X_1, X_2, \cdots, X_K]^{\mathrm{T}}$ 和一个有向非循环图 G，G 中的每个节点都对应一个随机变量，可以是观察变量，隐变量或是未知参数。G 中的每个连接 e_{ij} 表示两个随机变量 X_i 和 X_j 之间具有非独立的因果关系。$\mathbf{X}_{\pi_k}$ 表示变量 X_k 的所有父节点变量集合，每个随机变量的局部条件概率（local conditional probability distribution）为 $P(X_k|\mathbf{X}_{\pi_k})$。

如果 $\mathbf{X}$ 的联合概率分布可以分解为每个随机变量 X_k 的局部条件概率的连乘形式，即

$$p(\mathbf{x}) = \prod_{k=1}^{K} p(x_k|\mathbf{x}_{\pi_k}), \tag{11.9}$$

那么 $(G, \mathbf{X})$ 构成了一个贝叶斯网络。

条件独立性 在贝叶斯网络中，如果两个节点是直接连接的，它们肯定是非条件独立的，是直接因果关系。父节点是“因”，子节点是“果”。

如果两个节点不是直接连接的，但是它们之间有一条经过其他节点的路径连接互连接，它们之间的条件独立性就比较复杂。以三个节点的贝叶斯网络为例，给定三个节点 X_1, X_2, X_3，X_1 和 X_3 是不直接连接的，可以通过节点 x_2 连接。这三个节点之间可以有四种连接关系，如图11.3所示。

间接因果关系（图11.3a） 当 X_2 已知时，X_1 和 X_3 为条件独立，即 $X_1 \perp\!\!\!\perp X_3|X_2$；

间接果因关系（图11.3b） 当 X_2 已知时，X_1 和 X_3 为条件独立，即 $X_1 \perp\!\!\!\perp X_3|X_2$；

共因关系（图11.3c） 当 X_2 未知时，X_1 和 X_3 是不独立的；当 X_2 已知时，X_1 和 X_3 条件独立，即 $X_1 \perp\!\!\!\perp X_3|X_2$；

共果关系（图11.3d） 当 X_2 未知时，X_1 和 X_3 是独立的；当 X_2 已知时，X_1 和 X_3 不独立，即 $X_1 \not\perp\!\!\!\perp X_3|X_2$。

局部马尔可夫性质 对一个更一般的贝叶斯网络，其局部马尔可夫性质为：每个随机变量在给定父节点的情况下，条件独立于它的非后代节点。

$$X_k \perp\!\!\!\perp Z|X_{\pi_k}, \tag{11.10}$$

其中 Z 为 X_k 的非后代变量。

从公式 (11.3) 和 (11.9) 可得到。参见习题（11-3），第223页。

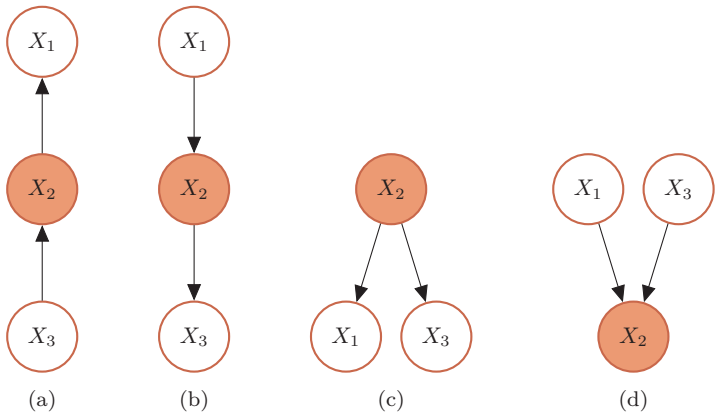


图 11.3: 三个变量的依赖关系示例。在 (a)(b) 中，$X_1 \not\perp\!\!\!\perp X_3|\emptyset$，但 $X_1 \perp\!\!\!\perp X_3|X_2$；在 (c) 中，$X_1 \not\perp\!\!\!\perp X_3|\emptyset$，但 $X_1 \perp\!\!\!\perp X_3|X_2$；在 (d) 中，$X_1 \perp\!\!\!\perp X_3|\emptyset$，但 $X_1 \not\perp\!\!\!\perp X_3|X_2$。

11.1.2 常见的有向图模型

很多经典的机器学习模型都可以使用有向图模型来描述，比如朴素贝叶斯分类器、隐马尔可夫模型、玻尔兹曼机、深度信念网络等。

玻尔兹曼机和深度信念网络的详细介绍见第十二章。

11.1.2.1 sigmoid 信念网络

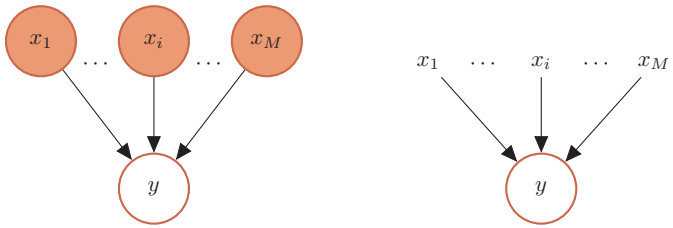
为了减少模型参数，可以使用参数化模型来建模有向图模型中的条件概率分布。一种简单的参数化模型为sigmoid 信念网络 [Neal, 1992]。*sigmoid* 信念网络（sigmoid belief network，SBN）中的变量取值为 $\{0, 1\}$。对于变量 X_k 和它的父节点集合 π_k，其条件概率分布表示为

更复杂的深度信念网络可以参见第12.3节，第237页。

$$P(X_k = 1|\mathbf{x}_{\pi_k}|\theta) = \sigma(\theta_0 + \sum_{x_i \in \mathbf{x}_{\pi_k}} \theta_i x_i), \tag{11.11}$$

其中 $\sigma(\cdot)$ 是 logistic sigmoid 函数，θ_i 是可学习的参数。假设变量 X_k 的父节点数量为 m，如果使用表格来记录条件概率需要 2^m 个参数，如果使用参数化模型只需要 $m + 1$ 个参数。如果对不同的变量的条件概率都共享使用一个参数化模型，其参数数量又可以大幅减少。

值得一提的是，sigmoid 信念网络与 logistic 回归模型都采用 logistic sigmoid 函数来计算条件概率。如果假设 sigmoid 信念网络中只有一个叶子节点，其所有的父节点之间没有连接，且取值为实数，那么 sigmoid 信念网络的网络结构和



(a) 只有一层的简单sigmoid信念网络 (b) logistic回归

图 11.4: sigmoid 信念网络和 logistic 回归模型的比较。

logistic 回归模型类似，如图11.4所示。但是，这两个模型区别在于 logistic 回归模型中的 $\mathbf{x}$ 作为一种确定性的参数，而非变量。因此，logistic 回归模型只建模条件概率 $p(y|\mathbf{x})$，是一种判别模型；而 sigmoid 信念网络建模 $p(\mathbf{x}, y)$，是一种生成模型。

logistic 回归模型也经常解释一种条件无向图模型。

11.1.2.2 朴素贝叶斯分类器

朴素贝叶斯分类器（naive Bayes classifier）是一类简单的概率分类器，在强（朴素）独立性假设的条件下运用贝叶斯公式来计算每个类别的后验概率。

给定一个有 d 维特征的样本 $\mathbf{x}$ 和类别 y，类别的后验概率为

$$p(y|\mathbf{x}, \theta) = \frac{p(x_1, \cdots, x_d|y)p(y)}{p(x_1, \cdots, x_d)} \tag{11.12}$$

$$\propto p(x_1, \cdots, x_d|y, \theta)p(y|\theta), \tag{11.13}$$

其中 θ 为概率分布的参数。

在朴素贝叶斯分类器中，假设在给定 Y 的情况下，X_i 之间是条件独立的，即 $X_i \perp\!\!\!\perp X_j|Y, \forall i \neq j$。图11.5给出了的图形表示。$p(y|\mathbf{x})$ 可以分解为

$$p(y|\mathbf{x}, \theta) \propto p(y|\theta_c) \prod_{i=1}^{d} p(x_i|y, \theta_{i,y}), \tag{11.14}$$

其中 θ_c 是 y 的先验概率分布的参数，$\theta_{i,y}$ 是条件概率分布 $p(x_i|y, \theta_{i,y})$ 的参数。如果 x_i 为连续值，$p(x_i|y, \theta_{i,y})$ 可以用高斯分布建模。如果 x_i 为离散值，$p(x_i|y, \theta_{i,y})$ 可以用多项分布建模。

虽然朴素贝叶斯分类器的条件独立性假设太强，但是在实际应用中，朴素

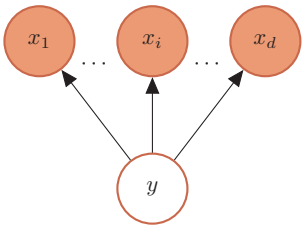


图 11.5: 朴素贝叶斯模型的图模型表示。

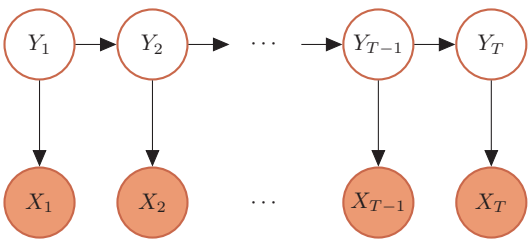


图 11.6: 隐马尔可夫模型。

贝叶斯分类器在很多任务上也能得到很好的结果，并且模型简单，可以有效防止过拟合。

11.1.2.3 隐马尔可夫模型

隐马尔可夫模型（hidden Markov model，HMM）[Baum and Petrie, 1966]是一种含有隐变量的马尔可夫过程。图11.6给出隐马尔可夫模型的图模型表示。

