# 学生实验报告

|  |  |  |  |
| --- | --- | --- | --- |
| 学号 | 1120182525 | 学院 | 徐特立学院 |
| 姓名 | 梁瑛平 | 专业 | 计算机科学与技术 |

## 聚类

## 实验简介

本实验采用三种聚类方法，对iris数据集进行聚类，其目的是提升学生应用聚类方法解决实际问题的能力。

## 实验目的

（1）帮助学生理解聚类方法在数据科学中的应用。

（2）帮助学生掌握聚类方法的原理。

## 相关理论与知识点

1. 聚类的原理：
   1. **Kmeans：算法流程：**

1、首先确定一个k值，即我们希望将数据集经过聚类得到k个集合。

2、从数据集中随机选择k个数据点作为质心。

3、对数据集中每一个点，计算其与每一个质心的距离（如欧式距离），离哪个质心近，就划分到那个质心所属的集合。

4、把所有数据归好集合后，一共有k个集合。然后重新计算每个集合的质心。

5、如果新计算出来的质心和原来的质心之间的距离小于某一个设置的阈值（表示重新计算的质心的位置变化不大，趋于稳定，或者说收敛），我们可以认为聚类已经达到期望的结果，算法终止。

6、如果新质心和原质心距离变化很大，需要迭代3~5步骤。

* 1. **GaussianMixture：**高斯混合模型（GMM）可以看做是k-means模型的一个优化。它既是一种工业界常用的技术手段，也是一种生成式模型。高斯混合模型试图找到多维高斯模型概率分布的混合表示，从而拟合出任意形状的数据分布。在最简单的场景中，GMM可以用与k-means相同的方式进行聚类。它使用EM算法进行迭代：

1. 选择位置和初始形状
2. 循环直至收敛：

E步骤：对于每个点，为每个点分别计算由该混合模型内的每个分量生成的概率。

M步骤：调整模型参数以最大化模型生成这些参数的可能性。

该算法保证该过程内的参数总会收敛到一个局部最优解。

* 1. **SpectralClustering：**谱聚类是从图论中演化出来的算法，后来在聚类中得到了广泛的应用。它的主要思想是把所有的数据看做空间中的点，这些点之间可以用边连接起来。距离较远的两个点之间的边权重值较低，而距离较近的两个点之间的边权重值较高，通过对所有数据点组成的图进行切图，让切图后不同的子图间边权重和尽可能的低，而子图内的边权重和尽可能的高，从而达到聚类的目的。

1. 聚类方法的应用：

常见业务应用场景如下

1.1 目标用户的群体分类

通过对特定运营目的和商业目的所挑选出的指标变量进行聚类分析，把目标群体划分成几个具有明显特征区别的细分群体，从而可以在运营活动中为这些细分群体采取精细化，个性化的运营和服务，最终提升运营的效率和商业效果（如把付费用户按照几个特定维度，如利润贡献，用户年龄，续费次数等聚类分析后得到不同特征的群体）1.2不同产品的价值组合

企业可以按照不同的商业目的，并依照特定的指标标量来为众多的产品种类进行聚类分析，把企业的产品体系进一步细分成具有不同价值，不同目的的多维度的产品组合，并且在此基础分别制定和相应的开发计划，运营计划和服务规划（如哪些产品畅销毛利率又高，哪些产品滞销且毛利又低）1.3探测，发现离群点，异常值

这里的离群点指相对于整体数据对象而言的少数数据对象，这些对象的行为特征与整体的数据行为特征很不一致（如某B2C电商平台上，比较昂贵，频繁的交易，就有可能隐含欺诈的风险尘封，需要风控部门提前关注，监控）

聚类分析的其他应用场景

1. 聚类产生的类别作为一个新的字段加入其他的模型搭建过程中，作为细分群体的建模依据。

2. 通过对变量进行聚类，可以检查数据的共线性，对同一分组内的变量相关性较高，通过数据变换或筛选精简变量

（3）聚类方法评价指标：

已知真实簇的标签分配，可以使用条件熵分析来定义一些直观的度量。

Rosenberg 和 Hirschberg (2007) 为任何簇分配定义了以下两个理想的目标:

