智能数据挖掘作业6

1920030004 黄铭瑞

# 实验目的

生成双层正方形数据集，实现对双层正方形数据的聚类。

# 实验原理

## DBSCAN聚类

### 定义

DBSCAN（Density-Based Spatial Clustering of Applications with Noise，具有噪声的基于密度的聚类方法）是一种基于密度的空间聚类算法。该算法将具有足够密度的区域划分为簇，并在具有噪声的空间数据库中发现任意形状的簇，DBSCAN算法将“簇”定义为密度相连的点的最大集合。

### 三类数据点

* **核心点（core point）：**若样本的epsilon邻域内至少包含了MinPts个样本，则称为核心点。
* **边界点（border point）：**若样本的epsilon邻域内包含的样本数目小于MinPts，但他在其他核心点的邻域内，则称为边界点。
* **噪声点（noise）：**既不是核心点也不是边界点的点。噪声点是不被聚类纳入的点。

### 密度相关

* **密度直达（directly density reachable）**：如果满足，那么样本点是由样本点对于参数密度直达。
* **密度可达（density reachable）：**如果存在一系列样本点，，任意两个相邻对象之间是直接密度可达，则是关于参数密度可达。
* **密度相连（density connected）：**如果在样本集中存在一个样本点，使得和均由样本点密度可达，那么称与对于参数密度相连。

### 算法步骤

**Step1:** 任选一个点，找到和它距离小于等于的所有点。如果找到的点数小于，标记为噪声点；如果点的个数大于，这个点被标为核心样本，并分配新的簇标签。

**Step2:** 访问点的所有邻居点，如果他们还未被分配到一个簇，那么将刚创建的簇标签分配给他们，如果他们是核心点，就依次访问他们的邻居，以此进行簇扩张，直到内没有更多核心样本。

**Step3:** 重复**step1**、**step2**，直到没有新的簇添加，结束。

# 实验过程

## 生成数据

使用random.random()-0.5生成落在(-0.5,0.5)上的随机数，再对x和y 的范围进行限制，得到x和y的list，使用zip()把x和y合并为二维array。在经过类型转换，保存。

## DBSCAN算法部分

### 找出epslion范围内的点

定义dist()，计算欧氏距离。

定义eps\_neighbor()，判断两个数据点之间的距离是否在eps范围内。

定义region\_query()，遍历数据集里所有点，把在eps范围内的点找出。

### 判断是否进行簇扩张

定义expand\_cluster()，dbscan()，不满足MinPts条件的列为噪声点，满足的点划分到该簇，如果是核心点，继续访问他们的邻居进行扩张。遍历所有的点，重复操作，返回聚类簇id和聚类簇数目。