隐马尔可夫模型的联合概率可以分解为

$$p(\mathbf{x}, \mathbf{y}, \theta) = \prod_{t=1}^{T} p(y_t|y_{t-1}, \theta_s)p(x_t|y_t, \theta_t), \tag{11.15}$$

这里$p(y_1|y_0)$一般简化为$p(y_1)$。

其中$p(x_t|y_t, \theta_t)$为输出概率，$p(y_t|y_{t-1}, \theta_s)$为转移概率，θ_s, θ_t分别表示两类条件概率的参数。

11.1.3 无向图模型

无向图模型，也称为马尔可夫随机场（Markov random field）或马尔可夫网络（Markov network），是一类用无向图来描述一组具有局部马尔可夫性质的随机向量$\mathbf{X}$的联合概率分布的模型。

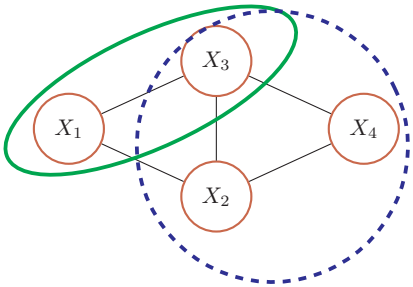


图 11.7: 无向图模型中的团和最大团。

定义 11.2 – 马尔可夫随机场： 给定一个随机向量$\mathbf{X} = [X_1, \cdots, X_K]^{\mathrm{T}}$和一个$K$个节点的无向图$G(\mathcal{V}, \mathcal{E})$（可以存在循环），图中的节点$k$表示随机变量$X_k$，$1 \leq k \leq K$。如果$(G, \mathbf{X})$满足局部马尔可夫性质，即一个变量$X_k$在给定它的邻居的情况下独立于所有其它变量，

$$p(x_k|\mathbf{x}_{-k}) = p(x_k|\mathbf{x}_{N(k)}), \tag{11.16}$$

其中$N(k)$为变量X_k的邻居集合，$-k$为除X_k外其它变量的集合，那么$(G, \mathbf{X})$就构成了一个马尔可夫随机场。

无向图的马尔可夫性 无向图中的马尔可夫性可以表示为

$$X_k \perp\!\!\!\perp \mathbf{X}_{-\mathrm{N}(k),-k} \mid \mathbf{X}_{\mathrm{N}(k)},$$

其中$\mathbf{X}_{-\mathrm{N}(k),-k}$为表示除$\mathbf{X}_{N(k)}$和$X_k$外的其它变量。

对于图11.2b中的4个变量，根据马尔可夫性质，可以得到$X_1 \perp\!\!\!\perp X_4|X_2, X_3$和$X_2 \perp\!\!\!\perp X_3|X_1, X_4$。

11.1.4 无向图模型的概率分解

团 由于无向图模型并不提供一个变量的拓扑顺序，因此无法用链式法则对$p(\mathbf{x})$进行逐一分解。无向图模型的联合概率一般以全连通子图为单位进行分解。

无向图中的一个全连通子图，称为团（clique），即团内的所有节点之间都

连边。图11.7中共有7个团，包括$\{X_1, X_2\}$，$\{X_1, X_3\}$，$\{X_2, X_3\}$，$\{X_3, X_4\}$，$\{X_2, X_4\}$，$\{X_1, X_2, X_3\}$，$\{X_2, X_3, X_4\}$。

在所有团中，如果一个团不能被其它的团包含，这个团就是一个最大团（maximal clique）。

因子分解 无向图中的的联合概率可以分解为一系列定义在最大团上的非负函数的乘积形式。

定理 11.1 – Hammersley-Clifford 定理： 如果一个分布$p(\mathbf{x}) > 0$满足无向图G中的局部马尔可夫性质，当且仅当$p(\mathbf{x})$可以表示为一系列定义在最大团上的非负函数的乘积形式，即

$$p(\mathbf{x}) = \frac{1}{Z} \prod_{c \in \mathcal{C}} \phi_c(\mathbf{x}_c), \tag{11.17}$$

其中$\mathcal{C}$为G中的最大团集合，$\phi_c(\mathbf{x}_c) \geq 0$是定义在团$c$上的势能函数（potential function），Z是配分函数（partition function），用来将乘积归一化为概率形式。

$$Z = \sum_{\mathbf{x} \in \mathcal{X}} \prod_{c \in \mathcal{C}} \phi_c(\mathbf{x}_c), \tag{11.18}$$

其中$\mathcal{X}$为随机向量X的取值空间。

Hammersley-Clifford定理的证明可以参考[Koller and Friedman, 2009]。无向图模型与有向图模型的一个重要区别是有配分函数Z。配分函数的计算复杂度是指数级的，因此在推断和参数学习时都需要重点考虑。

配分函数的计算参见第??节，第??页。

吉布斯分布 公式(11.17)中定义的分布形式也称为吉布斯分布（Gibbs distribution）。根据Hammersley-Clifford定理，无向图模型和吉布斯分布是一致的。吉布斯分布一定满足马尔可夫随机场的条件独立性质，并且马尔可夫随机场的概率分布一定可以表示成吉布斯分布。

由于势能函数必须为正的，因此我们一般定义为

$$\phi_c(\mathbf{x}_c) = \exp(-E_c(\mathbf{x}_c)), \tag{11.19}$$

其中$E(\mathbf{x}_c)$为能量函数（energy function）。

这里的负号是遵从物理上习惯，即能量越低意味着概率越高。

因此，无向图上定义的概率分布可以表示为：

$$P(\mathbf{x}) = \frac{1}{Z}\prod_{c\in\mathcal{C}}\exp(-E_c(\mathbf{x}_c)) \tag{11.20}$$

$$= \frac{1}{Z}\exp(\sum_{c\in\mathcal{C}} -E_c(\mathbf{x}_c)) \tag{11.21}$$

这种形式的分布又称为**玻尔兹曼分布**（Boltzmann distribution）。任何一个无向图模型都可以用公式(11.21)来表示其联合概率。

玻尔兹曼分布参见第227页。

11.1.5 常见的无向图模型

11.1.5.1 对数线性模型

势能函数的一般定义为

$$\phi_c(\mathbf{x}_c|\theta_c) = \exp\Big(\theta_c^{\mathrm{T}} f_c(\mathbf{x}_c)\Big), \tag{11.22}$$

其中函数$f_c(\mathbf{x}_c)$为定义在$\mathbf{x}_c$上的特征向量，θ_c为权重向量。这样联合概率$p(\mathbf{x})$的对数形式为

$$\log p(\mathbf{x}|\theta) = \sum_{c\in C}\theta_c^{\mathrm{T}} f_c(\mathbf{x}_c) - \log Z(\theta), \tag{11.23}$$

其中θ代表所有势能函数中的参数θ_c。这种形式的无向图模型**对数线性模型**（log-linear model）或**最大熵模型**（maximum entropy model）[Berger et al., 1996, Della Pietra et al., 1997]。

如果用对数线性模型来建模条件概率$p(y|\mathbf{x})$，

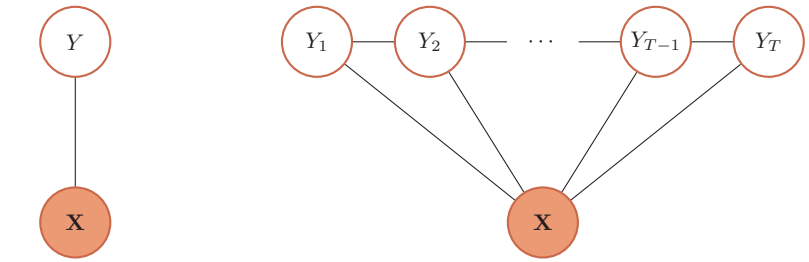
$$p(y|\mathbf{x},\theta) = \frac{1}{Z(\mathbf{x},\theta)}\exp\Big(\theta^{\mathrm{T}} f(\mathbf{x},y)\Big), \tag{11.24}$$

其中$Z(\mathbf{x},\theta) = \sum_y \exp(\theta^{\mathrm{T}} f_y(\mathbf{x},y))$。这种对数线性模型也称为**条件最大熵模型**或*softmax*回归模型。

softmax回归模型参见第3.2节，第46页。

11.1.5.2 条件随机场

条件随机场（conditional random field，CRF）[Lafferty et al., 2001]是一种直接建模条件概率的无向图模型。



(a) 最大熵模型 (b) 线性链的条件随机场

图 11.8: 最大熵模型和线性链的条件随机场。

和最大熵模型不同，条件随机场建模的条件概率 $p(\mathbf{y}|\mathbf{x})$ 中，$\mathbf{y}$ 一般为随机向量，因此需要对 $p(\mathbf{y}|\mathbf{x})$ 进行因子分解。假设条件随机场的最大团集合为 $\mathcal{C}$，其条件概率为

$$p(\mathbf{y}|\mathbf{x},\theta) = \frac{1}{Z(\mathbf{x},\theta)} \exp\Big(\sum_{c\in\mathcal{C}} \theta_c^{\mathrm{T}} f_c(\mathbf{x},\mathbf{y}_c)\Big), \tag{11.25}$$

其中 $Z(\mathbf{x},\theta) = \sum_y \exp(\sum_{c\in\mathcal{C}} f_c(\mathbf{x},\mathbf{y}_c)^{\mathrm{T}}\theta_c)$ 为归一化项。

