華中科技大学

课 程 实 验 报 告

课程名称： 大数据分析

专业班级： 物联网 1801 班

学　　号： U201814500

姓　　名： 王英嘉

指导教师： 崔金华

报告日期： 2020.12.16

计算机科学与技术学院

目录

实验四 kmeans 算法及其实现

4.1 实验目的

1、加深对聚类算法的理解,进一步认识聚类算法的实现；
2、分析 kmeans 流程,探究聚类算法原理；
3、掌握 kmeans 算法核心要点；
4、将 kmeans 算法运用于实际，并掌握其度量好坏方式。

4.2 实验内容

提供葡萄酒识别数据集，数据集已经被归一化。同学可以思考数据集为什么被归一化，如果没有被归一化，实验结果是怎么样的，以及为什么这样。

同时葡萄酒数据集中已经按照类别给出了 1、2、3 种葡萄酒数据，在 csv 文件中的第一列标注了出来，大家可以将聚类好的数据与标的数据做对比。

编写 kmeans 算法，算法的输入是葡萄酒数据集，葡萄酒数据集一共 13 维数据，代表着葡萄酒的 13 维特征，请在欧式距离下对葡萄酒的所有数据进行聚类，聚类的数量 K 值为 3。

在本次实验中，最终评价 kmeans 算法的精准度有两种，第一是葡萄酒数据集已经给出的三个聚类，和自己运行的三个聚类做准确度判断。第二个是计算所有数据点到各自质心距离的平方和。请各位同学在实验中计算出这两个值。

实验进阶部分：在聚类之后，任选两个维度，以三种不同的颜色对自己聚类的结果进行标注，最终以二维平面中点图的形式来展示三个质心和所有的样本点。效果展示图可如图 1 所示。

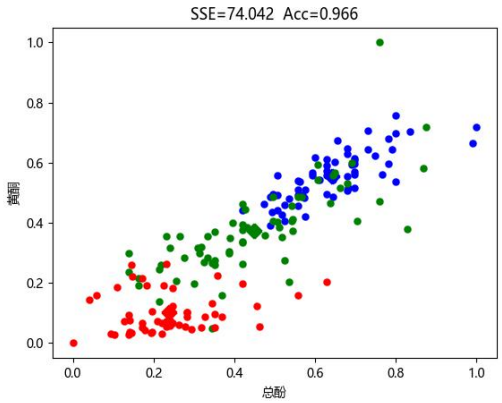

图 1 葡萄酒数据集在黄酮和总酚维度下聚类图像（SSE 为距离平方和，Acc 为准确率）

4.3 实验过程

4.3.1 编程思路

葡萄酒数据集需要被归一化，这是因为不同维度属性的度量范围不同，而在分类中默认要均衡每一个属性的影响（当然不同的任务中不同属性也可能权重不同，具体任务具体分析）如果一个属性值范围过大或过小的话，会对分类产生偏差。

Kmeans 算法流程说明如下：

首先随机选择 n 个初始簇心，给它们分配编号 1-n，然后进行循环迭代，计算每个点对每个簇心的距离，将其归类在最近的簇心的编号下。对数据进行每一轮循环后，重新计算不同簇的簇心位置，计算方式为取该簇所有点位置的平均值；循环在所有点的归类均不发生变化或者达到最大循环次数时退出。

Kmeans 算法原理比较简单，但也有很多地方需要注意。例如：初始簇心的选取非常重要，不同的初始化方式对分类结果影响较大；计算时注意代码的效率问题，尽量使用 numpy 进行矩阵加速;结果中很可能分类是比较正确的，但标签给的和原始数据中不匹配，这个也需要后期处理才能获得正确的分类准确率。

作图方面使用 matplotlib.pyplot 模块作散点图展示即可。

4.3.2 遇到的问题及解决方式

如 4.3.1 中所说，有的时候分类是正确的，但是赋的标签不匹配，这个在 n_cluster=2 时比较好处理，在分类数大于 2 时需要对分类结果的标签重新分配，找到最大的准确率即为所求。

4.3.3 实验测试与结果分析

如图 2 所示，在 6 轮循环之后，分类结束，分类准确率达到 94.94%。

```
(torch) yingjia@yingjia-Vostro-5370:~/文档/vscode/python/lab4$ python kmeans
.py
After 6 rounds...
Acc = 94.94%
```

