模式识别与机器学习

Pattern Recognition & Machine Learning

第五讲 聚类

- 本讲学习目标

  ✓理解聚类的两大类方法
  ✓掌握*K*-均值聚类方法，理解模糊*K*-均值聚类的原理
  ✓掌握谱聚类方法
  ✓掌握高斯混合模型聚类方法，了解无限高斯混合模型

目录

- *K*-均值聚类
  - 算法介绍
  - 模糊*K*-均值聚类
- 谱聚类
- 高斯混合模型聚类
  - 模型表示
  - 模型推理与参数估计
  - 无限高斯混合模型

聚类任务:
① 在相同簇中的数据尽可能相似
② 在不同簇中的数据尽可能不同

聚类方法:
① 基于数据间相似度的方法
② 基于密度估计的方法

***K*-均值（*K*-means）：**
将*K*个聚类簇的中心作为簇的代表，希望所有数据点与其所在聚类中心的距离总和最小

给定数据集$X=\{\mathbf{x}_1,\mathbf{x}_2,\cdots,\mathbf{x}_N\},\mathbf{x}_n\in R^D$，假设将数据集聚类为$K$个簇，数据点$\mathbf{x}_n$的类别记为$z_n,n\in\{1,2,\cdots,N\},\boldsymbol{\mu}=\{\boldsymbol{\mu}_1,\boldsymbol{\mu}_2,\cdots,\boldsymbol{\mu}_K\}$表示$K$个簇的中心。$K$-均值聚类算法的优化目标是最小化簇内误差平方和

$$\arg\min_{\boldsymbol{\mu},\mathbf{z}}\sum_{k=1}^{K}\sum_{n=1}^{N}I(z_n=k)\|\mathbf{x}_n-\boldsymbol{\mu}_k\|^2,$$

假设初始误差为

$$J=\sum_{k=1}^{K}\sum_{n=1}^{N}I(z_n=k)\|\mathbf{x}_n-\boldsymbol{\mu}_k\|^2.$$

数据$\mathbf{x}$从簇i移入簇j中时，各个簇的均值$\boldsymbol{\mu}_i$和$\boldsymbol{\mu}_i$分别变为$\tilde{\mathbf{u}}_i$和$\tilde{\mathbf{u}}_j$

$$\tilde{\boldsymbol{\mu}}_i=\frac{N_i\boldsymbol{\mu}_i-\mathbf{x}}{N_i-1}=\frac{N_i\boldsymbol{\mu}_i-\boldsymbol{\mu}_i+\boldsymbol{\mu}_i-\mathbf{x}}{N_i-1}=\boldsymbol{\mu}_i+\frac{\boldsymbol{\mu}_i-\mathbf{x}}{N_i-1},$$

$$\tilde{\boldsymbol{\mu}}_j=\frac{N_j\boldsymbol{\mu}_j+\mathbf{x}}{N_j+1}=\frac{N_j\boldsymbol{\mu}_j+\boldsymbol{\mu}_j-\boldsymbol{\mu}_j+\mathbf{x}}{N_j-1}=\boldsymbol{\mu}_j+\frac{\mathbf{x}-\boldsymbol{\mu}_j}{N_j+1},$$

其中，N_i和N_j分别表示当前簇i和簇j中的数据的数目，即

$$N_i=\sum_{n=1}^{N}I(z_n=i),\ N_j=\sum_{n=1}^{N}I(z_n=j)$$

假设移动之后误差J_i和J_j变为$\tilde{J}_i$和$\tilde{J}_j$

$$\widetilde{J}_i = J_i - \frac{N_i}{N_i - 1} \| \mathbf{x} - \boldsymbol{\mu}_i \|^2,$$

$$\widetilde{J}_j = J_j + \frac{N_j}{N_j + 1} \| \mathbf{x} - \boldsymbol{\mu}_j \|^2 .$$

结合$\tilde{J} = \tilde{J}_i + \tilde{J}_j + \alpha$可以得到，如果想要更新簇之后$\tilde{J} \le J$，需满足

$$\frac{N_i}{N_i - 1} \| \mathbf{x} - \boldsymbol{\mu}_i \|^2 \ge \frac{N_j}{N_j + 1} \| \mathbf{x} - \boldsymbol{\mu}_j \|^2$$

