

# EM 算法

1. 如果概率模型的变量都是观测变量，则给定数据之后，可以直接用极大似然估计法或者贝叶斯估计法来估计模型参数。

但是当模型含有隐变量时，就不能简单的使用这些估计方法。此时需要使用 EM 算法。

- EM 算法是一种迭代算法。
  - EM 算法专门用于含有隐变量的概率模型参数的极大似然估计，或者极大后验概率估计。
2. EM 算法的每次迭代由两步组成：

- E 步求期望。
- M 步求极大。

所以 EM 算法也称为期望极大算法。

## 一、示例

### 1.1 身高抽样问题

1. 假设学校所有学生中，男生身高服从正态分布  $\mathcal{N}(\mu_1, \sigma_1^2)$ ，女生身高服从正态分布  $\mathcal{N}(\mu_2, \sigma_2^2)$ 。

现在随机抽取200名学生，得到这些学生的身高  $\{x_1, x_2, \dots, x_n\}$ ，求参数  $\{\mu_1, \sigma_1^2, \mu_2, \sigma_2^2\}$  的估计。

2. 定义隐变量为  $z$ ，其取值为  $\{0, 1\}$ ，分别表示 男生、女生。

- 如果隐变量是已知的，即已知每个学生是男生还是女生  $\{z_1, z_2, \dots, z_n\}$ ，则问题很好解决：

- 统计所有男生的身高的均值和方差，得到  $\{\mu_1, \sigma_1^2\}$ ：

$$\mu_1 = \text{avg}\{x_i \mid z_i = 0\} \quad \sigma_1^2 = \text{var}\{x_i \mid z_i = 0\}$$

其中  $\{x_i \mid z_i = 0\}$  表示满足  $z_i = 0$  的  $x_i$  构成的集合。avg, var 分别表示平均值和方差。

- 统计所有女生的身高的均值和方差，得到  $\{\mu_2, \sigma_2^2\}$ ：

$$\mu_2 = \text{avg}\{x_i \mid z_i = 1\} \quad \sigma_2^2 = \text{var}\{x_i \mid z_i = 1\}$$

其中  $\{x_i \mid z_i = 1\}$  表示满足  $z_i = 1$  的  $x_i$  构成的集合。avg, var 分别表示平均值和方差。

- 如果已知参数  $\{\mu_1, \sigma_1^2, \mu_2, \sigma_2^2\}$ ，则任意给出一个学生的身高  $x$ ，可以知道该学生分别为男生/女生的概率。

$$p_1 = \frac{1}{\sqrt{2\pi} \times \sigma_1} \exp\left(-\frac{(x - \mu_1)^2}{2\sigma_1^2}\right)$$

$$p_2 = \frac{1}{\sqrt{2\pi} \times \sigma_2} \exp\left(-\frac{(x - \mu_2)^2}{2\sigma_2^2}\right)$$

则有： $p(z=0 \mid x) = \frac{p_1}{p_1+p_2}$ ,  $p(z=1 \mid x) = \frac{p_2}{p_1+p_2}$ 。因此也就知道该学生更可能为男生，还是更可能为女生。

因此：参数  $\{\mu_1, \sigma_1^2, \mu_2, \sigma_2^2\} \Leftrightarrow$  学生是男生/女生，这两个问题是相互依赖，相互纠缠的。

3. 为解决该问题，通常采取下面步骤：

- 先假定参数的初始值： $\{\mu_1^{<0>} , \sigma_1^{2<0>} , \mu_2^{<0>} , \sigma_2^{2<0>}\}$ 。
- 迭代： $i = 0, 1, \dots$ 
  - 根据  $\{\mu_1^{<i>} , \sigma_1^{2<i>} , \mu_2^{<i>} , \sigma_2^{2<i>}\}$  来计算每个学生更可能是属于男生，还是属于女生。  
这一步为 **E** 步 (**Expectation**)，用于计算隐变量的后验分布  $p(z | x)$ 。
  - 根据上一步的划分，统计所有男生的身高的均值和方差，得到  $\{\mu_1^{<i+1>} , \sigma_1^{2<i+1>}\}$ ；统计所有女生的身高的均值和方差，得到  $\{\mu_2^{<i+1>} , \sigma_2^{2<i+1>}\}$ 。  
这一步为 **M** 步 (**Maximization**)，用于通过最大似然函数求解正态分布的参数。
  - 当前后两次迭代的参数变化不大时，迭代终止。

## 1.2 三硬币模型

1. 已知三枚硬币 **A**, **B**, **C**，这些硬币正面出现的概率分别为  $\pi, p, q$ 。进行如下试验：

- 先投掷硬币 **A**，若是正面则选硬币 **B**；若是反面则选硬币 **C**。
- 然后投掷被选出来的硬币，投掷的结果如果是正面则记作 **1**；投掷的结果如果是反面则记作 **0**。
- 独立重复地  $N$  次试验，观测结果为：**1, 1, 0, 1, 0, ..., 0, 1**。

现在只能观测到投掷硬币的结果，无法观测投掷硬币的过程，求估计三硬币正面出现的概率。

2. 设：

- 随机变量  $Y$  是观测变量，表示一次试验观察到的结果，取值为 **1** 或者 **0**
- 随机变量  $Z$  是隐变量，表示未观测到的投掷 **A** 硬币的结果，取值为 **1** 或者 **0**
- $\theta = (\pi, p, q)$  是模型参数

则：

$$\begin{aligned} P(Y; \theta) &= \sum_Z P(Y, Z; \theta) = \sum_Z P(Z; \theta)P(Y | Z; \theta) \\ &= \pi p^Y (1-p)^{1-Y} + (1-\pi)q^Y (1-q)^{1-Y} \end{aligned}$$

