EMMGMM 推导

Soul Walker

2020 年 7 月 22 日

输入：训练数据集

$$T = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n\} = \{\boldsymbol{x}_i\}_{i=1}^n$$

其中，$\boldsymbol{x}_i \in \chi \in R^n$ 为实例的特征向量，$\boldsymbol{x}_i = (\boldsymbol{x}_i^{(1)}, \boldsymbol{x}_i^{(2)}, ..., \boldsymbol{x}_i^{(n)})^T$ ，$i = 1, 2, \cdots, n$，实例特征向量 $\boldsymbol{x}$，$\boldsymbol{X}$ 为训练样本集，形状为 (n, m)，我们希望将样本分为 K 个簇，$\boldsymbol{\mu}$ 为样本均值，$\boldsymbol{\Sigma}$ 为样本协方差矩阵。
输出：簇的划分。

0.1 背景

为了解决高斯模型的单峰性的问题，我们引入多个（这里假设有 K 个）高斯模型的加权平均来拟合多峰数据

$$p(x) = \sum_{k=1}^{K} p_k \mathcal{N}(\mu_k, \Sigma_k) \tag{0.1}$$

引入隐变量 z 表示对应的样本 x 是属于哪一个高斯分布

$$p(z=i) = p_i, \quad \sum_{i=1}^{K} p(z=i) = 1 \tag{0.2}$$

0.2 多个高斯分布（极大似然估计）

假设有 k 个高斯分布而且数据独立同分布,假设其参数集合 $\theta = \{p_1, \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1, \cdots, p_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}$，那么我们有

$$L(\theta|\boldsymbol{X}) = p(\boldsymbol{X}|\theta) \tag{0.3}$$

其中 $L(\theta|\boldsymbol{X})$ 为似然函数，那么参数估计为

$$\theta = \arg\max_{\theta} L(\theta|\boldsymbol{X}) \tag{0.4}$$

此处有

$$p(\boldsymbol{x}) = \sum_{z} p(\boldsymbol{x}, z) = \sum_{k=1}^{K} p(\boldsymbol{x}, z=k) = \sum_{k=1}^{K} p(z=k) p(\boldsymbol{x}|z=k) \tag{0.5}$$

因此可得如下公式：

$$p(\boldsymbol{x}|\theta) = \sum_{k=1}^{K} p_k \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{0.6}$$

接下来进行对 θ 的极大似然估计如下

$$\begin{aligned} \theta_{MLE} &= \arg\max_{\theta} logp(\boldsymbol{X}|\theta) \\ &= \arg\max_{\theta} \sum_{i=1}^{n} logp(\boldsymbol{x}_i|\theta) \\ &= \arg\max_{\theta} \sum_{i=1}^{n} log \sum_{k=1}^{K} p_k \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \end{aligned} \tag{0.7}$$

对于上式，我们直接进行求导，但是我们并不能得到解析解，因此极大似然估计算法走不通，那么我们将进而使用 EM 算法。

0.3 EM 算法

首先我们引入 Jensen 不等式（概率论版）

$$f(E[X]) \leq E(f[X]) \tag{0.8}$$

其中 $f(x)$ 为凸函数。

我们想要引入隐变量 z，那么我们对于 $p(\boldsymbol{x}|\theta)$ 问题的讨论就变成对 $p(\boldsymbol{x}, z|\theta)$ 的讨论，满足 $\int p(\boldsymbol{x}, z|\theta) = p(\boldsymbol{x}|\theta)$，也就是 $p(\boldsymbol{x}|\theta) = \sum_{k=1}^{K} p_k \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$，

其中可得 $p(z=k)=p_k$ 为隐变量的先验分布。那么似然函数有

$$
\begin{aligned}
L(\theta|\boldsymbol{X}) &= logp(\boldsymbol{X}|\theta) \\
&= log\prod_{i=1}^{n} p(\boldsymbol{x}_i, z_i|\theta) \\
&= \sum_{i=1}^{n} log[\sum_{k=1}^{K} p(\boldsymbol{x}_i|z_i=k,\theta)p(z_i=k)] \\
&= \sum_{i=1}^{n} log[\sum_{k=1}^{K} p(z_i=k|\boldsymbol{x}_i,\theta)\frac{p(\boldsymbol{x}_i|z_i=k,\theta)p(z_i=k)}{p(z_i=k|\boldsymbol{x}_i,\theta)}]
\end{aligned}
\tag{0.9}
$$

