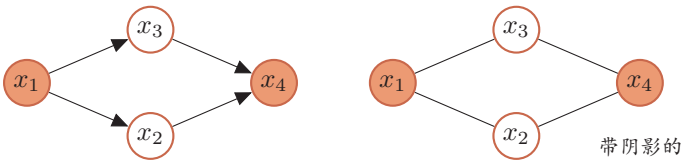
第十一章　概率图模型

在概率论和统计学中，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。图结构可以将概率模型的结构可视化，以一种直观、简单的方式描述了随机变量之间的条件独立性的性质，并可以将一个复杂的概率模型分解为一些简单模型的组合。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。很多机器学习模型都可以很作是概率模型（probabilistic model），将学习任务归结于计算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。

一个图由节点和节点之间的边组成。在概率图模型中，每一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。有向图模型也叫做贝叶斯网络，为有向非循环图，边的方向代表了因果关系。无向图模型，也叫做马尔可夫随机场，每条边代表来两个变量之间有概率依赖关系，但是并不表明任何因果关系。对于一个非全连接的图模型，都可以根据条件独立性将联合概率分布进行因子分解，表示为一组局部的条件概率分布的乘积。图11.1给出了两个代表性图模型的例子：有向图和无向图，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　贝叶斯网络

贝叶斯网络（bayesian network），又称**信念网络**（belief network），或有向图模型（directed graphical model），是指用有向图来表示概率分布的图模



(a) 有向图：贝叶斯网络 (b) 无向图：马尔可夫随机场

带阴影的节点表示可观测变量，不带阴影的节点表示不可观测的变量，连边表示两变量间的条件依赖关系。

图 11.1: 有向图和无向图示例

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，每个节点对应一个随机变量$X_k$，边集合$\mathcal{E}$中的每个连接表示两个变量之间的因果关系。

在本章后文中，“节点”与“随机变量”、“变量”的概念会经常混用。每个节点对应一个随机变量。

我们用x_i表示变量X_i的一个取值，K个变量的联合概率分布可以分解为K个条件概率的乘积。

$$p(x_1,x_2,\cdots,x_k) \triangleq P(X_1=x_1,X_2=x_2,\cdots,X_K=x_K) \tag{11.1}$$

$$= p(x_1)p(x_2|x_1)\cdots p(x_K|x_1,\cdots,x_{K-1}), \tag{11.2}$$

$$= \prod_{k=1}^{K} p(x_k|x1,\cdots,x_{k-1}), \tag{11.3}$$

其中，条件概率分布$p(x_k|x_1,\cdots,x_{k-1})$可以用图结构来表示，将在$v_1,\cdots,v_{k-1}$中的每个节点和节点$v_k$用一个有向边来连接。那么，整个联合概率分布对应的图为一个全连接的有向图。

更一般地，对任何结构（非全连接）的有向非循环图，比如图11.1a，我们也需要将其对应的概率模型进行分解。

定义 11.1 – 贝叶斯网络： 对于K随机变量$\{X_1,X_2,\cdots,X_K\}$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等；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。我们定义X_{π_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$P(X_k|X_{\pi_k})$。

如果$X=X_1,X_2,\cdots,X_K$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

$$p(x_1, x_2, \cdots, x_k) = \prod_{k=1}^{K} p(x_k|x_{\pi_k}), \tag{11.4}$$

那么 (G, X) 构成了一个**贝叶斯网络**。

局部马尔可夫性质 贝叶斯网络具有局部马尔可夫性质：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。

从公式(11.3)和(11.4)要得到习题(11-2)，第211页。

$$X_k \perp\!\!\!\perp Z|X_{\pi_k}, \tag{11.5}$$

其中，Z 为 X_k 的非后代节点的变量。

利用贝叶斯网络的局部马尔可夫性，我们可以对多元变量的联合概率进行简化，从而降低建模的复杂度。以图11.1a中的4个变量（假设每个 X_i 为二值变量）为例，如果我们不进行分解，直接用联合概率表来保存每一种的 X 的取值，则独立参数为 $2^4 - 1 = 15$ 个。

如果按图11.1a的结构进行分解，其联合概率为

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

是4个局部条件概率的乘积，这样只需要 $1 + 2 + 2 + 4 = 9$ 个独立参数。

11.1.1 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。我们先来看一个三个节点的简单例子。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11.2所示。