同质性(homogeneity): 每个簇只包含一个类的成员

完整性(completeness): 给定类的所有成员都分配给同一个簇。

这两个分别有对应的homogeneity\_score和completeness\_score作为分数去衡量。

同质性和完整性两个分数的调和平均成为V-measure，最新的sklearn中还支持带beta参数的v\_measure\_score

## 实验条件与环境

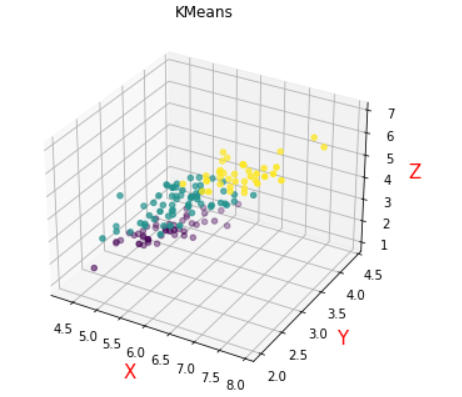
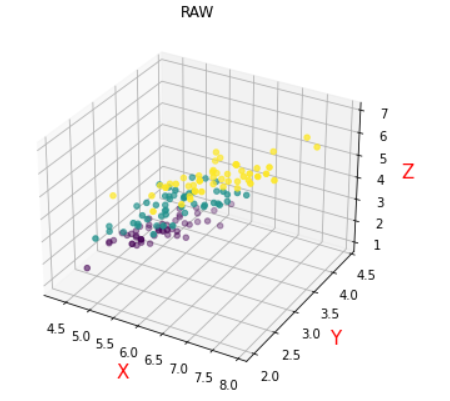
|  |  |  |  |
| --- | --- | --- | --- |
| 要求 | 名称 | 版本要求 | 备注 |
| **编程语言** | python | 3.6以上 |  |
| **开发环境** | Jupter | 无要求 |  |
| **第三方工具包/库/插件** | sklearn | 0.23.1 |  |
| **第三方工具包/库/插件** | Pandas | 1.0.5 |  |
| **第三方工具包/库/插件** | numpy | 1.16.2 |  |
| **其他工具** | 无 | 无要求 |  |
| **硬件环境** | 台式机、笔记本均可 | 无要求 |  |

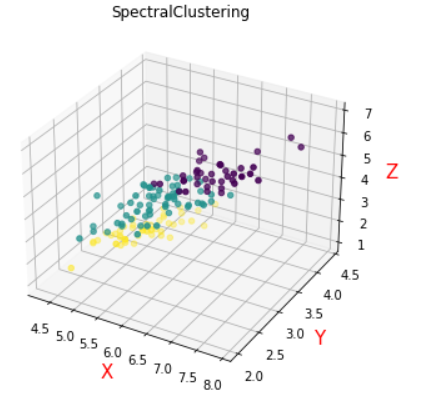
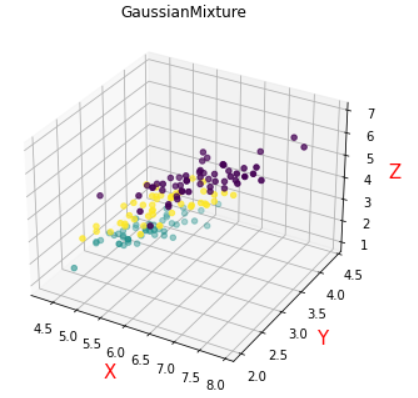
## 实验任务

|  |  |  |
| --- | --- | --- |
| **序号** | **任务名称** | **任务具体要求** |
| **1** | 数据理解 | 理解数据集背景以及数据含义。 |
| **2** | 数据读入 | 使用iris数据集 |
| **3** | 提取特征值 | 不使用数据集的类别标签信息，使用以下三种方法进行聚类实验 |
| **4** | Kernel K-means | 使用Kernel K-means方法对数据进行聚类 |
| **5** | EM聚类 | 使用EM聚类方法对数据进行聚类 |
| **6** | 谱聚类 | 使用谱聚类方法对数据进行聚类 |
| **7** | 评价 | 选择评价指标对聚类结果进行评价 |

## 实验结果及分析

|  |  |
| --- | --- |
| 聚类算法 | Homogeneity score |
| KMeans | 0.7514854021988338 |
| GaussianMixture | 0.8983263672602775 |
| SpectralClustering | 0.7603645798041669 |





## 收获与体会

（1）学会了三种经典聚类算法；

（2）了解了聚类算法的应用场景；

（3）实践了聚类算法，提高了代码能力；

## 备注及其他

无