CSCI 5523 Homework 5 Due Dec 14, 2011

Only a subset of questions will be graded.

Q1. [6 points] For this question, please refer to Fig 9.1 in the book. This figure shows the membership scores computed by fuzzy C-means clusters of a two dimensional data set.
Suppose you defined the core and border points based on the membership assignment of Fuzzy C-Means clustering using the following function:

if $\max_{i \; C}(m_i) >=$ theta, then core point
else border point.

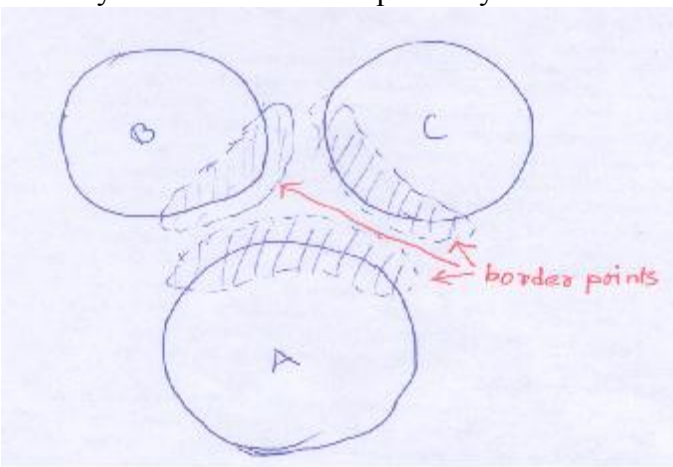
Here m_i is the membership function of a point to i-th cluster and C denotes the total number of cluster. For example, a point with membership scores (0.6,0.3,0.1) will be a border point for theta=0.8 and point with membership scores (0.8,0.1,0.1) will be a core point.

a) Is the above scheme able to find all the core points that are found by DBSCAN for a particular eps, mincount? You can vary theta to find them.

b) Is the above scheme able to find all the border points that are found by DBSCAN for a particular eps, mincount? You can vary theta to find them.

Justify your answer in both cases.

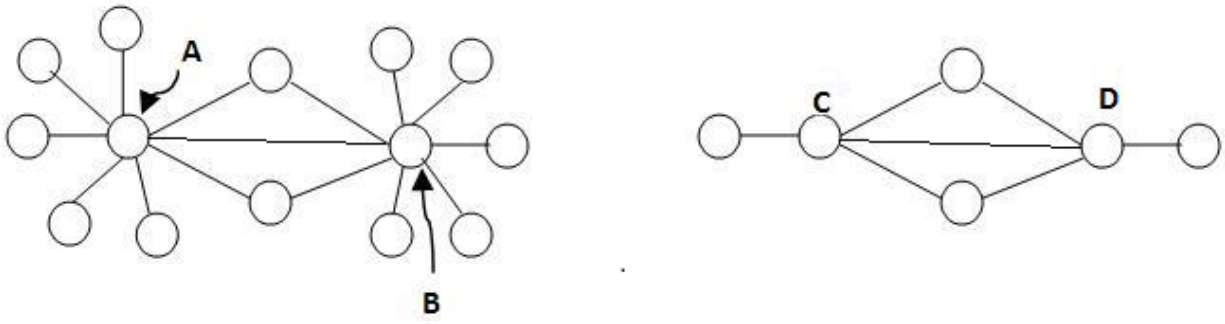
Answer: First, we choose a reasonable theta and mincount. So, all the points that are closer to the centroids of the three circle will be identified as core, and all the points toward the periphery will be identified as border points. If you vary the parameters, the number of core and border points will vary. But, no matter how you choose your parameters for DBSCAN, the question is whether you can identify the core and border points by this naïve scheme for any theta.



a) Core points will be identified by reasonable theta.

b) Some border points will be identified. But, the rest will not be. For example, in the figure, among all the border points you can identify only the points that are close to other clusters(marked by shaded regions). The reason is that the membership score for these points are distributed among three clusters, so the max of them will be less. But, for other points max is 1. So, they cannot be identified.

