Data Science

Assignment #3

|  |  |
| --- | --- |
| Course name | Data science |
| Assignment | Clustering\_dbscan |
| Major | Computer |
| Student ID | 2011004028 |
| Name | Seonghun Lee |

1. Summary of my algorithm

In this program, it can make clusters for the input file which only includes data points. Data points are constituted of data number, x axis, y axis, and I add one more value which is clustering number. So, this program make a few number of clusters as many as you want according to parameters such as epsilon and minimum points which means density of data points and clustering number you can choose the numbers.

This program has two main function. First, DBSCAN function is the main function make cluster, so you should input dataset, epsilon, and minimum points as parameters. This function find neighbors and judge whether data points are in same cluster or not. The second function is expand cluster function. This function is important part of this program, which expand the cluster size using neighbor’s neighbor.

1. Detailed description of my codes (for each functions)

|  |
| --- |
| DBSCAN function |
| * This function is most important function which make clusters * Seed[-1] is the last value of seed list and it show the state of the data points. If this value is -1, it means unvisited. if it is -2, it means outlier. And if it is more than and same as 0, it is clustering number. * Len of neighbor points is smaller than minimum points, it regards outlier. |
| Expand function |
| * This function is in the DBSCAN function, and it is also important part of this program. * It expand cluster size according to the neighbor’s neighbor state. If neighbor data points has enough neighbors it could expand the size of cluster. * NeighborPts add Next\_NeighborPts if it is not already in the neighbor list. |
| Region query function |
| * This function calculate the distance of two data points and find neighbors of seed point and return neighbors. * Epsilon is the maximum distance, so if it is neighbor, it should inside epsilon. |
| Main function |
| * Dbscan function return the number of clusters which is made and the cluster number saved in the dataset. * Cluster maintain the data data points as same cluster and sorting clusters as descending order. |

1. Instructions for compiling my codes at TA’s computer (e.g. screenshot)

|  |
| --- |
| 1. Put the clustering.py and input1.txt, input2.txt files into same folder |
|  |
| 1. Execute command prompt and go to this folder |
|  |
| 1. Execute clustering.py with command line arguments |
|  |
| 1. Check the output file in your folder |
|  |

1. Any other specification of my implementations and testing

|  |
| --- |
| Test result |
| * Result : input1 is almost 99, input2 is almost 95, input3 is almost 100. |