这里我们令

$$
f(\boldsymbol{x}_i) = \frac{p(\boldsymbol{x}_i|z_i=k,\theta)p(z_i=k)}{p(z_i=k|\boldsymbol{x}_i,\theta)} \tag{0.10}
$$

那么

$$
E_{z_i} = E[f(\boldsymbol{x}_i)] = \sum_{k=1}^{K} p(z_i=k|\boldsymbol{x}_i,\theta)f(x_i) \tag{0.11}
$$

使用 Jensen 不等式可得

$$
L(\theta|\boldsymbol{x}) \geq \sum_{i=1}^{n}\sum_{k=1}^{K} p(z_i=k|\boldsymbol{x}_i,\theta)log\frac{p(\boldsymbol{x}_i|z_i=k,\theta)p(z_i=k)}{p(z_i=k|\boldsymbol{x}_i,\theta)} \tag{0.12}
$$

那么我们的最大化原函数问题就可以转化为最大化原函数下界问题 ($L(\theta^{(t)}|\boldsymbol{x}) \geq L(\theta^{(t-1)}|\boldsymbol{x})$)。那么对于 EM 算法的描述可以为：求解 θ^{t-1}，然后确定 z_i（的分布），然后确定 n 个样本点的归属，然后确定 θ^t，然后确定 z_i（的分布），等等，如此迭代下去，直到算法收敛。
首先构造下界 $\beta(\theta,\theta^{t-1})$

$$
\sum_{i=1}^{n}\sum_{i=1}^{K} p(z_i=k|\boldsymbol{x}_i,\theta^{(t-1)})log\frac{p(\boldsymbol{x}_i|z_i=k,\theta^{(t-1)})p(z_i=k)}{p(z_i=k|\boldsymbol{x}_i,\theta^{(t-1)})} \tag{0.13}
$$

然后进行最大化

$$
\theta^{(t)} = \arg\max_{\theta} \beta(\theta,\theta^{(t-1)}) \tag{0.14}
$$

由下界前提可得

$$L(\theta^{(t)}|\boldsymbol{x}) \geq \beta(\theta^{(t)},\theta^{(t-1)}) \geq \beta(\theta^{(t-1)},\theta^{(t-1)}) \tag{0.15}$$

其中

$$\begin{aligned}
\beta(\theta^{(t-1)},\theta^{(t-1)}) &= \sum_{i=1}^{n}\sum_{i=1}^{K} p(z_i=k|\boldsymbol{x}_i,\theta^{(t-1)}) log \frac{p(\boldsymbol{x}_i|z_i=k,\theta^{(t-1)})p(z_i=k)}{p(z_i=k|\boldsymbol{x}_i,\theta^{(t-1)})} \\
&= \sum_{i=1}^{n}\sum_{i=1}^{K} p(z_i=k|\boldsymbol{x}_i,\theta^{(t-1)}) log \frac{p(z_i=k|\boldsymbol{x}_i,\theta^{(t-1)})p(\boldsymbol{x}|\theta^{(t-1)})}{p(z_i=k|\boldsymbol{x}_i,\theta^{(t-1)})} \\
&= \sum_{i=1}^{n}\sum_{i=1}^{K} p(z_i=k|\boldsymbol{x}_i,\theta^{(t-1)}) log p(\boldsymbol{x}_i|\theta^{(t-1)}) \\
&= L(\theta^{(t-1)}|\boldsymbol{x})
\end{aligned} \tag{0.16}$$

那么式 (15) 可以转化为

$$L(\theta^{(t)}|\boldsymbol{x}) \geq \beta(\theta^{(t)},\theta^{(t-1)}) \geq \beta(\theta^{(t-1)},\theta^{(t-1)}) = L(\theta^{(t-1)}|\boldsymbol{x}) \tag{0.17}$$

