高斯混合模型

Li Liang*

1 高斯混合模型

高斯混合模型（Gaussian Mixed Model）指的是多个高斯分布函数的线性组合，理论上 GMM 可以拟合出任意类型的分布，通常用于解决同一集合下的数据包含多个分布的情况，如图 1。

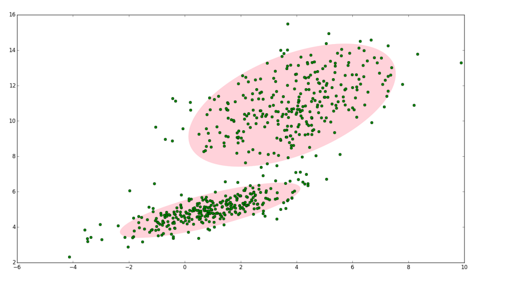

图 1: 两个二维高斯分布的数据集

对数据集 $D = \{x_1, x_2, ..., x_m\}$，高斯混合模型表示如下：

$$p(x) = \sum_{i=1}^{k} \alpha_i \, \mathcal{N}(x \mid \mu_i, \Sigma_i) \tag{1}$$

该分布由 k 个分量组成，其中 α_i 为混合系数，是样本 x 符合第 k 个分量分布的先验概率。第 k 个分布可采用多维高斯分布描述：

$$\mathcal{N}(x \mid \mu_i, \Sigma_i) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)} \tag{2}$$

2 高斯混合聚类

假设数据服从混合高斯分布，计算样本 x 符合第 i 个分量分布的后验概率，将其划分为后验概率最大的分量分布。怎么计算后验概率呢？

*https://github.com/leeliang/

引入一个随机变量 $z_j \in 1, 2, 3, ..., k$，$z_j = i$ 表示样本 x_j 符合第 i 个分量分布，则先验分布 $p(x_j = i) = \alpha_i$，根据贝叶斯定理，样本 x_j 符合第 i 个分量分布的后验概率为：

$$\begin{aligned} p(z_j = i \mid x_j) &= \frac{p(x_j = i)\ p(x_j \mid z_j = i)}{p(x_j)} \\ &= \frac{\alpha_i\ \mathcal{N}(x \mid \mu_i, \Sigma_i)}{\sum\limits_{l=1}^{k} \mathcal{N}(x \mid \mu_l, \Sigma_l)} \end{aligned} \tag{3}$$

后验概率的计算依赖于参数 (α, μ, Σ)，需要先计算这写参数才能计算后验概率，从而根据后验概率分类。

因为我们假设数据服从混合高斯分布，最大化混合高斯分布就可以求取参数。

$$\begin{aligned} \ell(\alpha, \mu, \Sigma) &= \log \prod_{j=1}^{m} p(x_j) \\ &= \sum_{j=1}^{m} \log[\sum_{i=1}^{k} \alpha_i\ \mathcal{N}(x_j \mid \mu_i, \Sigma_i)] \end{aligned} \tag{4}$$

若是上式最大化，则有：

$$\begin{aligned} \frac{\partial \ell(\alpha, \mu, \Sigma)}{\mu_i} &= 0 \\ \frac{\partial \ell(\alpha, \mu, \Sigma)}{\Sigma_i} &= 0 \\ \frac{\partial \ell(\alpha, \mu, \Sigma) + \lambda(\sum \alpha_i - 1)}{\alpha_i} &= 0 \quad (\text{拉格朗日乘子法}) \end{aligned} \tag{5}$$

根据上式，记后验概率 $\gamma_j(i) = p(z_j = i \mid x_j)$，有：

$$\begin{aligned} \mu_i &= \frac{\sum\limits_{j=1}^{m} \gamma_j(i) x_j}{\sum\limits_{j=1}^{m} \gamma_j(i)} \\ \Sigma_i &= \frac{\sum\limits_{j=1}^{m} \gamma_j(i)(x_j - \mu_i)(x_j - \mu_i)^T}{\sum\limits_{j=1}^{m} \gamma_j(i)} \\ \alpha_i &= \frac{1}{m} \sum_{j=1}^{k} \gamma_j(i) \end{aligned} \tag{6}$$

可采用 EM 算法计算上述参数，初始化参数，根据当前参数计算 $\gamma_j(i)$ (E-Step); 根据上式更新参数 (M-Step，即最大化似然函数)。