一个最常用的条件随机场为图11.8b中所示的链式结构，其条件概率为

$$p(\mathbf{y}|\mathbf{x},\theta) = \frac{1}{Z(\mathbf{x},\theta)} \exp\Big(\sum_{t=1}^{T} \theta_1^{\mathrm{T}} f_1(\mathbf{x},y_t) + \sum_{t=1}^{T-1} \theta_2^{\mathrm{T}} f_2(\mathbf{x},y_t,y_{t+1})\Big), \tag{11.26}$$

其中 $f_1(\mathbf{x},y_t)$ 为*状态特征*，一般和位置 t 相关，$f_2(\mathbf{x},y_t,y_{t+1})$ 为*转移特征*，一般可以简化为 $f_2(y_t,y_{t+1})$ 并使用状态转移矩阵来表示。

11.1.6 有向图和无向图之间的转换

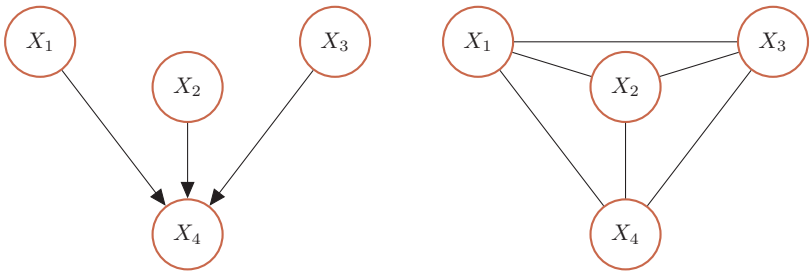
无向图模型可以表示有向图模型无法表示的一些依赖关系，比如循环依赖；但它不能表示有向图模型能够表示的某些关系，比如因果关系。

以图11.9a中的有向图为例，其联合概率分布可以分解为

$$p(\mathbf{x}) = p(x_1)p(x_2)p(x_3)p(x_4|x_1,x_2,x_3), \tag{11.27}$$

其中 $p(x_4|x_1,x_2,x_3)$ 和四个变量都相关。如果要转换为无向图，需要将这四个变量都归属于一个团中。因此需要将 x_4 的三个父节点之间都加上连边，如图11.9b所示。这个过程称为*道德化*（moralization）。转换后的无向图称为*道德图*（moral

道德化的名称来源是：有共同儿子的父节点都必须结婚（即有连边）。



(a) 有向图 (b) 道德图

图 11.9: 具有共果关系的有向图的道德化示例。

graph）。在道德化的过程中，原来有向图的一些独立性会丢失，比如上面例子中 $X_1 \perp\!\!\!\perp X_2 \perp\!\!\!\perp X_2|\emptyset$ 在道德图中不再成立。

11.2 推断

在图模型中，*推断*（inference）是指在观测到部分变量 $\mathbf{e} = \{e_1, e_2, \cdots, e_m\}$ 时，计算其它变量的某个子集 $\mathbf{q} = \{q_1, q_2, \cdots, q_n\}$ 的后验概率 $p(\mathbf{q}|\mathbf{e})$。

假设一个图模型中，除了变量 $\mathbf{e}$、$\mathbf{q}$ 外，其余变量表示为 $\mathbf{z}$。根据贝叶斯公式有

不失一般性，这里假设所有变量都为离散变量。

$$p(\mathbf{q}|\mathbf{e}) = \frac{p(\mathbf{q}, \mathbf{e})}{p(\mathbf{e})} \tag{11.28}$$

$$= \frac{\sum_{\mathbf{z}} p(\mathbf{q}, \mathbf{e}, \mathbf{z})}{\sum_{\mathbf{q},\mathbf{z}} p(\mathbf{q}, \mathbf{e}, \mathbf{z})}. \tag{11.29}$$

因此，图模型的推断问题可以转换为求任意一个变量子集的边际概率分布问题。

在图模型中，常用的推断方法可以分为*精确推断*和*近似推断*两类。虽然

本节介绍两种精确推断算法，下一节介绍近似推断算法。

11.2.1 变量消除法

以图11.2a的有向图为例，假设推断问题为计算后验概率 $p(x_1|x_4)$，需要计算两个边际概率 $p(x_1, x_4)$ 和 $p(x_4)$。

根据条件独立性假设，有

$$p(x_1, x_4) = \sum_{x_2, x_3} p(x_1)p(x_2|x_1)p(x_3|x_1)p(x_4|x_2, x_3), \tag{11.30}$$

假设每个变量取 K 个值，计算上面的边际分布需要 K^2 次加法以及 $K^2 \times 4$ 次乘法。

根据乘法的分配律，

$$ab + ac = a(b + c), \tag{11.31}$$

边际概率 $p(x_1, x_4)$ 可以写为

$$p(x_1, x_4) = p(x_1) \sum_{x_3} p(x_3|x_1) \sum_{x_2} p(x_2|x_1)p(x_4|x_2, x_3). \tag{11.32}$$

这样计算量可以减少到 $K^2 + K$ 次加法和 $K^2 + K + 1$ 次乘法。

这种方法是利用动态规划的思想，每次消除一个变量，来减少计算边际分布的计算复杂度，称为**变量消除法**（variable elimination algorithm）。随着图模型规模的增长，变量消除法的收益越大。

变量消除法可以按照不同的顺序来消除变量。比如上面的推断问题也可以按照 x_3，x_2 的消除顺序进行计算。

同理，边际概率 $p(x_4)$ 可以通过以下方式计算：

$$p(x_4) = \sum_{x_3} p(x_3|x_1) \sum_{x_2} p(x_4|x_2, x_3) \sum_{x_1} p(x_2|x_1)p(x_1). \tag{11.33}$$

变量消除法的一个缺点是在计算多个边际分布时存在很多重复的计算。比如在上面的图模型中，如果计算边际概率 $p(x_4)$ 和 $p(x_3)$ 时很多局部的求和计算是一样的。

11.2.2 信念传播

信念传播（belief propagation），也称为 *sum-product* **算法**或**消息传递算法**（message passing），是将变量消除法中的**和积**（sum-product）操作看作是消息，并保存起来，这样可以节省大量的计算资源。

本节以无向图为例来介绍信念传播，但其同样适用于有向图。

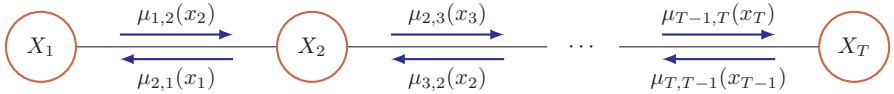


图 11.10: 无向马尔科夫链的消息传递过程。

11.2.2.1 链式结构上的的信念传播

以图11.10所示的无向马尔可夫链为例，其联合概率 $p(\mathbf{x})$ 为

$$p(\mathbf{x}) = \frac{1}{Z}\prod_{c\in\mathcal{C}}\phi_c(\mathbf{x}_c) \tag{11.34}$$

$$= \frac{1}{Z}\prod_{t=1}^{T-1}\phi(x_t, x_{t+1}) \tag{11.35}$$

其中 $\phi(x_t, x_{t+1})$ 是定义在团 (x_t, x_{t+1}) 的势能函数。

第 t 个变量的边际概率 $p(x_t)$ 为

$$p(x_t) = \sum_{x_1}\cdots\sum_{x_{t-1}}\sum_{x_{t+1}}\cdots\sum_{x_T} p(\mathbf{x}) \tag{11.36}$$

$$= \frac{1}{Z}\sum_{t_1}\cdots\sum_{x_{t-1}}\sum_{x_{t+1}}\cdots\sum_{x_T}\prod_{t=1}^{T-1}\phi(x_t, x_{t+1}). \tag{11.37}$$

假设每个变量取 K 个值，不考虑归一化项，通过公式 (11.37) 计算边际分布需要 K^{T-1} 次加法以及 $K^{T-1}\times(T-1)$ 次乘法。

根据乘法的分配律，边际概率 $p(x_t)$ 可以通过下面方式进行计算：

$$\begin{aligned}
p(x_t) &= \frac{1}{Z}\left(\sum_{x_1}\cdots\sum_{x_{t-1}}\prod_{j=1}^{t-1}\phi(x_j, x_{j+1})\right)\cdot\left(\sum_{x_{t+1}}\cdots\sum_{x_T}\prod_{j=t}^{T-1}\phi(x_j, x_{j+1})\right) \\
&= \frac{1}{Z}\left(\sum_{x_{t-1}}\phi(x_{t-1}, x_t)\cdots\left(\sum_{x_2}\phi(x_2, x_3)\Big(\sum_{x_1}\phi(x_1, x_2)\Big)\right)\right)\cdot \\
&\quad\left(\sum_{x_{t+1}}\phi(x_t, x_{t+1})\cdots\left(\sum_{x_{T-1}}\phi(x_{T-2}, x_{T-1})\Big(\sum_{x_T}\phi(x_{T-1}, x_T)\Big)\right)\right) \\
&= \frac{1}{Z}\mu_{t-1,t}(x_t)\mu_{t+1,t}(x_t),
\end{aligned} \tag{11.38}$$

其中 $\mu_{t-1,t}(x_t)$ 定义为变量 X_{t-1} 向变量 X_t 传递的消息（message）。$\mu_{t-1,t}(x_t)$ 是关于变量 X_t 的函数，可以递归计算：

$$\mu_{t-1,t}(x_t) \triangleq \sum_{x_{t-1}} \phi(x_{t-1}, x_t)\mu_{t-2,t-1}(x_{t-1}). \tag{11.39}$$

