# 高斯混合聚类模型

杨朝辉　PB17071433

2020-05-24

## 算法简介

### 高斯混合模型

高斯混合模型可以看作 KMeans 算法的拓展，其核心是采用带有隐变量的概率模型来表达聚类原型，混合模型$P(x|\theta)=\sum_{k=1}^{K}\alpha_k\phi(x|\theta_k)$，$\phi(x|\theta_k)$为随机向量的多元高斯函数，$\alpha_k$为归一化概率系数。即对于特性样本集将其看作是由该概率模型所生成的，算法目标便是希望在极大化似然函数$L=\prod_{n=1}^{N}P(x_n|\theta)$过程中寻找该模型的隐变量。相比 KMeans 方法硬性圈定地对样本聚类，GMM 更复合实际的数据分布，尤其样本量较大时，聚类效果的提升由中心极限定理保证。

### Expectation-Maximum 算法

对于含有隐变量的生成模型常使用 EM 算法迭代计算，E 步求 Q 函数，M 步求使得 Q 函数最大化的参数。EM 算法应用到的 GMM 中简述如下：（更详细的说明可参照相关教材）

1） 初始化隐变量参数$\mu_k$，$\Sigma_k$，$\alpha_k$，$k=1,2,\cdots,K$

2） E 步——计算后验概率（响应度）：

$$\gamma_{jk}=\frac{\alpha_k\phi(x_j|\theta_k)}{\sum_k\alpha_k\phi(x_j|\theta_k)},\ j=1,2,\cdots,N;\ k=1,2,\cdots,K$$

3） M 步——重新计算模型参数：

$$\mu_k = \frac{\sum_j \gamma_{jk} x_j}{\sum_j \gamma_{jk}},\ k = 1, 2, \cdots, K$$

$$\Sigma_k = \frac{\sum_j \gamma_{jk} (x_j - \mu_k)(x_j - \mu_k)^T}{\sum_j \gamma_{jk}},\ k = 1, 2, \cdots, K$$

$$\alpha_k = \frac{\sum_j \gamma_{jk}}{N},\ k = 1, 2, \cdots, K$$

4） 重复 EM 步直至收敛或迭代至一定次数

实际上，EM 算法的迭代是为了极大化目标函数的下界以做到与极大似然估计类似的效果，但它所学习的是局部最优解，不能保证全局最优，对参数初值的选取较为敏感。

## 实验条件

采用 python 语言及其相关的 numpy，matplotlib，sklearn 等支持，方便数据的预处理、数值计算及数据分布的可视化。构造 GMMCluster 聚类器对象，对鸢尾花和表情数据聚类分析，并于内置的 KMeans 聚类方法的结果作比较。

### 数据简介

采用鸢尾花数据，共计 150 个样本，包含三类品种鸢尾花的花瓣长度、花瓣宽带、花萼长度、花萼宽度四个属性列。

### 评估参数

内部指标

内部指标采用类内距离与类间距离度量，为方便比较，本实验采用距离平方和（sum of square）：

$$withinss = \sum_{k} \sum_{x \in C_k} \|x - \mu_k\|^2$$

$$totalss = \sum_{x} \|x - \mu\|^2$$

$$betweenss = totalss - withinss$$

对于任意样本集，总是希望聚类结果 withinss 尽可能大、betweenss 尽可能大。

外部指标

外部指标采取与某一外部模型（本实验中采用数据标签）的聚类结果向比较得出的 Rand Index 统计量作为度量指标，其取值在 0 和 1 之间，越大越好：

$$a = |SS|,\ SS = \{(\boldsymbol{x_i}, \boldsymbol{x_j}) | \lambda_i = \lambda_j, \lambda_i^* = \lambda_j^*, i < j\}$$

$$d = |DD|,\ DD = \{(\boldsymbol{x_i}, \boldsymbol{x_j}) | \lambda_i \neq \lambda_j, \lambda_i^* \neq \lambda_j^*, i < j\}$$