算法 10-1　K-均值聚类

输入：数据 $\{\mathbf{x}_n\}_{n=1}^N$，聚类数目 K

1:　将数据随即划分为 K 个簇，即初始化 $\mathbf{z}$，并计算初始聚类中心 $\boldsymbol{\mu}_1,\boldsymbol{\mu}_2,...,\boldsymbol{\mu}_K$ 以及总簇内误差平方和 J。

2:　REPEAT

3:　从某个簇内数据数目大于 1 的簇 i 中任选一个数据点 $\mathbf{x}_n$，即 $z_n = i$ 且 $N_i > 1$;

4:　对于 $j = 1,2,...,K$，计算

$$\rho_j = \begin{cases} \dfrac{N_i}{N_i - 1} \| \mathbf{x}_n - \boldsymbol{\mu}_i \|^2 & j = i, \\ \dfrac{N_j}{N_j + 1} \| \mathbf{x}_n - \boldsymbol{\mu}_j \|^2 & j \neq i. \end{cases}$$

对于所有的 j，满足 $\rho_k < \rho_j$，则把数据 $\mathbf{x}_n$ 从簇 i 移入簇 j 中;

5:　重新计算聚类中心 $\boldsymbol{\mu}_1,\boldsymbol{\mu}_2,...,\boldsymbol{\mu}_K$ 以及总簇内误差平方和 J

6:　UNTIL 总簇内误差平方和 J 保持不变

输出：聚类指示变量 $\mathbf{z}$，聚类中心 $\boldsymbol{\mu}_1,\boldsymbol{\mu}_2,...,\boldsymbol{\mu}_K$，总簇内误差平方和 J

算法 10-2 批处理的 K-均值聚类

输入：数据 $\{\mathbf{x}_n\}_{n=1}^N$，聚类数目 K

1: 设置初始聚类中心 $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, ..., \boldsymbol{\mu}_K$。例如，随机从数据集中挑选 K 个数据点作为初始聚类中心；

2: REPEAT

3: 把每个数据点 $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N$ 划分到距离（可以是欧氏距离或其他度量方法）最近的中心所在的簇：

$$z_n \leftarrow \arg\min_k \| \mathbf{x}_n - \boldsymbol{\mu}_k \|^2 . \tag{10.7}$$

4: 根据聚类指示变量 $\{z_n\}$，重新计算每个聚类中心 $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, ..., \boldsymbol{\mu}_K$：

$$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{n=1}^{N} I(z_n = k) \mathbf{x}_n, \tag{10.8}$$

其中，$N_k = \sum_{n=1}^{N} I(z_n = k).$

5: UNTIL 聚类指示变量 $\{z_n\}$ 保持不变

输出：指示变量 $\{z_n\}$，聚类中心 $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, ..., \boldsymbol{\mu}_K$

模糊K-均值聚类的优化目标是最小化簇内误差平方和，即

$$\arg\min_{\{\boldsymbol{\mu}_k, d_{nk}\}} J = \sum_{k=1}^{K}\sum_{n=1}^{N}(d_{nk})^m \| \mathbf{x}_n - \boldsymbol{\mu}_k \|^2,$$

$$s.t. \sum_{k=1}^{K} d_{nk} = 1,\ n = 1, 2, ..., N,$$

其中$m > 1$，是控制聚类结果模糊程度的参数。对上述优化问题使用拉格朗日乘子法，可得到非约束优化问题

$$\arg\min_{\{\boldsymbol{\mu}_k, d_{nk}\}} J = \sum_{k=1}^{K}\sum_{n=1}^{N}(d_{nk})^m \| \mathbf{x}_n - \boldsymbol{\mu}_k \|^2 + \sum_{n=1}^{N}\alpha_n(\sum_{k=1}^{K} d_{nk} - 1),$$