$\mu_{t+1,t}(x_t)$ 是变量 X_{t+1} 向变量 X_t 传递的消息，定义为

$$\mu_{t+1,t}(x_t) \triangleq \sum_{x_{t+1}} \phi(x_t, x_{t+1})\mu_{t+2,t+2}(x_{t+1}). \tag{11.40}$$

边际概率 $p(x_t)$ 的计算复杂度减少为 $O(TK^2)$。如果要计算整个序列上所有变量的边际概率，不需要将消息传递的过程重复 T 次，因为其中每两个相邻节点上的消息是相同的。

链式结构图模型的信念传播过程为

1. 依次计算前向传递的消息 $\mu_{t-1,t}(x_t)$，$t = 1, \cdots, T-1$；
2. 依次计算反向传递的消息 $\mu_{t+1,t}(x_t)$，$t = T-1, \cdots, 1$；
3. 在任意节点 t 上计算配分函数 Z，

$$Z = \sum_{x_t} \mu_{t-1,t}(x_t)\mu_{t+1,t}(x_t). \tag{11.41}$$

这样就可以通过公式（11.38）计算所有变量的边际概率了。

11.2.2.2 树结构上的信念传播

信念传播也可以对推广到树结构的图模型上。如果一个有向图满足任意两个变量只有 条路径（忽略方向），且只有 个没有 结点的结点，那么这个有向图为树结构，其中唯一没有父节点的节点称为根节点。如果一个无向图满足任意两个变量只有 条路径，那么这个无向图也为树结构。在树结构的无向图中，任意一个节点都可以作为根节点。

树结构图模型的信念传播过程为：1）从叶子节点到根节点依次计算并传递消息；2）从根据节点开始到叶子节点，依次计算并传递消息；3）在每个节点上计算所有接受的消息的乘积（无向图是还需要归一化），就得到了对于变量的边际概率。

如果图结构中存在环路，可以使用联合树算法（junction tree algorithm）[**?**] 来将图结构转换为无环图。

11.3 近似推断

在实际应用中，精确推断一般用于结构比较简单的推断问题上。当图模型的结构比较复杂时，精确推断的计算开销会比较大。此外，如果图模型中存在连续变量，并且这些连续变量的没有闭型（closed-form）的积分函数时，也无法使用精确推断。因此，在很多情况下也常常采用近似的方法来进行推断。

近似推断（approximate inference）主要有以下三种方法：

1. 环路信念传播： 当图模型中存在环路时，使用 sum-product 算法时，消息会在环路中一直传递，可能收敛或不收敛。环路信念传播（loopy belief propagation）是在具有环路的图上依然使用 sum-product 算法，即使得到不精确解，也某些任务上也可以近似精确解。
2. 变分法： 图模型中有些变量的局部条件分布可能非常复杂，或其积分无法计算。变方法（variational method）是引入一个变分分布（通常是比较简单的分布）来近似这些条件概率，然后通过迭代的方法进行计算。首先是更新变分分布的参数来最小化变分分布和真实分布的差异（比如交叉熵或 KL 距离），然后在根据变分分布来进行推断。
3. 采样法： 采样法（sampling method）是通过模拟的方式来采集符合某个分布 $p(x)$ 的一些样本，并通过这些样本来估计和这个分布有关的运算，比如期望等。

在本章中，我们主要介绍基于采样法的近似推断。

采样法（sampling method），也叫蒙特卡罗方法（Monte Carlo method）或统计模拟方法（,）是 1940 年代中期提出的一种通过随机采样的方法来近似估计一些计算问题的数值解。由于计算机的出现和快速发展，很多难以计算的问题都可以通过随机模拟的方法来进行估计。

蒙特卡罗方法诞生于上个世纪 40 年代美国的“曼哈顿计划”，其名字来源于摩纳哥的一个以赌博业闻名的城市**蒙特卡罗**，象征概率。

一个最简单的应用蒙特卡罗方法的例子是计算圆周率 π。我们知道半径为 r 的圆的面积为 πr^2，而直径为 $2r$ 的正方形的面积为 $4r^2$。当我们用正方形去嵌套一个相切的圆时，它们的面积之比是 $\frac{1}{4}\pi$。当不知道 π 时，我们无法计算圆的面积。因此，需要通过模拟的方法来进行近似估计。首先在正方形内部按均值采样的方式随机生成若干点，计算它们与圆心点的距离，从而判断是否落在圆的内部。然后去统计落子圆内部的点占到所有点的比例。当有足够的点时，这个比例应该接近于 $\frac{1}{4}\pi$，而从近似的估算出 π 的值。

蒙特卡罗方法有很多具体的实现方式，但其基本思想可以归结为根据一个已知分布的 $p(x)$ 来计算函数 $f(x)$ 的期望

$$\mathbb{E}[f(x)] = \int_x f(x)p(x)dx, \tag{11.42}$$

这里假设 x 为连续变量，如果 x 是离散变量，可以将积分替换为求和。

当 $p(x)$ 比较复杂时，很难用解析的方法来计算这个期望。为了计算 $\mathbb{E}[f(x)]$，我们可以通过数值解法的方法来近似计算。首先从 $p(x)$ 中独立抽取的 N 个样本 $x^{(1)}, x^{(2)}, \cdots, x^{(N)}$，$f(x)$ 的期望可以用这 N 个样本的均值 $\hat{f}_N$ 来近似。

$$\hat{f}_N = \frac{1}{N}\left(f(x^{(1)}) + \cdots + f(x^{(N)})\right). \tag{11.43}$$

根据大数定律，当 N 趋向于无穷大时，样本均值收敛于期望值。

$$\hat{f}_N \xrightarrow{P} \mathbb{E}_p[f(x)] \qquad \text{当 } N \to \infty. \tag{11.44}$$

这就是蒙特卡罗方法的理论依据。

采样 蒙特卡罗方法的难点是如何从 $p(x)$ 中抽取的一组变量，即采样（sampling）。采样是统计模拟中的核心问题，即如何让计算机生成满足概率分布 $p(x)$ 的样本。

采样也叫抽样。

cdf

如果我们希望生成一些符合分布 $p(x)$ 的样本，但是有时会难以直接对 $p(x)$ 进行采样。比如 $p(x)$ 非常复杂，其累积分布函数的逆函数难以计算，或者我们不知道 $p(x)$ 的精确值，只知道未归一化的分布 $\hat{p}(x)$

$$p(x) = \frac{1}{Z}\hat{p}(x), \tag{11.45}$$

其中 Z 为配分函数。

这种情况下，我们可以采用一种间接采样的方法。先根据一个比较容易采样的分布进行采样，然后通过一些策略来间接得到符合 $p(x)$ 分布的样本。常用的采样算法有拒绝采样、重要性采样、马尔可夫链蒙特卡罗采样等。

11.3.1 拒绝采样

拒绝采样（rejection sampling），也叫接受-拒绝采样（acceptance-rejection sampling）。

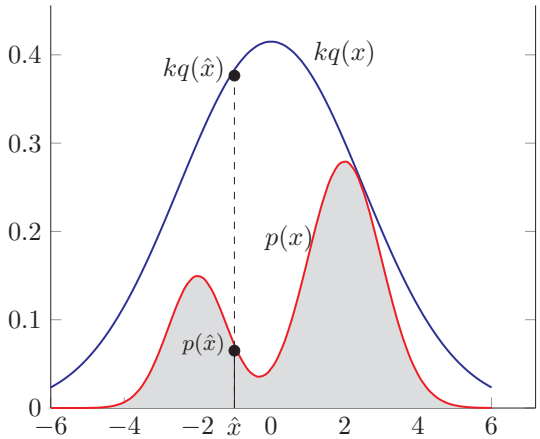


图 11.11: 拒绝采样。

因为原始分布 $p(x)$ 难以直接采样，我们可以引入一个容易采样的分布 $q(x)$，一般称为**提议分布**（proposal distribution），然后以某个标准来拒绝一部分的样本使得最终采集的样本服从分布 $p(x)$。

提议分布在很多文献中也翻译为**参考分布**。

在**拒绝采样**中，已知未归一化的分布 $\hat{p}(x)$，我们需要构建一个提议分布 $q(x)$ 和一个常数 k，使得 $kq(x)$ 可以覆盖函数 $\hat{p}(x)$，即 $kq(x) \geq \hat{p}(x), \forall x$。如图11.11所示。对于每次抽取的样本 $\hat{x}$，计算**接受概率**（acceptance probability）：

$$\alpha(\hat{x}) = \frac{\hat{p}(\hat{x})}{kq(\hat{x})}, \tag{11.46}$$

并以概率 $\alpha\hat{x}$ 来接受样本 $\hat{x}$。

拒绝采样的采样过程如下算法11.1所示。

判断**拒绝采样**的好坏就是看起采样效率，即总体的接受率。如果函数 $kq(x)$ 远大于原始分布函数 $\hat{p}(x)$，拒绝率会比较高，采样效率会非常不理想。但要找到一个和 $\hat{p}(x)$ 比较接近的提议分布往往比较困难。特别是在高维空间中，其采样率会非常低，导致很难应用的实际问题中。

11.3.2 重要性采样

重要性采样是一种非均匀采样。

算法 11.1: 拒绝采样

输入: 提议分布 $q(\mathbf{x})$;
常数 k;
样本集合 $\mathcal{V} = \varnothing$;

```
repeat
    根据 q(x) 随机生成一个样本 x̂;
    计算接受概率 α(x̂);
    从 (0,1) 的均匀分布中随机生成一个值 z;
    if z ≤ α(x̂) then                    /* 以 α(x̂) 的概率接受 x̂ */
        V = V ∪ {x̂};
    end
until 直到获得 N 个样本 (|V| = N);
```

