Gaussian Mixture Model 02 Maximum Likelihood Estimation

Chen Gong

24 December 2019

本节我们想使用极大似然估计来求解 Gaussian Mixture Model (GMM) 的最优参数结果。首先，我们明确一下参数的意义：

X：Observed data，$X=(x_1,x_2,\cdots,x_N)$。

(X,Z)：Complete data，$(X,Z)=\{(x_1,z_1),(x_2,z_2),\cdots,(x_N,z_N)\}$。

θ：parameter，$\theta=\{P_1,\cdots,P_k,\mu_1,\cdots,\mu_k,\Sigma_1,\cdots,\Sigma_k\}$。

1 Maximum Likelihood Estimation 求解参数

$$
\begin{aligned}
P(x) &= \sum_Z P(X,Z) \\
&= \sum_{k=1}^K P(X,z=C_k) \\
&= \sum_{k=1}^K P(z=C_k)\cdot P(X|z=C_k) \\
&= \sum_{k=1}^K P_k\cdot \mathcal{N}(X|\mu_k,\Sigma_k)
\end{aligned} \tag{1}
$$

其中，P_k 也就是数据点去第 k 个高斯分布的概率。下面我们开始使用 MLE 来求解 θ：

$$
\begin{aligned}
\hat{\theta}_{MLE} &= \arg\max_\theta \log P(X) \\
&= \arg\max_\theta \log\prod_{i=1}^N P(x_i) \\
&= \arg\max_\theta \sum_{i=1}^N \log P(x_i) \\
&= \arg\max_\theta \sum_{i=1}^N \log\sum_{k=1}^K P_k\cdot\mathcal{N}(x_i|\mu_k,\Sigma_k)
\end{aligned} \tag{2}
$$

我们想要求的 θ 包括，$\theta=\{P_1,\cdots,P_k,\mu_1,\cdots,\mu_k,\Sigma_1,\cdots,\Sigma_k\}$。

2 MLE 的问题

按照之前的思路，我们就要分布对每个参数进行求偏导来计算最终的结果。但是问题马上就来了，大家有没有看到 log 函数里面是一个求和的形式，而不是一个求积的形式。这意味着计算非常的困难。甚至可以说，我们根本就求不出解析解。如果是单一的 Gaussian Distribution：

$$\log P(x_i) = \log \frac{1}{\sqrt{2\pi}\sigma} exp\left\{-\frac{(x_i-\mu)^2}{2\sigma}\right\} \tag{3}$$

根据 log 函数优秀的性质，这个问题是可以解的。但是，很不幸后面是一个求和的形式。所以，直接使用 MLE 求解 GMM，无法得到解析解。