聚类

2018211568号 2018211316班 杜明欣

任务定义

> 对数据集进行聚类任务，采用kmeans和GMM相结合方法

输入输出

> 输入：cluster.dat聚类数据集

方法描述

采用 Kmeans 聚类方法进行预训练

- 观察不同K类别数对聚类结果的影响，选择合适的类别数区间
- 将Kmeans聚类后的中心点作为GMM模型初始各混合成份的均值

使用GMM模型训练

在合适类别数区间，采用EM算法进行多次实验，其中：

$$E-step：估计后验概率$$

$$Z_{ik}=\frac{\pi_k N(x_i|\mu_k,\sigma_k)}{\sum_{j=1}^{K}\pi_j N(x_i|\mu_j,\sigma_j)}$$

$$M-step：更新参数$$

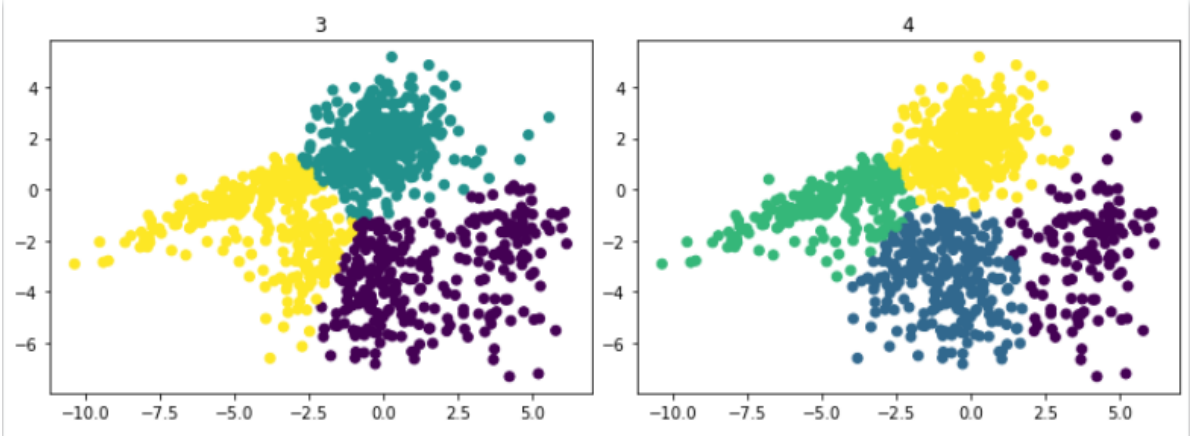
$$\mu_k=\frac{1}{N_k}\sum_{i=1}^{N}Z_{ik}x_i$$

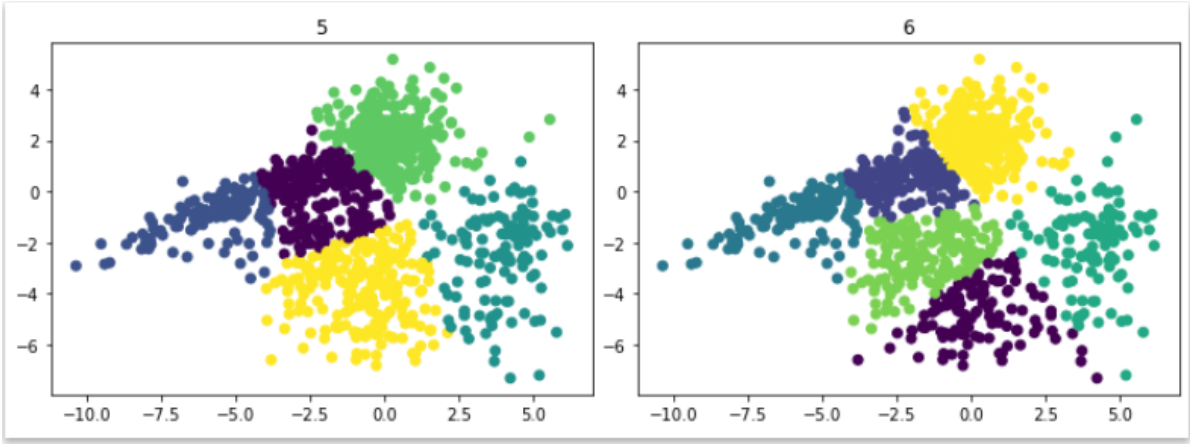
$$\sigma_k=\frac{1}{N_k}\sum_{i=1}^{N}Z_{ik}(x_i-\mu_k)(x_i-\mu_k)^T$$

