

Gaussian Mixture Model 01 Model Introduction

Chen Gong

23 December 2019

这一章开始，我们将进入到 Gaussian Mixture Model (GMM) 的学习。而为什么要学习 GMM 呢？这是因为单峰分布已经不能准备的反映数据的分布了。正如下面的一个分布：

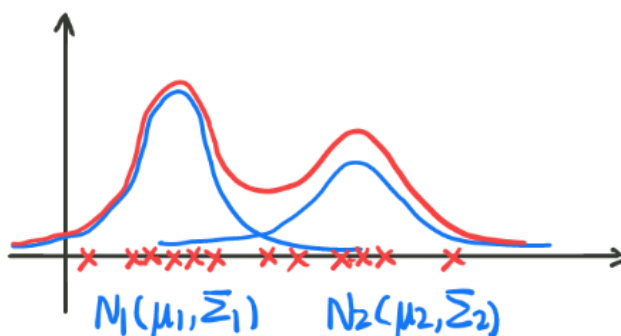


图 1: 数据分布举例

对于如上的数据分布来说，如果强行用单峰的 Gaussian Distribution 来表示这个分布，显然是可以的。但是，很明显是不合适的。会造成较大的误差，不能较好的表示整个数据的分布特征。

1 从几何角度来看

从几何角度来看比较的简单，也就是多个高斯分布来取加权平均值。也就是一个混合高斯分布就是多个高斯分布叠加而成的。那么，概率密度函数，可以被我们写成：

$$p(x) = \sum_{k=1}^K \alpha_k \mathcal{N}(\mu_k, \Sigma_k), \quad \sum_{k=1}^K \alpha_k = 1 \quad (1)$$

2 从混合模型角度来看 (生成模型)

如果当输入变量的维度高于一维的时候，我们就不能使用简单的加权来看了。因为，这时，我们已经无法简单的用加权平均来计算了，正如下图所示。其中， X 是 Observable Variable， Z 是 Latent Variable。这个 Z 是个什么意思呢？我们先举一个小例子。看到图 2 中那个打了红圈圈的数据点。它既属于 C_1 的分布，并且也属于 C_2 的分布，我们可以写作：

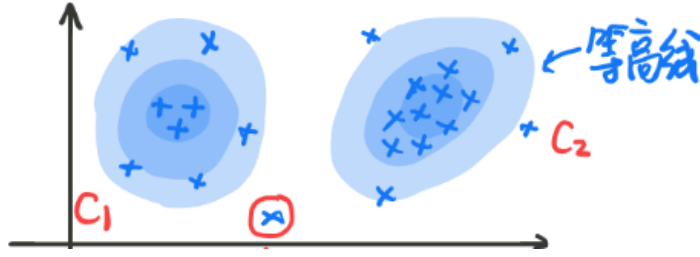


图 2: 二维数据分布举例

$$\begin{cases} X \sim C_1 \\ X \sim C_2 \end{cases} \quad (2)$$

这样写太麻烦了，我们可以直接写成 $X \sim Z$ ，这里的 Z 就是一个离散的随机变量，它包含了 C_1, C_2, \dots, C_N 的概率分布， Z 服从的是类别分布，其实就是看对应的样本 X 是属于哪一个高斯分布的概率。可以被我们写成：

| Z | C_1 | C_2 | \dots | C_k |
|--------|-------|-------|---------|-------|
| $P(Z)$ | P_1 | P_2 | \dots | P_k |

表 1: 隐变量 Z 的离散概率分布

并且， $\sum_k P_k = 1$ 。接下来，我们来说一说，如何来生成 N 个样本点， x_1, x_2, \dots, x_N 。

我们假设有一个骰子，有 K 个面，每个面都是不均匀的，假设我们可以控制每一个面的质量，那么这个骰子的面出现的概率会符合某个分布。有 K 个面，就有 K 个高斯分布。那么每次我们就投一下这个骰子，根据出现的面 K ，选择在第 K 个高斯分布中进行采样，生成一个样本点 x_i 。