输出: 样本集合 $\mathcal{V}$

函数 $f(x)$ 在分布 $p(x)$ 的期望可以写为

$$\mathbb{E}_p[f(x)] = \int_x f(x)p(x)dx \tag{11.47}$$

$$= \int_x f(x)\frac{p(x)}{q(x)}q(x)dx \tag{11.48}$$

$$= \int_x f(x)w(x)q(x)dx \tag{11.49}$$

$$= \mathbb{E}_q[f(x)w(x)]. \tag{11.50}$$

其中 $w(x)$ 称为重要性权重。

重要性采样（importance sampling）是通过引入重要性权重，将分布 $p(x)$ 下 $f(x)$ 的期望就变为在分布 $q(x)$ 下 $f(x)w(x)$ 的期望，从而可以近似为

$$\hat{f}_N = \frac{1}{N}\left(f(x^{(1)})w(x^{(1)}) + \cdots + f(x^{(N)})w(x^{(N)})\right), \tag{11.51}$$

其中 $x^{(1)}, \cdots, x^{(N)}$ 为独立从 $q(x)$ 中随机抽取的点。

重要性采样也可以在只知道未归一化的分布 $\hat{p}(x)$ 的情况下计算函数 $f(x)$ 的期望。

$p(x) = \frac{\hat{p}(x)}{Z}$，$Z$ 为配分函数。

$$\mathbb{E}_p[f(x)] = \int_x f(x)\frac{\hat{p}(x)}{Z}dx \tag{11.52}$$

$$= \frac{\int_x \hat{p}(x) f(x) dx}{\int_x \hat{p}(x) dx} \tag{11.53}$$

$$\approx \frac{\sum_{n=1}^{N} f(x^{(n)}) \hat{w}(x^{(n)})}{\sum_{n=1}^{N} \hat{w}(x^{(n)})}, \tag{11.54}$$

其中 $\hat{w}(x) = \frac{\hat{p}(x)}{q(x)}$，$x^{(1)}, \cdots, x^{(N)}$ 为独立从 $q(x)$ 中随机抽取的点。。

11.3.3 马尔可夫链蒙特卡罗采样

在高维空间中，拒绝采样和重要性采样的效率一般比较低。马尔可夫链蒙特卡罗（Markov chain Monte Carlo，MCMC）方法是一种更好的采样方法，可以很容易地对高维变量进行采样。

MCMC方法也有很多不同的具体采样方法，但其核心思想是将采样过程看作是一个马尔可夫链。如果这个马尔可夫链的平稳分布为 $p(\mathbf{x})$，那么在平稳状态时抽取的样本就服从 $p(\mathbf{x})$ 的分布。

马尔可夫链参见第??节，第??页。

11.3.3.1 Metropolis-Hastings 算法

Metropolis-Hastings 算法，简称MH算法，是一种应用广泛的MCMC方法。在MH算法中，马尔可夫链状态转移的提议分布为 $q(\mathbf{x}|\mathbf{x}')$，但这个分布的平稳分布不一定是 $p(\mathbf{x})$。因此，MH算法引入拒绝采样的思想来修正提议分布，使得最终采样的分布为 $p(\mathbf{x})$。

在MH算法中，假设第 t 次采样的样本为 $\mathbf{x}_t$，首先根据提议分布 $q(\mathbf{x}|\mathbf{x}_t)$ 抽取一个样本 $\hat{\mathbf{x}}$，并以概率 $A(\hat{\mathbf{x}}, \mathbf{x}_t)$ 来接受 $\hat{\mathbf{x}}$ 作为第 $t+1$ 次的采样样本 $\mathbf{x}_{t+1}$，

$$A(\hat{\mathbf{x}}, \mathbf{x}_t) = \min\left(1, \frac{p(\hat{\mathbf{x}}) q(\mathbf{x}_t|\hat{\mathbf{x}})}{p(\mathbf{x}_t) q(\hat{\mathbf{x}}|\mathbf{x}_t)}\right). \tag{11.55}$$

MH算法的采样过程如算法11.2所示。因为每次 $q(\mathbf{x}|\mathbf{x}_t)$ 随机生成一个样本 $\hat{\mathbf{x}}$，并以概率 $A(\hat{\mathbf{x}}, \mathbf{x}_t)$ 的方式接受，因此马尔可夫链的转移概率为 $q(\hat{\mathbf{x}}|\mathbf{x}_t)A(\hat{\mathbf{x}}, \mathbf{x}_t)$，该马尔可夫链可以达到平稳状态，且平稳分为 $p(\mathbf{x})$。

证明. 根据马尔可夫链的细致平稳条件，有

细致平稳条件参见定理(D.1)，第352页。

$$p(\mathbf{x}_t) q(\hat{\mathbf{x}}|\mathbf{x}_t) A(\hat{\mathbf{x}}, \mathbf{x}_t) = p(\mathbf{x}_t) q(\hat{\mathbf{x}}|\mathbf{x}_t) \min\left(1, \frac{p(\hat{\mathbf{x}}) q(\mathbf{x}_t|\hat{\mathbf{x}})}{p(\mathbf{x}_t) q(\hat{\mathbf{x}}|\mathbf{x}_t)}\right) \tag{11.56}$$

$$= \min\Big(p(\mathbf{x}_t) q(\hat{\mathbf{x}}|\mathbf{x}_t), p(\hat{\mathbf{x}}) q(\mathbf{x}_t|\hat{\mathbf{x}})\Big) \tag{11.57}$$

算法 11.2: Metropolis-Hastings 算法

输入: 提议分布 $q(\mathbf{x}|\mathbf{x}')$;
采样间隔 M;
样本集合 $\mathcal{V} = \varnothing$;
随机初始化 $\mathbf{x}_0$;
$t = 0$;
repeat
  // 预热过程
  根据 $q(\mathbf{x}|\mathbf{x}_t)$ 随机生成一个样本 $\hat{\mathbf{x}}$;
  计算接受概率 $A(\hat{\mathbf{x}}, \mathbf{x}_t)$;
  从 $(0, 1)$ 的均匀分布中随机生成一个值 z;
  if $z \leq \alpha$ **then** /* 以 $A(\hat{\mathbf{x}}, \mathbf{x}_t)$ 的概率接受 $\hat{\mathbf{x}}$ */
    $\mathbf{x}_{t+1} = \hat{\mathbf{x}}$;
  else /* 拒绝接受 $\hat{\mathbf{x}}$ */
    $\mathbf{x}_{t+1} = \mathbf{x}_t$;
  end
  t++;
  if 未到平稳状态 **then**
    continue;
  end
  // 采样过程，每隔 M 次采一个样本
  if $t \mod M = 0$ **then**
    $\mathcal{V} = \mathcal{V} \cup \{\mathbf{x}_t\}$;
  end
until 直到获得 N 个样本 $(|\mathcal{V}| = N)$;
输出: 样本集合 $\mathcal{V}$

$$= p(\hat{\mathbf{x}})q(\mathbf{x}_t|\hat{\mathbf{x}}) \min\left(\frac{p(\mathbf{x}_t)q(\hat{\mathbf{x}}|\mathbf{x}_t)}{p(\hat{\mathbf{x}})q(\mathbf{x}_t|\hat{\mathbf{x}})}, 1\right) \tag{11.58}$$

$$= p(\hat{\mathbf{x}})q(\mathbf{x}_t|\hat{\mathbf{x}})A(\mathbf{x}_t, \hat{\mathbf{x}}). \tag{11.59}$$

因此，$p(\mathbf{x})$ 是转移概率为 $q(\hat{\mathbf{x}}|\mathbf{x}_t)A(\hat{\mathbf{x}}, \mathbf{x}_t)$ 的马尔可夫链的的平稳分布。 □

11.3.3.2 Metropolis 算法

如果 MH 算法中的提议分布是对称的，即 $q(\hat{\mathbf{x}}|\mathbf{x}_t) = q(\mathbf{x}_t|\hat{\mathbf{x}})$，第 $t+1$ 次采样的接受率可以简化为

$$A(\hat{\mathbf{x}}, \mathbf{x}_t) = \min\left(1, \frac{p(\hat{\mathbf{x}})}{p(\mathbf{x}_t)}\right). \tag{11.60}$$

这种 MCMC 方法称为 *Metropolis 算法*。

11.3.3.3 吉布斯采样

吉布斯采样（Gibbs sampling）是一种有效地对高维空间中的分布进行采样的 MCMC 方法，可以看作是 Metropolis-Hastings 算法的特例。吉布斯采样使用*全条件概率*作为提议分布来依次对每个维度进行采样，并设置接受率为 $A = 1$。

对于一个 M 维的随机向量 $\mathbf{X} = [X_1, X_2, \cdots, X_M]^{\mathrm{T}}$，其第 i 个变量 X_i 的*全条件概率*（full conditional probability）为

$$p(x_i|\mathbf{x}_{-i}) \triangleq P(X_i = x_i|\mathbf{x}_{-i}) \tag{11.61}$$

$$= p(x_i|x_1, x_2, \cdots, x_{i\text{-}1}, x_{i+1}, \cdots, x_M), \tag{11.62}$$

其中 $\mathbf{x}_{-i} = [x_1, x_2, \cdots, x_{i\text{-}1}, x_{i+1}, \cdots, x_M]^{\mathrm{T}}$ 表示除 X_i 外其它变量的取值。