结果分析

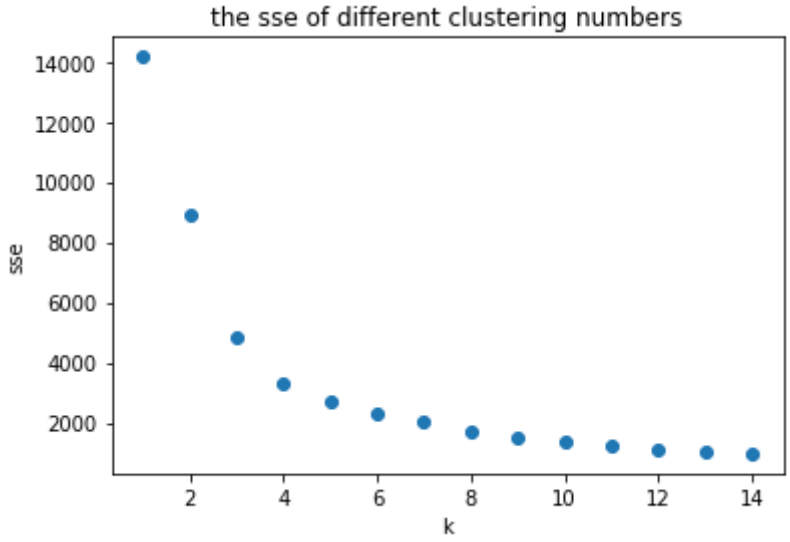
Kmeans 预训练结果

不同k值对应聚类效果图





SSE平方误差和作为聚类任务内部评价指标

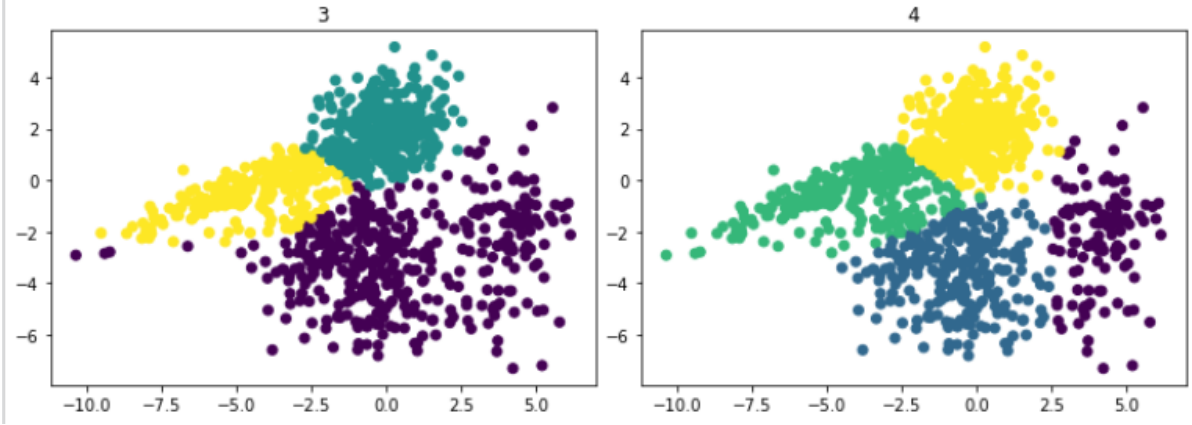


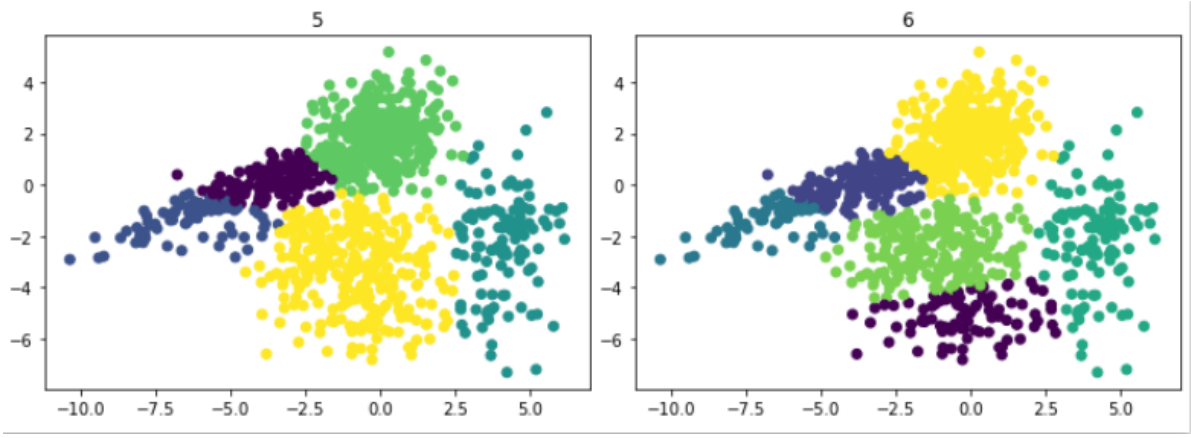
> 如图所示k={1，2，3}时SSE下降很大，k>8后SSE下降缓慢，同时分类数过多会出现过拟合情况，将一些本应属于1簇的样本划分为更小簇。故选择k={3，4，5，6，7}为合适类别数区间。