$$RI = \frac{2(a+d)}{N(N-1)}$$

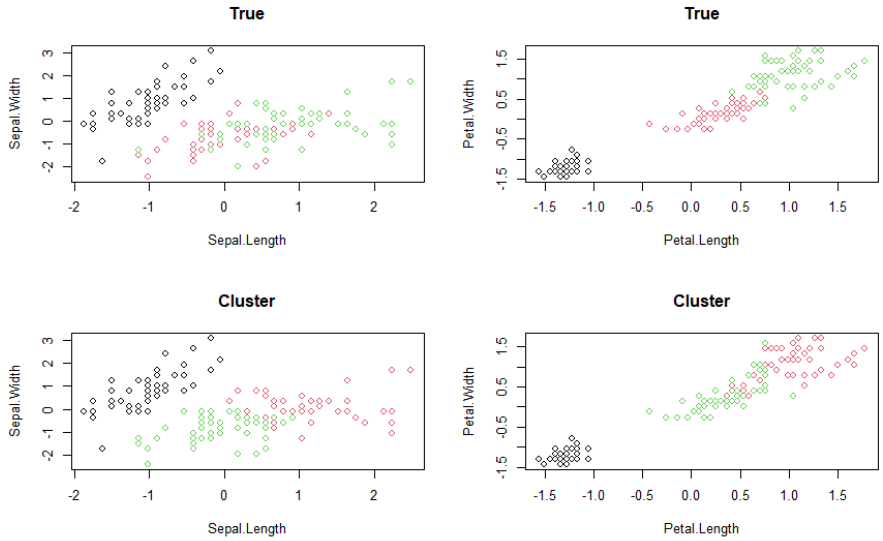
其中 N 为样本数量，$\lambda$ 为某一样本聚类所属类别标号。

# 结果分析

**聚类效果**

为便于聚类可视化，首先在维度较低的鸢尾花数据集（四个属性类）上用所实现的聚类器进行聚类分析，并与内置的 KMeans 算法作比较，结果如下：

内置 KMeans 算法聚类效果

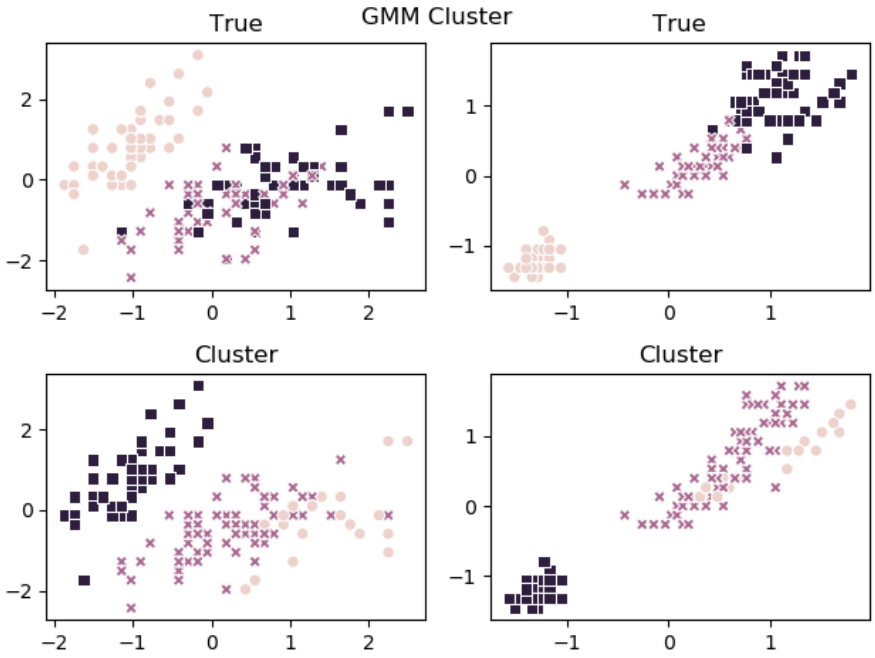



类内总距离平方和平方：196

类间总距离平方和：400

外部指标 RI：0.663

自实现 GMM 聚类效果（横纵坐标所代表数据同上）




类内总距离平方和平方：197

类间总距离平方和：402.5

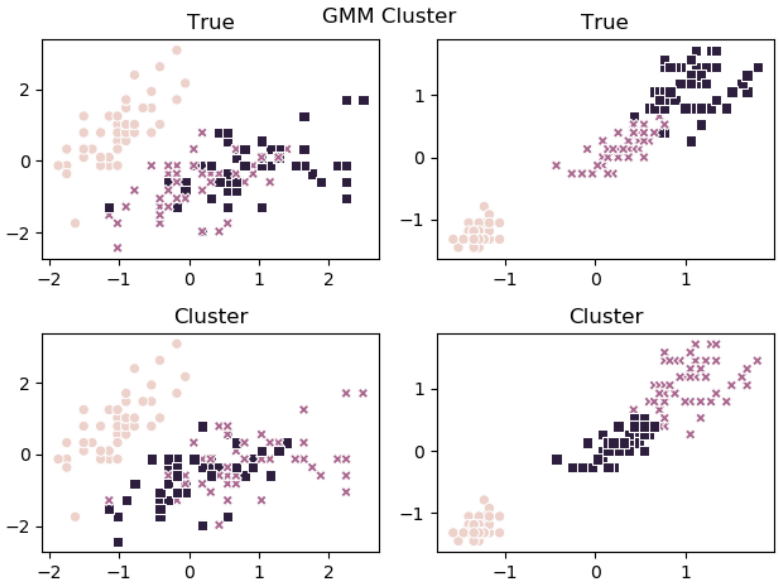
外部指标 RI：0.8737

从以上指标可见该模型在鸢尾花数据集上的聚类效果优于内置的 KMeans 聚类方法。

**模型限制与提升**

（1） 高斯混合分布假设了特征独立，这是无法保证的。比如上面鸢尾花数据中鸢尾花的花萼、花瓣的长度、宽度或许能够视作近似独立的，但对于较高维的数据如上面有24个特征维度的人脸表情数据，属性间的独立性假设是难以接受的。当然这可以进一步使用主成分分析或其他降维方式加以改善。

（2） 高斯模型的期望初值选择非常重要，不同初值结果都不一样，甚至初始不合适时不收敛，本实验中随机选取样本初值是一种较为普遍并且实际表现还可以的方式。更进一步，一般可以把kmeans的结果的聚类中心作为模型均值向量初始点，或者选取多组初值观察。如下是选取内置kmeans聚类结果的中心作为初值，结果确实有一定的提升。




类内总距离平方和平方：169（较之之前的197）

类间总距离平方和：431（较之之前的402）

RI指标：0.9575（较之之前的0.87）