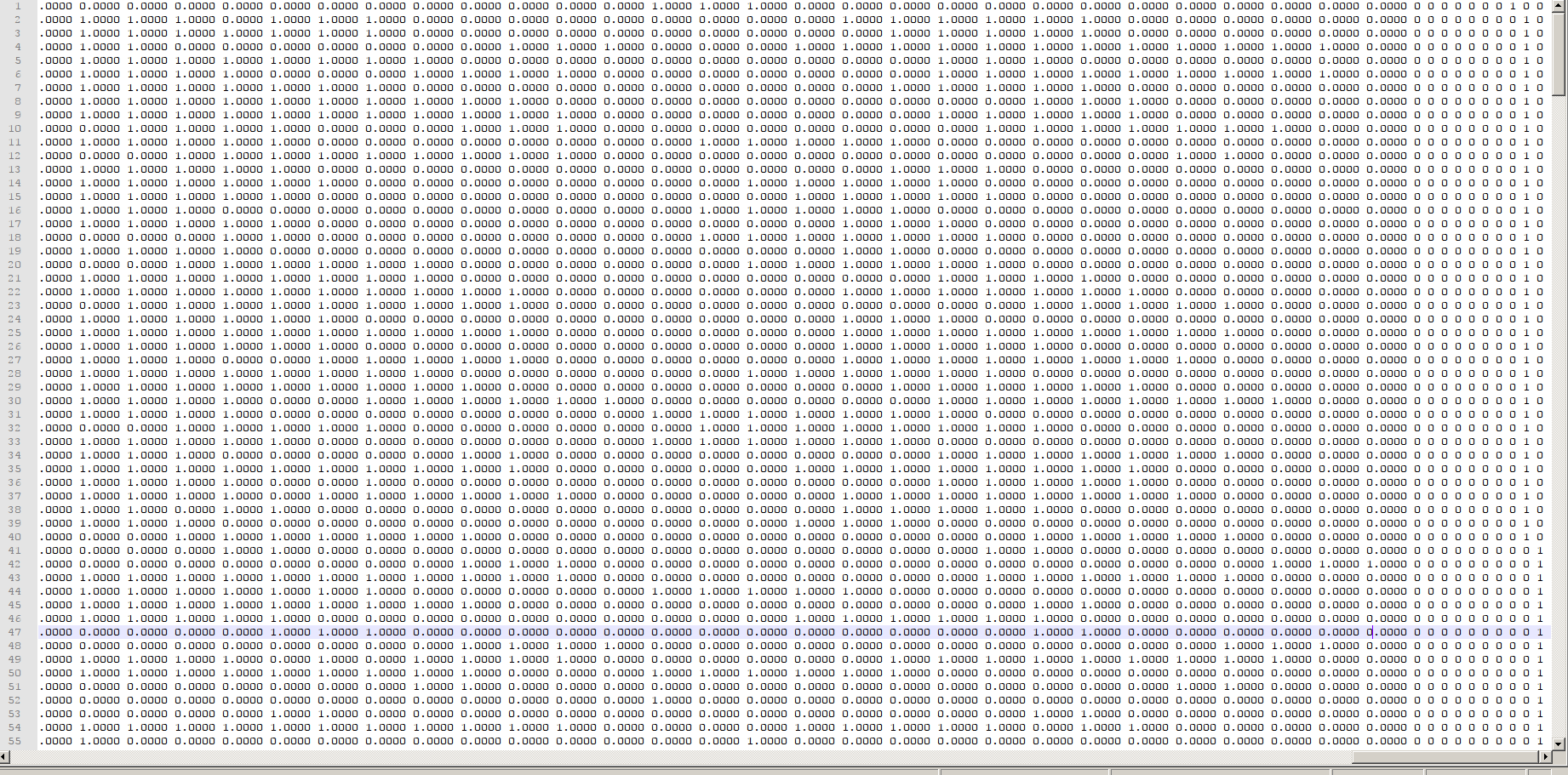
实验一：KNN分类器

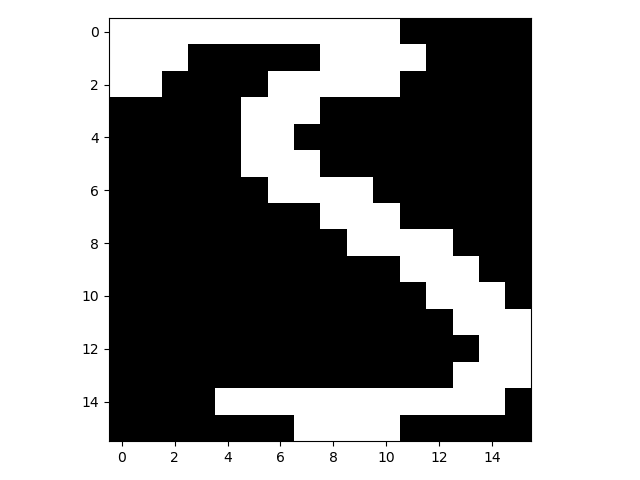
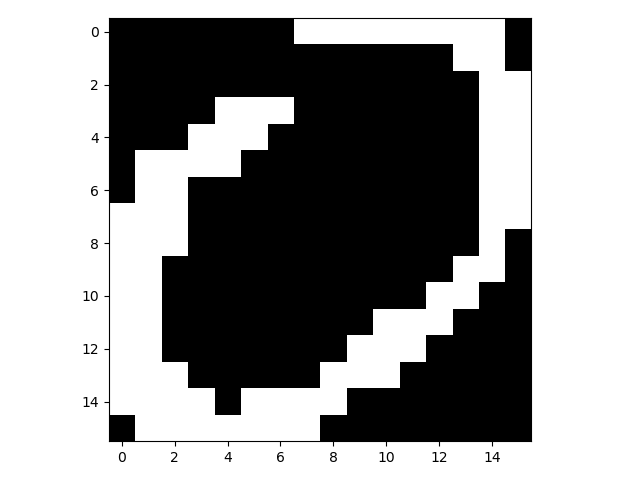
问题描述

1. 近邻算法是机器学习领域中一种用于分类和回归的非参数统计方法。简单来说，在分类问题中，K-近邻算法的思想可概括为如下：给定一个训练数据集，其中的实例类别已定。分类是对于新的类别，根据其k个最近邻的训练实例的类别，通过多数表决法等方式进行预测。由以上可以看出，K-近邻算法不具有显性的学习过程，K-近邻算法实际上利用训练数据集对特征空间进行划分，并作为其分类的模型。

本次实验的数据集是一个二进制手写数字图像的分类问题，训练集为1115张大小为16×16的二进制图像，测试集为478张。数据保存在semeion\_train.csv和semeion\_test.csv两个文件中，其保存格式如图1所示，csv文件中是一个1115×266的矩阵，其中前256列是手写数字图像的一维存储形式，1代表该点有着色，0代表无，后十列是每张图片对应的one-hot形式存储的标签。



