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x_i的 epsilon 邻域内至少包含了 MinPts 个样本，则称x_i为核心点。
- **边界点（border point）：**若样本x_i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x_i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$p \in N_{eps}(q), and \left|N_{eps}(q)\right| \geq MinPts$，那么样本点$p$是由样本点$q$对于参数$\{eps, MinPts\}$密度直达。
- **密度可达（density reachable）：**如果存在一系列样本点，$q \to a \to b \to c \to d \to p$，任意两个相邻对象之间是直接密度可达，则$p$是$q$关于参数$\{eps, MinPts\}$密度可达。

- **密度相连（density connected）：**如果在样本集D中存在一个样本点o，使得p和q均由样本点o密度可达，那么称p与q对于参数$\{eps, MinPts\}$密度相连。

算法步骤

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

Step2: 访问点q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eps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

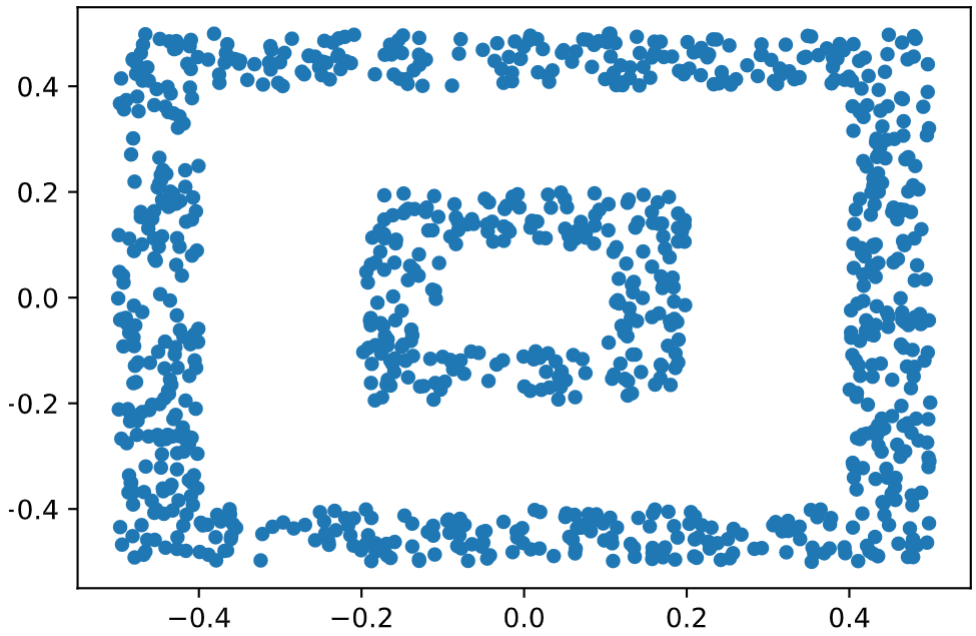
定义 plotFeature(), 用于对就聚类结果的绘图。每个独特的簇 id 用同一种颜

色标出。有几个不同的聚类簇，就有多少种颜色显示。

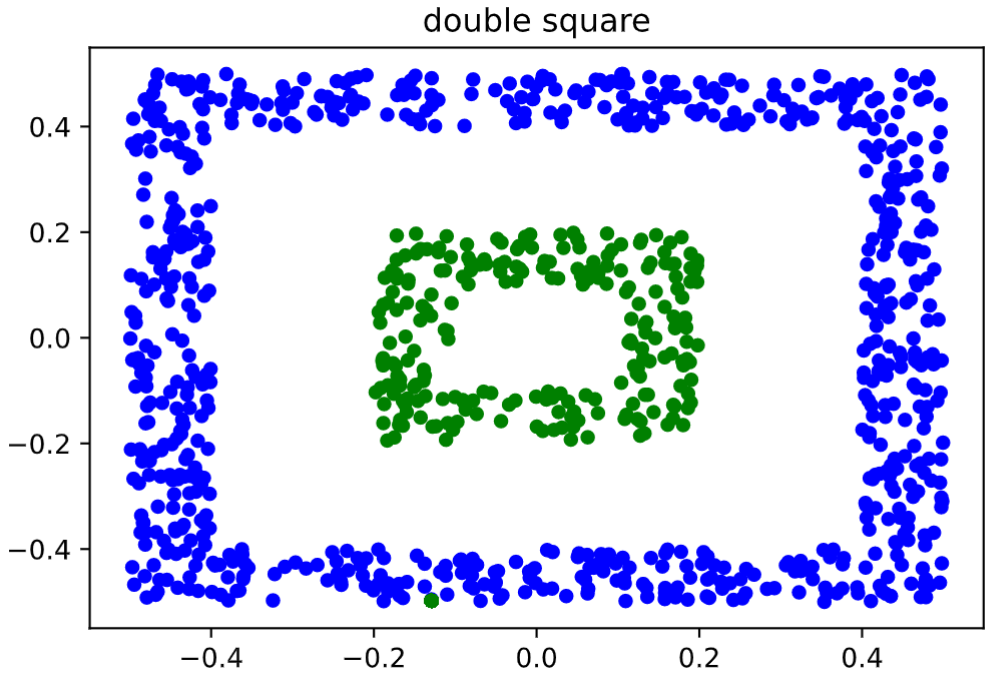
读取保存的数据文件，DBSCAN 聚类，绘制聚类图。

实验结果

原始数据：



DBSCAN 聚类结果：



附录

double square.html