吉布斯采样可以按照任意的顺序根据全条件分布依次对每个变量进行采样。假设从一个随机的初始化状态 $\mathbf{x}^{(0)} = [x_1^{(0)}, x_2^{(0)}, \cdots, x_M^{(0)}]^{\mathrm{T}}$ 开始，按照*下标顺序*依次对 M 个变量进行采样。

$$x_1^{(1)} \sim P(X_1|x_2^{(0)}, x_3^{(0)}, \cdots, x_M^{(0)}), \tag{11.63}$$

$$x_2^{(1)} \sim P(X_2|x_1^{(1)}, x_3^{(0)} \cdots, x_M^{(0)}), \tag{11.64}$$

$$\vdots$$

$$x_M^{(1)} \sim P(X_M|x_1^{(1)}, x_2^{(1)} \cdots, x_{M-1}^{(0)}), \tag{11.65}$$

$$\vdots$$

$$x_1^{(t)} \sim P(X_1|x_2^{(t-1)}, x_3^{(t-1)}, \cdots, x_M^{(t-1)}), \tag{11.66}$$

$$x_2^{(t)} \sim P(X_2|x_1^{(t)}, x_3^{(t-1)} \cdots, x_M^{(t-1)}), \tag{11.67}$$

$$\vdots$$

$$x_M^{(t)} \sim P(X_M|x_1^{(t)}, x_2^{(t)} \cdots, x_{M-1}^{(t)}), \tag{11.68}$$

其中 $x_i^{(t)}$ 是第 t 次迭代时变量 X_i 的采样。随着迭代次数 t 的增加，变量 $\mathbf{x}^{(t)} = [x_1^{(t)}, x_2^{(t)} \cdots, x_M^{(t)}]^{\mathrm{T}}$ 将收敛于概率分布 $p(\mathbf{x})$。

11.4 学习

图模型的学习可以分为两部分：一是网络结构学习，即寻找最优的网络结构。网络结构学习一般比较困难，一般是由领域专家来构建；二是网络参数估计，即已知网络结构，估计每个条件概率分布的参数。

本节只讨论在给定网络结构条件下的参数估计问题。图模型的参数估计问题又分为包含隐变量时的参数估计问题和包含隐变量时的参数估计问题。

11.4.1 不含隐变量的参数估计

如果图模型中不包含隐变量，即所有变量都是可观测的，那么网络参数一般可以直接通过最大似然来进行估计。

有向图模型 在有向图模型中，所有变量 $\mathbf{x}$ 的联合概率分布可以分解为每个随机变量 x_k 的局部条件概率 $p(x_k|x_{\pi_k}, \theta_k)$ 的连乘形式，其中 θ_k 为第 k 个变量的局部条件概率的参数。

给定 N 个训练样本 $\mathcal{D} = \{\mathbf{x}^{(i)}\}, 1 \leq i \leq N$，其对数似然函数为

$$\mathcal{L}(\mathcal{D}|\theta) = \frac{1}{N} \sum_{i=1}^{N} \log p(\mathbf{x}^{(i)}, \theta) \tag{11.69}$$

$$= \frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{K} \log p(x_k^{(i)}|x_{\pi_k}^{(i)}, \theta_k), \tag{11.70}$$

其中 θ_k 为模型中的所有参数。

因为所有变量都是可观测的，最大化对数似然 $\mathcal{L}(\mathcal{D}|\theta))$，只需要分别地最大

化每个变量的条件似然来估计其参数。

$$\theta_k = \arg\max \sum_{i=1}^{N} \log p(x_k^{(i)} | x_{\pi_k}^{(i)}, \theta_k). \tag{11.71}$$

如果变量$\mathbf{x}$是离散的，直接简单的方式是在训练集上统计每个变量的条件概率表。但是条件概率表需要的参数比较多。假设条件概率$p(x_k|x_{\pi_k})$的父节点数量为M，所有变量为二值变量，其条件概率表需要2^M个参数。为了减少参数数量，可以使用参数化的模型，比如*sigmoid*信念网络。如果变量$\mathbf{x}$是连续的，可以使用高斯函数来表示条件概率分布，称为高斯信念网络。在此基础上，还可以通过让所有的条件概率分布共享使用同一组参数来进一步减少参数的数量。

无向图模型 在无向图模型中，所有变量$\mathbf{x}$的联合概率分布可以分解为定义在最大团上的势能函数的连乘形式。以对数线性模型为例，

$$p(\mathbf{x}|\theta) = \frac{1}{Z(\theta)} \exp\Big(\sum_{c\in\mathcal{C}} \theta_c^{\mathrm{T}} f_c(\mathbf{x}_c) \Big), \tag{11.72}$$

其中$Z(\theta) = \sum_{\mathbf{x}} \exp(\sum_{c\in\mathcal{C}} \theta_c^{\mathrm{T}} f_c(\mathbf{x}_c))$。

给定N个训练样本$\mathcal{D} = \{\mathbf{x}^{(i)}\}, 1 \leq i \leq N$，其对数似然函数为

$$\mathcal{L}(\mathcal{D}|\theta) = \frac{1}{N} \sum_{i=1}^{N} \log p(\mathbf{x}^{(i)}, \theta) \tag{11.73}$$

$$= \frac{1}{N} \sum_{i=1}^{N} \Big(\sum_{c\in C} \theta_c^{\mathrm{T}} f_c(\mathbf{x}_c^{(i)}) \Big) - \log Z(\theta), \tag{11.74}$$

其中θ_c为定义在团c上的势能函数的参数。

如果采用梯度上升方法进行最大似然估计，$\mathcal{L}(\mathcal{D}|\theta)$关于参数$\theta_c$的偏导数为

$$\frac{\partial \mathcal{L}(\mathcal{D}|\theta)}{\partial \theta_c} = \frac{1}{N} \sum_{i=1}^{N} \Big(f_c(\mathbf{x}_c^{(i)}) \Big) - \frac{\log Z(\theta)}{\partial \theta_c} \tag{11.75}$$

其中

$$\frac{\log Z(\theta)}{\partial \theta_c} = \sum_{\mathbf{x}} \frac{1}{Z(\theta)} \cdot \exp\Big(\sum_{c\in\mathcal{C}} \theta_c^{\mathrm{T}} f_c(\mathbf{x}_c) \Big) \cdot f_c(\mathbf{x}_c) \tag{11.76}$$

$$= \sum_{\mathbf{x}} p(\mathbf{x}|\theta) f_c(\mathbf{x}_c) \triangleq \mathbb{E}_{\mathbf{x}\sim p(\mathbf{x}|\theta)} \Big[f_c(\mathbf{x}_c) \Big]. \tag{11.77}$$

因此，

$$\frac{\partial \mathcal{L}(\mathcal{D}|\theta)}{\partial \theta_c} = \frac{1}{N}\sum_{i=1}^{N} f_c(\mathbf{x}_c^{(i)}) - \mathbb{E}_{\mathbf{x}\sim p(\mathbf{x}|\theta)}\Big[f_c(\mathbf{x}_c)\Big] \tag{11.78}$$

$$= \mathbb{E}_{\mathbf{x}\sim \tilde{p}(\mathbf{x})}\Big[f_c(\mathbf{x}_c)\Big] - \mathbb{E}_{\mathbf{x}\sim p(\mathbf{x}|\theta)}\Big[f_c(\mathbf{x}_c)\Big], \tag{11.79}$$

其中 $\tilde{p}(\mathbf{x})$ 定义为经验数据分布。由于在最优点时梯度为0，因此无向图的最大似然估计的优化目标等价于：对于每个团 c 上的特征 $f_c(\mathbf{x}_c)$，其在经验分布 $\tilde{p}(\mathbf{x})$ 下的期望等于模型分布 $p(\mathbf{x}|\theta)$ 下的期望。

对比公式 (11.71) 和公式 (11.79) 可以看出，无向图模型的参数估计要比有向图更为复杂。在有向图中，每个局部条件概率的参数是独立的；而在无向图中，所有的参数都是相关的，无法分解。

对于一般的无向图模型，公式 (11.79) 中的 $\mathbb{E}_{\mathbf{x}\sim p(\mathbf{x}|\theta)}[f_c(\mathbf{x}_c)]$ 往往很难计算，因为涉及到在联合概率空间 $p(\mathbf{x}|\theta)$ 计算期望。当模型变量比较多时，这个计算往往无法实现。因此，无向图的参数估计通常采用近似的方法。一种近似方法是利用采样来近似计算这个期望。另一种近似方法是*坐标上升法*，即固定其它参数，来优化一个势能函数的参数。

11.4.2 含隐变量的参数估计

如果图模型中包含隐变量，即有部分变量是不可观测的，就需要用EM算法进行参数估计。

11.4.2.1 EM 算法

边际似然也称为**证据**(evidence)。

在一个包含隐变量的图模型中，令 $\mathbf{X}$ 定义可观测变量集合，令 $\mathbf{Z}$ 定义隐变量集合，一个样本 $\mathbf{x}$ 的*边际似然函数*（marginal likelihood）为

$$p(\mathbf{x}|\theta) = \sum_{\mathbf{z}} p(\mathbf{x}, \mathbf{z}|\theta), \tag{11.80}$$

其中 θ 为模型参数。图11.12给出了带隐变量的贝叶斯网络的图模型结构。

给定 N 个训练样本 $\mathcal{D} = \{\mathbf{x}^{(i)}\}, 1 \leq i \leq N$，其训练集的*对数边际似然*为

$$\mathcal{L}(\mathcal{D}|\theta) = \frac{1}{N}\sum_{i=1}^{N} \log p(\mathbf{x}^{(i)}, \theta) \tag{11.81}$$