其中$\{\alpha_n\}$是拉格朗日乘子。对上式分别关于$\boldsymbol{\mu}_k, d_{nk}$求偏导数并设置为零，可以得到$\boldsymbol{\mu}_k, d_{nk}$的计算表达式如下：

$$\boldsymbol{\mu}_k = \frac{\sum_{n=1}^{N}(d_{nk})^m \mathbf{x}_n}{\sum_{n=1}^{N}(d_{nk})^m},\ k = 1, 2, ..., K,$$

$$d_{nk} = 1 \Big/ \sum_{j=1}^{K}\left(\frac{\| \mathbf{x}_n - \boldsymbol{\mu}_k \|}{\| \mathbf{x}_n - \boldsymbol{\mu}_j \|}\right)^{2/(m-1)},\ n = 1, 2, ..., N,\ k = 1, 2, ..., K.$$

算法 10-3　模糊 K -均值聚类

输入：数据 $\{\mathbf{x}_n\}_{n=1}^N$，聚类数目 K

1:　设置初始隶属度矩阵 $D=[d_{nk}]$。

2:　REPEAT

3:　根据当前隶属度 $\{d_{nk}\}$ 计算各个聚类中心：

$$\boldsymbol{\mu}_k=\frac{\sum_{n=1}^{N}(d_{nk})^m\mathbf{x}_n}{\sum_{n=1}^{N}(d_{nk})^m}, \tag{10.13}$$

4:　根据聚类中心重新计算每个数据点的隶属度：

$$d_{nk}=1\Big/\sum_{j=1}^{K}\left(\frac{\|\mathbf{x}_n-\boldsymbol{\mu}_k\|}{\|\mathbf{x}_n-\boldsymbol{\mu}_j\|}\right)^{2/(m-1)}. \tag{10.14}$$

5:　UNTIL　隶属度 $\{d_{nk}\}$ 保持不变

6:　若需要对聚类结果去模糊化，可以根据隶属度矩阵得到每个数据点明确所属的簇 z_n 为：

$$z_n=\arg\max_k d_{nk}. \tag{10.15}$$

输出：隶属度 $\{d_{nk}\}$ 或聚类指示变量 $\{z_n\}$，聚类中心 $\boldsymbol{\mu}_1,\boldsymbol{\mu}_2,\ldots,\boldsymbol{\mu}_K$

目录

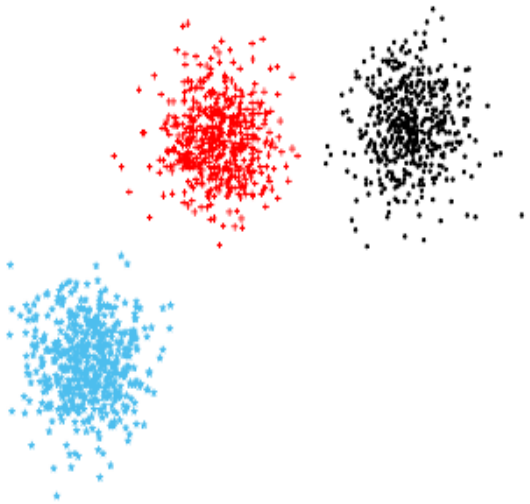
- K-均值聚类
  - 算法介绍
  - 模糊K-均值聚类
- 谱聚类
- 高斯混合模型聚类
  - 模型表示
  - 模型推理与参数估计
  - 无限高斯混合模型