那么收敛性可得证。

0.4 EM 算法求解 GMM

对于给定数据集 $\boldsymbol{x} \in R^d$，$\boldsymbol{\mu} \in R^d$，$\boldsymbol{\Sigma} \in R^{d\times d}$ 的高斯分布

$$\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^d det(\boldsymbol{\Sigma})}} exp[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})] \tag{0.18}$$

第一步，已知 $\theta^{(t-1)}$（可通过之前的迭代可得）。
第二步，计算 z_i 的后验概率 $p(z_i=k|\boldsymbol{x}_i,\theta^{(t-1)})$

$$\begin{aligned}
p(z_i=k|\boldsymbol{x}_i,\theta^{(t-1)}) &= \frac{p(\boldsymbol{x}_i|z_i=k,\theta^{(t-1)})p(z_i=k|\theta^{(t-1)})}{\sum_{k=1}^{K} p(\boldsymbol{x}_i|z_i=k,\theta^{(t-1)})p(z_i=k|\theta^{(t-1)})} \\
&= \frac{\mathcal{N}(\boldsymbol{x}_i|\theta_k^{(t-1)})p_k^{(t-1)}}{\sum_{k=1}^{K}\mathcal{N}(\boldsymbol{x}_i|\theta_k^{(t-1)})p_k^{(t-1)}} \\
&\overset{令}{=} \boldsymbol{q}_{ik}^{(t-1)}
\end{aligned} \tag{0.19}$$

这里 q_{ik} 的意义是第 i 个样本点属于第 k 个后验概率，且 $\boldsymbol{q}_{ik} \in R^{n\times k}$。

第三步，E 步

$$\begin{aligned}
&\beta(\theta,\theta^{(t-1)})\\
=&\sum_{i=1}^{n}\sum_{k=1}^{K}\boldsymbol{q}_{ik}^{(t-1)}log(\frac{p(\boldsymbol{x}_i,z_i|\theta)}{\boldsymbol{q}_{ik}^{(t-1)}})\\
=&\sum_{i=1}^{n}\sum_{k=1}^{K}\boldsymbol{q}_{ik}^{(t-1)}log(\frac{p(\boldsymbol{x}_i|z_i=k,\theta)p(z_i=k|\theta)}{\boldsymbol{q}_{ik}^{(t-1)}})\\
=&\sum_{i=1}^{n}\sum_{k=1}^{K}(\boldsymbol{q}_{ik}^{(t-1)}logp(\boldsymbol{x}_i|z_i=k,\theta)+\boldsymbol{q}_{ik}^{(t-1)}logp_k-\boldsymbol{q}_{ik}^{(t-1)}log\boldsymbol{q}_{ik}^{(t-1)})\\
=&\sum_{i=1}^{n}\sum_{k=1}^{K}\boldsymbol{q}_{ik}^{(t-1)}(logp(\boldsymbol{x}_i|z_i=k,\theta)+logp_k-log\boldsymbol{q}_{ik}^{(t-1)})\\
=&\sum_{i=1}^{n}\sum_{k=1}^{K}\boldsymbol{q}_{ik}^{(t-1)}(logp_k-log\boldsymbol{q}_{ik}^{(t-1)}-\frac{d}{2}log\sqrt{(2\pi)^d}-\frac{1}{2}logdet(\boldsymbol{\Sigma}_k)-\frac{1}{2}(\boldsymbol{x}_i-\boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}_i-\boldsymbol{\mu}_k))
\end{aligned} \tag{0.20}$$

第四步，M 步

对于 $\boldsymbol{\mu}_k$：

$$\boldsymbol{\mu}_k=\arg\max_{\boldsymbol{\mu}_k}\beta(\theta,\theta^{(t-1)}) \tag{0.21}$$