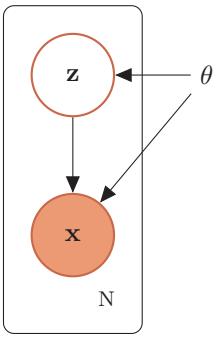


图 11.12: 带隐变量的贝叶斯网络。图中的矩形表示其中的变量重复 N 次。这种表示方法称为*盘子表示法*（plate notation），是图模型中表示重复变量的方法。

$$= \frac{1}{N}\sum_{i=1}^{N}\log\sum_{\mathbf{z}} p(\mathbf{x}^{(i)},\mathbf{z}|\theta). \tag{11.82}$$

通过最大化整个训练集的对数边际似然 $\mathcal{L}(\mathcal{D}|\theta)$，可以估计出最优的参数 θ^*。然而计算边际似然函数时涉及 $p(x)$ 的*推断*问题，需要在对数函数内部进行积分。除非 $p(\mathbf{x},\mathbf{z}|\theta)$ 的形式非常简单，否则这个积分难以直接计算。

为了计算 $\log p(\mathbf{x}|\theta)$，我们引入一个额外的*变分函数* $q(\mathbf{z})$，$q(\mathbf{z})$ 为定义在隐变量 $\mathbf{Z}$ 上的分布。这样，样本 $\mathbf{x}$ 的边际对数似然函数为

$$\log p(\mathbf{x}|\theta) = \log\sum_{\mathbf{z}} q(\mathbf{z})\frac{p(\mathbf{x},\mathbf{z}|\theta)}{q(\mathbf{z})} \tag{11.83}$$

$$\geq \sum_{\mathbf{z}} q(\mathbf{z})\log\frac{p(\mathbf{x},\mathbf{z}|\theta)}{q(\mathbf{z})} \tag{11.84}$$

利用 Jensen 不等式。

$$\triangleq ELBO(q,\mathbf{x}|\theta), \tag{11.85}$$

其中 $ELBO(q,\mathbf{x}|\theta)$ 为对数边际似然函数 $\log p(\mathbf{x}|\theta)$ 的下界，称为*证据下界*（evidence lower bound，ELBO）。公式 (11.84) 使用了 Jensen 不等式。由 Jensen 不等式的性质可知，仅当 $q(\mathbf{z}) = p(\mathbf{z}|\mathbf{x},\theta)$ 时，对数边际似然函数 $\log p(\mathbf{x}|\theta)$ 和其下界 $ELBO(q,\mathbf{x}|\theta)$ 相等，

Jensen 不等式参见第D.2.6.1节，第350页。

$$\log p(\mathbf{x}|\theta) = ELBO(q,\mathbf{x}|\theta).$$

参见习题（11-4），第223页。

这样最大化对数边际似然函数 $\log p(\mathbf{x}|\theta)$ 的过程可以分解为两个步骤：（1）先找到近似分布 $q(\mathbf{z})$ 使得 $\log p(\mathbf{x}|\theta) = ELBO(q,\mathbf{x}|\theta)$；（2）再寻找参数 θ 最大化 $ELBO(q,\mathbf{x}|\theta)$。这就是*期望最大化*（expectation-maximum，EM）算法。

*EM 算法*是含隐变量图模型的常用参数估计方法，通过迭代的方法来最大化边际似然。EM 算法具体分为两个步：E 步和 M 步。这两步不断重复，直到收敛到某个局部最优解。在第 t 步时，

1. E 步（expectation step）：固定参数 θ_t，找到一个分布

$$q_{t+1}(\mathbf{z}) = \arg\max_{q} ELBO(q, \mathbf{x}|\theta_t). \tag{11.86}$$

   根据 Jensen 不等式的性质，$q(\mathbf{z}) = p(\mathbf{z}|\mathbf{x}, \theta_t)$ 时，$ELBO(q, \mathbf{x}|\theta_t)$ 最大。因此，这一步可以看作是一种推断问题，计算后验概率 $p(\mathbf{z}|\mathbf{x}, \theta_t)$。

2. M 步（maximization step）：固定 $q(\mathbf{z})$，找到一组参数使得证据下界最大，即

$$\theta_{t+1} = \arg\max_{\theta} ELBO(q_{t+1}, \mathbf{x}|\theta). \tag{11.87}$$

   这一步可以看作是全观测变量图模型的参数估计问题，可以使用第11.4.1节中方法进行参数估计。

收敛性证明 假设在第 t 步时参数为 θ_t，在 E 步时找到一个变分分布 $q_{t+1}(\mathbf{z})$ 使得 $\log p(\mathbf{x}|\theta_t) = ELBO(q, \mathbf{x}|\theta_t)$。在 M 步时固定 $q_{t+1}(\mathbf{z})$ 找到一组参数，$ELBO(q_{t+1}, \mathbf{x}|\theta_{t+1}) \geq ELBO(q_{t+1}, \mathbf{x}|\theta_t)$。因此有

$$\log p(\mathbf{x}|\theta_{t+1}) \geq ELBO(q_{t+1}, \mathbf{x}|\theta_t) \geq ELBO(q_t, \mathbf{x}|\theta_t) = \log p(\mathbf{x}|\theta_t), \tag{11.88}$$

即每经过一次迭代对数边际似然增加，$\log p(\mathbf{x}|\theta_{t+1}) \geq \log p(\mathbf{x}|\theta_t)$。

在 E 步中，最理想的变分分布 $q(\mathbf{z})$ 是等于后验分布 $p(\mathbf{z}|\mathbf{x}, \theta)$。而后验分布 $p(\mathbf{z}|\mathbf{x}, \theta)$ 是一个推断问题。如果 $\mathbf{z}$ 是有限的一维离散变量（比如混合高斯模型），计算起来还比较容易。否则，$p(\mathbf{z}|\mathbf{x}, \theta)$ 一般情况下很难计算的。因此需要通过近似推断的方法来进行估计，比如变分自编码器。

变分自编码器参见第13.1节，第246页。

信息论的视角 对数边际似然可以通过下面方式进行分解：

$\sum_{\mathbf{z}} q(\mathbf{z}) = 1.$

$p(\mathbf{x}, \mathbf{z}|\theta) = p(\mathbf{z}|\mathbf{x}, \theta)p(\mathbf{x}|\theta).$

$$\log p(\mathbf{x}|\theta) = \sum_{\mathbf{z}} q(\mathbf{z}) \log p(\mathbf{x}|\theta) \tag{11.89}$$

$$= \sum_{\mathbf{z}} q(\mathbf{z}) \Big(\log p(\mathbf{x}, \mathbf{z}|\theta) - \log p(\mathbf{z}|\mathbf{x}, \theta) \Big) \tag{11.90}$$

$$= \sum_{\mathbf{z}} q(\mathbf{z}) \log \frac{p(\mathbf{x}, \mathbf{z}|\theta)}{q(\mathbf{z})} - \sum_{\mathbf{z}} q(\mathbf{z}) \log \frac{p(\mathbf{z}|\mathbf{x}, \theta))}{q(\mathbf{z})} \tag{11.91}$$

$$= ELBO(q, \mathbf{x}|\theta) + D_{\mathrm{KL}}(q(\mathbf{z}) \| p(\mathbf{z}|\mathbf{x}, \theta)), \tag{11.92}$$

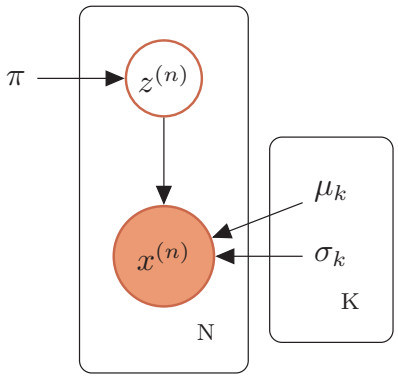


图 11.13: 高斯混合模型。

其中 $D_{\mathrm{KL}}(q(\mathbf{z})\|p(\mathbf{z}|\mathbf{x},\theta))$ 为分布 $q(\mathbf{z})$ 和后验分布 $p(\mathbf{z}|\mathbf{x},\theta)$ 的KL散度。

参见第E.3.2节，第357页。

由于 $D_{\mathrm{KL}}(q(\mathbf{z})\|p(\mathbf{z}|\mathbf{x},\theta)) \geq 0$，并当且仅当 $q(\mathbf{z}) = p(\mathbf{z}|\mathbf{x},\theta)$ 为0，因此 $ELBO(q,\mathbf{x}|\theta)$ 为 $\log p(\mathbf{x}|\theta)$ 的一个下界。

11.4.2.2 高斯混合模型

图11.13给出了高斯混合模型的图模型表示。

本节介绍一个EM算法的应用例子：高斯混合模型。高斯混合模型（Gaussian mixture model，GMM）是由多个高斯分布组成的模型，其密度函数为多个高斯密度函数的加权组合。

不失一般性，这里考虑一维的情况。假设样本 x 从 K 个高斯分布中生成的。每个高斯分布为

$$\mathcal{N}(x|\mu_k,\sigma_k) = \frac{1}{\sqrt{2\pi}\sigma_k}\exp\Big(-\frac{(x-\mu_k)^2}{2\sigma_k^2}\Big), \tag{11.93}$$

其中 μ_k 和 σ_k 分别为第 k 个高斯分布的均值和方差。

高斯混合模型的概率密度函数为

$$p(x) = \sum_{k=1}^{K}\pi_k\mathcal{N}(x|\mu_k,\sigma_k), \tag{11.94}$$

其中 π_k 表示第 K 个高斯分布的权重系数并满足 $\pi_k \leq 0, \sum_{k=1}^{K}\pi_k = 1$，即样本 x 由第 K 个高斯产生的概率。