间接因果关系（图11.2a） 在已知 x_2 时，x_1 和 x_3 为条件独立；
间接果因关系（图11.2b） 在已知 x_2 时，x_1 和 x_3 为条件独立；
共因关系（图11.2c） x_1 和 x_3 是不独立的，在已知 x_2 时，x_1 和 x_3 条件独立；
共果关系（图11.2d） x_1 和 x_3 是独立的，在已知 x_2 时，x_1 和 x_3 不独立

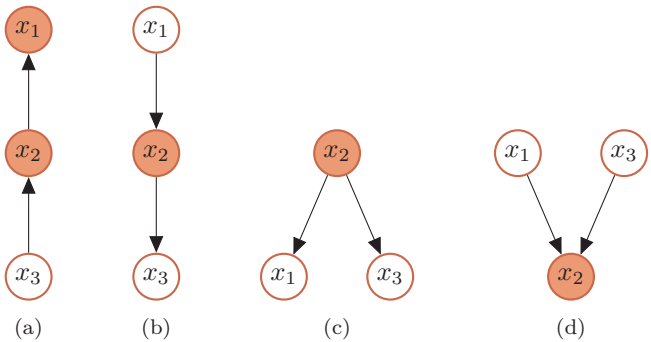


图 11.2: 由 x_1 到 x_3 并经过 x_2 的四种路径类型。在 (a)(b) 中，在已知 x_2 时，x_1 和 x_3 为条件独立；在 (c) 中，x_1 和 x_3 是不独立的，在已知 x_2 时，x_1 和 x_3 条件独立；在 (d) 中，x_1 和 x_3 是独立的，在已知 x_2 时，x_1 和 x_3 不独立。

11.1.2 常用的贝叶斯网络模型

很多经典的机器学习模型都可以看做是贝叶斯网络，比如模型朴素贝叶斯、隐马尔可夫模型、神经网络等。

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

$$p(y|\mathbf{x}) = p(y|x_1, \cdots, x_d) \tag{11.7}$$

$$= \frac{p(x_1, \cdots, x_d|y)p(y)}{p(x_1, \cdots, x_d)} \tag{11.8}$$

$$\propto p(x_1, \cdots, x_d|y)p(y). \tag{11.9}$$

$$\tag{11.10}$$

我们假设在给定 y 的情况下，x_i 之间是条件独立的，即 $x_i \perp\!\!\!\perp x_j|y, \forall i \neq j$。那么，$p(y|\mathbf{x})$ 可以简化为

$$p(y|\mathbf{x}) \propto p(y) \prod_{i=1}^{d} p(x_i|y). \tag{11.11}$$

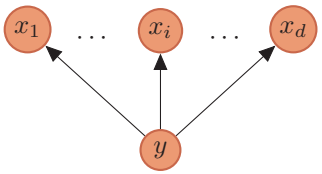


图 11.3: 朴素贝叶斯模型的图模型表示。

图11.3给出了朴素贝叶斯模型的图形表示。条件概率分布 $p(x_i|y)$ 可以是带有可学习参数的密度函数。如果 x_i 为连续值，$p(x_i|y)$ 可以为高斯分布。如果 x_i 为离散值，$p(x_i|y)$ 可以为多项分布。

显然，这个条件独立性假设太严格了。实际任务中的样本特征应该不满足这样的理想假设。但是，朴素贝叶斯分类器在很多任务上也能得到很好的结果。

Sigmoid 信念网络

Sigmoid 信念网络（Sigmoid belief network，SBN）[Neal, 1992] 是用 logistic sigmoid 函数来建模有向图中的条件概率分布。

Sigmoid 信念网络网络中的变量为二值变量，取值为 $\{0, 1\}$。对于变量 x_k 和它的父节点集合 π_k，其条件概率分布为

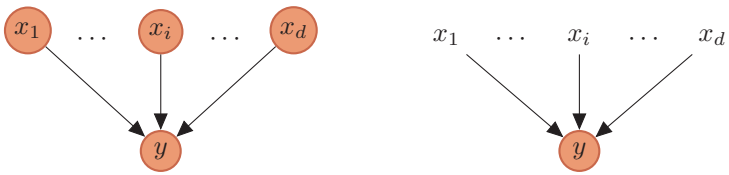
$$p(x_k = 1|x_{\pi_k}) = \sigma(w_0 + \sum_{x_i \in \pi_k} w_i x_i), \tag{11.12}$$