Q2. [10 points] Suppose you create the following graph based on the connectivity in facebook. Compute the SNN and Jaccard similarity of two pairs (A, B) and (C,D). Among these two pairs, which is more intuitively similar?(4) Does this pair have higher score by SNN also over the other? (2+2) Why or why not? If not, how can you modify alg 9.10 for handling this issue? (2)

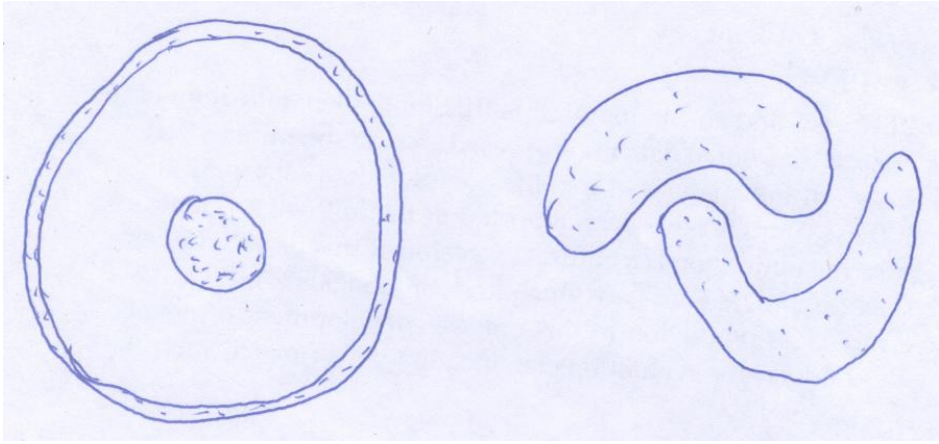


Answer: Jaccard: A and B: 2/6+6+2. C and D: 2/6. SNN: 2 for both.
Intuitively C and D are more similar, since they have less neighbors. On the other hand, A and B have many neighbors(friends). So, having 2 common friends can be more spurious in nature. SNN does not take the individual similarities of these points into account unlike Jaccard.

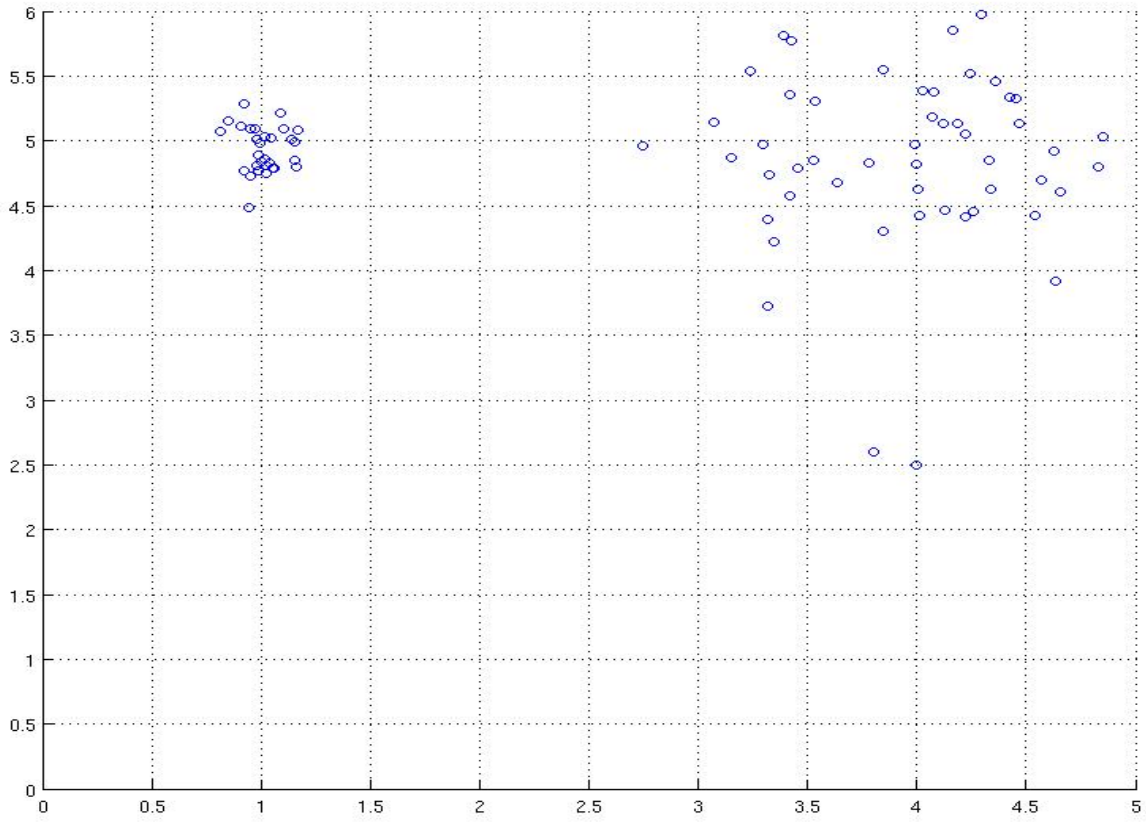
The modified algorithm is can be
if two points x and y are not among the knewarest neighbors of each other then
    similarity(x,y) = 0
else
    similarity(x,y) = number of shared neighbors/total number of points.
end

