

## 2. 用EM算法求解高斯混合模型

### 2.1 不含隐变量的高斯分布参数估计---极大似然估计

找到与样本的分布最接近的概率分布模型。例如，给定一组样本  $X = \{x_1, x_2, \dots, x_n\}$ ，已知它们来自于高斯分布  $N(\mu, \sigma)$ ，估计参数  $\mu, \sigma$ 。

方法是：已知高斯分布的概率密度函数为

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

则  $X$  的似然函数为

$$L(x) = \prod_{i=1}^N \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

对似然函数取对数并且进行化简

$$\begin{aligned} l(x) &= \log \prod_i^n \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} \\ &= \sum_i^n \log \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} \\ &= \left( \sum_i^n \log \frac{1}{\sqrt{2\pi}\sigma} \right) + \left( \sum_i^n -\frac{(x_i-\mu)^2}{2\sigma^2} \right) \\ &= -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_i^n (x_i - \mu)^2 \end{aligned}$$

将对数似然函数  $l(x)$  对参数  $\mu$  和  $\sigma$  分别求偏导数，令它们等于0。

$$\begin{aligned} \frac{\partial l}{\partial \mu} &= \frac{1}{\sigma^2} \sum_{i=1}^n (x_i - \mu) = 0 \\ \frac{\partial l}{\partial \sigma} &= -\frac{n}{2} \frac{4\pi\sigma}{2\pi\sigma^2} + \frac{1}{\sigma^3} \sum_i^n (x_i - \mu)^2 = 0 \end{aligned}$$

整理后可以得到

$$\begin{aligned} \mu &= \frac{1}{n} \sum_i^n x_i \\ \sigma^2 &= \frac{1}{n} \sum_i^n (x_i - \mu)^2 \end{aligned}$$

### 2.2 高斯混合模型GMM

随机变量 $X$ 的分布服从由 $K$ 个高斯分布混合而成的分布。定义如下变量。 $\pi_1, \pi_2, \dots, \pi_K$ 为选取各个高斯分布的概率。第 $k$ 个高斯分布的均值为 $\mu_k$ ，标准差为 $\Sigma_k$ 。所以 $x$ 的概率分布可以表示为

$$p(x_i|\theta) = \sum_{k=1}^K \pi_k N(x_i|\mu_k, \Sigma_k)$$

$\theta$ 表示 $x$ 的概率分布的参数，在这里表示 $\pi, \mu, \Sigma$ 。

给定一组样本 $X = \{x_1, x_2, \dots, x_n\}$ ，试估计 $\pi, \mu, \Sigma$ 。如果样本 $x_i$ 是标量，标准差 $\Sigma$ 是一个数，样本 $x_i$ 是向量，标准差 $\Sigma$ 是方阵。

首先定义对数似然函数

$$\begin{aligned} l_{\pi, \mu, \Sigma}(x) &= \log \prod_i \left( \sum_{k=1}^K \pi_k N(x_i|\mu_k, \Sigma_k) \right) \\ &= \sum_{i=1}^N \log \left( \sum_{k=1}^K \pi_k N(x_i|\mu_k, \Sigma_k) \right) \end{aligned}$$

随机初始化参数 $\pi, \mu, \Sigma$ ，计算第 $i$ 个样本 $x_i$ 是来自于第 $k$ 个高斯分布生成的概率是

$$\gamma(i, k) = \frac{\pi_k N(x_i|\mu_k, \Sigma_k)}{\sum_{j=1}^K \pi_j N(x_i|\mu_j, \Sigma_j)}$$

对于第 $k$ 个高斯分布，可以看成是生成了 $\{\gamma(1, k)x_1, \gamma(2, k)x_2, \dots, \gamma(n, k)x_n\}$ 样本集。利用上面极大似然估计法得到的高斯分布的均值和方差，

$$\begin{cases} \mu = \frac{1}{n} \sum_i x_i \\ \sigma^2 = \frac{1}{n} \sum_i (x_i - \mu)^2 \end{cases}$$

代入样本集中得到新的 $\pi, \mu, \Sigma$

$$\begin{cases} N_k = \sum_{i=1}^N \gamma(i, k) \\ \mu_k = \frac{1}{N_k} \sum_{i=1}^N \gamma(i, k) x_i \\ \Sigma_k = \frac{1}{N_k} \sum_{i=1}^N \gamma(i, k) (x_i - \mu_k) (x_i - \mu_k)^T \\ \pi_k = \frac{N_k}{N} = \frac{1}{N} \sum_{i=1}^N \gamma(i, k) \end{cases}$$

其中 $N_k$ 表示第 $k$ 个高斯分布生成的样本点的个数。 $\pi_k$ 表示大小为 $N$ 的样本集中，有多少样本点来自 $k$ 的占比，来表示概率。值得注意的是， $\sum_{k=1}^K N_k = N$ 。对两个过程重复迭代至收敛，得到 $\pi, \mu, \Sigma$ 的估计值。