谱聚类：

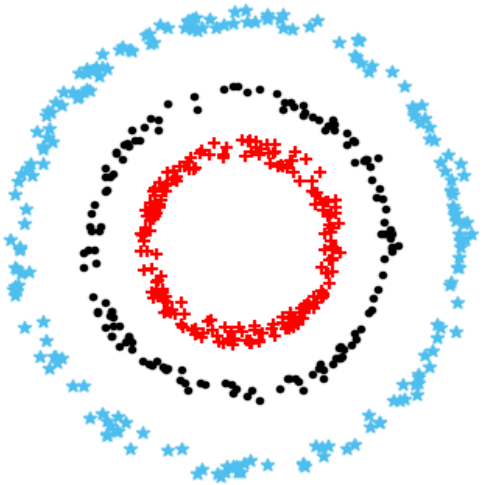
① 预处理：构建代表数据集的无向图并计算相似度矩阵

② 谱表示：构造相应的拉普拉斯矩阵，并且计算拉普拉斯矩阵的特征值和特征向量，其中一个或多个特征向量构成了所有数据点在新的空间中的表示

③ 聚类：使用聚类算法（如K-均值）对新的数据表示进行聚类

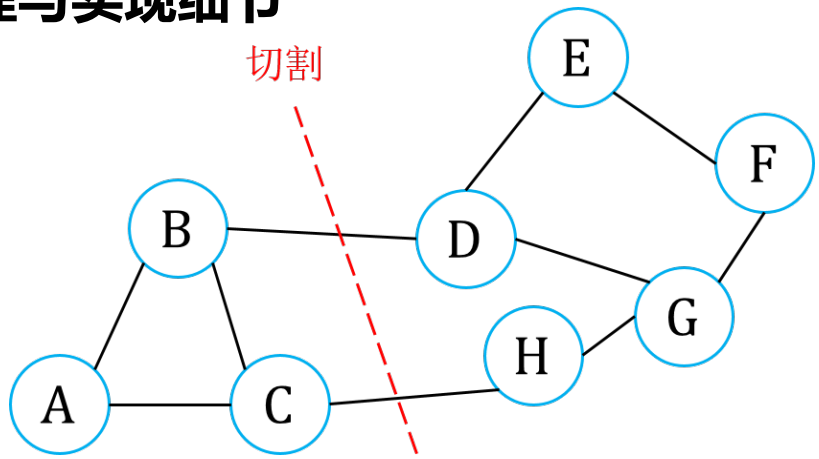


（a）具有紧凑结构的数据



（b）具有连接结构的数据

谱聚类算法的原理与实现细节



首先，利用相似度矩阵构建带权重的无向图$G(V,E)$，计算拉普拉斯矩阵。
常见的高斯相似度：

$$w_{ij} = w_{ji} = \exp(\frac{-\| \mathbf{x}_i - \mathbf{x}_j \|^2}{2\sigma^2}).$$

拉普拉斯矩阵由相似度矩阵W和度矩阵D计算得出。
度矩阵是测量每个顶点与其他顶点关联度的对角矩阵

$$d_{ii} = \sum_{j}^{N} w_{ij}$$

拉普拉斯矩阵表示为$L = D - W$.

拉普拉斯矩阵性质：

① 对于任意的向量$\mathbf{a} \in \boldsymbol{R}^D$，都满足

$$\mathbf{a}^\top L\mathbf{a} = \frac{1}{2}\sum_{i,j=1}^{N} w_{ij}(a_i - a_j)^2$$

② 拉普拉斯矩阵是对称半正定矩阵。

③ 矩阵的最小特征值是**0**，对应的特征向量的元素都为**1**。

④ 拉普拉斯矩阵具有N个非负特征值。

除了传统拉普拉斯矩阵，还有如归一化的拉普拉斯矩阵 $L_N = D^{-1/2}LD^{-1/2}$

其次，定义最优切割的优化目标。
对于任意两个子图A和B，满足$A, B \subset \mathcal{G}$，且$A \cap B = \emptyset$，A和B之间的权重定义为

$$W(A,B) = \sum_{i \in A, j \in B} w_{i,j}$$

无向图$\mathcal{G}(V,E)$在切割之后得到所有子图$A_1, A_2, \cdots, A_K$之间的权重之和为

$$W(A_1, A_2, ..., A_K) = \frac{1}{2}\sum_{k=1}^{K} W(A_k, \mathcal{G} \setminus A_k),$$

其中$\mathcal{G} \setminus A_i$是子图$A_i$的补集。

图$\mathcal{G}$的最优切割的目标是最小化$W(A_1, A_2, \cdots, A_K)$.

对于N个顶点和K个子图，引入指示矩阵$H \in R^{N \times K}$，且$h_{n,k} \neq 0$表示顶点n被划分到子图k中，否则$h_{n,k} = 0$.

