聚类——认识GMM算法

作者：凯鲁嘎吉 - 博客园 http://www.cnblogs.com/kailugaji/

一、GMM概述

定义 9.2（高斯混合模型） 高斯混合模型是指具有如下形式的概率分布模型：

$$P(y|\theta)=\sum_{k=1}^{K}\alpha_k\phi(y|\theta_k) \tag{9.24}$$

其中，α_k 是系数，$\alpha_k\geqslant 0$，$\sum_{k=1}^{K}\alpha_k=1$；$\phi(y|\theta_k)$ 是高斯分布密度，$\theta_k=(\mu_k,\sigma_k^2)$，

$$\phi(y|\theta_k)=\frac{1}{\sqrt{2\pi}\sigma_k}\exp\left(-\frac{(y-\mu_k)^2}{2\sigma_k^2}\right) \tag{9.25}$$

称为第 k 个分模型.

二、GMM算法步骤

算法 9.2（高斯混合模型参数估计的 EM 算法）

输入：观测数据 $y_1, y_2, \cdots, y_N$，高斯混合模型；

输出：高斯混合模型参数.

（1）取参数的初始值开始迭代

（2）E 步：依据当前模型参数，计算分模型 k 对观测数据 y_j 的响应度

$$\hat{\gamma}_{jk} = \frac{\alpha_k \phi(y_j \mid \theta_k)}{\sum_{k=1}^{K} \alpha_k \phi(y_j \mid \theta_k)}, \qquad j = 1, 2, \cdots, N; \quad k = 1, 2, \cdots, K$$

（3）M 步：计算新一轮迭代的模型参数

$$\hat{\mu}_k = \frac{\sum_{j=1}^{N} \hat{\gamma}_{jk} y_j}{\sum_{j=1}^{N} \hat{\gamma}_{jk}}, \qquad k = 1, 2, \cdots, K$$

$$\hat{\sigma}_k^2 = \frac{\sum_{j=1}^{N} \hat{\gamma}_{jk} (y_j - \mu_k)^2}{\sum_{j=1}^{N} \hat{\gamma}_{jk}}, \qquad k = 1, 2, \cdots, K$$

$$\hat{\alpha}_k = \frac{\sum_{j=1}^{N} \hat{\gamma}_{jk}}{N}, \qquad k = 1, 2, \cdots, K$$

（4）重复第（2）步和第（3）步，直到收敛.

三、具体推导参考文献

1. 李航. 统计学习方法[M]. 清华大学出版社, 2012.

2. Bishop C M. Pattern Recognition and Machine Learning (Information Science and Statistics)[M]. Springer-Verlag New York, Inc. 2006.

注：GMM数学公式推导用到了贝叶斯公式、条件期望公式、拉格朗日乘数法、极大似然估计、参数估计。概率论与数理统计的内容居多，事先应掌握概率论与数理统计基本内容。

四、总结

1. GMM算法中间参数估计部分用到了EM算法，EM算法分为两步：

（1）E步：求目标函数期望，更多的是求目标函数取对数之后的期望值。

（2）M步：使期望最大化。用到极大似然估计，拉格朗日乘数法，对参数求偏导，最终确定新的参数。

2.K-means，FCM与GMM算法参数估计的数学推导思路大体一致，都先确立目标函数，然后使目标函数最大化的参数取值就是迭代公式。

3.三个算法都需要事先指定k。K-means与FCM中的k指的是要聚的类的个数，GMM算法中的k指的是k个单高斯混合模型。

4.三个算法流程一致：

（1）通过一定的方法初始化参数（eg:随机，均值……）

（2）确立目标函数

（3）通过一定的方法使目标函数最大化，更新参数迭代公式（eg:EM，粒子群……）

（4）设置一定的终止条件，使算法终止。若不满足条件，转向（3）

补充：GMM的MATLAB代码：https://github.com/kailugaji/Gaussian_Mixture_Model_for_Clustering

Clustering - Mixture of Gaussians