## 2.3 高斯混合模型的EM算法

下面对上述过程进行推导

首先定义隐变量 $z_i$

$$z_{i,k} = \begin{cases} 1, & \text{第 } i \text{ 个样本属于第 } k \text{ 个高斯分布} \\ 0, & \text{其他} \end{cases}$$

这是一个 $K$ 维的one-hot的向量，下标 $i$ 表示第 $i$ 个观测的隐变量。假设 $z_{i,k}$ 之间是独立同分布的，因此 $z_i$ 的概率分布是

$$p(z_i) = p(z_{i,1}) p(z_{i,2}) \dots p(z_{i,K}) = \prod_{k=1}^K \pi_k^{z_{i,k}}$$

其次计算样本 $x_i$ 在给定隐变量下的条件分布 $P(x_i|z_i)$

因为单个维度的隐变量为条件， $x_i$ 的条件概率为 $p(x_i|z_{i,k}=1) = N(x_i|\mu_k, \Sigma_k)$  因此隐变量向量 $z_i$ 为条件的 $x_i$ 的条件概率为

$$p(x_i|z_i) = \prod_{k=1}^K N(x_i|\mu_k, \Sigma_k)^{z_{i,k}}$$

利用条件概率和隐变量的概率，可以求得观测变量的概率

$$\begin{aligned} p(x_i) &= \sum_{z_i} p(z_i) p(x_i|z_i) \\ &= \sum_{z_i} \left( \prod_{k=1}^K \pi_k^{z_{i,k}} N(x_i|\mu_k, \Sigma_k)^{z_{i,k}} \right) \\ &= \sum_{k=1}^K \pi_k N(x_i|\mu_k, \Sigma_k) \end{aligned}$$

这个式子也说明了为什么在2.2节一开始，就表明了 $x$ 如果服从混合高斯分布，它的写法是长这样的。上式中，对 $z_i$ 求和 $\sum_{z_i}$ ，实际上是 $\sum_{k=1}^K$ ，并且括号里面的连乘部分，其实仅仅只有样本真正对应的第 $k$ 个高斯分布，即第 $k$ 项是有意义的，其余项的 $z_k = 0$ ，即连乘的都是1。因此在第三个等式外层的 $\sum_{k=1}^K$ 循环是这么得来的。

进而样本集的对数似然函数为

$$\begin{aligned} l_{\pi, \mu, \Sigma}(X) &= \log \prod_i^N \left( \sum_{k=1}^K \pi_k N(x_i|\mu_k, \Sigma_k) \right) \\ &= \sum_{i=1}^N \log \left( \sum_{k=1}^K \pi_k N(x_i|\mu_k, \Sigma_k) \right) \end{aligned}$$

接下来介绍 $\gamma(i, k)$ ，表示给定第 $i$ 个样本 $x_i$ 来自第 $k$ 个高斯分布生成的概率，是关于隐变量的条件概率。因此

$$\begin{aligned}
\gamma(i, k) &= \frac{P(z_{i,k} = 1, x_i | \theta)}{\sum_{l=1}^K P(z_{i,l} = 1, x_i | \theta)} = \left( \frac{P(z_{i,k} = 1, x_i | \theta)}{P(x_i | \theta)} = P(z_{i,k} = 1 | x_i, \theta) \right) \\
&= \frac{P(x_i | z_{i,k} = 1, \theta) P(z_{i,k} = 1 | \theta)}{\sum_{l=1}^K P(x_i | z_{i,l} = 1, \theta) P(z_{i,l} = 1 | \theta)} \\
&= \frac{\pi_k N(x_i | \mu_k, \Sigma_k)}{\sum_{l=1}^K \pi_l N(x_i | \mu_l, \Sigma_l)}
\end{aligned}$$

分子是第*i*个样本，同时又属于第*k*个高斯分布生成的概率。因为 $\gamma(i, k)$ 是概率，所以需要分母进行归一化。第二个等式是写成了条件概率的形式。因此，根据定义， $P(z_{i,k} = 1 | \theta) = \pi_k$ 表示第*i*个样本属于第*k*个高斯分布生成的概率，这里忽略了下标*i*，因为对于任意一个样本点*i*，属于第*k*个高斯分布生成的概率都为 $\pi_k$ ，不用区分样本。其次， $P(x_i | z_{i,k} = 1, \theta)$ 表示给定了样本 $x_i$ 是来自第*k*个高斯分布作为已知条件，因此使用第*k*个高斯分布的参数来生成样本 $x_i$ ，因此 $P(x_i | z_{i,k} = 1, \theta) = N(x_i | \mu_k, \Sigma_k)$ 。因此分子变成了 $\pi_k N(x_i | \mu_k, \Sigma_k)$ 。分母的变化和分子一致。 $\gamma(i, k)$ 可以计算出来了。

