****

****

**QG中期考核详细报告书**

**题目：机器学习算法**

**学院：计算机**

**专业：信息安全**

**姓名：刘栋濠**

**学号：3119005422**

**2020年04月 24**

一、k-nn算法（基于iris数据集）

1.算法思想：

（1）计算已知类别数据集中的点与当前点之间的距离；

　 （2）按照距离递增次序排序；

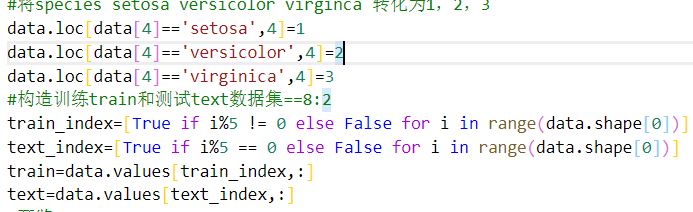
　 （3）选取与当前点距离最小的k个点；

　 （4）确定前k个点所在类别的出现频率；

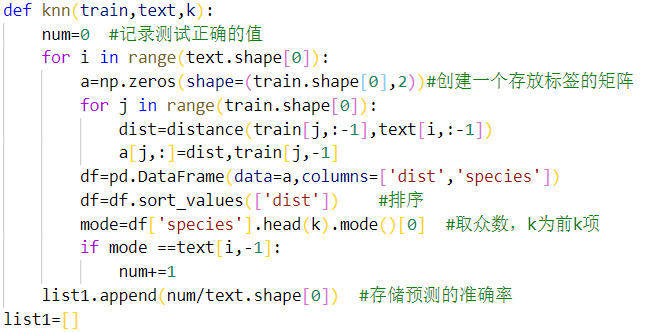
　　（5）返回前k个点出现频率最高的类别作为当前点的预测分类。

2.对数据的预处理：

（1）首先导入iris数据集，对三种花类编号1，2，3，并按8：2创建训练集和测试集

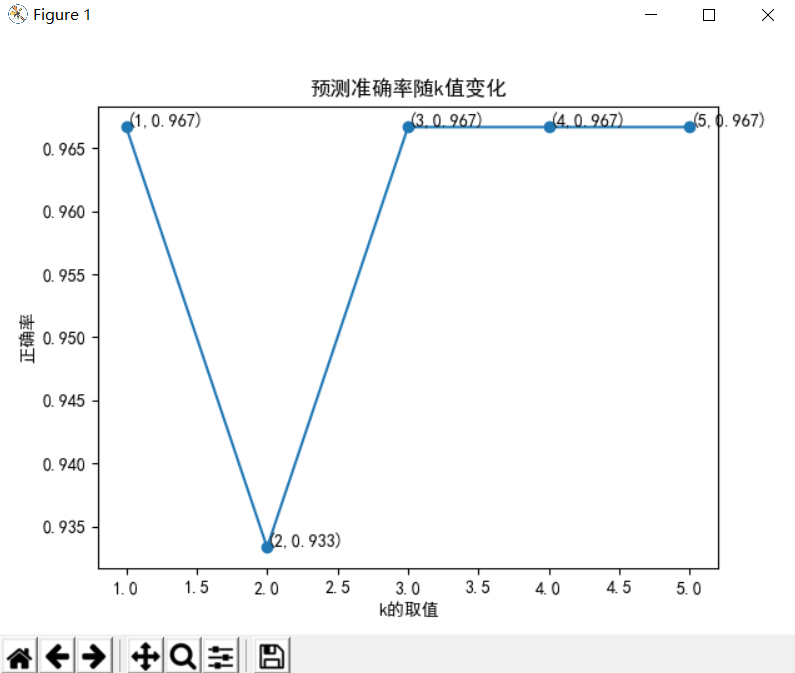


3.算法核心步骤：



计算测试集中每个数据与训练集的欧式距离，并排序，根据k值选取前k个项中的众数，记录预估正确的个数，算出准确率

4.结果可视化：



K-nn算法：

优点：算法简单，准确率高，受异常点干扰小

缺点：计算复杂性高，对大数据处理效率低

二．k-means聚类iris数据集：

1.算法思想：

1、首先确定一个k值，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。(这里采用均值法求解)