概率图可以被我们描述为如下形式：

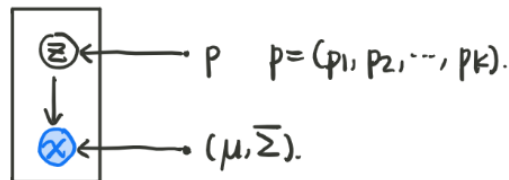


图 3: GMM 的概率图表达形式

我们根据一个离散的随机变量 Z 来选择是选取那个高斯分布，利用这个高斯分布 $\mathcal{N}(\mu, \Sigma)$ 来采样得到我们想要的样本点。而且，离散随机变量 Z 符合一个离散分布 $p = (p_1, p_2, \dots, p_k)$ 。

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, $X = (x_1, x_2, \dots, x_N)$ 。

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

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

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 ：

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

我们要求的 θ 包括， $\theta = \{P_1, \dots, P_k, \mu_1, \dots, \mu_k, \Sigma_1, \dots, \Sigma_k\}$ 。

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：

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

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。

Gaussian Mixture Model 03 Expectation Maximization

Chen Gong

25 December 2019

上一小节中，我们看到了使用极大似然估计的方法，我们根本就求不出最优参数 θ 的解析解。所以，我们使用迭代的方法来求近似解。

EM 算法的表达式，可以被我们写为：

$$\theta^{(t+1)} = \arg \max_{\theta} \underbrace{\mathbb{E}_{P(Z|X, \theta^{(t)})} [\log P(X, Z|\theta)]}_{Q(\theta, \theta^{(t)})} \quad (1)$$

经过一系列的迭代，我们可以得到 $\theta^0, \theta^1, \dots, \theta^{(t)}$ ，迭代到一定次数以后我们得到的 $\theta^{(N)}$ 就是我们想要得到的结果。EM 算法大体上可以分成两个部分，E-step 和 M-step，

1 E-Step

$$\begin{aligned} Q(\theta, \theta^{(t)}) &= \int_Z \log P(X, Z|\theta) \cdot P(Z|X, \theta^{(t)}) dZ \\ &= \sum_Z \log \prod_{i=1}^N P(x_i, z_i|\theta) \cdot \prod_{i=1}^N P(z_i|x_i, \theta^{(t)}) \\ &= \sum_{z_1, \dots, z_N} \sum_{i=1}^N \log P(x_i, z_i|\theta) \cdot \prod_{i=1}^N P(z_i|x_i, \theta^{(t)}) \\ &= \sum_{z_1, \dots, z_N} [\log P(x_1, z_1|\theta) + \log P(x_2, z_2|\theta) + \dots + \log P(x_N, z_N|\theta)] \cdot \prod_{i=1}^N P(z_i|x_i, \theta^{(t)}) \end{aligned} \quad (2)$$

为了简化推导，我们首先只取第一项来化简一下，

$$\begin{aligned} &\sum_{z_1, \dots, z_N} \log P(x_1, z_1|\theta) \cdot \prod_{i=1}^N P(z_i|x_i, \theta^{(t)}) dZ \\ &= \sum_{z_1, \dots, z_N} \log P(x_1, z_1|\theta) \cdot P(z_1|x_1, \theta^{(t)}) \cdot \prod_{i=2}^N P(z_i|x_i, \theta^{(t)}) \\ &= \sum_{z_1} \log P(x_1, z_1|\theta) \cdot P(z_1|x_1, \theta^{(t)}) \cdot \sum_{z_2, \dots, z_N} \prod_{i=2}^N P(z_i|x_i, \theta^{(t)}) \end{aligned} \quad (3)$$

而：

$$\begin{aligned}
\sum_{z_2, \dots, z_N} \prod_{i=2}^N P(z_i | x_i, \theta^{(t)}) &= \sum_{z_2, \dots, z_N} P(z_2 | x_2, \theta^{(t)}) \cdot P(z_3 | x_3, \theta^{(t)}) \cdots P(z_N | x_N, \theta^{(t)}) \\
&= \sum_{z_2} P(z_2 | x_2, \theta^{(t)}) \cdot \sum_{z_3} P(z_3 | x_3, \theta^{(t)}) \cdots \sum_{z_N} P(z_N | x_N, \theta^{(t)}) \\
&= 1 \cdot 1 \cdots 1 \\
&= 1
\end{aligned} \tag{4}$$