为了对数据有更直观的理解，我们可以将每一条数值形式存储的图片转换为二进制图像来进行查看，如图2所示，此时我们便可以清晰地看出每一条数据的含义。



解决方法

不难看出，此问题是一个典型的分类问题，类别共有十项，分别是0-9的每一个数字，每个数字在二维平面上的结构特征具有明显不同，我们运用K-NN的思想，将训练数据分为10个不同的簇，然后将测试数据逐条输入，通过K-NN算法即可判别出其所属的种类。

简单来说，我所采用的解决方法是基于K-NN算法的核心思想进行设计的，具体流程如下：

* 计算待分类点与已知类别的点之间的欧氏距离
* 按照距离递增次序排序
* 选取与待分类点距离最小的k个点
* 确定前k个点所在类别的出现次数
* 返回前k个点出现次数最高的类别作为待分类点的预测分类

此外实验中我同样尝试了基于距离加权的K-NN算法，流程如下：

* 计算待分类点与已知类别的点之间的欧氏距离
* 按照距离递增次序排序
* 选取与待分类点距离最小的k个点
* 设置一个大小为10的权值数组，将每一项的值置为-1
* 将k个点与待分类点的距离加到权值数组对应的项
* 返回权值数组中非负的最小的类别作为待分类点的预测分类

实验及代码分析