概率图模型

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。一种有效减少参数量的方法是独立性假设。

- 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。
- 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。
- 学习问题：图模型的学习包括图结构的学习和参数的学习，即参数估计问题。

模型表示

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。

有向图模型（Directed Graphical model）

定义 11.1－贝叶斯网络： 对于一个随机向量$\mathbf{X}=[X_1,X_2,\cdots,X_K]^{\mathrm{T}}$和一个有$K$个节点的有向非循环图$G$，$G$中的每个节点都对应一个随机变量，可以是可观测的变量，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分布（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。

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。间接因果关系、间接果因关系、共因关系、共果关系。

常见的有向图模型

朴素贝叶斯分类器、隐马尔可夫模型、深度信念网络。

sigmoid 信念网络（sigmoid belief network， SBN）

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，网络中的变量取值为$\{0,1\}$。对于变量x_k的父节点集合π_k，其条件概率分布表示为

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

Sigmoid 信念网络与 Logistic 回归模型都采用 Logistic 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 Logistic 回归模型类似。这两个模型区别在于 Logistic 回归模型中的 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|x)，是一种判别模型；而 sigmoid 信念网络建模 p(x,y)，是一种生成模型。

朴素贝叶斯分类器（Naive Bayes Classifier， NB）

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 x 和类别 y，类别的后验概率为

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

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

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，条件概率分布 p(y|x)可以分解为

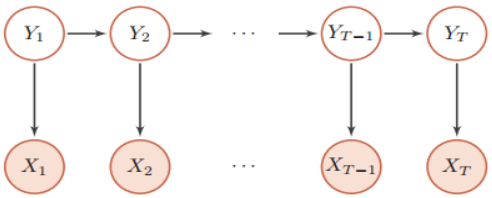
$$p(y|\mathbf{x},\theta) \propto p(y|\theta_c)\prod_{i=1}^{d} p(x_i|y,\theta_{i,y}), \tag{11.14}$$

若输入为连续值，可以用高斯分布建模；如果是离散值，可以用多项分布建模。模型简单，可以有效防止过拟合。

隐马尔可夫模型（Hidden Markov Model， HMM）

含有隐变量的马尔可夫过程。隐马尔可夫模型的联合概率可以分解为

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



无向图模型

用无向图来描述一组具有局部马尔可夫性质的随机向量 X 的联合概率分布的模型。

定义 11.2－马尔可夫随机场： 对于一个随机向量 $\mathbf{X} = [X_1,\cdots,X_K]^{\mathrm{T}}$ 和一个有 K 个节点的无向图 $G(\mathcal{V},\mathcal{E})$（可以存在循环），图 G 中的节点 k 表示随机变量 X_k，$1 \le k \le K$。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，

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

其中 $N(k)$ 为变量 X_k 的邻居集合，$\setminus k$ 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。

无向图模型的概率分解

无向图模型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。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。

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。

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

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$\phi_c(\mathbf{x}_c) = \exp(-E_c(\mathbf{x}_c))$正的，因此我们一般定义为

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

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

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来表示其联合概率。

常见的无向图模型

对数线性模型（也叫最大熵模型）、条件随机场、玻尔兹曼机、受限玻尔兹曼机。

对数线性模型（Log-Linear Model）

势能函数一般定义为

联合概率p(x)的对数形为 $\phi_c(\mathbf{x}_c|\theta_c) = \exp\left(\theta_c^{\mathrm{T}} f_c(\mathbf{x}_c)\right)$

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

如果用对数线性模型来建模条件概率$p(y|\mathbf{x})$，

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

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。

条件随机场（Conditional Random Field， CRF）

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

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

有向图和无向图之间的转换

推断（inference）

指在观测到部分变量e = $\{e_1, e_2, \cdots, e_m\}$时，计算其它变量的某个子集q = $\{q_1, q_2, \cdots, q_m\}$的后验概率p(q|e)。

精确推断和近似推断。

变量消除法（variable elimination algorithm）

每次消除一个变量，按照不同的顺序来减少计算边际分布的计算复杂度。

信念传播（Belief Propagation， BP）

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

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

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

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

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

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

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

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

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

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，

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

树结构上的信念传播算法

如果一个有向图满足任意两个变量只有一条路径（忽略方向），且只有一个没有父节点的节点，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一条路径，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收消息的乘积（如果是无向图还需要归一化），就得到了所有变量的边际概率。

近似推断（Approximate Inference）

- **环路信念传播**：当图模型中存在环路时，使用和积算法时，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，LBP）是在具有环路的图上依然使用和积算法，即使得到不精确解，在某些任务上也可以近似精确解。
- **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。变分法（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），然后再根据变分分布来进行推断。
- **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。