注意：随机变量  $Y$  的数据可以观测，随机变量  $Z$  的数据不可观测

3. 将观测数据表示为  $\mathbb{Y} = \{y_1, y_2, \dots, y_N\}$ ，未观测数据表示为  $\mathbb{Z} = \{z_1, z_2, \dots, z_N\}$ 。

由于每次试验之间都是独立的，则有：

$$P(\mathbb{Y}; \theta) = \prod_{j=1}^N P(Y = y_j; \theta) = \prod_{j=1}^N [\pi p^{y_j} (1-p)^{1-y_j} + (1-\pi)q^{y_j} (1-q)^{1-y_j}]$$

4. 考虑求模型参数  $\theta = (\pi, p, q)$  的极大似然估计，即：

$$\hat{\theta} = \arg \max_{\theta} \log P(\mathbb{Y}; \theta)$$

这个问题没有解析解，只有通过迭代的方法求解，**EM** 算法就是可以用于求解该问题的一种迭代算法。

5. **EM** 算法求解：

首先选取参数的初值，记作  $\theta^{<0>} = (\pi^{<0>} , p^{<0>} , q^{<0>})$ ，然后通过下面的步骤迭代计算参数的估计值，直到收敛为止：

设第  $i$  次迭代参数的估计值为： $\theta^{<i>} = (\pi^{<i>} , p^{<i>} , q^{<i>})$ ，则 **EM** 算法的第  $i+1$  次迭代如下：

- E 步：计算模型在参数  $\theta^{<i>} = (\pi^{<i>}, p^{<i>}, q^{<i>})$  下，观测数据  $y_j$  来自于投掷硬币 B 的概率：

$$\mu_j^{<i+1>} = \frac{\pi^{<i>} (p^{<i>})^{y_j} (1 - p^{<i>})^{1-y_j}}{\pi^{<i>} (p^{<i>})^{y_j} (1 - p^{<i>})^{1-y_j} + (1 - \pi^{<i>}) (q^{<i>})^{y_j} (1 - q^{<i>})^{1-y_j}}$$

它其实就是  $P(Z = 1 | Y = y_j)$ ，即：已知观测变量  $Y = y_j$  的条件下，观测数据  $y_j$  来自于投掷硬币 B 的概率。

- M 步：计算模型参数的新估计值：

$$\begin{aligned}\pi^{<i+1>} &= \frac{1}{N} \sum_{j=1}^N \mu_j^{<i+1>} \\ p^{<i+1>} &= \frac{\sum_{j=1}^N \mu_j^{<i+1>} y_j}{\sum_{j=1}^N \mu_j^{<i+1>}} \\ q^{<i+1>} &= \frac{\sum_{j=1}^N (1 - \mu_j^{<i+1>}) y_j}{\sum_{j=1}^N (1 - \mu_j^{<i+1>})}\end{aligned}$$

- 第一个式子：通过后验概率  $P(Z | Y)$  估计值的均值作为先验概率  $\pi$  的估计。
- 第二个式子：通过条件概率  $P(Y | Z = 1)$  的估计来求解先验概率  $p$  的估计。
- 第三个式子：通过条件概率  $P(Y | Z = 0)$  的估计来求解先验概率  $q$  的估计。

#### 6. EM 算法的解释：

- 初始化：随机选择三枚硬币 A, B, C 正面出现的概率  $\pi, p, q$  的初始值  $\pi^{<0>}, p^{<0>}, q^{<0>}$ 。
- E 步：在已知概率  $\pi, p, q$  的情况下，求出每个观测数据  $y_j$  是来自于投掷硬币 B 的概率。即：  
 $p(z_j | y_j = 1)$ 。  
 于是对于  $N$  次实验，就知道哪些观测数据是由硬币 B 产生，哪些是由硬币 C 产生。
- M 步：在已知哪些观测数据是由硬币 B 产生，哪些是由硬币 C 产生的前提下：
  - $\pi$  就等于硬币 B 产生的次数的频率。
  - $p$  就等于硬币 B 产生的数据中，正面向上的频率。
  - $q$  就等于硬币 C 产生的数据中，正面向上的频率。

## 二、EM算法原理

### 2.1 观测变量和隐变量

- 令  $Y$  表示观测随机变量， $\mathbb{Y} = \{y_1, y_2, \dots, y_N\}$  表示对应的数据序列；令  $Z$  表示隐随机变量， $\mathbb{Z} = \{z_1, z_2, \dots, z_N\}$  表示对应的数据序列。  
 $\mathbb{Y}$  和  $\mathbb{Z}$  连在一起称作完全数据，观测数据  $\mathbb{Y}$  又称作不完全数据。
- 假设给定观测随机变量  $Y$ ，其概率分布为  $P(Y; \theta)$ ，其中  $\theta$  是需要估计的模型参数，则不完全数据  $\mathbb{Y}$  的似然函数是  $P(\mathbb{Y}; \theta)$ ，对数似然函数为  $L(\theta) = \log P(\mathbb{Y}; \theta)$ 。  
假定  $Y$  和  $Z$  的联合概率分布是  $P(Y, Z; \theta)$ ，完全数据的对数似然函数是  $\log P(\mathbb{Y}, \mathbb{Z}; \theta)$ ，则根据每次观测之间相互独立，有：

$$\begin{aligned}\log P(\mathbb{Y}; \theta) &= \sum_i \log P(Y = y_i; \theta) \\ \log P(\mathbb{Y}, \mathbb{Z}; \theta) &= \sum_i \log P(Y = y_i, Z = z_i; \theta)\end{aligned}$$