GMM模型训练

训练结果

> 使用**kmeans**聚类后的**中心点**作为GMM模型初始u值，**加快收敛速率，使聚类更为高效**
>
> 以下是不同k值对应GMM训练结果

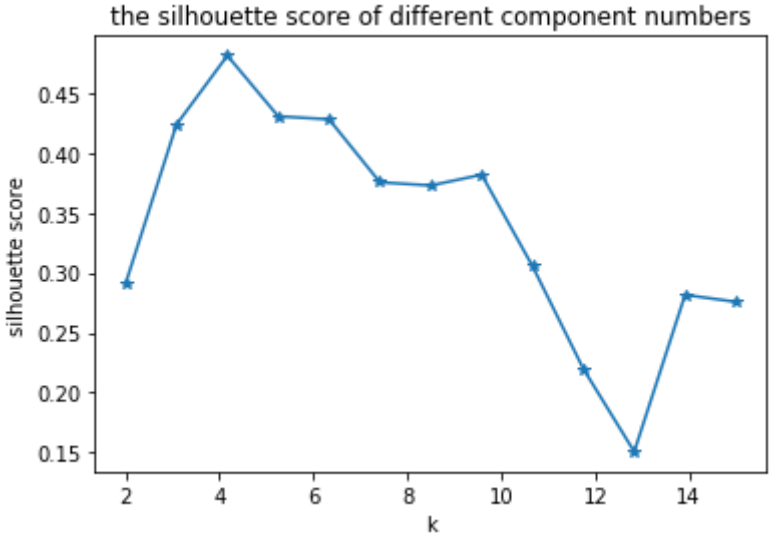




评价指标

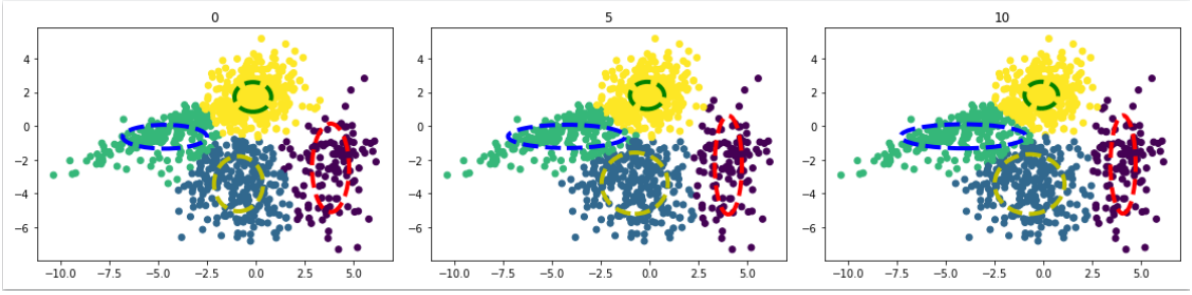
> **Silhouette Coefficient(平均轮廓系数):**
>
> 兼顾聚类的凝聚度Cohesion和分离度Separation
>
> 计算方法:
>
> 设样本到类内其他点平均距离为a, 样本到类外其他簇样本点最小平均平均为b
>
> `(b - a) / max(a, b)`
>
> 平均轮廓系数**越大**，聚类效果**越好**