蒙特卡罗方法（Monte Carlo Method）

（大数定理为理论依据）通过随机采样的方法来近似估计一些计算问题的数值解。随机采样指从给定概率密度函数 p(x)中抽取出符合其概率分布的样本。

拒绝采样（Rejection Sampling）

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)$\geqslant\hat{p}(x),\forall$x。

对于每次抽取的样本$\hat{x}$，计算接受概率（acceptance probability），并以概率$\alpha\hat{x}$来接受样本$\hat{x}$。

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

算法 11.1: 拒绝采样

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

1 **repeat**
2　根据$q(x)$随机生成一个样本$\hat{x}$;
3　计算接受概率$\alpha(\hat{x})$;
4　从$(0,1)$的均匀分布中随机生成一个值z;
5　**if** $z\leq\alpha(\hat{x})$ **then**　　/* 以$\alpha(\hat{x})$的概率接受$\hat{\mathbf{x}}$ */
6　　$\mathcal{V}=\mathcal{V}\cup\{\hat{x}\}$;
7　**end**
8 **until** 直到获得 N 个样本$(|\mathcal{V}|=N)$;

输出: 样本集合$\mathcal{V}$

判断一个拒绝采样方法的好坏就是看其采样效率，即总体的接受率。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，导致很难应用到实际问题中。

重要性采样

函数 f(x)在分布 p(x)下的期望可以写为:

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

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)的期望变为在分布 q(x)下 f(x)w(x)的期望，从而可以近似为

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

$$= \int_x f(x)w(x)q(x)dx$$

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

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

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

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

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

下计算函数 f(x)的期望。

马尔可夫链蒙特卡罗方法（Markov Chain Monte Carlo, MCMC）

在高维空间中，拒绝采样和重要性采样的效率随空间维数的增加而指数降低。

很容易地对高维变量进行采样,将采样过程看作是一个马尔可夫链。MCMC 方法的关键是如何构造出平稳分布为 p(x)的马尔可夫链，并且该马尔可夫链的状态转移分布$q(x|x')$一般为比较容易采样的分布。

一是马尔可夫链需要经过一段时间的随机游走才能达到平稳状态，这段时间称为预烧期（Burn-in Period）。预烧期内的采样点并不服从分布 p(x)，需要丢弃；二是基于马尔可夫链抽取的相邻样本是高度相关的。

Metropolis-Hastings 算法

马尔可夫链的状态转移分布$q(x|x')$，其平稳分布往往不是$p(x)$。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(x)。

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

$$A(\hat{\mathbf{x}}, \mathbf{x}_t) = \min\left(1, \frac{p(\hat{\mathbf{x}})q(\mathbf{x}_t|\hat{\mathbf{x}})}{p(\mathbf{x}_t)q(\hat{\mathbf{x}}|\mathbf{x}_t)}\right). \tag{11.54}$$