所以，式 (3) 也就等于：

$$\sum_{z_1, \dots, z_N} \log P(x_1, z_1 | \theta) \cdot \prod_{i=1}^N P(z_i | x_i, \theta^{(t)}) dZ = \sum_{z_1} \log P(x_1, z_1 | \theta) \cdot P(z_1 | x_1, \theta^{(t)}) \tag{5}$$

将式 (5) 中得到的结果，代入到式 (2) 中，我们就可以得到：

$$\begin{aligned}
Q(\theta, \theta^{(t)}) &= \sum_{z_1} \log P(x_1, z_1 | \theta) \cdot P(z_1 | x_1, \theta^{(t)}) + \cdots + \sum_{z_N} \log P(x_N, z_N | \theta) \cdot P(z_N | x_N, \theta^{(t)}) \\
&= \sum_{i=1}^N \sum_{Z_i} \log P(x_i, z_i | \theta) \cdot P(z_i | x_i, \theta^{(t)})
\end{aligned} \tag{6}$$

那么，下一步我们就是要找到， $P(x_i, z_i | \theta)$ 和 $P(z_i | x_i, \theta^{(t)})$ 的表达方式了。其中：

$$P(X, Z) = P(Z)P(X|Z) = P_Z \cdot \mathcal{N}(X | \mu_Z, \Sigma_Z) \tag{7}$$

$$P(Z|X) = \frac{P(X, Z)}{P(X)} = \frac{P_Z \cdot \mathcal{N}(X | \mu_Z, \Sigma_Z)}{\sum_{i=1}^K P_{Z_i} \cdot \mathcal{N}(X | \mu_{Z_i}, \Sigma_{Z_i})} \tag{8}$$

所以，我们将式 (8) 代入到式 (6) 中，就可以得到：

$$Q(\theta, \theta^{(t)}) = \sum_{i=1}^N \sum_{Z_i} \log P_{Z_i} \cdot \mathcal{N}(X | \mu_{Z_i}, \Sigma_{Z_i}) \cdot \frac{P_{Z_i}^{\theta^{(t)}} \cdot \mathcal{N}(x_i | \mu_{Z_i}^{\theta^{(t)}}, \Sigma_{Z_i}^{\theta^{(t)}})}{\sum_{k=1}^K P_k^{\theta^{(t)}} \cdot \mathcal{N}(x_i | \mu_k^{\theta^{(t)}}, \Sigma_k^{\theta^{(t)}})} \tag{9}$$

2 M-Step

根据我们在 E-Step 中的推导，我们可以得到：

$$\begin{aligned}
Q(\theta, \theta^{(t)}) &= \sum_{i=1}^N \sum_{Z_i} \log P_{Z_i} \cdot \mathcal{N}(X | \mu_{Z_i}, \Sigma_{Z_i}) \cdot \underbrace{\frac{P_{Z_i}^{\theta^{(t)}} \cdot \mathcal{N}(x_i | \mu_{Z_i}^{\theta^{(t)}}, \Sigma_{Z_i}^{\theta^{(t)}})}{\sum_{k=1}^K P_k^{\theta^{(t)}} \cdot \mathcal{N}(x_i | \mu_k^{\theta^{(t)}}, \Sigma_k^{\theta^{(t)}})}}_{P(Z_i | X_i, \theta^{(t)})} \\
&= \sum_{i=1}^N \sum_{k=1}^K \log (P_k \cdot \mathcal{N}(X | \mu_k, \Sigma_k)) \cdot P(Z_i = C_k | X_i, \theta^{(t)}) \\
&= \sum_{k=1}^K \sum_{i=1}^N \log (P_k \cdot \mathcal{N}(X | \mu_k, \Sigma_k)) \cdot P(Z_i = C_k | X_i, \theta^{(t)}) \\
&= \sum_{k=1}^K \sum_{i=1}^N (\log P_k + \log \mathcal{N}(X_i | \mu_k, \Sigma_k)) \cdot P(Z_i = C_k | X_i, \theta^{(t)})
\end{aligned} \tag{10}$$

我们的目的也就是进行不断迭代，从而得出最终的解，用公式表达也就是：

$$\theta^{(t+1)} = \arg \max_{\theta} Q(\theta, \theta^{(t)}) \quad (11)$$

我们需要求解的参数也就是， $\theta^{(t+1)} = \{P_1^{(t+1)}, \dots, P_k^{(t+1)}, \mu_1^{(t+1)}, \dots, \mu_k^{(t+1)}, \Sigma_1^{(t+1)}, \dots, \Sigma_k^{(t+1)}\}$ 。

