廣東工業大學

QG 中期考核详细报告书

题　　目　数据挖掘组中期考核

学　　院　计算机学院

专　　业　软件工程

年级班别　19 级 3 班

学　　号　3119005120

学生姓名　吴佳杨

2020 年 04 月

任务要求：

首先恭喜大家坚持到了 QG 工作室数据挖掘组中期考核，坚持到这里不容易。下面就是我们简简单单的中期考核了，具体考核要求如下：

（1）从下列 10 个算法中任选两个实现，有余力的同学可以实现更多算法（基础任务：实现两个；建议任务：实现四个）

（2）算法需实现两种形式：1. 只允许使用 numpy,pandas,matplotlib 三个主要库进行算法的编写实现；不允许调用 Sklearn 库。2.使用 Sklearn 库中的模型实现算法。将自己实现的算法和调用库的算法进行对比评估。

（3）部分数据集放在了 data 文件夹中，或者附带了下载链接。对数据集信息不了解的可以自行查看相关的数据集介绍文件，或者自行百度查找相关数据集介绍。

（4）这次中期考核要求实现详细文档。详细文档内容应包括数据集的处理，算法步骤和思想，算法实现结果评估，不足和优化之处等（可以自行编写）。详细文档封面放在附件中。

（5）中期考核还需要制作答辩 ppt，答辩时间为 3 分钟，大家把握好时间制作 ppt。

注意：中期考核时限为一星期，各位需要在下星期五大组培训之前将自己的代码、详细报告文档和中期答辩 ppt 放到 github 并且发给对应的导师。文档需要 word 格式和 pdf 格式。

一、K-means 算法

1. 考核要求

1）基本任务：

内容一：理解 k-means 算法的思想

内容二：使用 python 实现 k-means 算法

内容三：使用 Iris 数据集进行测试

内容四：对参数 k 进行调整，记录结果

2）进阶任务：

实现二分 k-means 代码并测试，并进行比较

2. 数据集

数据集名称：Iris 数据集

3.数据集的处理

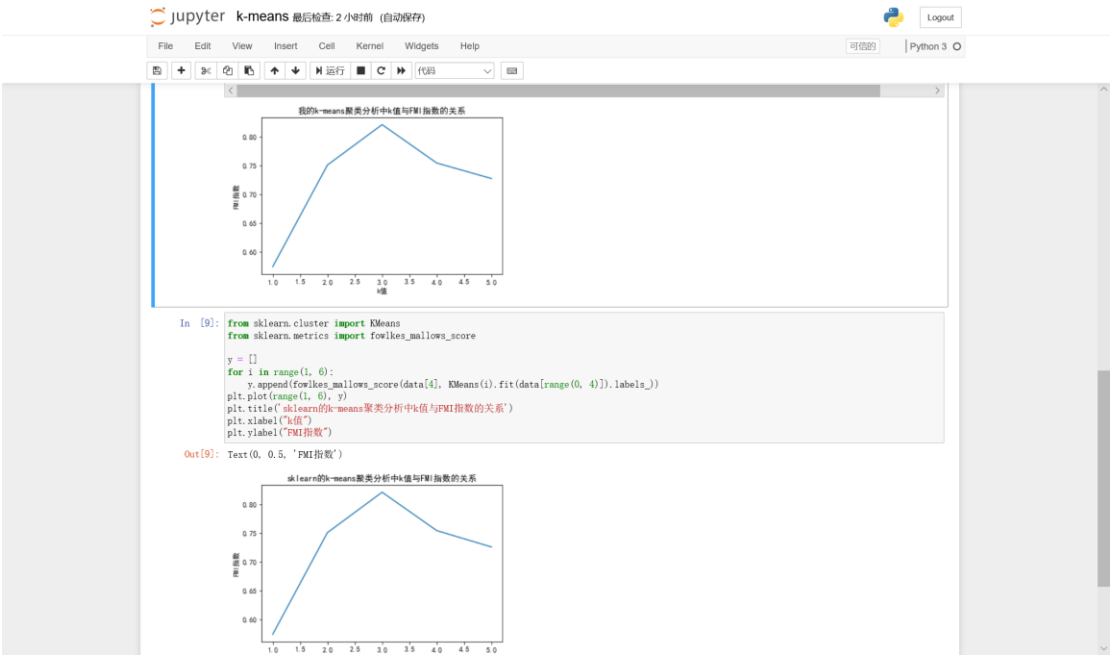
用 pandas 库的 read_csv 函数读取数据并转化为 Dataframe 形式，取除了最后一列的所有列作为训练集。