每次$q(x|x_t)$随机生成一个样本$\hat{x}$，并以概率 A($\hat{x}$, x_t)的方式接受，因此修正马尔可夫链的状态转移概率为

$$q'(\hat{x}|x_t) = q(\hat{x}|x_t)A(\hat{x}|x_t)$$

该修正马尔可夫链可以达到平稳状态，且平稳分为 p(x)。

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 ≤ α then                          /* 以 A(x̂, x_t) 的概率接受 x̂ */
        x_{t+1} = x̂;
    else                                   /* 拒绝接受 x̂ */
        x_{t+1} = 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
        V = V ∪ {x_t};
    end
until 直到获得 N 个样本 (|V| = N);
输出: 样本集合 V
```

Metropolis 算法

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

$$A(\hat{\mathbf{x}}, \mathbf{x}_t) = \min\left(1, \frac{p(\hat{\mathbf{x}})}{p(\mathbf{x}_t)}\right). \tag{11.62}$$

吉布斯采样（Gibbs Sampling）

一种有效地对高维空间中的分布进行采样，使用全条件概率（Full Conditional Probability）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 X=$[X_1, X_2, \cdots, X_M]^T$，其第 i 个变量$X_i$的全条件概率为

$$\begin{aligned} p(x_i|\mathbf{x}_{-i}) &\triangleq P(X_i = x_i|X_{-i} = \mathbf{x}_{-i}) \\ &= p(x_i|x_1, x_2, \cdots, x_{i-1}, x_{i+1}, \cdots, x_M) \end{aligned}$$

可以按任意的顺序根据全条件分布依次对每个变量进行采样。

吉布斯采样的每单步采样也构成一个马尔可夫链。假设每个单步（采样维度为第 i 维）的状态转移概率 $q(\mathbf{x}|\mathbf{x}')$ 为

$$q(\mathbf{x}|\mathbf{x}') = \begin{cases} \frac{p(\mathbf{x})}{p(\mathbf{x}'_{-i})} & \text{if } \mathbf{x}_{-i} = \mathbf{x}'_{-i} \\ 0 & \text{otherwise,} \end{cases} \tag{11.71}$$

其中边际分布 $p(\mathbf{x}'_{-i}) = \sum_{x'_i} p(\mathbf{x}')$，等式 $\mathbf{x}_{-i} = \mathbf{x}'_{-i}$ 表示 $x_j = x'_j, \forall j \neq i$，因此有 $p(\mathbf{x}'_{-i}) = p(\mathbf{x}_{-i})$，并可以得到

$$p(\mathbf{x}')q(\mathbf{x}|\mathbf{x}') = p(\mathbf{x}')\frac{p(\mathbf{x})}{p(\mathbf{x}'_{-i})} = p(\mathbf{x})\frac{p(\mathbf{x}')}{p(\mathbf{x}_{-i})} = p(\mathbf{x})q(\mathbf{x}'|\mathbf{x}). \tag{11.72}$$

根据细致平稳条件，公式 (11.71) 中定义的状态转移概率 $q(\mathbf{x}|\mathbf{x}')$ 的马尔可夫链的平稳分布为 $p(\mathbf{x})$。随着迭代次数 t 的增加，样本 $\mathbf{x}^{(t)} = [x_1^{(t)}, x_2^{(t)} \cdots, x_M^{(t)}]^{\mathrm{T}}$ 将收敛于概率分布 $p(\mathbf{x})$。

学习

一是网络结构学习，即寻找最优的网络结构；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。

不含隐变量的参数估计

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 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。

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

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

$$=\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K}\log p(x_k^{(i)}|x_{\pi_k}^{(i)},\theta_k), \tag{11.74}$$

因为所有变量都是可观测的，最大化对数似然$L(D|\theta)$，只需要分别地最大化每个变量的条件似然来估计其参数。

$$\theta_k=\arg\max\sum_{i=1}^{N}\log p(x_k^{(i)}|x_{\pi_k}^{(i)},\theta_k). \tag{11.75}$$

无向图模型 在无向图模型中，所有变量 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。

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。

无向图的参数估计通常采用近似的方法。一是利用采样来近似计算这个期望；二是坐标上升法，即固定其它参数，来优化一个势能函数的参数。

含隐变量的参数估计

EM 算法

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

$$p(x|\theta)=\sum_z p(x,z|\theta)$$

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

$$\mathcal{L}(\mathcal{D}|\theta)=\frac{1}{N}\sum_{i=1}^{N}\log p(\mathbf{x}^{(i)},\theta) \tag{11.85}$$

$$=\frac{1}{N}\sum_{i=1}^{N}\log\sum_{\mathbf{z}}p(\mathbf{x}^{(i)},\mathbf{z}|\theta). \tag{11.86}$$

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。样本$\mathbf{x}$的对数边际似然函数为

(1) 先找到近似分布$q(z)$使得$log\ p(x|\theta)\ =\ ELBO(q,x|\theta)$；（2）再寻找参数$\theta$最大化

$$\log p(\mathbf{x}|\theta) = \log \sum_{\mathbf{z}} q(\mathbf{z}) \frac{p(\mathbf{x},\mathbf{z}|\theta)}{q(\mathbf{z})} \tag{11.87}$$

$$\geq \sum_{\mathbf{z}} q(\mathbf{z}) \log \frac{p(\mathbf{x},\mathbf{z}|\theta)}{q(\mathbf{z})} \tag{11.88}$$

利用 Jensen 不等式。

$$\triangleq ELBO(q,\mathbf{x}|\theta), \tag{11.89}$$

$ELBO(q,x|\theta)$。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骤：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 步更新时，E 步和 M 步分布为：