3. 由于  $\mathbb{Y}$  发生，根据最大似然估计，则需要求解对数似然函数：

$$\begin{aligned} L(\theta) &= \log P(\mathbb{Y}; \theta) = \sum_{i=1} \log P(Y = y_i; \theta) = \sum_{i=1} \log \sum_Z P(Y = y_i, Z; \theta) \\ &= \sum_{i=1} \log \left[ \sum_Z P(Y = y_i | Z; \theta) P(Z; \theta) \right] \end{aligned}$$

的极大值。其中  $\sum_Z P(Y = y_i, Z; \theta)$  表示对所有可能的  $Z$  求和，因为边缘分布  $P(Y) = \sum_Z P(Y, Z)$ 。  
该问题的困难在于：该目标函数包含了未观测数据的分布的积分和对数。

## 2.2 EM算法

### 2.2.1 原理

1. **EM** 算法通过迭代逐步近似极大化  $L(\theta)$ 。

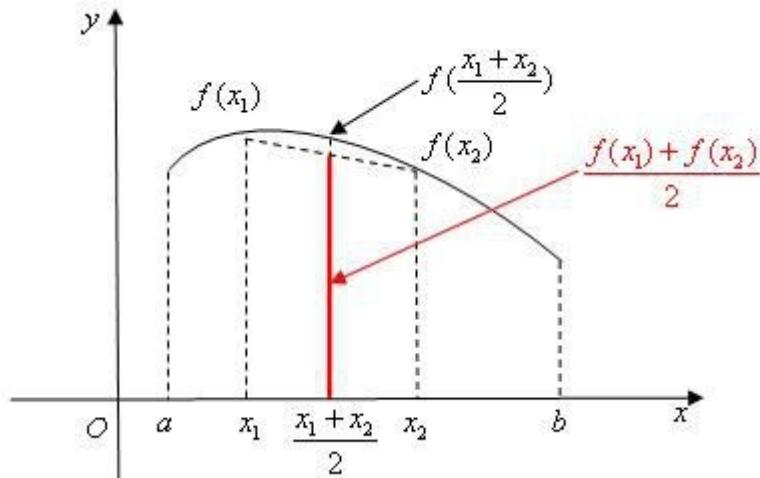
假设在第  $i$  次迭代后， $\theta$  的估计值为： $\theta^{<i>}$ 。则希望  $\theta$  新的估计值能够使得  $L(\theta)$  增加，即：  
 $L(\theta) > L(\theta^{<i>})$ 。

为此考虑两者的差： $L(\theta) - L(\theta^{<i>}) = \log P(\mathbb{Y}; \theta) - \log P(\mathbb{Y}; \theta^{<i>})$ 。

这里  $\theta^{<i>}$  已知，所以  $\log P(\mathbb{Y}; \theta^{<i>})$  可以直接计算得出。

2. **Jensen 不等式**：如果  $f$  是凸函数， $x$  为随机变量，则有： $\mathbb{E}[f(x)] \leq f(\mathbb{E}[x])$ 。

- 如果  $f$  是严格凸函数，当且仅当  $x$  是常量时，等号成立。



- 当  $\lambda_j$  满足  $\lambda_j \geq 0, \sum_j \lambda_j = 1$  时，将  $\lambda_j$  视作概率分布。

设随机变量  $y$  满足概率分布  $p(y = y_j) = \lambda_j$ ，则有： $\log \sum_j \lambda_j y_j \geq \sum_j \lambda_j \log y_j$ 。

3. 考虑到条件概率的性质，则有  $\sum_Z P(Z | Y; \theta) = 1$ 。因此有：

$$\begin{aligned}
L(\theta) - L(\theta^{<i>}) &= \sum_j \log \sum_Z P(Y = y_j, Z; \theta) - \sum_j \log P(Y = y_j; \theta^{<i>}) \\
&= \sum_j \left[ \log \sum_Z P(Z | Y = y_j; \theta^{<i>}) \frac{P(Y = y_j, Z; \theta)}{P(Z | Y = y_j; \theta^{<i>})} - \log P(Y = y_j; \theta^{<i>}) \right] \\
&\geq \sum_j \left[ \sum_Z P(Z | Y = y_j; \theta^{<i>}) \log \frac{P(Y = y_j, Z; \theta)}{P(Z | Y = y_j; \theta^{<i>})} - \log P(Y = y_j; \theta^{<i>}) \right] \\
&= \sum_j \left[ \sum_Z P(Z | Y = y_j; \theta^{<i>}) \log \frac{P(Y = y_j | Z; \theta) P(Z; \theta)}{P(Z | Y = y_j; \theta^{<i>})} - \log P(Y = y_j; \theta^{<i>}) \times 1 \right] \\
&= \sum_j \left[ \sum_Z P(Z | Y = y_j; \theta^{<i>}) \log \frac{P(Y = y_j | Z; \theta) P(Z; \theta)}{P(Z | Y = y_j; \theta^{<i>})} \right. \\
&\quad \left. - \log P(Y = y_j; \theta^{<i>}) \times \sum_Z P(Z | Y = y_j; \theta^{<i>}) \right] \\
&= \sum_j \left[ \sum_Z P(Z | Y = y_j; \theta^{<i>}) \log \frac{P(Y = y_j | Z; \theta) P(Z; \theta)}{P(Z | Y = y_j; \theta^{<i>}) P(Y = y_j; \theta^{<i>})} \right]
\end{aligned}$$

等号成立时，需要满足条件：

$$\begin{aligned}
P(Z | Y = y_j; \theta^{<i>}) &= \frac{1}{n_Z} \\
\frac{P(Y = y_j, Z; \theta)}{P(Z | Y = y_j; \theta^{<i>})} &= \text{const}
\end{aligned}$$