最优切割的优化问题通常有两种表达方式：

① 比率切割（ratio cut）

$$\arg\min_{H} \frac{1}{2}\sum_{k=1}^{K}\frac{W(A_i,\mathcal{G}\setminus A_k)}{|A_k|},$$

$$h_{i,k}=\begin{cases}\dfrac{1}{\sqrt{|A_k|}} & \text{if } v_i\in A_k,\\ 0 & \text{otherwise,}\end{cases}$$

其中$|A_i|$表示集合A_i的大小，即子图A_i中的顶点个数。

② 归一化切割（normalized cut）

$$\arg\min_{H} \frac{1}{2}\sum_{k=1}^{K}\frac{W(A_k,\mathcal{G}\setminus A_k)}{vol(A_k)},$$

$$h_{i,k}=\begin{cases}\dfrac{1}{\sqrt{vol(A_k)}} & \text{if } v_i\in A_k,\\ 0 & \text{otherwise,}\end{cases}$$

其中$vol(A_i)$表示集合A_i中所有边的权重的和，即 $vol(A_i)=\sum_{j\in A_i} d_{jj}$

以**比率切割**为例，优化目标的每一项可以写为

$$\frac{W(A_k,\mathcal{G}\setminus A_k)}{|A_k|}=\frac{1}{2}\Big(\sum_{i\in A_k,j\notin A_k} w_{ij}\frac{1}{|A_k|}+\sum_{i\notin A_k,j\in A_k} w_{ij}\frac{1}{|A_k|}\Big)$$

$$=\frac{1}{2}\left(\sum_{i\in A_k,j\notin A_k} w_{ij}\Big(\frac{1}{\sqrt{|A_k|}}-0\Big)^2+\sum_{i\notin A_k,j\in A_k} w_{ij}\Big(0-\frac{1}{\sqrt{|A_k|}}\Big)^2\right)$$

$$=\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}(h_{ik}-h_{jk})^2$$

$$=\mathbf{h}_k^{\top}L\mathbf{h}_k,$$

因此优化问题可以表示为

$$\arg\min_{H}\operatorname{Tr}(H^{\top}LH),$$

$$h_{i,k}=\begin{cases}\dfrac{1}{\sqrt{|A_k|}} & \text{if } v_i\in A_k,\\ 0 & \text{otherwise.}\end{cases}$$

根据$h_{i,k}$的定义可得 $H^{\top}H=\mathbf{I}$，因此约束 $H^{\top}H=\mathbf{I}$，得到优化问题表示为

$$\arg\min_{H} \mathrm{Tr}(H^{\top}LH),$$

$$s.t.\ H^{\top}H=\mathbf{I}.$$

近似解可以通过拉格朗日乘子法获得，对优化问题引入K个拉格朗日乘子$\{\lambda_1,\lambda_2,\cdots,\lambda_K\}$，记为向量$\boldsymbol{\lambda}$，可得如下无约束优化问题

$$\arg\min_{H} \mathrm{Tr}(H^{\top}LH)+\mathrm{Tr}\left[\mathrm{diag}(\boldsymbol{\lambda})(\mathbf{I}-H^{\top}H)\right],$$

其中，$\mathrm{diag}(\boldsymbol{\lambda})$表示对角元素分别为$\{\lambda_1,\lambda_2,\cdots,\lambda_K\}$的对角阵。
对上式得优化目标关于H求导并设置为零，可得最优解H满足

$$LH=H\mathrm{diag}(\boldsymbol{\lambda}),$$

且对应得目标值为

$$\mathrm{Tr}(H^{\top}LH)=\sum_{k=1}^{K}\lambda_k.$$

因此H的解为拉普拉斯矩阵L的前K个最小特征值对应的特征向量构成的矩阵。

归一化切割与比率切割类似，优化目标得每一项可以表示为

$$\frac{W(A_k,\mathcal{G}\setminus A_k)}{vol(A_k)}=\frac{1}{2}\Big(\sum_{i\in A_k,j\notin A_k}w_{ij}\frac{1}{vol(A_k)}+\sum_{i\notin A_k,j\in A_k}w_{ij}\frac{1}{vol(A_k)}\Big)$$

$$=\frac{1}{2}\left(\sum_{i\in A_k,j\notin A_k}w_{ij}\Big(\frac{1}{\sqrt{vol(A_k)}}-0\Big)^2+\sum_{i\notin A_k,j\in A_k}w_{ij}\Big(0-\frac{1}{\sqrt{vol(A_k)}}\Big)^2\right)$$

$$=\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}w_{ij}(h_{ik}-h_{jk})^2$$

