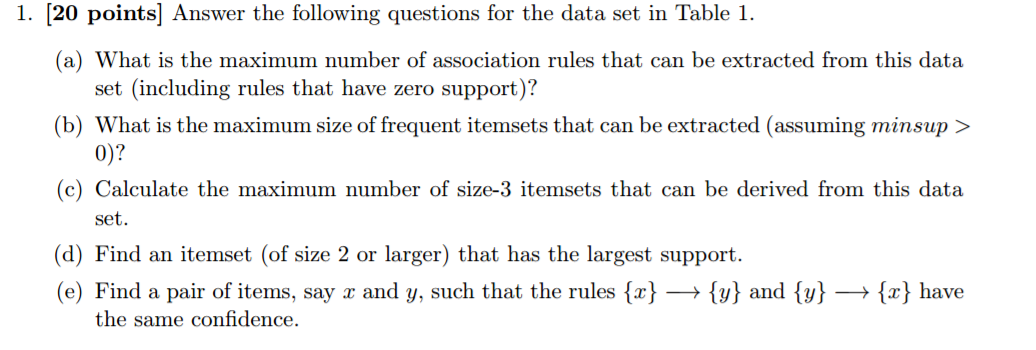
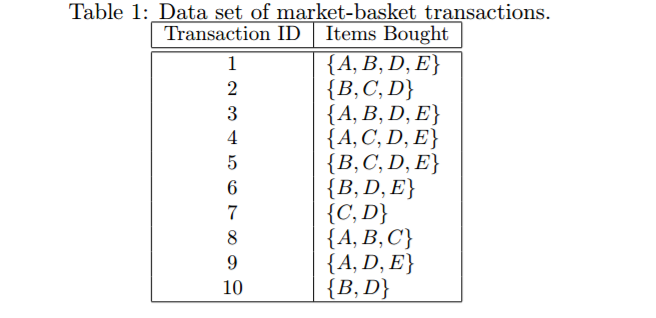
**ASSINGMENT NO- 3 DATA MINING Aparna Paul | RUID: 174009860**

**Question No: 1**





**ANSWER:**

1. D=5

R=3^5-2^ (5+1) +1

R=180

1. Assuming minsup > 0

All the item set qualifies as frequent item set

So, maximum size of frequent item sets is 4

1. 5C3= 5! /3! (5-3)!

= (5\*4\*3\*2\*1)/ {(3\*2\*1) (2\*1)}

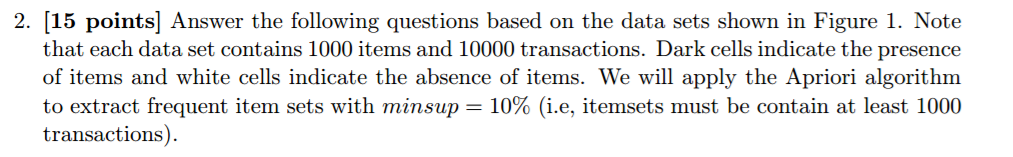
=10

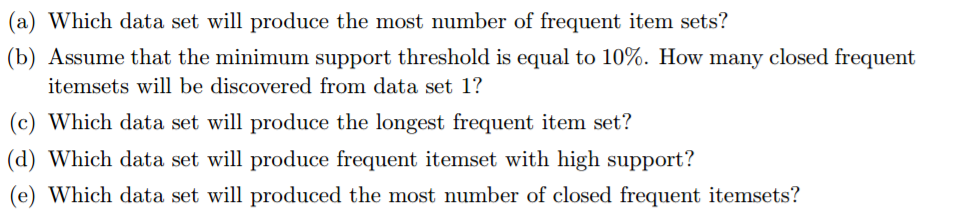
1. {B,D} {D,E}

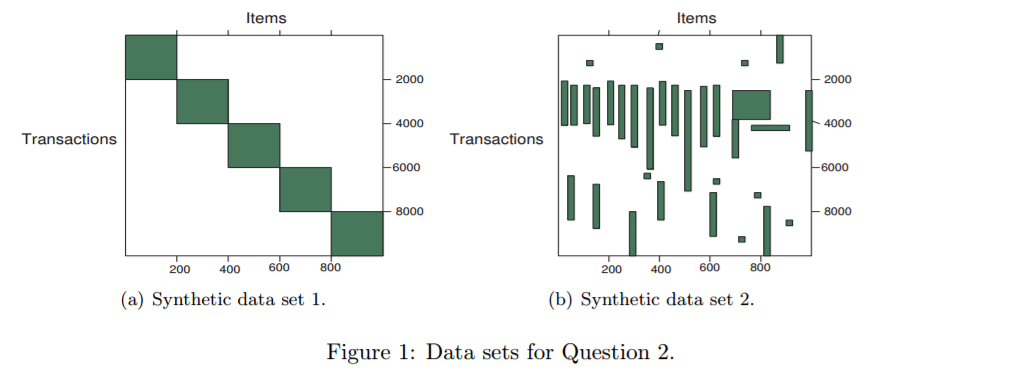
|  |  |
| --- | --- |
| **Item** | **Support** |
| A | 5/10 |
| B | 7/10 |
| C | 5/10 |
| D | 9/10 |
| E | 6/10 |