其中  $n_Z$  为随机变量  $Z$  的取值个数。

4. 令：

$$B(\theta, \theta^{<i>}) = L(\theta^{<i>}) + \sum_j \left[ \sum_Z P(Z | Y = y_j; \theta^{<i>}) \log \frac{P(Y = y_j | Z; \theta) P(Z; \theta)}{P(Z | Y = y_j; \theta^{<i>}) P(Y = y_j; \theta^{<i>})} \right]$$

则有： $L(\theta) \geq B(\theta, \theta^{<i>})$ ，因此  $B(\theta, \theta^{<i>})$  是  $L(\theta^{<i>})$  的一个下界。

- 根据定义有： $L(\theta^{<i>}) = B(\theta^{<i>}, \theta^{<i>})$ 。因为此时有：

$$\frac{P(Y = y_j | Z; \theta^{<i>}) P(Z; \theta^{<i>})}{P(Z | Y = y_j; \theta^{<i>}) P(Y = y_j; \theta^{<i>})} = \frac{P(Y = y_j, Z; \theta^{<i>})}{P(Y = y_j, Z; \theta^{<i>})} = 1$$

- 任何可以使得  $B(\theta, \theta^{<i>})$  增大的  $\theta$ ，也可以使  $L(\theta)$  增大。

为了使得  $L(\theta)$  尽可能增大，则选择使得  $B(\theta, \theta^{<i>})$  取极大值的  $\theta$ ： $\theta^{<i+1>} = \arg \max_{\theta} B(\theta, \theta^{<i>})$

5. 求极大值：

$$\begin{aligned}
& \theta^{<i+1>} = \arg \max_{\theta} B(\theta, \theta^{<i>}) \\
&= \arg \max_{\theta} \left[ L(\theta^{<i>}) + \sum_j \left( \sum_Z P(Z | Y = y_j; \theta^{<i>}) \log \frac{P(Y = y_j | Z; \theta) P(Z; \theta)}{P(Z | Y = y_j; \theta^{<i>}) P(Y = y_j; \theta^{<i>})} \right) \right] \\
&= \arg \max_{\theta} \sum_j \left( \sum_Z P(Z | Y = y_j; \theta^{<i>}) \log \frac{P(Y = y_j | Z; \theta) P(Z; \theta)}{P(Z | Y = y_j; \theta^{<i>}) P(Y = y_j; \theta^{<i>})} \right) \\
&= \arg \max_{\theta} \sum_j \left( \sum_Z P(Z | Y = y_j; \theta^{<i>}) \log P(Y = y_j | Z; \theta) P(Z; \theta) \right) \\
&= \arg \max_{\theta} \sum_j \left( \sum_Z P(Z | Y = y_j; \theta^{<i>}) \log P(Y = y_j, Z; \theta) \right)
\end{aligned}$$

其中： $L(\theta^{<i>})$ ,  $P(Z | Y = y_j; \theta^{<i>}) P(Y = y_j; \theta^{<i>})$  与  $\theta$  无关，因此省略。

## 2.2.2 算法

### 1. EM 算法：

◦ 输入：

- 观测变量数据  $\mathbb{Y} = \{y_1, y_2, \dots, y_N\}$
- 联合分布  $P(Y, Z; \theta)$ ，以及条件分布  $P(Z | Y; \theta)$

◦ 联合分布和条件分布的形式已知（比如说高斯分布等），但是参数未知（比如均值、方差）

◦ 输出：模型参数  $\theta$

◦ 算法步骤：

▪ 选择参数的初值  $\theta^{<0>}$ ，开始迭代。

▪ E 步：记  $\theta^{<i>}$  为第  $i$  次迭代参数  $\theta$  的估计值，在第  $i + 1$  步迭代的 E 步，计算：

$$\begin{aligned}
Q(\theta, \theta^{<i>}) &= \sum_{j=1}^N \mathbb{E}_{P(Z|Y=y_j; \theta^{<i>})} \log P(Y = y_j, Z; \theta) \\
&= \sum_{j=1}^N \left( \sum_Z P(Z | Y = y_j; \theta^{<i>}) \log P(Y = y_j, Z; \theta) \right)
\end{aligned}$$

其中  $\mathbb{E}_{P(Z|Y=y_j; \theta^{<i>})} \log P(Y = y_j, Z; \theta)$  表示：对于观测点  $Y = y_j$ ， $\log P(Y = y_j, Z; \theta)$  关于后验概率  $P(Z | Y = y_j; \theta^{<i>})$  的期望。

▪ M 步：求使得  $Q(\theta, \theta^{<i>})$  最大化的  $\theta$ ，确定  $i + 1$  次迭代的参数的估计值  $\theta^{<i+1>}$

$$\theta^{<i+1>} = \arg \max_{\theta} Q(\theta, \theta^{<i>})$$

▪ 重复上面两步，直到收敛。

2. 通常收敛的条件是：给定较小的正数  $\varepsilon_1, \varepsilon_2$ ，满足： $\|\theta^{<i+1>} - \theta^{<i>}\| < \varepsilon_1$  或者  $\|Q(\theta^{<i+1>}, \theta^{<i>}) - Q(\theta^{<i>}, \theta^{<i>})\| < \varepsilon_2$ 。

3.  $Q(\theta, \theta^{<i>})$  是算法的核心，称作  $Q$  函数。其中：

- 第一个符号表示要极大化的参数（未知量）。
- 第二个符号表示参数的当前估计值（已知量）。

4. EM 算法的直观理解：EM 算法的目标是最大化对数似然函数  $L(\theta) = \log P(\mathbb{Y})$ 。

- 直接求解这个目标是有问题的。因为要求解该目标，首先要得到未观测数据的分布  $P(Z | Y; \theta)$ 。如：身高抽样问题中，已知身高，需要知道该身高对应的是男生还是女生。  
但是未观测数据的分布就是待求目标参数  $\theta$  的解的函数。这是一个“鸡生蛋-蛋生鸡”的问题。
- EM 算法试图多次猜测这个未观测数据的分布  $P(Z | Y; \theta)$ 。  
每一轮迭代都猜测一个参数值  $\theta^{<i>}$ ，该参数值都对应着一个未观测数据的分布  $P(Z | Y; \theta^{<i>})$ 。如：已知身高分布的条件下，男生/女生的分布。
- 然后通过最大化某个变量来求解参数值。这个变量就是  $B(\theta, \theta^{<i>})$  变量，它是真实的似然函数的下界。
  - 如果猜测正确，则  $B$  就是真实的似然函数。
  - 如果猜测不正确，则  $B$  就是真实似然函数的一个下界。