图 2 终端运行结果

所有数据点到各自质心距离的平方和 SSE 以及绘制的散点图（包括质心）如下图中所示，比较了 dim1 和 dim2、dim3 和 dim4、dim5 和 dim6 不同维度组合间的分类效果，可以看出 dim5 和 dim6 的分类最明显。

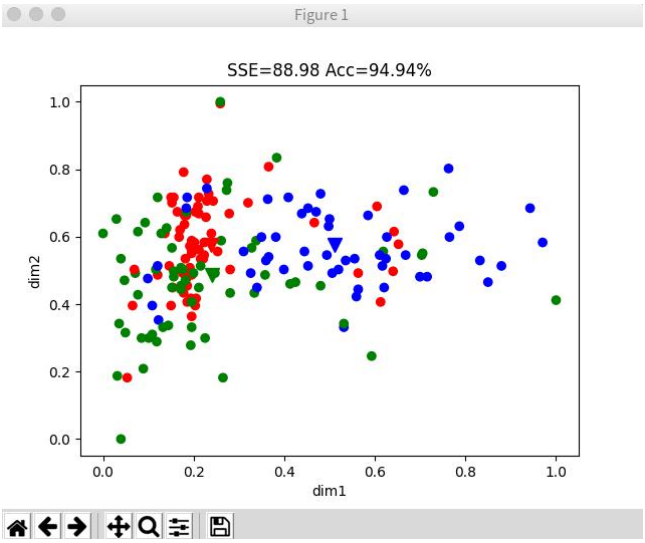

图 3 dim1 和 dim2 分类效果

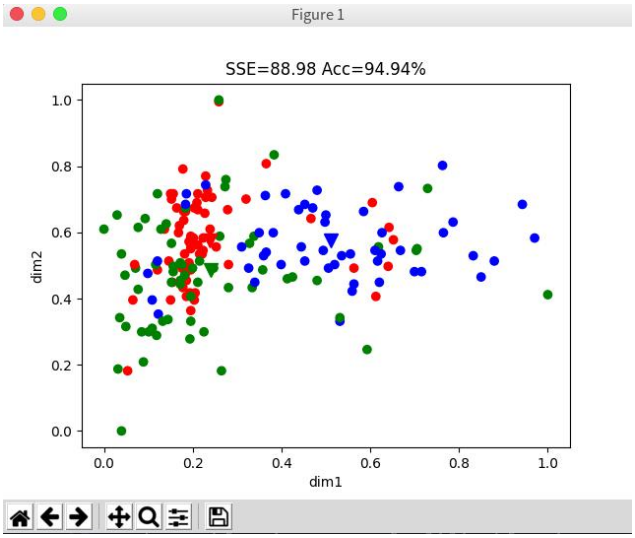

图 4 dim3 和 dim4 分类效果

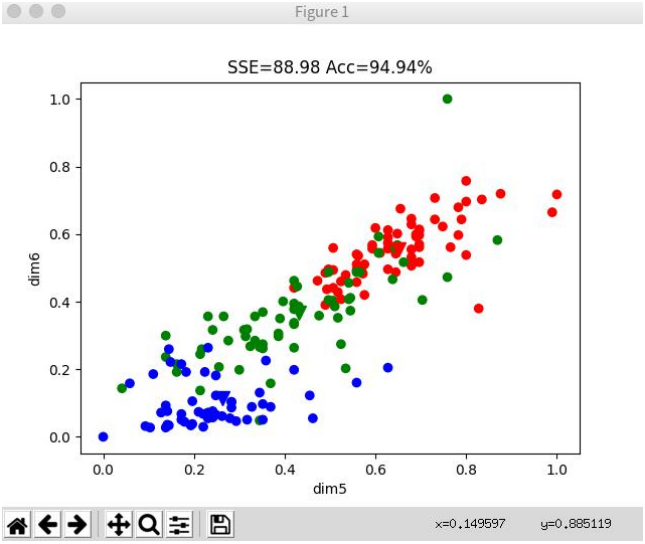

图 5 dim5 和 dim6 分类效果

4.4 实验总结

本次实验难度较小，没有遇到太多问题。在实验中掌握了基本 kmeans 算法的基本原理与评价指标，也意识到该算法在很多策略上都需要优化才能发挥出更好的分类效果，在此基础上进一步了解了 kmeans 衍生的很多优化版本，如 kmeans++、mini batch kmeans 等等，在聚类算法方面收获很大。