$$=\mathbf{h}_k^{\top}L\mathbf{h}_k.$$

根据$h_{i,k}$的定义，可得$H^{\top}DH=\mathbf{I}$. 不约束H为指示矩阵，但仍约束$H^{\top}DH=\mathbf{I}$，归一化切割的优化问题可以近似表示为

$$\arg\min_{F}\mathrm{Tr}(F^{\top}D^{-1/2}LD^{-1/2}F),$$

$$s.t.\ F^{\top}F=\mathbf{I},$$

其中$F=D^{1/2}H$. 对上述优化问题引入K个拉格朗日乘子，记为$\boldsymbol{\lambda}$

$$\arg\min_{H}\mathrm{Tr}(F^{\top}D^{-1/2}LD^{-1/2}F)+\mathrm{Tr}\big[\mathrm{diag}(\boldsymbol{\lambda})(F^{\top}F-\mathbf{I})\big].$$

算法 10-4 谱聚类

输入：数据 $\{x_n\}_{n=1}^N$，聚类数目 K

1: 定义一个相似度矩阵 W，如根据 $w_{ij}=w_{ji}=\exp(\frac{-\|\mathbf{x}_i-\mathbf{x}_j\|^2}{2\sigma^2})$ 构造高斯相似度矩阵；

2: 根据公式 $d_{ii}=\sum_{j}^{N} w_{ij}$ 和 $L=D-W$ 由相似度矩阵 W 构造图的拉普拉斯矩阵 L；

3: 求解特征值问题，如 $L\mathbf{h}=\lambda\mathbf{h}$（归一化谱聚类求解 $D^{-1/2}LD^{-1/2}\mathbf{h}=\lambda\mathbf{h}$）；

4: 选择前 K 个最小特征值对应的特征向量 $\{\mathbf{h}_k\}$ 来构造数据在 K 维新空间的表示 $H=[\mathbf{h}_1,\mathbf{h}_2,...,\mathbf{h}_H]$，并对 H 按照行归一化，其中 H 的每一行代表每一个数据点的表示；

5: 对新的数据表示 H 使用 K-均值等方法进行聚类。

输出：新的数据表示 H 和 K-均值等方法的聚类结果

目录

- K-均值聚类
  - 算法介绍
  - 模糊K-均值聚类
- 谱聚类
- 高斯混合模型聚类
  - 模型表示
  - 模型推理与参数估计
  - 无限高斯混合模型

- 模型表示

核心思想是假设数据可能来自不同的高斯分布，来自同一个高斯分布的数据点最可能属于同一个簇。
模型中引入潜变量$\mathbf{z}$，用于指示数据所属的成分。
$z_k = 1$表示该数据所属的成分是$k, k = 1,2,\cdots,K$.指示向量的先验分布为

$$p(\mathbf{z}) = \prod_{k=1}^{K} \pi_k^{z_k},$$

$$\sum_{k=1}^{K} \pi_k = 1,\ 0 \le \pi_k \le 1,$$

其中π_k是模型参数。根据模型假设，每个成分都是高斯分布，可得模型的似然分布为

$$p(\mathbf{x} \mid \mathbf{z}) = \prod_{k=1}^{K} \mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}_k, \Sigma_k)^{z_k}.$$

高斯混合模型的边缘似然表示为

$$p(\mathbf{x}) = \sum_{\mathbf{z}} p(\mathbf{z}) p(\mathbf{x} \mid \mathbf{z}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}_k, \Sigma_k).$$

- 模型推理与参数估计

定义$X = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N\}$和$Z = \{\mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_N\}$分别表示所有观测数据和潜变量，可以得到对数联合分布为

$$\ln p(X, Z \mid \boldsymbol{\mu}, \Sigma, \boldsymbol{\pi}) = \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} [\ln \pi_k + \ln \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \Sigma_k)],$$

其中$\boldsymbol{\mu} = \{\boldsymbol{\mu}_k\}_{k=1}^{K}$, $\Sigma = \{\Sigma_k\}_{k=1}^{K}$, $\boldsymbol{\pi} = [\pi_1, \pi_2, ..., \pi_K]^\top$