$\beta(\theta,\theta^{(t-1)})$ 对 $\boldsymbol{\mu}_k$ 求偏导，并令其等于 0 得

$$\begin{aligned}
\frac{\partial\beta(\theta,\theta^{(t-1)})}{\partial\boldsymbol{\mu}_k}=\sum_{i=1}^{n}\boldsymbol{q}_{ik}\Sigma_k^{-1}(\boldsymbol{x}_i-\boldsymbol{\mu}_k)=0\\
\sum_{i=1}^{n}\boldsymbol{q}_{ik}\boldsymbol{x}_i=\sum_{i=1}^{n}\boldsymbol{q}_{ik}\boldsymbol{\mu}_k\\
\boldsymbol{\mu}_k=\frac{\sum_{i=1}^{n}\boldsymbol{q}_{ik}\boldsymbol{x}_i}{\sum_{i=1}^{n}\boldsymbol{q}_{ik}}
\end{aligned} \tag{0.22}$$

那么

$$\boldsymbol{\mu}_k^{(t)}=\frac{\sum_{i=1}^{n}\boldsymbol{q}_{ik}^{(t-1)}\boldsymbol{x}_i}{\sum_{i=1}^{n}\boldsymbol{q}_{ik}^{(t-1)}} \tag{0.23}$$

对于 $\boldsymbol{\Sigma}_k^{(t)}$：

$$\boldsymbol{\Sigma}_k^{(t)} = \arg\max_{\boldsymbol{\Sigma}_k} \beta(\theta, \theta^{(t-1)}) \tag{0.24}$$

$\beta(\theta, \theta^{(t-1)})$ 对 $\boldsymbol{\Sigma}_k^{-1}$ 求偏导，并令其等于 0 得

$$\frac{\partial \beta(\theta, \theta^{(t-1)})}{\partial \boldsymbol{\Sigma}_k^{-1}} = \frac{1}{2} \sum_{i=1}^{n} \boldsymbol{q}_{ik} [\boldsymbol{\Sigma}_k - (\boldsymbol{x}_i - \boldsymbol{\mu}_k^t)^T (\boldsymbol{x}_i - \boldsymbol{\mu}_k^t)] = 0$$
$$\Sigma_k = \frac{\sum_{i=1}^n \boldsymbol{q}_{ik} (\boldsymbol{x}_i - \boldsymbol{\mu}_k)^T (\boldsymbol{x}_i - \boldsymbol{\mu}_k)}{\sum_{i=1}^n \boldsymbol{q}_{ik}} \tag{0.25}$$

那么

$$\boldsymbol{\Sigma}_k^{(t)} = \frac{\sum_{i=1}^n \boldsymbol{q}_{ik}^{(t-1)} (\boldsymbol{x}_i - \boldsymbol{\mu}_k^{(t)})^T (\boldsymbol{x}_i - \boldsymbol{\mu}_k^{(t)})}{\sum_{i=1}^n \boldsymbol{q}_{ik}^{(t-1)}} \tag{0.26}$$

对于 p_k：

$$p_k = \begin{cases} \arg\max\limits_{p_k} \sum\limits_{i=1}^{n} \sum\limits_{k=1}^{K} log p_k \\ s.t. \quad \sum\limits_{k=1}^{K} p_k = 1 \end{cases} \tag{0.27}$$

使用拉格朗日乘子法解决如下

$$L(p_k, \lambda) = \sum_{i=1}^{n} \sum_{k=1}^{K} \boldsymbol{q}_{ik} log p_k - \lambda (\sum_{k=1}^{K} p_k - 1) \tag{0.28}$$

上式对 p_k 求偏导，并令其等于 0 得

$$\frac{\partial L(p_k, \lambda)}{\partial p_k} = \sum_{i=1}^{n} \boldsymbol{q}_{ik} \frac{1}{p_k} + \lambda = 0$$
$$p_k = \frac{\sum_{i=1}^n \boldsymbol{q}_{ik}}{\lambda} \tag{0.29}$$

那么

$$
\begin{aligned}
\sum_{k=1}^{K} p_k &= \frac{\sum_{k=1}^{K}\sum_{i=1}^{n} \boldsymbol{q}_{ik}}{\lambda} \\
1 &= \sum_{i=1}^{n} \frac{1}{\lambda} \\
1 &= \frac{n}{\lambda} \\
\lambda &= n
\end{aligned} \tag{0.30}
$$

带入式 (29) 得

$$
p_k^{(t)} = \frac{\sum_{i=1}^{n} \boldsymbol{q}_{ik}^{t-1}}{n} \tag{0.31}
$$

至此，以上算法推导结束。