聚类算法

手动实现k-means算法或者使用sklearn的k-means api来对数据进行聚类。

数据集描述

1、sklearn中的dataset中的iris数据集，包含150条数据，每条数据有4个属性。

2、`data.csv`数据集共10000条数据。尝试分别在前200条、前1000条、前10000条数据，利用k-means算法或者k-means算法的改进实现所有数据的聚类，并分别画出对应的数据可视化图。

代码实现

简单的代码实现：

https://github.com/case-smart-data-engineering/4.4.1-1

更一般化代码实现：

k_means.py文件

目录结构

```
/
|--data
|  |--data1.csv
|--main.py # 调用sklearn实现的k-means
|--k_means.py # 手动实现的k-means
```

参考资料

- Sklearn库的使用视频（共30分钟，请勿上传到网络上），包括数据集的预处理（如数据集的查看，构造训练集、验证集和测试集）和朴素贝叶斯、决策树、支持向量机等传统分类算法的调用：百度网盘链接:https://pan.baidu.com/s/1Ywl3qWGDha7K-i7aG3TRLA 提取码: v786
- 机器学习实验的背景知识：
  - Python标准库的使用视频（约50分钟，请勿上传到网络上），包括collections、itertools、time和random：百度网盘链接: https://pan.baidu.com/s/1wCY2w_-RM5PBHgDHy9gLyg 提取码: furv
  - Numpy库的使用视频（约45分钟，请勿上传到网络上），包括四大运算、创建数组和数组的性质等：百度网盘链接: https://pan.baidu.com/s/1CUD-ZrrINEtzw-yNPk87kg 提取码: a194
  - Pandas库的使用视频（约1个小时20分钟，请勿上传到网络上），包括数值运算及统计分析、处理缺失值和合并数据：百度网盘链接: https://pan.baidu.com/s/1WiWLez3lhsWDYX7zor3Ovw 提取码: w9hr