Q3. [5 points] To obtain a measure of cluster goodness, the silhouette coefficient combines a measure of cluster coherence and a measure of cluster separation. Give an example of a set of clusters for which the clusters are well defined according to one of the standard definitions (e.g., center-based, contiguity-based, density-based), but the silhouette coefficient doesn't work well, i.e., the value of the silhouette coefficient does not indicate a good clustering.

Answer: Silhouette co-efficient is first computed for individual points of a cluster and then they are averaged to get the final score for a cluster. Here, we consider the score for **clusters,** not the individual points(although partial points were given for those cases). In that case, some contiguity-based clusters or the case where one cluster is completely contained by other. Tow such examples are shown below. For first example, the points of B that are closer to A will have poor silhouette co-efficient, although they can be identified by agglomerative and density-based clusters. Similarly, the outer cluster will have poor score in second example.



Q4. **[8 points]** Consider the 2-dimensional data set shown in the figure below. There are 90 points that belong to two well-defined clusters.



1. Which three points would you consider outliers by visual inspection? Mark these points on a printout of the figure.

X	y	Distance to 1st nearest neighbor	Distance to 2nd nearest neighbor	Distance to 10th nearest neighbor	LOF score
1.01	4.86	0.02	0.04	0.120	1.06
1.02	4.81	0.02	0.04	0.110	0.780
0.940	4.48	0.230	0.270	0.360	4.58
4.63	3.92	0.520	0.650	0.900	2.35
4.83	4.80	0.240	0.240	0.660	1.29
4.36	5.46	0.130	0.140	0.430	1.11
4.54	4.43	0.210	0.270	0.520	1.20
3.31	3.73	0.490	0.660	1.13	2.69
4	2.50	0.220	1.41	2.00	3.97
3.80	2.60	0.220	1.23	1.91	3.54

1. Using the following table, state which three points would be considered outliers using the distance to kth-nearest neighbor algorithm (assume k=1,2,10).
   k = 1: (4.63, 3.92), (3.31, 3.73) and (4.83, 4.80)
   k = 2: (3.31, 3.73), (4, 2.50) and (3.80, 2.60)
   k = 10: (4, 2.50), (3.80, 2.60) and (3.31, 3.73)

2. Using the same table, state which three points would be considered outliers using the LOF algorithm. (0.940, 4.48), (4, 2.50) and (3.80, 2.60)

3. Based on your answers to parts 1-3, what can you say about the relative performance of the distance to kth-nearest neighbor and LOF algorithms?
KNN based approach is sensitive to the choice of k. LOF based approach pefroms better as it is a density based approach and it suits the outliers given in the data set.

Q5. (4 points) State one difference and one similarity between clustering and proximity based anomaly detection techniques.

Similarity: Both of them work with the notion of similarity and the similarity could be density, contiguity etc.
Difference: Proximity based anamaly detection treats all points equally when an anamaly score is computed, clustering based techniques use the notion of cluster centroids or a cluster representative to assign scores.

Q6. (8 points) Suppose you came up with a new clustering based anomaly detection technique. You first cluster the data with K=2 using any clustering algorithm and then, you declare all members of the smaller cluster as anomalies.

a) Describe one advantage and one disadvantage of this scheme over the K-Mean based anomaly detection technique in the book.
Advantage: The described approach is suitable when the anamaly points form a smaller cluster, as opposed to being spread around true clusters.
Disadvantage: It is not suitable when the anamaly points are separated out and cannot form a cluster.

b) How can you modify this new algorithm further to compute anomaly scores?
One way is to compute the distance for each point in the smaller cluster from the closest centroid of the other clusters.