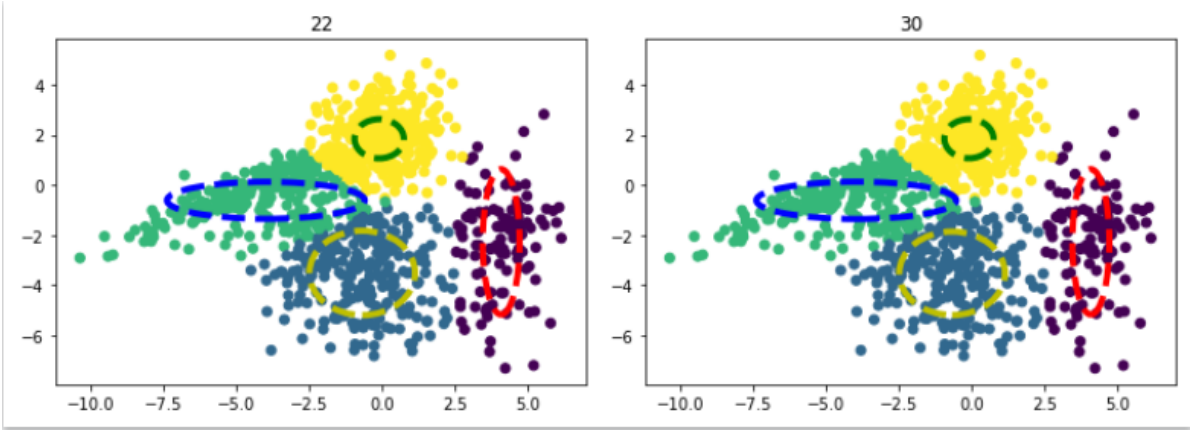
下图为不同k值对应GMM模型聚类结果平均轮廓系数



由此可见K=4时，平均轮廓系数达到峰值，聚合效果最好

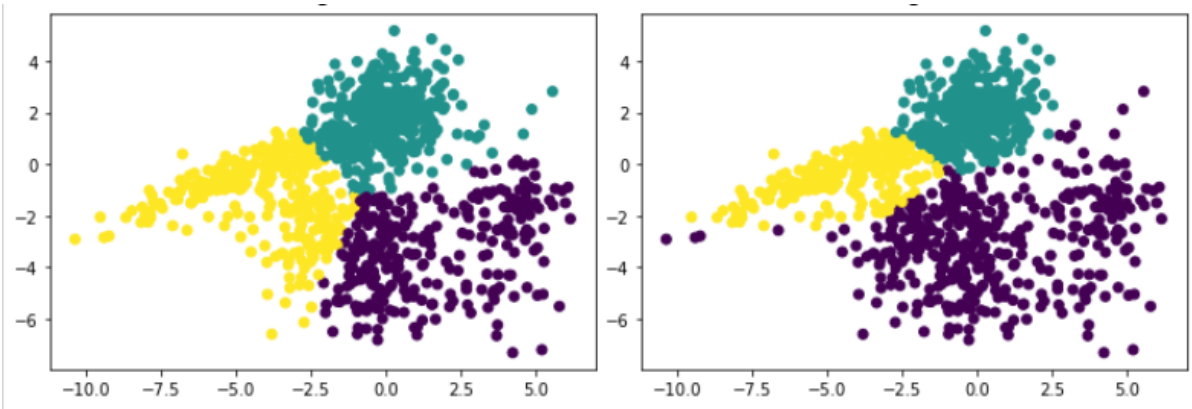
故以K=4为例观察迭代中各成分参数变化

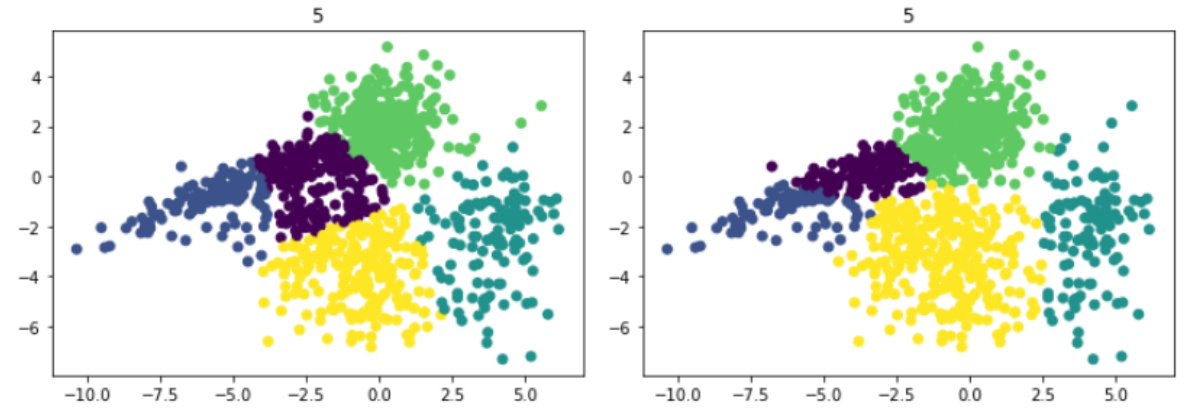




模型对比

模型对比左侧为KMeans模型、右侧GMM模型





> 上图所示，kmeans聚类结果多呈圆形，由于其计算样本点到簇中心距离选择最小归类导致
>
> GMM则可以解决此问题，如上图中右上子图GMM划分黄类时呈椭圆型细长区域