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。或者迭代多次。

2.算法核心步骤：

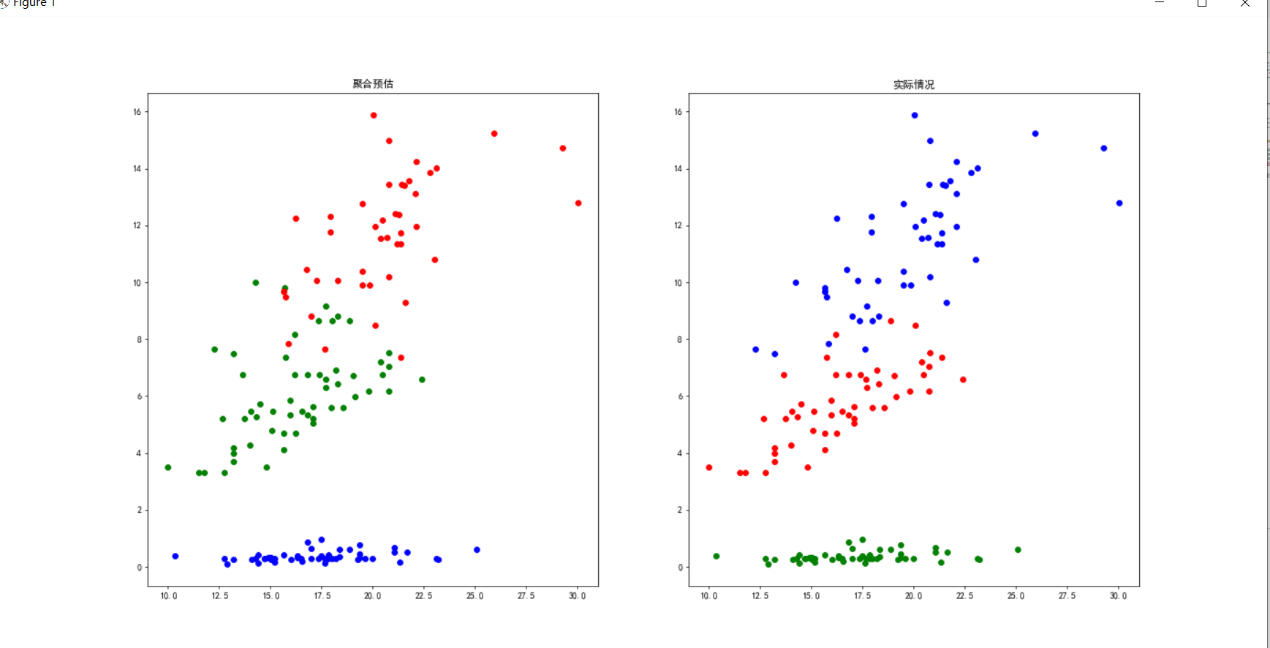
（1）先从数据中任意选取三个数据作为中心簇；

（2）遍历所有数据并计算与三个簇的距离，将数据归入距离最近的簇中；

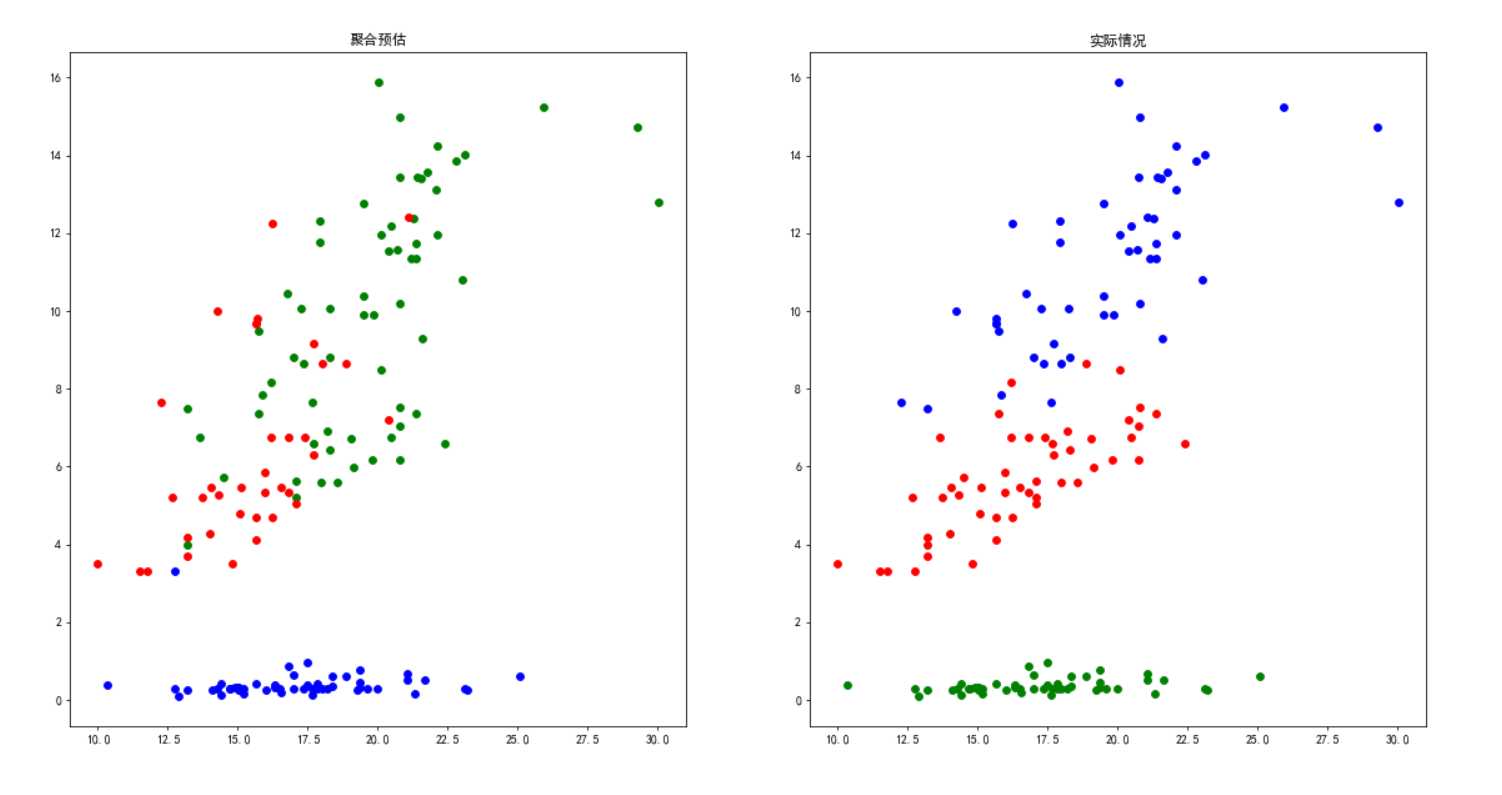
（3）对每个簇中的数据按列求均值，并更新到原来的中心簇

（4）迭代多次

3.数据可视化：



4.评估与优化：

对于初始中心簇的选取随机时，会出现聚散效果差，例如：

这样的情况；

此时对于随机选取：

修改；

修改思想：

将所有数据与零向量求距离，并排序，按三类选取150个数据时选取第25，75，125的数据作为中心簇，聚类效果得到优化

k-means：

优点：

原理简单，算法易实现，收敛速度快

簇密集时聚类效果好

缺点：

未知数据k值选定困难

对于随机选定的的初始中心簇影响较大

对于异常点不敏感