5. 隐变量估计问题也可以通过梯度下降法等算法求解，但由于求和的项数随着隐变量的数目以指数级上升，因此代价太大。
- EM 算法可以视作一个非梯度优化算法。
  - 无论是梯度下降法，还是 EM 算法，都容易陷入局部极小值。

### 2.2.3 收敛性定理

1. 定理一：设  $P(Y; \theta)$  为观测数据的似然函数， $\theta^{<i>}$  为 EM 算法得到的参数估计序列， $P(Y; \theta^{<i>})$  为对应的似然函数序列，其中  $i = 1, 2, \dots$ 。  
则： $P(Y; \theta^{<i>})$  是单调递增的，即： $P(Y; \theta^{<i+1>}) \geq P(Y; \theta^{<i>})$ 。
2. 定理二：设  $L(\theta) = \log P(Y; \theta)$  为观测数据的对数似然函数， $\theta^{<i>}$  为 EM 算法得到的参数估计序列， $L(\theta^{<i>})$  为对应的对数似然函数序列，其中  $i = 1, 2, \dots$ 。
  - 如果  $P(Y; \theta)$  有上界，则  $L(\theta^{<i>})$  会收敛到某一个值  $L^*$ 。
  - 在函数  $Q(\theta, \theta^{<i>})$  与  $L(\theta)$  满足一定条件下，由 EM 算法得到的参数估计序列  $\theta^{<i>}$  的收敛值  $\theta^*$  是  $L(\theta)$  的稳定点。
    - 关于“满足一定条件”：大多数条件下其实都是满足的。
3. 定理二只能保证参数估计序列收敛到对数似然函数序列的稳定点  $L^*$ ，不能保证收敛到极大值点。
4. EM 算法的收敛性包含两重意义：
  - 关于对数似然函数序列  $L(\theta^{<i>})$  的收敛。
  - 关于参数估计序列  $\theta^{<i>}$  的收敛。

前者并不蕴含后者。
5. 实际应用中，EM 算法的参数的初值选择非常重要。
  - 参数的初始值可以任意选择，但是 EM 算法对初值是敏感的，选择不同的初始值可能得到不同的参数估计值。
  - 常用的办法是从几个不同的初值中进行迭代，然后对得到的各个估计值加以比较，从中选择最好的（对数似然函数最大的那个）。
6. EM 算法可以保证收敛到一个稳定点，不能保证得到全局最优点。其优点在于：简单性、普适性。

## 三、EM算法与高斯混合模型

### 3.1 高斯混合模型

1. 高斯混合模型(Gaussian mixture model, GMM)：指的是具有下列形式的概率分布模型：

$$P(y; \theta) = \sum_{k=1}^K \alpha_k \phi(y; \theta_k)$$

其中  $\alpha_k$  是系数，满足：

- $\alpha_k \geq 0, \sum_{k=1}^K \alpha_k = 1$ 。
- $\phi(y; \theta_k)$  是高斯分布密度函数，称作第  $k$  个分模型， $\theta_k = (\mu_k, \sigma_k^2)$ ：

$$\phi(y; \theta_k) = \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left(-\frac{(y - \mu_k)^2}{2\sigma_k^2}\right)$$

2. 如果用其他的概率分布密度函数代替上式中的高斯分布密度函数，则称为一般混合模型。

## 3.2 参数估计

1. 假设观察数据  $\mathbb{Y} = \{y_1, y_2, \dots, y_N\}$  由高斯混合模型  $P(y; \theta) = \sum_{k=1}^K \alpha_k \phi(y; \theta_k)$  生成，其中  $\theta = (\alpha_1, \alpha_2, \dots, \alpha_K; \theta_1, \theta_2, \dots, \theta_K)$ 。

可以通过 EM 算法估计高斯混合模型的参数  $\theta$ 。

2. 可以设想观察数据  $y_j$  是这样产生的：

- 首先以概率  $\alpha_k$  选择第  $k$  个分模型  $\phi(y; \theta_k)$ 。
- 然后以第  $k$  个分模型的概率分布  $\phi(y; \theta_k)$  生成观察数据  $y_j$ 。

这样，观察数据  $y_j$  是已知的，观测数据  $y_j$  来自哪个分模型是未知的。

对观察变量  $y$ ，定义隐变量  $z$ ，其中  $p(z = k) = \alpha_k$ 。

3. 完全数据的对数似然函数为：

$$P(y = y_j, z = k; \theta) = \alpha_k \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left(-\frac{(y_j - \mu_k)^2}{2\sigma_k^2}\right)$$

其对数为：

$$\log P(y = y_j, z = k; \theta) = \log \alpha_k - \log \sqrt{2\pi}\sigma_k - \frac{(y_j - \mu_k)^2}{2\sigma_k^2}$$