4.算法步骤和思想

从数据集中随机选取 k 个点作为聚类中心，将数据集中每个点划分到最近的聚类中并重新计算聚类中心，重复执行上一步直到不再有点被划分至不同的聚类中。

同时要注意函数的复用。

5.算法实现结果评估



6.不足之处

与 sklearn 的 k-means 比起来，我这个耗时更久，而且代码不是很简洁。

二、k 近邻算法

1.考核要求

1）基本任务：

内容一：理解 KNN 算法的思想

内容二：使用 Python 实现 KNN 算法

内容三：使用 Iris 数据集进行测试

内容四：记录测试结果

2）进阶任务：

自行优化 KNN 算法并进行代码测试

2. 数据集

数据集名称：Iris 数据集

3.数据集的处理

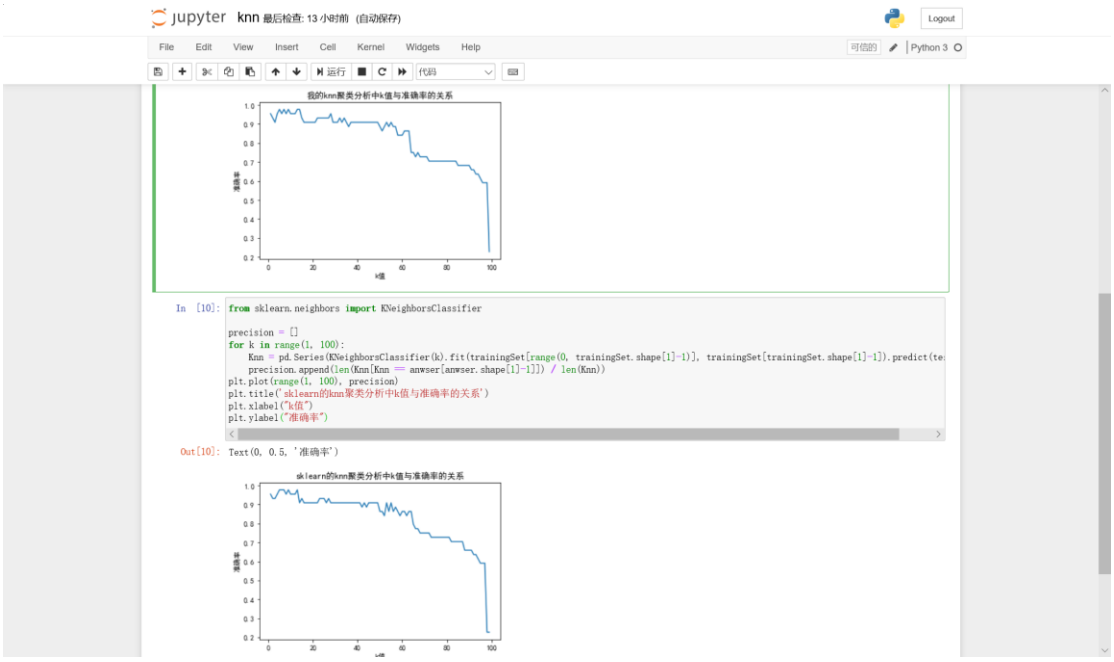
用 pandas 库的 read_csv 函数读取数据并转化为 Dataframe 形式，先剔除重复值，再随机选取 70%作为训练集，将剩下的 30%剔除答案作为测试集。

4.算法步骤和思想

针对测试集中的每条数据集，在训练集中找出离他最近的 k 个点，将该数据划分给这些点中数量最多的类。

同时要注意函数的复用。

5.算法实现结果评估

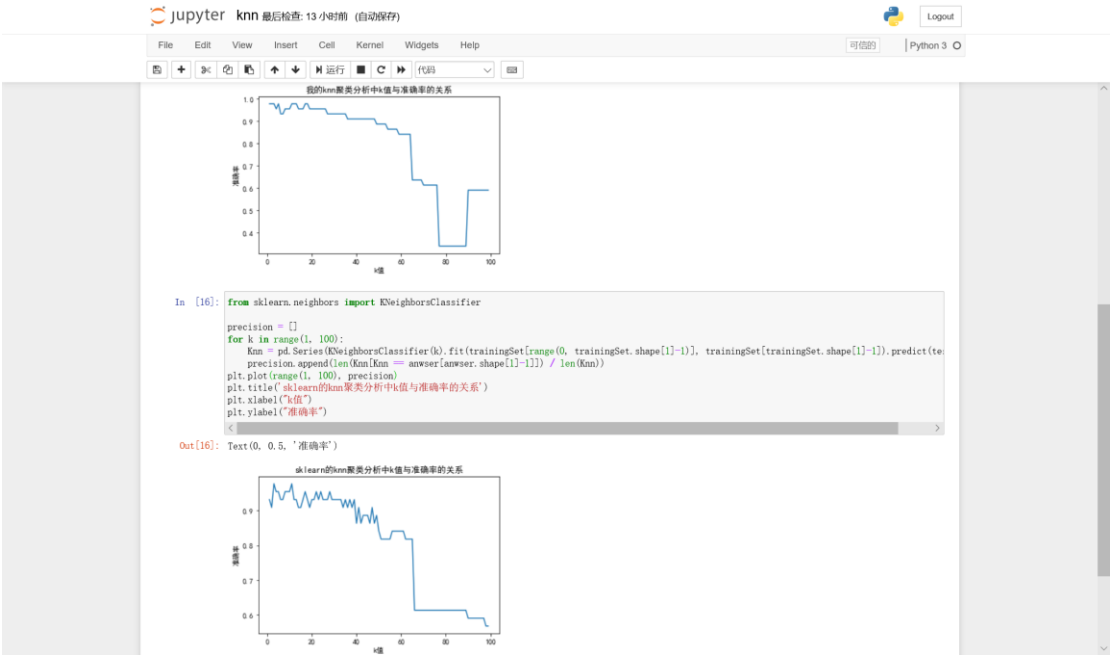


6.不足之处

与 sklearn 的 knn 比起来，我这个耗时更久，而且代码不是很简洁。

7.尝试优化

尝试动态扩充训练集，即：将分类完毕的测试集数据并入训练集，并对之后的测试集产生效果。



似乎。。好像。。k 在[1,50]之间的时候，准确率真的上升了一些？