其中 $\sigma(\cdot)$ 是 logistic sigmoid 函数，w_i 是可学习的参数。

图11.4a给出了一个只有一层的简单 sigmoid 信念网络。

Sigmoid 信念网络与 logistic 回归模型 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。如果将 sigmoid 信念网络简化为只有一个叶子节点，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 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 回归模型也可以看做是一种条件无向图模型。



(a) 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。

11.1.3 参数估计

贝叶斯网络的学习可以分为两部分：一是网络参数估计，即给定网络结构，估计每个条件概率分布的参数；二是寻找最优的网络结构。由于后者的优化一般比较困难，因此贝叶斯网络的结构一般是由领域专家来构建。在给定网络结构的条件下，网络的参数一般通过最大似然来进行估计。

在贝叶斯网络中，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 的局部条件概率的连乘形式。假设每个局部条件概率 $p(x_k|x_{\pi_k})$ 的参数为 θ_k，则 $\mathbf{x}$ 的对数似然函数为

$$\log p(\mathbf{x}, \Theta) = \sum_{k=1}^{K} \log p(x_k|x_{\pi_k}, \theta_k), \tag{11.13}$$

其中，Θ 为模型中的所有参数。

最大化 $\mathbf{x}$ 的对数似然，只需要分别地最大化每个条件似然 $\log p(x_k|x_{\pi_k}, \theta_k)$。如果 $\mathbf{x}$ 中所有变量都是可观测的并且是离散的，只需要在训练集上统计每个变量的条件概率表即可。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。有时为了减少参数数量，我们可以使用参数化的模型，比如 logistic sigmoid 函数。如果所有变量是连续的，我们可以使用高斯函数来表示条件概率分布。前者就是 sigmoid 信念网络，后者就是高斯信念网络。在此基础上，我们可以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数量。

如果 $\mathbf{x}$ 中所有变量都是可观测的，对于每个对数条件似然函数 $\log p(x_k|x_{\pi_k}, \theta_k)$，我们可以将 x_{π_k} 作为输入变量，x_k 作为输出变量，并使用监督学习的方法来需要每个参数 θ_k。如果变量中有一部分变量为隐变量，就需要使用 EM 算法来进行参数估计。

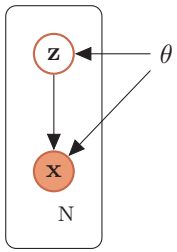


图 11.5: 带隐变量的贝叶斯网络。

11.1.4 EM 算法

假设有一组变量，有部分变量是是不可观测的，如何进行参数估计呢？

图11.5给出了带隐变量的贝叶斯网络的图模型结构。

边际似然（marginal likelihood）也称为**证据**（evidence）。

对于一个样本$\mathbf{x}$，令其“缺失”的变量为$\mathbf{z}$，样本$\mathbf{x}$的对数边际似然函数（log marginal likelihood）为

$$l(\theta; \mathbf{x}) = \log p(\mathbf{x}; \theta) \tag{11.14}$$

$$= \log \int_{\mathbf{z}} p(\mathbf{x}, \mathbf{z}; \theta) d\mathbf{z}, \tag{11.15}$$

其中θ为模型参数。

通过最大化整个训练集上所有样本的对数边际似然，可以估计出最优的参数θ^*。然而，在边际似然函数中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，计算这个积分十分麻烦。