后验概率为：

$$P(z = k | y = y_j; \theta^{<i>}) = \frac{\alpha_k \frac{1}{\sqrt{2\pi}\sigma_k^{<i>}} \exp\left(-\frac{(y_j - \mu_k^{<i>})^2}{2\sigma_k^{<i>2}}\right)}{\sum_{t=1}^K \alpha_t \frac{1}{\sqrt{2\pi}\sigma_t^{<i>}} \exp\left(-\frac{(y_j - \mu_t^{<i>})^2}{2\sigma_t^{<i>2}}\right)}$$

$$\text{即} : P(z = k | y = y_j; \theta^{<i>}) = \frac{P(y = y_j, z = k; \theta^{<t>})}{\sum_{t=1}^K P(y = y_j, z = t; \theta)}.$$

则  $Q$  函数为：

$$\begin{aligned}
Q(\theta, \theta^{<i>}) &= \sum_{j=1}^N \left( \sum_z P(z \mid y = y_j; \theta^{<i>}) \log P(y = y_j, z; \theta) \right) \\
&= \sum_{j=1}^N \sum_{k=1}^K P(z = k \mid y = y_j; \theta^{<i>}) \left( \log \alpha_k - \log \sqrt{2\pi} \sigma_k - \frac{(y_j - \mu_k)^2}{2\sigma_k^2} \right)
\end{aligned}$$

求极大值： $\theta^{<i+1>} = \arg \max_{\theta} Q(\theta, \theta^{<i>})$ 。

根据偏导数为 0，以及  $\sum_{k=1}^K \alpha_k = 1$  得到：

- $\alpha_k$  :

$$\alpha_k^{<i+1>} = \frac{n_k}{N}$$

其中： $n_k = \sum_{j=1}^N P(z = k \mid y = y_j; \theta^{<i>})$ ，其物理意义为：所有的观测数据  $\mathbb{Y}$  中，产生自第  $k$  个分模型的观测数据的数量。

- $\mu_k$  :

$$\mu_k^{<i+1>} = \frac{\overline{Sum}_k}{n_k}$$

其中： $\overline{Sum}_k = \sum_{j=1}^N y_j P(z = k \mid y = y_j; \theta^{<i>})$ ，其物理意义为：所有的观测数据  $\mathbb{Y}$  中，产生自第  $k$  个分模型的观测数据的总和。

- $\sigma^2$  :

$$\sigma_k^{<i+1>2} = \frac{\overline{Var}_k}{n_k}$$

其中： $\overline{Var}_k = \sum_{j=1}^N (y_j - \mu_k^{<i>})^2 P(z = k \mid y = y_j; \theta^{<i>})$ ，其物理意义为：所有的观测数据  $\mathbb{Y}$  中，产生自第  $k$  个分模型的观测数据，偏离第  $k$  个模型的均值 ( $\mu_k^{<i>}$ ) 的平方和。

#### 4. 高斯混合模型参数估计的 EM 算法：

- 输入：

- 观察数据  $\mathbb{Y} = \{y_1, y_2, \dots, y_N\}$
- 高斯混合模型的分量数  $K$

- 输出：高斯混合模型参数  $\theta = (\alpha_1, \alpha_2, \dots, \alpha_K; \mu_1, \mu_2, \dots, \mu_K; \sigma_1^2, \sigma_2^2, \dots, \sigma_K^2)$

- 算法步骤：

- 随机初始化参数  $\theta^{<0>}$ 。
- 根据  $\theta^{<i>}$  迭代求解  $\theta^{<i+1>}$ ，停止条件为：对数似然函数值或者参数估计值收敛。

$$\alpha_k^{<i+1>} = \frac{n_k}{N}, \mu_k^{<i+1>} = \frac{\overline{Sum}_k}{n_k}, \sigma_k^{<i+1>2} = \frac{\overline{Var}_k}{n_k}$$

其中：

- $n_k = \sum_{j=1}^N P(z = k \mid y = y_j; \theta^{<i>})$ 。

其物理意义为：所有的观测数据  $\mathbb{Y}$  中，产生自第  $k$  个分模型的观测数据的数量。

- $\overline{Sum}_k = \sum_{j=1}^N y_j P(z = k \mid y = y_j; \theta^{<i>})$ 。

其物理意义为：所有的观测数据  $\mathbb{Y}$  中，产生自第  $k$  个分模型的观测数据的总和。

■  $\overline{Var}_k = \sum_{j=1}^N (y_j - \mu_k^{<i>})^2 P(z = k \mid y = y_i; \theta^{<i>})$ 。

其物理意义为：所有的观测数据  $\mathbb{Y}$  中，产生自第  $k$  个分模型的观测数据，偏离第  $k$  个模型的均值 ( $\mu_k^{<i>}$ ) 的平方和。

## 四、EM 算法与 kmeans 模型

1. **kmeans** 算法：给定样本集  $\mathbb{D} = \{\vec{x}_1, \vec{x}_2, \dots, \vec{x}_N\}$ ，针对聚类所得簇划分  $\mathcal{C} = \{\mathbb{C}_1, \mathbb{C}_2, \dots, \mathbb{C}_K\}$ ，最小化平方误差：

$$\min_{\mathcal{C}} \sum_{k=1}^K \sum_{\vec{x}_i \in \mathbb{C}_k} \|\vec{x}_i - \vec{\mu}_k\|_2^2$$

其中  $\vec{\mu}_k = \frac{1}{|\mathbb{C}_k|} \sum_{\vec{x}_i \in \mathbb{C}_k} \vec{x}_i$  是簇  $\mathbb{C}_k$  的均值向量。