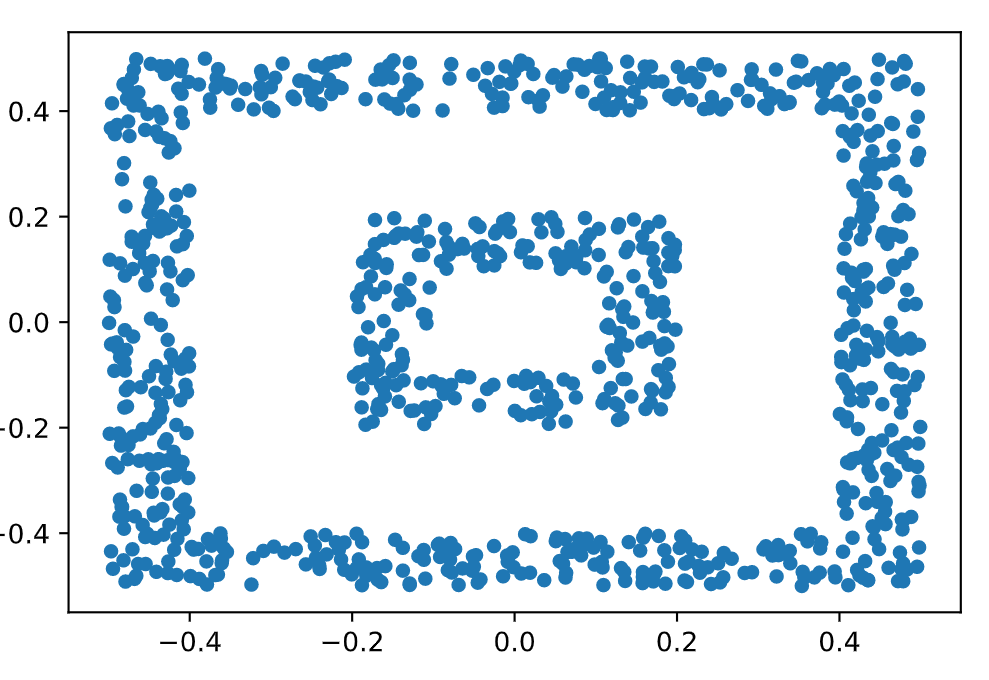
## 绘图部分

定义plotFeature(), 用于对就聚类结果的绘图。每个独特的簇id用同一种颜色标出。有几个不同的聚类簇，就有多少种颜色显示。

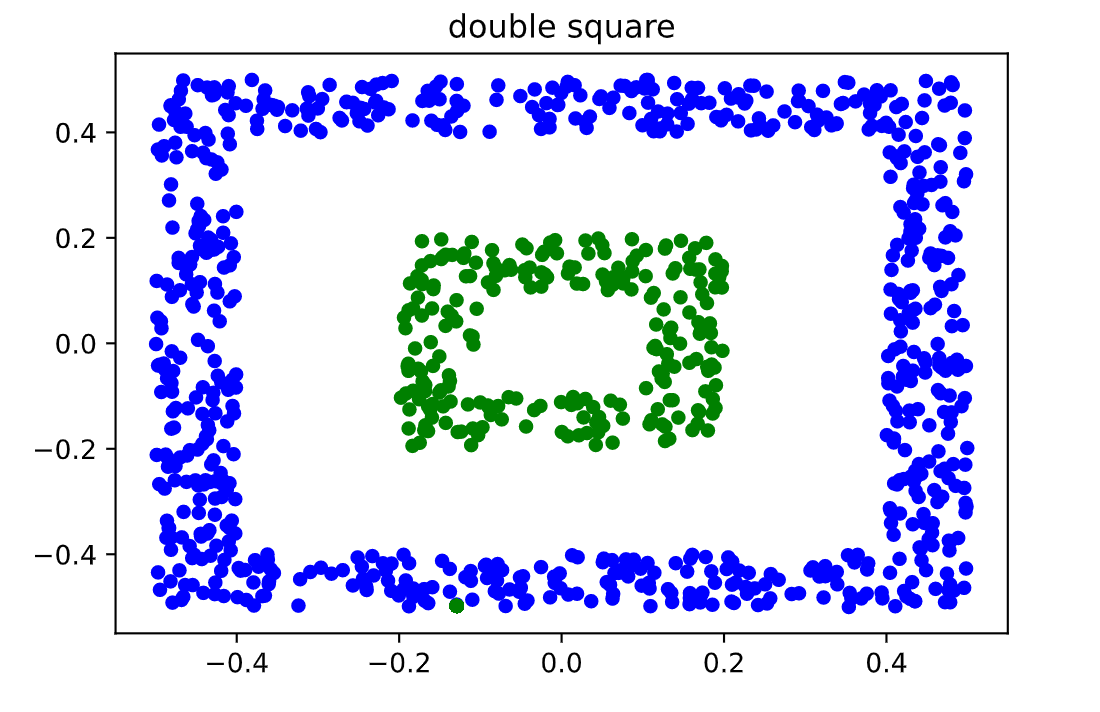
读取保存的数据文件，DBSCAN聚类，绘制聚类图。

# 实验结果

原始数据：



DBSCAN聚类结果：



# 附录

[double square.html](double%20square.html)