为了可以计算$\log p(\mathbf{x}; \theta)$，我们引入一个额外的变分函数$q(\mathbf{z})$，样本$\mathbf{x}$的边

数学小知识 | Jensen 不等式

如果 X 是随机变量，g 是凸函数，则

$$g\left(\mathbb{E}[X]\right) \leq \mathbb{E}\left[g(X)\right].$$

等式当且仅当 X 是一个常数或 g 是线性时成立。

际对数似然函数为

$$
\begin{aligned}
l(\theta;\mathbf{x}) &= \log \int_{\mathbf{z}} q(\mathbf{z}) \frac{p(\mathbf{x},\mathbf{z};\theta)}{q(\mathbf{z})} d\mathbf{z} && (11.16)\\
&\geq \int_{\mathbf{z}} q(\mathbf{z}) \log \frac{p(\mathbf{x},\mathbf{z};\theta)}{q(\mathbf{z})} d\mathbf{z} && \text{利用 Jensen 不等式} \quad (11.17)\\
&= \int_{\mathbf{z}} q(\mathbf{z}) \log \frac{p(\mathbf{x}|\mathbf{z};\theta)p(\mathbf{z};\theta)}{q(\mathbf{z})} d\mathbf{z} && (11.18)\\
&= \int_{\mathbf{z}} q(\mathbf{z}) \log p(\mathbf{x}|\mathbf{z};\theta) d\mathbf{z} + \int_{\mathbf{z}} q(\mathbf{z}) \log \frac{p(\mathbf{z};\theta)}{q(\mathbf{z})} d\mathbf{z} && (11.19)\\
&= \int_{\mathbf{z}} q(\mathbf{z}) \log p(\mathbf{x}|\mathbf{z};\theta) d\mathbf{z} - \int_{\mathbf{z}} q(\mathbf{z}) \log \frac{q(\mathbf{z})}{p(\mathbf{z};\theta)} d\mathbf{z} && (11.20)\\
&= \mathbb{E}_{\mathbf{z}\sim q(\mathbf{z})}[\log p(\mathbf{x}|\mathbf{z};\theta)] - D_{\mathrm{KL}}(q(\mathbf{z})||p(\mathbf{z};\theta)) && (11.21)\\
&\triangleq ELBO(q(\mathbf{z}),\theta;\mathbf{x}) && (11.22)
\end{aligned}
$$

其中，$ELBO(q(\mathbf{z}),\theta;\mathbf{x})$ 为对数边际似然函数 $l(\theta;\mathbf{x})$ 的下界，称为证据下界（evidence lower bound，ELBO）。

参见习题（11-3），第211页。

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 $l(\theta;\mathbf{x})$ 和其下界 $ELBO(q(\mathbf{z}),\theta;\mathbf{x})$ 相等，

$$l(\theta;\mathbf{x}) = ELBO(q,\theta;\mathbf{x}).$$

这样，最大化对数边际似然函数 $l(\theta;\mathbf{x})$ 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 $l(\theta;\mathbf{x}) = ELBO(q,\theta;\mathbf{x})$；（2）再寻找参数 θ 最大化 $ELBO(q,\theta;\mathbf{x})$。这就是期望最大化（expectation-maximum，EM）算法。

*EM 算法*是常用的含隐变量的参数估计方法，通过迭代的方法来最大化边际似然。其具体的过程可以分为两步：

- E 步（expectation step）：固定参数 θ，找到一个分布 $q(\mathbf{z}) = p(\mathbf{z}|\mathbf{x};\theta)$，使得 $l(\theta;\mathbf{x}) = ELBO(q,\theta;\mathbf{x})$ 最大；
- M 步（maximization step）：固定 $q(\mathbf{z})$，找到一组参数使得证据下界最大，即 $\theta^* = \arg\max_\theta ELBO(q(\mathbf{z}),\theta;\mathbf{x})$。

这两个步骤（E 步和 M 步）不断重复，直到收敛到某个局部最优解。

假设在第 t 步时参数为 θ_t，在 E 步时找到一个变分分布 $q_t(\mathbf{z})$ 使得 $l(\theta_t;\mathbf{x}) = ELBO(q_t(\mathbf{z}),\theta_t;\mathbf{x})$。在 M 步时固定 $q_t(\mathbf{z})$ 找到一组参数，$ELBO(q_t(\mathbf{z}),\theta_{t+1};\mathbf{x}) \geq ELBO(q_t(\mathbf{z}),\theta_t;\mathbf{x})$。因此有

$$l(\theta_{t+1};\mathbf{x}) \geq ELBO(q_t(\mathbf{z}),\theta_{t+1};\mathbf{x}) \geq ELBO(q_t(\mathbf{z}),\theta_t;\mathbf{x}) = l(\theta_t;\mathbf{x}), \quad (11.23)$$

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

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。

变分自编码器参见第12.1节，第213页。

高斯混合模型

本节中，我们来看一个 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

$$\mathcal{N}(x|\mu_k,\sigma_k) = \frac{1}{\sqrt{2\pi}\sigma_k}\exp(-\frac{(x-\mu_k)^2}{2\sigma_k^2}), \quad (11.24)$$