2. 定义观测随机变量为  $\vec{x}$ ，观测数据为  $\mathbb{D}$ 。定义隐变量为  $z$ ，它表示  $\vec{x}$  所属的簇的编号。设参数  $\theta = (\vec{\mu}_1, \vec{\mu}_2, \dots, \vec{\mu}_K)$ ，则考虑如下的生成模型：

$$P(\vec{x}, z \mid \theta) \propto \begin{cases} \exp(-\|\vec{x} - \vec{\mu}_z\|_2^2) & \|\vec{x} - \vec{\mu}_z\|_2^2 = \min_{1 \leq k \leq K} \|\vec{x} - \vec{\mu}_k\|_2^2 \\ 0 & \|\vec{x} - \vec{\mu}_z\|_2^2 > \min_{1 \leq k \leq K} \|\vec{x} - \vec{\mu}_k\|_2^2 \end{cases}$$

其中  $\min_{1 \leq k \leq K} \|\vec{x} - \vec{\mu}_k\|_2^2$  表示距离  $\vec{x}$  最近的中心点所在的簇编号。即：

- 若  $\vec{x}$  最近的簇就是  $\vec{\mu}_z$  代表的簇，则生成概率为  $\exp(-\|\vec{x} - \vec{\mu}_z\|_2^2)$ 。
- 若  $\vec{x}$  最近的簇不是  $\vec{\mu}_z$  代表的簇，则生成概率等于 0。

3. 计算后验概率：

$$P(z \mid \vec{x}, \theta^{<i>}) \propto \begin{cases} 1 & \|\vec{x}_i - \vec{\mu}_z\|_2^2 = \min_{1 \leq k \leq K} \|\vec{x} - \vec{\mu}_k^{<i>}\|_2^2 \\ 0 & \|\vec{x}_i - \vec{\mu}_z\|_2^2 > \min_{1 \leq k \leq K} \|\vec{x} - \vec{\mu}_k^{<i>}\|_2^2 \end{cases}$$

即：

- 若  $\vec{x}$  最近的簇就是  $\vec{\mu}_z$  代表的簇，则后验概率为 1。
- 若  $\vec{x}$  最近的簇不是  $\vec{\mu}_z$  代表的簇，则后验概率为 0。

4. 计算  $Q$  函数：

$$Q(\theta, \theta^{<i>}) = \sum_{j=1}^N \left( \sum_z P(z \mid \vec{x} = \vec{x}_j; \theta^{<i>}) \log P(\vec{x} = \vec{x}_j, z; \theta) \right)$$

设距离  $\vec{x}_j$  最近的聚类中心为  $\vec{\mu}_{t_j}^{<i>}$ ，即它属于簇  $t_j$ ，则有：

$$Q(\theta, \theta^{<i>}) = \text{const} - \sum_{j=1}^N \|\vec{x}_j - \vec{\mu}_{t_j}\|_2^2$$

则有：

$$\theta^{<i+1>} = \arg \max_{\theta} Q(\theta, \theta^{<i>}) = \arg \min_{\theta} \sum_{j=1}^N \|\vec{x}_j - \vec{\mu}_{t_j}\|_2^2$$

定义集合  $\mathbb{I}_k = \{j \mid t_j = k\}$ ， $k = 1, 2, \dots, K$ ，它表示属于簇  $k$  的样本的下标集合。则有：

$$\sum_{j=1}^N \|\vec{x}_j - \vec{\mu}_{t_j}\|_2^2 = \sum_{k=1}^K \sum_{j \in \mathbb{I}_k} \|\vec{x}_j - \vec{\mu}_k\|_2^2$$

则有：

$$\theta^{<i+1>} = \arg \min_{\theta} \sum_{k=1}^K \sum_{j \in \mathbb{I}_k} \|\vec{x}_j - \vec{\mu}_k\|_2^2$$

这刚好就是 **k-means** 算法的目标：最小化平方误差。

5. 由于求和的每一项都是非负的，则当每一个内层求和  $\sum_{j \in \mathbb{I}_k} \|\vec{x}_j - \vec{\mu}_k\|_2^2$  都最小时，总和最小。即：

$$\vec{\mu}_k^{<i+1>} = \arg \min_{\vec{\mu}_k} \sum_{j \in \mathbb{I}_k} \|\vec{x}_j - \vec{\mu}_k\|_2^2$$

得到： $\vec{\mu}_k^{<i+1>} = \frac{1}{|\mathbb{I}_k|} \sum_{j \in \mathbb{I}_k} \vec{x}_j$ ，其中  $|\mathbb{I}_k|$  表示集合  $|\mathbb{I}_k|$  的大小。

这就是求平均值来更新簇中心。

## 五、EM 算法的推广

### 5.1 F 函数

1. **F** 函数：假设隐变量  $Z$  的概率分布为  $\tilde{P}(Z)$ ，定义分布  $\tilde{P}(Z)$  与参数  $\theta$  的函数  $F(\tilde{P}, \theta)$  为：

$$F(\tilde{P}, \theta) = \mathbb{E}_{\tilde{P}}[\log P(Y, Z; \theta)] + H(\tilde{P})$$

其中  $H(\tilde{P}) = -\mathbb{E}_{\tilde{P}} \log \tilde{P}$  是分布  $\tilde{P}(Z)$  的熵。

通常假定  $P(Y, Z; \theta)$  是  $\theta$  的连续函数，因此  $F(\tilde{P}, \theta)$  为  $\tilde{P}(Z)$  和  $\theta$  的连续函数。

2. 函数  $F(\tilde{P}, \theta)$  有下列重要性质：

- 对固定的  $\theta$ ，存在唯一的分布  $\tilde{P}_\theta(Z)$  使得极大化  $F(\tilde{P}, \theta)$ 。此时  $\tilde{P}_\theta(Z) = P(Z | Y; \theta)$ ，并且  $\tilde{P}_\theta$  随着  $\theta$  连续变化。

- 若  $\tilde{P}_\theta(Z) = P(Z | Y; \theta)$ ，则  $F(\tilde{P}, \theta) = \log P(Y; \theta)$ 。