高斯混合模型的生成过程可以分为两步：

1. 首先按 $\pi_1, \pi_2, \cdots, \pi_K$ 的分布，随机选取一个高斯分布；
2. 假设选中第 k 个高斯分布，再从高斯分布 $\mathcal{N}(x|\mu_k, \sigma_k)$ 中选取一个样本 x。

参数估计 给定 N 个由高斯混合模型生成的训练样本 $x^{(1)}, x^{(2)}, \cdots, x^{(N)}$，希望能学习其中的参数 $\pi_k, \mu_k, \sigma_k, 1 \leq k \leq K$。由于我们无法观测样本 $x^{(n)}$ 是从哪个高斯分布生成的，因此无法直接用最大似然来进行参数估计。我们引入一个隐变量 $z^{(n)} \in [1, K]$ 来表示其来自于哪个高斯分布，$z^{(n)}$ 服从多项分布，其多项分布的参数为 $\pi_1, \pi_2, \cdots, \pi_K$，即

$$p(z^{(n)} = k) = \pi_k. \tag{11.95}$$

对每个样本 $x^{(n)}$，其对数边际分布为

$$\log p(x^{(n)}) = \log \sum_{z^{(n)}} p(z^{(n)}) p(x^{(n)}|z^{(n)}) \tag{11.96}$$

$$= \log \sum_{k=1}^{K} \pi_k \mathcal{N}(x^{(n)}|\mu_k, \sigma_k). \tag{11.97}$$

根据 EM 算法，参数估计可以分为两步进行迭代：

E 步 先固定参数 μ, σ，计算后验分布 $p(z^{(n)}|x^{(n)})$

$$\gamma_{nk} \triangleq p(z^{(n)} = k|x^{(n)}) \tag{11.98}$$

$$= \frac{p(z^{(n)}) p(x^{(n)}|z^{(n)})}{p(x^{(n)})} \tag{11.99}$$

$$= \frac{\pi_k \mathcal{N}(x^{(n)}|\mu_k, \sigma_k)}{\sum_{k=1}^{K} \pi_k \mathcal{N}(x^{(n)}|\mu_k, \sigma_k)}, \tag{11.100}$$

其中 γ_{nk} 定义了样本 $x^{(n)}$ 属于第 k 个高斯分布的后验概率。

M 步 令 $q(z = k) = \gamma_{nk}$，训练集 $\mathcal{D}$ 的证据下界为

$$ELBO(\gamma, \mathcal{D}|\pi, \mu, \sigma) = \log \sum_{k=1}^{K} \gamma_{nk} \log \frac{p(x^{(n)}, z^{(n)} = k)}{\gamma_{nk}} \tag{11.101}$$

$$= \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma_{nk} \left(\log \mathcal{N}(x^{(n)}|\mu_k, \sigma_k) + \log \frac{\pi_k}{\gamma_{nk}} \right) \tag{11.102}$$

$$= \sum_{n=1}^{N}\sum_{k=1}^{K}\gamma_{nk}\left(\frac{-(x-\mu_k)^2}{2\sigma_k^2} - \log\sigma_k + \log\pi_k\right) + C, \tag{11.103}$$

其中C为和参数无关的常数。

因此，参数估计问题可以转为优化问题。

$$\begin{aligned} &\max_{\pi,\mu,\sigma} ELBO(\gamma, \mathcal{D}|\pi,\mu,\sigma), \\ &s.t. \sum_{k=1}^{K}\pi_k = 1. \end{aligned} \tag{11.104}$$

利用拉格朗日方法，分别求$ELBO(\gamma, \mathcal{D}|\pi,\mu,\sigma) + \lambda(\sum_{k=1}^{K}\pi_k - 1)$关于$\pi_k, \mu_k, \sigma_k$的偏导数，并令其等于0。可得，

$$\pi_k = \frac{N_k}{N}, \tag{11.105}$$

$$\mu_k = \frac{1}{N_k}\sum_{n=1}^{N}\gamma_{nk}x^{(n)}, \tag{11.106}$$

$$\sigma_k = \frac{1}{N_k}\sum_{n=1}^{N}\gamma_{nk}(x^{(n)} - \mu_k)^2, \tag{11.107}$$

其中

$$N_k = \sum_{n=1}^{N}\gamma_{nk}. \tag{11.108}$$

参见习题（11-5），第223页。

高斯混合模型的训练过程如算法11.3所示。

11.5 总结和深入阅读

概率图模型提供了一个概率描述框架，可能将很多机器学习问题都归结概率图模型的框架中。目前是概率图模型一个非常庞大的研究领域，涉及众多的模型和算法。图11.14给出了概率图模型所涵盖的内容。

要更全面深入地了解概率图模型，可以阅读《Probabilistic Graphical Models: Principles and Techniques》[Koller and Friedman, 2009]、《Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference》[Pearl, 2014]，或机器学习书籍中的相关章节[Bishop, 2007]。

算法 11.3: 高斯混合模型的参数学习算法。

输入: 训练样本：$x^{(1)}, x^{(2)}, \cdots, x^{(N)}$;

随机初始化参数：π_k, μ_k, σ_k，$1 \leq k \leq K$;

repeat

// E 步

固定参数，根据公式 (11.100) 计算 γ_{nk}，$1 \leq k \leq K$,，$1 \leq n \leq N$;

// M 步

固定 γ_{nk}，根据公式 (11.105)，(11.106) 和 (11.107)，计算 π_k, μ_k, σ_k，$1 \leq k \leq K$;

until 对数边际分布 $\sum_{n=1}^{N} \log p(x^{(n)})$ 收敛;

输出: π_k, μ_k, σ_k，$1 \leq k \leq K$

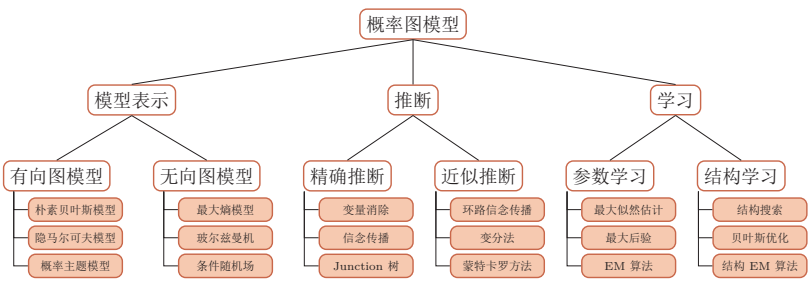


图 11.14: 概率图模型的总体框架。

概率图模型中最基本的假设是条件独立性。图形化表示直观地描述了随机变量之间的条件独立性，使我们可以更加容易地理解复杂模型的内在性质。

图模型与神经网络的关系 图模型和神经网络有着类似的网络结构，但两者也有很大的不同。图模型的节点是随机变量，其图结构的主要功能是用来描述变量之间的依赖关系，一般是稀疏连接。使用图模型的好处是可以有效进行统计推断。而神经网络中的节点是神经元，是一个计算节点。如果将神经网络中每个神经元看做是一个二值随机变量，那神经网络就变成一个 sigmoid 信念网络。

图模型中的每个变量一般有着明确的解释，变量之间依赖关系一般是人工来定义。而神经网络中的神经元则没有直观的解释。

图模型一般是生成模型，可以用生成样本，也可以通过贝叶斯公式用来做分类。而神经网络是判别模型，直接用来分类。

判别模型也可以用图模型来表示。

图模型的参数学习的目标函数为似然函数或条件似然函数，若包含隐变量则通常通过EM算法来求解。而神经网络参数学习的目标为交叉熵或平方误差等损失函数。

习题 11-1 证明公式(11.10)。

习题 11-2 在图11.2a的有向图，分析按不同的消除顺序计算边际概率$p(x_3)$时的计算复杂度。

习题 11-3 在树结构的图模型上应用信念传播时，推导其消息计算公式。

习题 11-4 证明仅当$q(\mathbf{z}) = p(\mathbf{z}|\mathbf{x}|\theta)$时，对数边际似然函数$l(\theta;\mathbf{x})$和其下界$L(q,\theta;\mathbf{x})$相等。

习题 11-5 在高斯混合分布的参数估计中，证明M步中的参数更新公式，即公式(11.105)，(11.106)和(11.107)。

习题 11-6 考虑一个伯努利混合分布，即

$$p(x|\mu,\pi) = \sum_{k=1}^{K} \pi_k p(x|\mu_k), \tag{11.109}$$

其中$p(x|\mu_k) = \mu_k^x(1-\mu_k)^{(1-x)}$为伯努利分布。

伯努利混合分布参见第D.2.1.1节，第343页。

给定一组训练集合$D = \{x^{(1)}, x^{(2)}, \cdots, x^{(N)}\}$，推导用EM算法进行参数估计的更新公式。

参考文献

Leonard E Baum and Ted Petrie. Statistical inference for probabilistic functions of finite state markov chains. *The annals of mathematical statistics*, 37(6): 1554–1563, 1966.

Adam L Berger, Vincent J Della Pietra, and Stephen A Della Pietra. A maximum entropy approach to natural language processing. *Computational linguistics*, 22(1):39–71, 1996.

Christopher M. Bishop. *Pattern recognition and machine learning, 5th Edition*. Information science and statistics. Springer, 2007. ISBN 9780387310732.

Stephen Della Pietra, Vincent Della Pietra, and John Lafferty. Inducing features of random fields. *IEEE transactions on pattern analysis and machine intelligence*, 19(4): 380–393, 1997.

Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, 2001.

Radford M Neal. Connectionist learning of belief networks. *Artificial intelligence*, 56(1):71–113, 1992.

Judea Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Elsevier, 2014.