- 模型推理与参数估计

EM算法交替执行两个步骤：求期望（E步）和解最大化（ M步）
在**E步**，首先计算潜变量Z的后验分布，即每个数据点所属各个成分的概率。根据贝叶斯公式，可得每个数据点对应的指示变量的后验概率：

$$p(z_{nk}=1|\mathbf{x}_n)=\frac{\pi_k\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k,\Sigma_k)}{\sum_{j=1}^{K}\pi_j\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j,\Sigma_j)}.$$

然后计算对数联合分布关于潜变量的后验分布的期望：

$$\mathbb{E}_Z[\ln p(X,Z|\boldsymbol{\mu},\Sigma,\boldsymbol{\pi})]=\sum_{n=1}^{N}\sum_{k=1}^{K}p(z_{nk}=1|\mathbf{x}_n)[\ln\pi_k+\ln\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k,\Sigma_k)].$$

在**M步**，需要求解使上式中期望最大的参数

$$\arg\max_{\{\pi_k\},\{\boldsymbol{\mu}_k\},\{\Sigma_k\}}\sum_{n=1}^{N}\sum_{k=1}^{K}p(z_{nk}=1|\mathbf{x}_n)[\ln\pi_k+\ln\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k,\Sigma_k)],$$

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

- 模型推理与参数估计

对带约束的优化问题引入拉格朗日乘子 λ，得到等价无约束优化问题：

$$\arg\max_{\{\pi_k\},\{\boldsymbol{\mu}_k\},\{\Sigma_k\}} \sum_{n=1}^{N}\sum_{k=1}^{K} p(z_{nk}=1\,|\,\mathbf{x}_n)[\ln \pi_k + \ln \mathcal{N}(\mathbf{x}_n\,|\,\boldsymbol{\mu}_k,\Sigma_k)] + \lambda\left(\sum_{k=1}^{K}\pi_k - 1\right).$$

对上式的优化目标远古参数求导并设置为零

$$\pi_k = \frac{1}{N}\sum_{n=1}^{N} p(z_{nk}=1|\mathbf{x}_n),$$

$$\boldsymbol{\mu}_k = \frac{\sum_{n=1}^{N} p(z_{nk}=1|\mathbf{x}_n)\mathbf{x}_n}{\sum_{n=1}^{N} p(z_{nk}=1|\mathbf{x}_n)},$$

$$\Sigma_k = \frac{\sum_{n=1}^{N} p(z_{nk}=1|\mathbf{x}_n)(\mathbf{x}_n-\boldsymbol{\mu}_k)(\mathbf{x}_n-\boldsymbol{\mu}_k)^{\top}}{\sum_{n=1}^{N} p(z_{nk}=1|\mathbf{x}_n)}.$$

- 无限高斯混合模型

中餐馆过程假设餐馆中有无限个桌子，第一位顾客坐在第一张桌子，之后第n个顾客会以 $n_k/(n-1+\alpha)$的概率坐在已经有人的第k个桌子上，以$\alpha/(n-1+\alpha)$的概率坐在没有人的新桌子上，其中n_k表示第k个桌子上已有的顾客数，$n-1$表示在这个顾客之前已有的顾客总数，α是狄利克雷过程的参数。

使用中餐馆过程，无限高斯过程混合模型的指示变量Z的先验概率满足

$$p(z_{nk}=1|Z_{\backslash n},\alpha)=\begin{cases}\dfrac{N_k}{N-1+\alpha}, & k\le K,\\ \dfrac{\alpha}{N-1+\alpha}, & k>K,\end{cases}$$

假设每个高斯成分的均值与方差的先验分布为Normal-inverse-Wishart（NIW）分布，即

$$p(\boldsymbol{\mu}_k \mid \Sigma_k, \boldsymbol{\mu}_0, \kappa_0) = \mathcal{N}(\boldsymbol{\mu}_0, \Sigma_k / \kappa_0),$$
$$p(\Sigma_k \mid \Lambda_0, \nu_0) = IW(\Sigma_k \mid \Lambda_0, \nu_0),$$