3. 定理一：设  $L(\theta) = \log P(Y; \theta)$  为观测数据的对数似然函数， $\theta^{<i>}$  为 **EM** 算法得到的参数估计序列，函数  $F(\tilde{P}, \theta) = \sum_Y \mathbb{E}_{\tilde{P}}[\log P(Y, Z; \theta)] + H(\tilde{P})$ ，则：

- 如果  $F(\tilde{P}, \theta)$  在  $\tilde{P}^*(Z)$  和  $\theta^*$  有局部极大值，那么  $L(\theta)$  也在  $\theta^*$  有局部极大值。
- 如果  $F(\tilde{P}, \theta)$  在  $\tilde{P}^*(Z)$  和  $\theta^*$  有全局极大值，那么  $L(\theta)$  也在  $\theta^*$  有全局极大值。

4. 定理二：**EM** 算法的一次迭代可由 **F** 函数的极大-极大算法实现：设  $\theta^{<i>}$  为第  $i$  次迭代参数  $\theta$  的估计， $\tilde{P}^{<i>}$  为第  $i$  次迭代函数  $\tilde{P}(Z)$  的估计。在第  $i+1$  次迭代的两步为：

- 对固定的  $\theta^{<i>}$ ，求  $\tilde{P}^{<i+1>}$  使得  $F(\tilde{P}, \theta^{<i>})$  极大化。
- 对固定的  $\tilde{P}^{<i+1>}$ ，求  $\theta^{<i+1>}$  使得  $F(\tilde{P}^{<i+1>}, \theta)$  极大化。

### 5.2 GEM算法1

1. **GEM** 算法1（**EM** 算法的推广形式）：

- 输入：

- 观测数据  $\mathbb{Y} = \{y_1, y_2, \dots\}$
- **F** 函数

- 输出：模型参数

- 算法步骤：

- 初始化参数  $\theta^{<0>} , \text{开始迭代。}$
- 第  $i + 1$  次迭代 :
  - 记  $\theta^{<i>} \text{ 为参数 } \theta \text{ 的估计值} , \tilde{P}^{<i>} \text{ 为函数 } \tilde{P} \text{ 的估计值。求 } \tilde{P}^{<i+1>} \text{ 使得 } F(\tilde{P}, \theta^{<i>}) \text{ 极大化。}$
  - 求  $\theta^{<i+1>} \text{ 使得 } F(\tilde{P}^{<i+1>}, \theta) \text{ 极大化。}$
  - 重复上面两步直到收敛。

2. 该算法的问题是，有时候求  $F(\tilde{P}^{<i+1>}, \theta)$  极大化很困难。

## 5.3 GEM算法2

1. GEM 算法2 ( EM 算法的推广形式 ) :

- 输入 :
  - 观测数据  $\mathbb{Y} = \{y_1, y_2, \dots\}$
  - $Q$  函数
- 输出 : 模型参数
- 算法步骤 :
  - 初始参数  $\theta^{<0>} , \text{开始迭代。}$
  - 第  $i + 1$  次迭代 :
    - 记  $\theta^{<i>} \text{ 为参数 } \theta \text{ 的估计值} , \text{计算}$
    - $$Q(\theta, \theta^{<i>}) = \sum_{j=1}^N \left( \sum_Z P(Z | Y = y_j; \theta^{<i>}) \log P(Y = y_j, Z; \theta) \right)$$
    - 求  $\theta^{<i+1>} \text{ 使得 } Q(\theta^{<i+1>}, \theta^{<i>}) > Q(\theta^{<i>}, \theta^{<i>})$
    - 重复上面两步，直到收敛。

2. 此算法不需要求  $Q(\theta, \theta^{<i>})$  的极大值，只需要求解使它增加的  $\theta^{<i+1>}$  即可。

## 5.4 GEM算法3

1. GEM 算法3 ( EM 算法的推广形式 ) :

- 输入 :
  - 观测数据  $\mathbb{Y} = \{y_1, y_2, \dots\}$
  - $Q$  函数
- 输出 : 模型参数
- 算法步骤 :
  - 初始参数  $\theta^{<0>} = (\theta_1^{<0>}, \theta_2^{<0>}, \dots, \theta_d^{<0>}) , \text{开始迭代}$
  - 第  $i + 1$  次迭代 :
    - 记  $\theta^{<i>} = (\theta_1^{<i>}, \theta_2^{<i>}, \dots, \theta_d^{<i>})$  为参数  $\theta = (\theta_1, \theta_2, \dots, \theta_d)$  的估计值，计算
    - $$Q(\theta, \theta^{<i>}) = \sum_{j=1}^N \left( \sum_Z P(Z | Y = y_j; \theta^{<i>}) \log P(Y = y_j, Z; \theta) \right)$$
    - 进行  $d$  次条件极大化：

- 首先在  $\theta_2^{<i>} , \dots , \theta_d^{<i>}$  保持不变的条件下求使得  $Q(\theta, \theta^{<i>})$  达到极大的  $\theta_1^{<i+1>}$
- 然后在  $\theta_1 = \theta_1^{<i+1>} , \theta_j = \theta_j^{<i>} , j = 3, \dots, d$  的条件下求使得  $Q(\theta, \theta^{<i>})$  达到极大的  $\theta_2^{<i+1>}$
- 如此继续，经过  $d$  次条件极大化，得到  $\theta^{<i+1>} = (\theta_1^{<i+1>} , \theta_2^{<i+1>} , \dots , \theta_d^{<i+1>})$ ，使得  $Q(\theta^{<i+1>} , \theta^{<i>}) > Q(\theta^{<i>} , \theta^{<i>})$
- 重复上面两步，直到收敛。

2. 该算法将 EM 算法的 M 步分解为  $d$  次条件极大化，每次只需要改变参数向量的一个分量，其余分量不变。