接着利用样本集的对数似然函数进行极大似然估计。已知对数似然函数为

$$\begin{aligned}
l_{\pi, \mu, \Sigma}(X) &= \sum_{i=1}^N \log \left( \sum_{k=1}^K \pi_k N(x_i | \mu_k, \Sigma_k) \right) \\
&= \sum_{i=1}^N \log \left( \sum_{k=1}^K Q(z_{i,k}) \frac{\pi_k N(x_i | \mu_k, \Sigma_k)}{Q(z_{i,k})} \right) \\
&\geq \sum_{i=1}^N \sum_{k=1}^K Q(z_{i,k}) \log \left( \frac{\pi_k N(x_i | \mu_k, \Sigma_k)}{Q(z_{i,k})} \right) \\
&= \sum_{i=1}^N \sum_{k=1}^K Q(z_{i,k}) \log \left( \frac{\frac{\pi_k}{\sqrt{2\pi}} \Sigma_k^{-1} e^{-\frac{(x_i - \mu_k)^T \Sigma_k^{-2} (x_i - \mu_k)}{2}}}{Q(z_{i,k})} \right) = B
\end{aligned}$$

因此利用对数似然函数的下界，对下界求偏导数。如果想要让下界B取到极大值，要让线段与log函数相交的点在同一点上，即线段汇聚成一点，因此要求

$$\frac{\pi_k N(x_i | \mu_k, \Sigma_k)}{Q(z_{i,k})} = c$$

因此

$$Q(z_{i,k}) \propto \pi_k N(x_i | \mu_k, \Sigma_k)$$

又因为 $\sum_{k=1}^K Q(z_{i,k}) = 1$ ，因此

$$Q(z_{i,k}) = \frac{\pi_k N(x_i | \mu_k, \Sigma_k)}{\sum_{l=1}^K \pi_l N(x_i | \mu_l, \Sigma_l)}$$

发现 $Q(z_{i,k})$ 刚好等于 $\gamma(i, k)$ ，表示给定的样本点 $x_i$ ，属于第 $k$ 个高斯分布生成的概率。

对第 $k$ 个高斯分布的均值 $u_k$ 求偏导，并令其等于0：

$$\begin{aligned}\frac{\partial B}{\partial u_k} &= -\nabla_{\mu_k} \sum_{i=1}^N \sum_{k=1}^K \gamma(i, k) \frac{(x_i - \mu_k)^T \Sigma_k^{-2} (x_i - \mu_k)}{2} \\ &= \sum_{i=1}^N \gamma(i, k) (x_i - \mu_k)^T \Sigma_k^{-2} = 0\end{aligned}$$

进而得到

$$\mu_k = \frac{\sum_{i=1}^N \gamma(i, k) x_i}{\sum_{i=1}^N \gamma(i, k)}$$

对第 $k$ 个高斯分布的概率 $\pi_k$ 求偏导，并令其等于0。需要注意的是， $\pi_k$ 需要满足两个条件

$$\begin{cases} \sum_{k=1}^K \pi_k = 1 \\ \pi_k \geq 0, \quad k = 1 \dots K \end{cases}$$

因此利用拉格朗日乘子法，引入拉格朗日乘子，并且去掉与 $\pi_k$ 无关的项，构造拉格朗日函数

$$L(\pi_k, \lambda) = \sum_{i=1}^N \sum_{k=1}^K \gamma(i, k) \log \pi_k + \lambda \left( \sum_{k=1}^K \pi_k - 1 \right)$$

构造这样的拉格朗日函数，可以保证求解出来的 $\pi_k$ 是非负的，因此不需要额外引入拉格朗日乘子/求拉格朗日函数对 $\pi_k$ 的偏导数

$$\frac{\partial L}{\partial u_k} = \sum_{i=1}^N \frac{\gamma(i, k)}{\pi_k} + \lambda = 0$$

求得 $\pi_k = -\frac{\sum_{i=1}^N \gamma(i, k)}{\lambda}$ 。又因为 $\sum_{k=1}^K \pi_k = 1$ ，代入上式，得到

$$\lambda = -\sum_{k=1}^K \sum_{i=1}^N \gamma(i, k)$$

因此

$$\pi_k = \frac{\sum_{i=1}^N \gamma(i, k)}{\sum_{k=1}^K \sum_{i=1}^N \gamma(i, k)} = \frac{N_k}{N}$$

其中分子 $N_k$ 表示第 $k$ 个高斯分布生成的样本点的个数，分母 $N$ 表示一共有多少个样本点。

对第 $k$ 个高斯分布的标准差 $\Sigma_k$ 求偏导，并令其等于0。这个过程推导比较复杂，但是最后可以得到

$$\Sigma_k^2 = \frac{\sum_{i=1}^N \gamma(i, k) (x_i - \mu_k) (x_i - \mu_k)^T}{\sum_{i=1}^N \gamma(i, k)} = \frac{1}{N_k} \sum_{i=1}^N \gamma(i, k) (x_i - \mu_k) (x_i - \mu_k)^T$$

至此完成了用EM算法求解GMM模型的参数估计的推导过程。

总结一下，在运行EM算法时，E step要求找到Q函数，即 $\gamma(i, k)$ 。M step计算 $\mu_k, \Sigma_k, \pi_k$ 。重复以上过程直到收敛。