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

$$p(x) = \sum_{k=1}^{K}\pi_k\mathcal{N}(x|\mu_k,\sigma_k), \quad (11.25)$$

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。

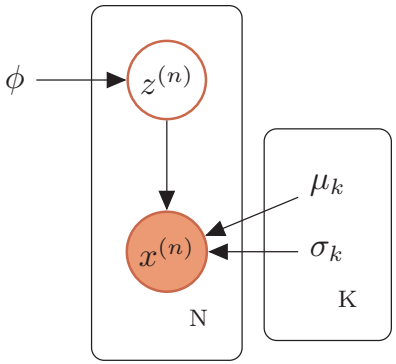


图 11.6: 高斯混合模型。

从高斯混合模型中采样的过程可以分为两步：

1. 首先以 $\pi_1, \pi_2, \cdots, \pi_K$ 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。

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

$$p(z^{(n)} = k) = \pi_k. \tag{11.26}$$

高斯混合模型的图结构如图11.6所示。

对每个样本 $x^{(n)}$，其对数边际分布为

$$\log p(x^{(n)}) = \log \sum_{z^{(n)}} p(z^{(n)}) p(x^{(n)}|z^{(n)}) \tag{11.27}$$

$$= \log \sum_{k=1}^{K} \pi_k \mathcal{N}(x^{(n)}|\mu_k, \sigma_k). \tag{11.28}$$

根据 EM 算法，先计算后验分布 $p(z^{(n)}|x^{(n)})$

$$\gamma_k(z^{(n)}) \triangleq p(z^{(n)} = k|x^{(n)}) \tag{11.29}$$

$$= \frac{p(z^{(n)})p(x^{(n)}|z^{(n)})}{p(x^{(n)})} \tag{11.30}$$

$$= \frac{\pi_k \mathcal{N}(x^{(n)}|\mu_k, \sigma_k)}{\sum_{k=1}^{K} \pi_k \mathcal{N}(x^{(n)}|\mu_k, \sigma_k)} \tag{11.31}$$

所有训练样本 $x^{(1)}, x^{(2)}, \cdots, x^{(N)}$ 的证据下界为

$$\begin{aligned} & ELBO(\pi, \mu, \sigma) \\ &= \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma_k(z^{(n)}) \left(\log \mathcal{N}(x^{(n)}|\mu_k, \sigma_k) - \log \frac{\pi_k}{\gamma_k(z^{(n)})} \right) \qquad (11.32) \\ &= \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma_k(z^{(n)}) \left(\frac{-(x-\mu_k)^2}{2\sigma_k^2} - \log \sigma_k - \log \pi_k \right) + C, \qquad (11.33) \end{aligned}$$

其中 C 为和参数无关的常数。

参数估计问题可以转为优化问题

$$\begin{aligned} & \max_{\pi, \mu, \sigma} ELBO(\pi, \mu, \sigma), \\ & s.t. \sum_{k=1}^{K} \pi_k = 1. \end{aligned} \tag{11.34}$$

分别求 $ELBO(\pi, \mu, \sigma)$ 关于 π_k, μ_k, σ_k 的偏导数，并令其等于 0。可得，

$$\pi_k = \frac{N_k}{N}, \tag{11.35}$$

$$\mu_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma_k(z^{(n)}) x^{(n)}, \tag{11.36}$$

$$\sigma_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma_k(z^{(n)}) (x^{(n)} - \mu_k)^2, \tag{11.37}$$

其中

$$N_k = \sum_{n=1}^{N} \gamma_k(z^{(n)}). \tag{11.38}$$

参见习题（11-4），第211页。

高斯混合模型的训练过程如算法11.1所示。

算法 11.1: 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 (11.31) 计算 $\gamma_k(z^{(n)})$，$1 \leq k \leq K$，，$1 \leq n \leq N$;

// M 步

固定 $\gamma_k(z^{(n)})$，根据公式 (11.35)，(11.36) 和 (11.37)，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，$1 \leq k \leq K$

11.2 马尔可夫随机场

马尔可夫随机场（Markov random fields），也叫无向图模型，或马尔可夫网络（Markov network)，是一类用无向图来表示一组具有马尔可夫性质的随机变量 X 的联合概率分布模型。

和贝叶斯网络类似，马尔可夫随机场也图结构来随机变量之间的依赖关系。但是，贝叶斯网络是有向非循环图，而马尔可夫随机场是一个无向图，并且可以存在循环。这样，马尔可夫随机场可以表示贝叶斯网络无法表示的一些依赖关系，如循环依赖；但它不能表示贝叶斯网络能够表示的某些关系，如推导关系。