其联合分布为

$$p(\boldsymbol{\mu}_k, \Sigma_k) = \mathrm{NIW}(\boldsymbol{\mu}_0, \kappa_0, \Lambda_0, \nu_0)$$
$$\propto |\Sigma_k|^{-((\nu_0+D)/2+1)} \exp\left\{-\frac{1}{2}\mathrm{Tr}(\Lambda_0\Sigma_k^{-1}) - \frac{\kappa_0}{2}(\boldsymbol{\mu}_k - \boldsymbol{\mu}_0)^\top \Sigma_k^{-1}(\boldsymbol{\mu}_k - \boldsymbol{\mu}_0)\right\}.$$

Gibbs采样

① 给定所有数据的所属成分，根据均值$\boldsymbol{\mu}_k$和协方差Σ_k的联合后验分布对$\boldsymbol{\mu}_k$和Σ_k进行采样。 $\boldsymbol{\mu}_k$和Σ_k的联合后验分布表示为

$$p(\boldsymbol{\mu}_k, \Sigma_k \mid X) = \mathrm{NIW}(\boldsymbol{\mu}_N^k, \kappa_N^k, \Lambda_N^k, \nu_N^k),$$

$$\boldsymbol{\mu}_N^k = \frac{\kappa_0 \boldsymbol{\mu}_0 + N_k \bar{X}_k}{\kappa_0 + N_k},$$

$$\kappa_N^k = \kappa_0 + N_k,$$

$$\Lambda_N^k = \Lambda_0 + \sum_{n=1}^{N_K} (\mathbf{x}_n^k - \bar{X}_k)(\mathbf{x}_n^k - \bar{X}_k)^\top + \frac{\kappa_0 N_k}{\kappa_0 + N_k} (\bar{X}_k - \boldsymbol{\mu}_0)(\bar{X}_k - \boldsymbol{\mu}_0)^\top,$$

$$\nu_N^k = \nu_0 + N_k,$$

其中， $\bar{X}_k$是属于第k个成分的数据的均值，$\mathbf{x}_n^k$是第k个成分中的数据点，N_k表示属于第k个成分的数据的数目。

Gibbs采样

② 对于每一个数据点$\mathbf{x}_n$，在给定其所属高斯分布的均值和协方差的情况下，对变量$\mathbf{z}_n$根据后验概率进行采样，$\mathbf{z}_n$的后验概率如下

$$p\left(z_{nk}=1 \mid Z_{\backslash n}, \mathbf{x}_n, \boldsymbol{\mu}_k, \Sigma_k, \alpha\right) \propto \begin{cases} \dfrac{N_k}{N-1+\alpha} p(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \Sigma_k) & k \le K, \\ \dfrac{\alpha}{N-1+\alpha} \int p(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \Sigma_k) p(\boldsymbol{\mu}_k, \Sigma_k) d\boldsymbol{\mu}_k d\Sigma_k & k > K. \end{cases}$$

交替执行上述两个步骤直至达到规定的迭代次数，假设得到S个指示变量Z的样本，记为$\{Z^{(s)}\}_{s=1}^{S}$，并且最终确定的聚类数目为K'，那么z_{nk}的后验分布可以通过如下计算得到

$$p(z_{nk}=1 \mid \mathbf{x}_n) = \frac{\sum_{s=1}^{S} I(z_{nk}^{(s)}=1)}{S},\ n=1,2,\ldots,N,\ k=1,2,\ldots,K'.$$

1. Xu R, Wunsch D. Survey of Clustering Algorithms[J]. IEEE Transactions on Neural Networks, 2005, 16(3): 645–678.
2. Duda R O, Hart P E, Stork D G. Pattern Classification[M]. New York: John Wiley & Sons, 2012.
3. 张学工. 模式识别[M]. 第三版. 北京：清华大学出版社, 2009.
4. Von Luxburg U. A Tutorial on Spectral Clustering[J]. Statistics and Computing, 2007, 17(4): 395-416.
5. Shi J, Malik J. Normalized Cuts and Image Segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2000, 22(8): 888-905.
6. Banfield J D, Raftery A E. Model-Based Gaussian and Non-Gaussian Clustering[J]. Biometrics, 1993, 49(3): 803-821.
7. Rasmussen C E. The Infinite Gaussian Mixture Model[C]//Advances in Neural Information Processing Systems. Cambridge, MA: MIT Press, 2000: 554-560.