Programming Assignment #2

Build a Decision Tree

2013011457 컴퓨터전공

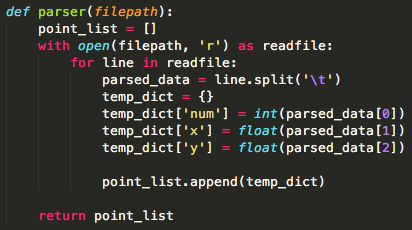
Jisoo Song

1. Summary of Algorithm

I implemented DBSCAN algorithm in python. I refer to pseudocode which written in powerpoint file that used in data mining class. But there’s one big difference from that pseudocode. In pseudocode, region query is called every iteration for each vector point, but I did region query on every vector point before clustering to avoid repeating region query for same vector point.

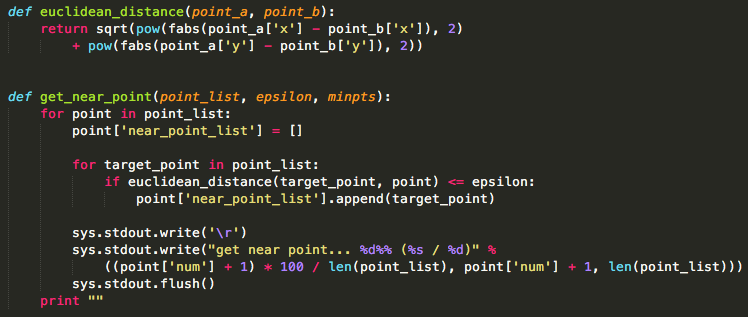
After that, the algorithm clusters each vector point following pseudocode. But I implemented postprocessing function because the number of clusters is given. It often generates many clusters more than given number of clusters. So, I handled each vector point that in extra clusters to assign it to right clusters.

1. Details about each Functions
2. Parser



parser function parses input data and returns parsed data. It reads the first line and gets vector point’s number, and it’s x, y value. It saves each point’s information on the list of dictionaries.

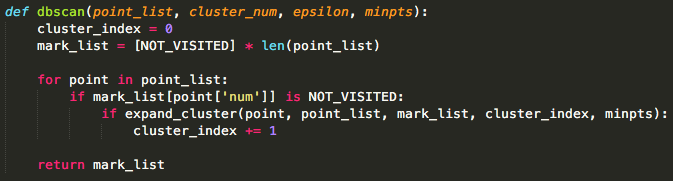
1. Euclidean\_distance and get\_near\_point



Euclidean\_distance function calculates Euclidean distance among given 2 vector points. It uses some basic math function in python like square root, and power.

Get\_near\_point function calculates what are the near points for a given point based on epsilon value. It calculates Euclidean distance for every point from given one point, and if the distance is smaller than epsilon value, that neighbor point is appended to near\_point\_list of each point.

1. DBSCAN



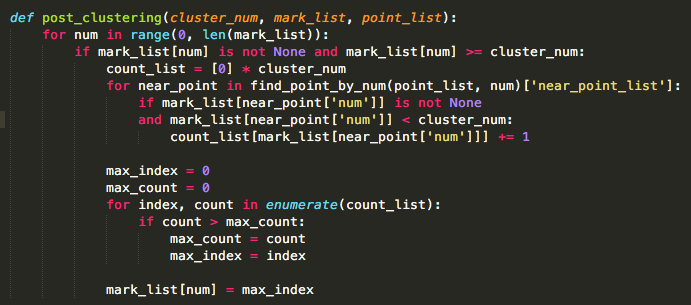
DBSCAN function follows pseudocode of DBSCAN algorithm almost samely. For every vector point, if that point is not visited, call expand\_cluster function and plus 1 to the cluster\_index value to increase the cluster’s number.

1. Expand\_cluster



expand\_cluster function expands the cluster using minpts value. It iterates on every near vector points for given vector point to find whether it has enough neighbor vector points. If it has enough vector points, it is appended to the same cluster with original given vector point.

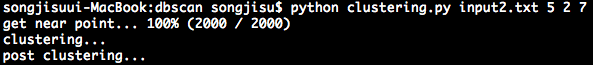
1. Postprocessing



It processes for vector points that assigned to extra clusters. This function finds other close clusters to assign for the other big clusters.

1. How to compile the codes

I implemented this code using python 2.7. You can compile this codes like this:



My main function’s name is clustering.py. I followed the rules of arguments that written on the notice file.