给定个有 K 个节点的无向图 $G(\mathcal{V}, \mathcal{E})$，其中 $\mathcal{V} = \{v_1, v_2, \cdots, v_K\}$ 表示节点集合。每个节点 v_k 表示一个随机变量 X_k。如果 (G, 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其它所有变量，那么 (G, X) 就构成了一个马尔可夫随机场。

局部马尔可夫性表示为 $X_v \perp\!\!\!\perp X_{\mathcal{V} \backslash \mathrm{N}[v]} \mid X_{\mathrm{N}(v)}$，即

$$P(X_v = x_v | X_u, u \neq v) = P(X_v = x_v | X_u, v \in N_u), \tag{11.39}$$

吉布斯（Gibbs）分布 如果无向图模型能够表示成一系列在G的最大团（们）上的非负函数乘积的形式，这个无向图模型的概率分布P(X)就称为*Gibbs*分布。

马尔可夫网络的联合分布可以表示为：

$$P(\mathbf{X}=\mathbf{x})=\frac{1}{Z}\prod_{c\in\mathcal{C}}\phi_c(\mathbf{x}_c), \tag{11.40}$$

其中，$\phi_c(\mathbf{x}_c)$是定义在团c上的势能函数（Potential Function），Z是配分函数（Partition Function），

$$Z=\sum_{\mathbf{x}\in\mathcal{X}}\prod_{c\in\mathcal{C}}\phi_c(\mathbf{x}_c). \tag{11.41}$$

乘积包括了图中的所有团。注意马尔可夫性质在团内的节点存在，在团之间是不存在依赖关系的。这里，实际上，马尔可夫网联络经常表示为对数线性模型。

由于势能函数必须为正的，因此我们一般定义为

$$\phi_c(\mathbf{x}_c)=\exp(-E(X_c)), \tag{11.42}$$

其中，$E(X_c)$为能量函数（energy function），这个表示叫做玻尔兹曼分布。

马尔可夫网络的联合分布可以表示为：

$$P(\mathbf{X}=\mathbf{x})=\frac{1}{Z}\prod_{c\in\mathcal{C}}\exp(-E(X_c)) \tag{11.43}$$

$$=\frac{1}{Z}\exp(\sum_{c\in\mathcal{C}}-E(X_c)) \tag{11.44}$$

Hammersley Clifford理论认为，马尔可夫随机场和Gibbs分布是一致的。即吉布斯分布一定满足由node separation导致的条件独立性，并且马尔可夫随机场的概率分布一定可以表示成最大团们上的非负函数乘积形式。

德维希·玻尔兹曼（Ludwig Boltzmann，1844 - 1906），奥地利物理学家、哲学家。主要贡献为分子动力学。

11.3 推断

给定一组变量，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$\mathbf{q}=\{q_1,q_2,\cdots,q_n\}$的后验概率分布$p(\mathbf{q}|\mathbf{e})$。

在图模型中，我们可以利用图结构来

数学小知识 | 玻尔兹曼分布

在统计力学中，玻尔兹曼分布（Boltzmann distribution）是描述粒子处于特定状态下的概率，是关于状态能量与系统温度的函数。

$$p_\alpha = \frac{1}{Z}\exp\left(\frac{-E_\alpha}{kT}\right), \tag{11.45}$$

其中，p_α 粒子处于为状态 α 的概率，E_α 为状态 α 的能量，k 为玻尔兹曼常量，T 为系统温度，$\exp(\frac{-E_\alpha}{kT})$ 称为玻尔兹曼因子（Boltzmann factor），是没有归一化的概率；Z 为归一化因子，是对系统所有状态进行总和。

在统计力学中，Z 一般称为配分函数（partition function），其定义为

$$Z = \sum_\alpha \exp\left(\frac{-E_\alpha}{kT}\right). \tag{11.46}$$

玻尔兹曼分布取自奥地利物理学家路德维希·玻尔兹曼（Ludwig Boltzmann），他在 1868 年研究热平衡气体的统计力学时首次提出了这一分布。

玻尔兹曼分布的一个性质是两个状态概率的比率仅仅依赖于两个状态能量的差值。

$$\frac{p_\alpha}{p_\beta} = \exp\left(\frac{E_\beta - E_\alpha}{kT}\right). \tag{11.47}$$

11.3.1 近似推断

11.4 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。

11.5 总结和深入阅读

习题 11-1 证明Jensen不等式。

习题 11-2 证明公式(11.5)。

习题 11-3 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 11-4 求解公式(11.35)，(11.36)和(11.37)。

参考文献

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