4.3 Finite Mixture Model

有限混合模型与聚类

定义一个联合分布

$$p(z,\boldsymbol{x}) = P(z)p(\boldsymbol{x}|z)$$

其中z是物体种类，是一个未观测到的隐变量，$z \in \{1,2,3,\dots,K\}$;

$\boldsymbol{x} = \{A_1, A_2, \dots, A_n\}$是物体的属性，可观测；

$P(z)$是z的分布，$\pi_k = P(z=k)$是类别 k 的大小；

$p(\boldsymbol{x}|z)$是给定类别时物体属性的条件分布，$p(\boldsymbol{x}|z=k)$意味着类别 k 下物体属性的条件分布。

以上联合分布的边际分布是

$$p(\boldsymbol{x}) = \sum_{k=1}^{K} P(z=k)p(\boldsymbol{x}|z=k) = \sum_{k=1}^{K} \pi_k p(\boldsymbol{x}|z=k)$$

以上是一个单独类别的混合分布，每个$p(\boldsymbol{x}|z=k)$都是 mixture 当中的一个 component，对应的π_k称为混合参数。

因此以上模型称之为有限混合模型 Finite Mixture Model/FMM.

FMM 学习过程：给定无标签数据集$\{\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_n\}$以及一个数字$K$，利用最大似然估计确定参数$P(z)$以及$p(\boldsymbol{x}|z=k)$.

FMM 分类问题：给定参数$P(z)$以及$p(\boldsymbol{x}|z=k)$，

1. 软分类 Soft Assignment 定义为: 当物体拥有属性值$\boldsymbol{x}_n$时，有以下的概率是属于类别 k:

$$P(z=k|\boldsymbol{x}_n) = \frac{P(z=k)p(\boldsymbol{x}_n|z=k)}{p(\boldsymbol{x}_n)} = \frac{\pi_k p(\boldsymbol{x}|z=k)}{\sum_{k=1}^{K} \pi_k p(\boldsymbol{x}|z=k)}$$

2. 硬分类 Hard Assignment 定义为：拥有属性值$\boldsymbol{x}_n$的物体属于类别 k，iff 满足以下条件：

$$P(z=k^*|\boldsymbol{x}_n) \geq P(z=k|\boldsymbol{x}_n), \forall k^* \neq k$$

连续数据空间的高斯混合模型 Gaussian Mixture Models/GMM

一维高斯分布：

$$p(x|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp[-\frac{(x-\mu)^2}{2\sigma^2}]$$

二维高斯分布：

$$p(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma})=\frac{1}{\sqrt{(2\pi)^d\det(\boldsymbol{\Sigma})}}\exp[-\frac{(\boldsymbol{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}{2}]$$

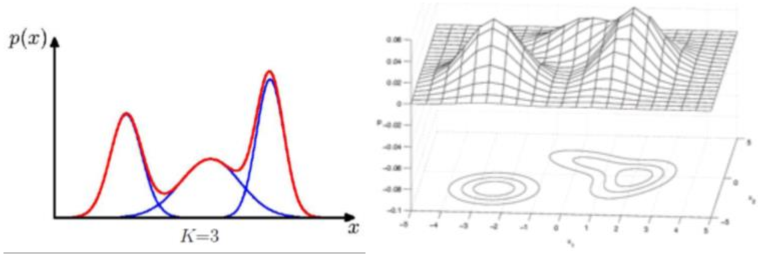
其中 d 是维数，$\boldsymbol{x}$代表数据，$\boldsymbol{\mu}$代表均值向量，$\boldsymbol{\Sigma}$代表协方差矩阵。

高斯混合模型 GMM (Soft Assignment): 令

$$p(\boldsymbol{x})=\sum_{k=1}^{K}\pi_k p(\boldsymbol{x}|z=k)$$

$$p(\boldsymbol{x}|z=k)\sim\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)$$

一维和二维高斯混合模型的一个例子：



从而问题变为：给定无标签数据集$\{\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_n\}$以及一个数字$K$，确定

1. 混合参数$\pi_k(k=1,2,\ldots,K)$
2. 成分参数 Component parameters: $\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k\ (k=1,2,\ldots,K)$

使 K 成分混合模型有最大似然估计，从而使得混合模型和数据匹配较好。

<u>Expectation Maximization/EM 算法</u>：

1. 初始化$\pi_k,\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k$
2. 重复以下过程直至收敛：

   Expectation: 对于每个训练样本$\boldsymbol{x}_n$

   a. from k = 1 to K, 计算

$$r_{nk}\equiv P(z=k|\boldsymbol{x}_n)=\frac{\pi_k p(\boldsymbol{x}|z=k)}{\sum_{k=1}^{K}\pi_k p(\boldsymbol{x}|z=k)}=\frac{\pi_k\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)}{\sum_{k=1}^{K}\pi_k\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)}$$

   b. 根据概率将其分成 K 个碎片样本 fractional examples：

$$\boldsymbol{x}_n[r_{nk}]\ (k=1,2,\ldots,K)$$

   c. 将每个碎片样本$\boldsymbol{x}_n[r_{nk}]$分类放入对应的聚类 k；

   Maximization: 重新估计$\pi_k,\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k$：

$$\pi_k^{new}=\frac{N_k}{N},\qquad \boldsymbol{\mu}_k^{new}=\frac{1}{N_k}\sum_{n-1}^{N}r_{nk}\boldsymbol{x}_n,$$

$$\boldsymbol{\Sigma}_k^{new}=\frac{1}{N_k}\sum_{n=1}^{N}r_{nk}(\boldsymbol{x}_n-\boldsymbol{\mu}_k^{new})(\boldsymbol{x}_n-\boldsymbol{\mu}_k^{new})'$$