E 步（Expectation step）：固定参数θ_t，找到一个分布使得$ELBO(q,x|\theta_t)$最大，即等于$log\ p(x|\theta_t)$。

$$q_{t+1}(\mathbf{z}) = \arg\max_{q} ELBO(q,\mathbf{x}|\theta_t). \tag{11.90}$$

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

M 步（Maximization step）：固定$q_{t+1}(z)$，找到一组参数使得证据下界最大，即

$$\theta_{t+1} = \arg\max_{\theta} ELBO(q_{t+1},\mathbf{x}|\theta). \tag{11.91}$$

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$\theta_{t+1}$，使得$ELBO(q_{t+1},\mathbf{x}|\theta_{t+1}) \geq ELBO(q_{t+1},\mathbf{x}|\theta_t)$。因此有

$$\log p(\mathbf{x}|\theta_{t+1}) \geq ELBO(q_{t+1},\mathbf{x}|\theta_{t+1}) \geq ELBO(q_{t+1},\mathbf{x}|\theta_t) = \log p(\mathbf{x}|\theta_t), \tag{11.92}$$

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

信息论的视角 对数边际似然可以通过下面方式进行分解：

$$\log p(\mathbf{x}|\theta) = \sum_{\mathbf{z}} q(\mathbf{z}) \log p(\mathbf{x}|\theta) \tag{11.93}$$

$$= \sum_{\mathbf{z}} q(\mathbf{z}) \Big(\log p(\mathbf{x},\mathbf{z}|\theta) - \log p(\mathbf{z}|\mathbf{x},\theta) \Big) \tag{11.94}$$

$$= \sum_{\mathbf{z}} q(\mathbf{z}) \log \frac{p(\mathbf{x},\mathbf{z}|\theta)}{q(\mathbf{z})} - \sum_{\mathbf{z}} q(\mathbf{z}) \log \frac{p(\mathbf{z}|\mathbf{x},\theta))}{q(\mathbf{z})} \tag{11.95}$$

$$= ELBO(q,\mathbf{x}|\theta) + D_{\mathrm{KL}}(q(\mathbf{z})\|p(\mathbf{z}|\mathbf{x},\theta)), \tag{11.96}$$

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。

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。

高斯混合模型（Gaussian Mixture Model， GMM）

高斯混合模型的概率密度函数为

$$p(x) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x|\mu_k,\sigma_k), \tag{11.98}$$

其中π_k表示第k个高斯分布的权重系数并满足$\pi_k \geq 0, \sum_{k=1}^{K} \pi_k = 1$，即样本$x$

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

$$p(z^{(n)} = k) = \pi_k. \tag{11.99}$$

对每个样本 $x^{(n)}$，其对数边际分布为

$$\log p(x^{(n)}) = \log \sum_{z^{(n)}} p(z^{(n)}) p(x^{(n)}|z^{(n)}) \tag{11.100}$$

$$= \log \sum_{k=1}^{K} \pi_k \mathcal{N}(x^{(n)}|\mu_k, \sigma_k). \tag{11.101}$$

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 $p(z^{(n)}|x^{(n)})$

$$\gamma_{nk} \triangleq p(z^{(n)} = k|x^{(n)}) \tag{11.102}$$

$$= \frac{p(z^{(n)}) p(x^{(n)}|z^{(n)})}{p(x^{(n)})} \tag{11.103}$$

$$= \frac{\pi_k \mathcal{N}(x^{(n)}|\mu_k, \sigma_k)}{\sum_{k=1}^{K} \pi_k \mathcal{N}(x^{(n)}|\mu_k, \sigma_k)}, \tag{11.104}$$

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

$$ELBO(\gamma, \mathcal{D}|\pi, \mu, \sigma) = \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma_{nk} \log \frac{p(x^{(n)}, z^{(n)} = k)}{\gamma_{nk}} \tag{11.105}$$

$$= \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma_{nk} \left(\log \mathcal{N}(x^{(n)}|\mu_k, \sigma_k) + \log \frac{\pi_k}{\gamma_{nk}} \right) \tag{11.106}$$

$$= \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma_{nk} \left(\frac{-(x - \mu_k)^2}{2\sigma_k^2} - \log \sigma_k + \log \pi_k \right) + C, \tag{11.107}$$

其中 C 为和参数无关的常数。

$$\max_{\pi, \mu, \sigma} ELBO(\gamma, \mathcal{D}|\pi, \mu, \sigma),$$

$$s.t. \sum_{k=1}^{K} \pi_k = 1. \tag{11.108}$$

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 π_k, μ_k, σ_k 的偏导数，并令其等于 0。可得，

$$N_k = \sum_{n=1}^{N} \gamma_{nk}. \tag{11.112}$$

$$\pi_k = \frac{N_k}{N}, \tag{11.109}$$

$$\mu_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma_{nk} x^{(n)}, \tag{11.110}$$

$$\sigma_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma_{nk} (x^{(n)} - \mu_k)^2, \tag{11.111}$$

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(11.104)计算γ_{nk}，$1 \leq k \leq K$,，$1 \leq n \leq N$;

// M步

固定γ_{nk}，根据公式(11.109)，(11.110)和(11.111)，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，$1 \leq k \leq K$