首先，我们来展示一下怎么求解 $P_K^{(t+1)}$ ：

由于在等式 (10), $\sum_{k=1}^K \sum_{i=1}^N (\log P_k + \log \mathcal{N}(X|\mu_k, \Sigma_k)) \cdot P(Z_i = C_k|X_i, \theta^{(t)})$ 中的 $\log \mathcal{N}(X|\mu_k, \Sigma_k)$ 部分和 P_k 并没有什么关系。所以，可以被我们直接忽略掉。所以，求解问题，可以被我们描述为：

$$\begin{cases} \arg \max_{P_k} \sum_{k=1}^K \sum_{i=1}^N \log P_k \cdot P(Z_i = C_k|X_i, \theta^{(t)}) \\ s.t. \quad \sum_{k=1}^K P_k = 1 \end{cases} \quad (12)$$

使用拉格朗日算子法，我们可以写成：

$$\mathcal{L}(P, \lambda) = \sum_{k=1}^K \sum_{i=1}^N \log P_k \cdot P(Z_i = C_k|X_i, \theta^{(t)}) + \lambda \left(\sum_{k=1}^K P_k - 1 \right) \quad (13)$$

$$\begin{aligned} \frac{\partial \mathcal{L}(P, \lambda)}{\partial P_k} &= \sum_{i=1}^N \frac{1}{P_k} \cdot P(Z_i = C_k|X_i, \theta^{(t)}) + \lambda = 0 \\ &\Rightarrow \sum_{i=1}^N P(Z_i = C_k|X_i, \theta^{(t)}) + P_k \lambda = 0 \\ &\xrightarrow{k=1, \dots, K} \sum_{i=1}^N \underbrace{\sum_{k=1}^K P(Z_i = C_k|X_i, \theta^{(t)})}_1 + \underbrace{\sum_{k=1}^K P_k \lambda}_1 = 0 \\ &\Rightarrow N + \lambda = 0 \end{aligned} \quad (14)$$

所以，我们可以轻易的得到 $\lambda = -N$ ，所以有

$$P_K^{(t+1)} = \frac{1}{N} \sum_{i=1}^N P(Z_i = C_k|X_i, \theta^{(t)}) \quad (15)$$

那么，我们所有想要求的参数也就是 $P^{(t+1)} = (P_1^{(t+1)}, P_2^{(t+1)}, \dots, P_k^{(t+1)})$ 。

求解 $P_k^{(t+1)}$ 是一个有约束的求最大值问题，由于带约束所以我们要使用拉格朗日乘子法。而且这里使用到了一个 track，也就是将从 1 到 k，所有的数据集做一个整合，非常的精彩，这样就直接消掉了 P_k 无法计算的问题。而至于 θ 的其他部分，也就是关于 $\{\mu_1^{(t+1)}, \dots, \mu_k^{(t+1)}, \Sigma_1^{(t+1)}, \dots, \Sigma_k^{(t+1)}\}$ 的计算，使用的方法也是一样的，这个问题就留给各位了。

为什么极大似然估计搞不定的问题，放在 EM 算法里面我们就可以搞定了呢？我们来对比一下两个方法中，需要计算极值的公式。

$$\sum_{k=1}^K \sum_{i=1}^N (\log P_k + \log \mathcal{N}(X_i|\mu_k, \Sigma_k)) \cdot P(Z_i = C_k|X_i, \theta^{(t)}) \quad (16)$$

$$\arg \max_{\theta} \sum_{i=1}^N \log \sum_{k=1}^K P_k \cdot \mathcal{N}(x_i|\mu_k, \Sigma_k) \quad (17)$$

极大似然估计一开始计算的就是 $P(X)$ ，而 EM 算法中并没有出现有关 $P(X)$ 的计算，而是全程计算都是 $P(X, Z)$ 。而 $P(X)$ 实际上就是 $P(X, Z)$ 的求和形式。所以，每次单独的考虑 $P(X, Z)$ 就避免了在 \log 函数中出现求和操作。