EM 算法收敛：

令

$\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_n\}, \boldsymbol{\pi} = \{\boldsymbol{\pi}_1, \boldsymbol{\pi}_2, \dots, \boldsymbol{\pi}_K\}, \boldsymbol{\mu} = \{\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \dots, \boldsymbol{\mu}_K\}, \boldsymbol{\Sigma} = \{\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2, \dots, \boldsymbol{\Sigma}_K\}$

则对数似然方程

$$l(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma} | \boldsymbol{X}) = \log p(\boldsymbol{X} | \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \log \prod_{n=1}^{N} p(\boldsymbol{x}_n | \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$$

EM 算法致力于计算模型参数的最大似然估计 MLE：

$$(\boldsymbol{\pi}^*, \boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*) = \arg\max_{\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}} \log p(\boldsymbol{X} | \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$$

给定一个阈值，当$l(t+1) - l(t) < \text{threshold}$时终止算法。

在实际计算当中，outliers 或者重复点处的最大似然可能为无穷，解决方法是为协方差矩阵的特征值设定一个范围；同时，可以采用多种初始值多次实验的方法来避免算法陷入局部最佳。

聚类质量分析：

1. held-out likelihood
2. <u>贝叶斯信息准则 Bayes Information Criterion/BIC</u>:

$$\max_{\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}} \log p(\boldsymbol{X} | \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) - \frac{d}{2} \log N$$

　其中 d 是模型参数个数。

可从 K = 2 开始慢慢增加 K，直至 held-out likelihood 或者 BIC 开始减小。

补充：K-mean 聚类算法：

- Select K points as the initial centroids.
- **repeat**:
  - For K clusters by assigning all points to the closest centroid.
  - Recompute the centroid of each cluster
- **until** the centroids don't change

Hard Assignment

离散数据空间的隐含类别模型 Latent Class Models/LCM

离散取值的属性集$\boldsymbol{x} = \{A_1, A_2, \dots, A_n\}$，混合模型为

$$P(\boldsymbol{x}) = \sum_{k=1}^{K} \pi_k P(\boldsymbol{x} | z = k)$$

其中

$$P(\boldsymbol{x}|z=k)=P(A_1,A_2,\dots,A_n|z=k)$$

假设属性相互独立，即有

$$P(A_1,A_2,\dots,A_n|z=k)=P(A_1|z=k)P(A_2|z=k)\dots P(A_n|z=k)$$

所以

$$P(\boldsymbol{x})=\sum_{k=1}^{K}\pi_k P(\boldsymbol{x}|z=k)=\sum_{k=1}^{K}\pi_k\prod_{i=1}^{n}P(A_i|z=k)$$

其与朴素贝叶斯方法相同，唯一不同的一点是其类别标签未知，是隐含的。

同理一个属性为$\boldsymbol{x}=\{a_1,a_2,\dots,a_n\}$的物体有概率

$$P(z=k|a_1,a_2,\dots,a_n)=\frac{\pi_k\prod_{i=1}^{n}P(a_i|z=k)}{\sum_{k=1}^{K}\pi_k\prod_{i=1}^{n}P(a_i|z=k)}$$

被判入类别 k。

LCM 学习过程：给定无标签离散数据集$\{\boldsymbol{x}_1,\boldsymbol{x}_2,\dots,\boldsymbol{x}_N\}$以及一个数字$K$，确定

1. 混合参数$\pi_k=P(\boldsymbol{x}|z=k)\ (k=1,2,\dots,K)$
2. 成分参数 Component parameters: $P(A_i|z=k)\ (k=1,2,\dots,K)$

使 K 成分混合模型有最大似然估计，从而使得混合模型和数据匹配较好。

学习算法依然为 EM 算法：

1. 初始化$\pi_k,P(A_i|z=k)$
2. 重复以下过程直至收敛：

   Expectation: 对于每个训练样本$\boldsymbol{x}_n$

   d. from k = 1 to K, 计算

   $$r_{nk}\equiv P(z=k|\boldsymbol{x}_n)=\frac{\pi_k\prod_{i=1}^{n}P(a_i|z=k)}{\sum_{k=1}^{K}\pi_k\prod_{i=1}^{n}P(a_i|z=k)}$$

   e. 根据概率将其分成 K 个碎片样本 fractional examples：

   $$\boldsymbol{x}_n[r_{nk}]\ (k=1,2,\dots,K)$$

   f. 将每个碎片样本$\boldsymbol{x}_n[r_{nk}]$分类放入对应的聚类 k；

   Maximization: 重新估计$\pi_k,P(A_i|z=k)$：

EM 算法在 LCM 上总是收敛的：令

$$\boldsymbol{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\dots,\boldsymbol{x}_n\},\theta=\{\pi_k,P(A_i|z=k)|k=1,\dots K,i=1,\dots n\}$$

则对数似然方程

$$l(\theta|\boldsymbol{X})=\log p(\boldsymbol{X}|\theta)=\log\prod_{j=1}^{N}p\left(\boldsymbol{x}_j|\theta\right)\le 0$$

令$\theta_1,\theta_2,\dots$为 EM 算法产生的参数，其单调递增，且上界为 0.