X=A and Y=C are the pair of items

**Question No: 2**



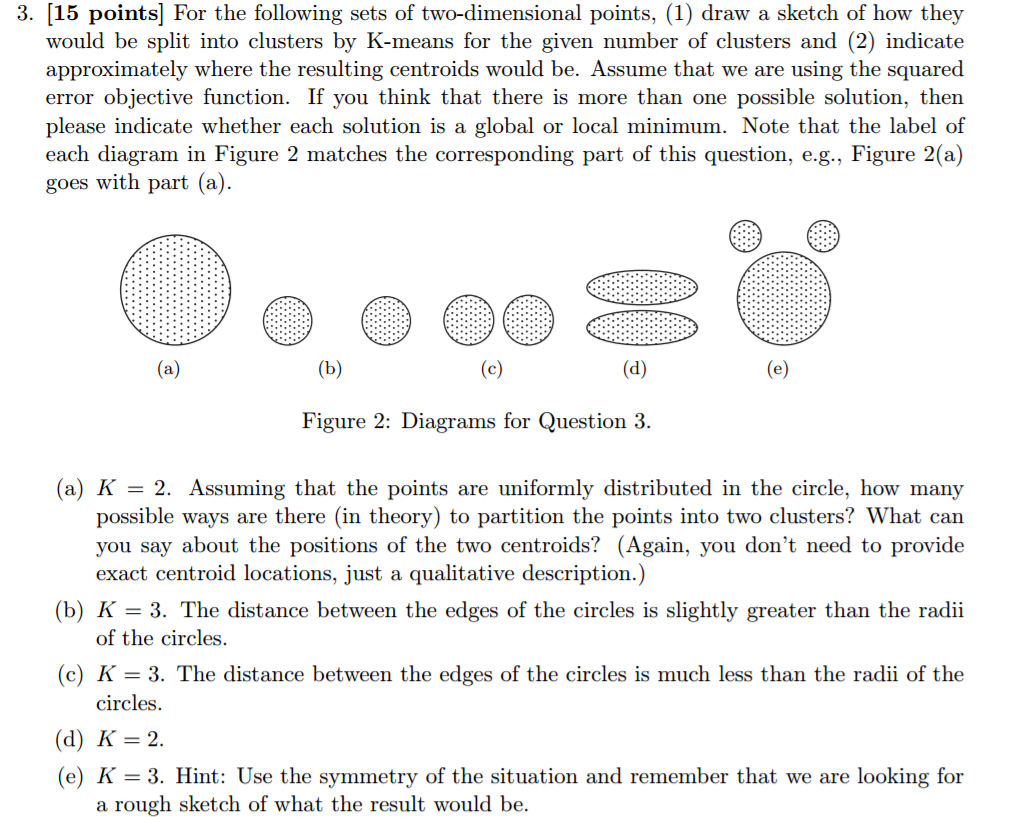




**Answer:**

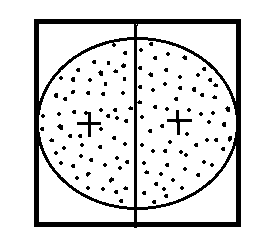
1. Set 2 will produce the most number of frequent item sets compared to the other because larger frequent item set has to be generated along its sub-sets.
2. 5 because frequent superset does not have same support as the subset.
3. Synthetic data set 2 will produce the longest frequent data because it’s more item set and long transactions.
4. Synthetic data set 2; because it has high frequent items with large transaction compared to data set 1
5. Synthetic dat set 2

**Question No: 3**

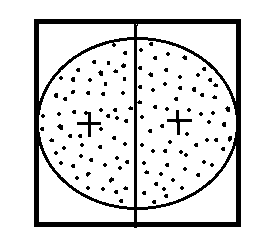


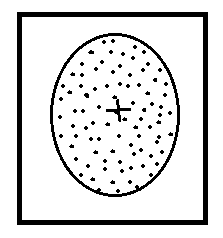
**Answer:**

1. There are an infinite number of ways to part the circle into 2 clusters. We can take any line that bisects the circle. The line can make any angle between (0, 180) x-axis. Centroids will lie on the perpendicular bisector of the line that splits the circle into 2 clusters and will be symmetrically placed. All these solution will have same, globally minimum error.

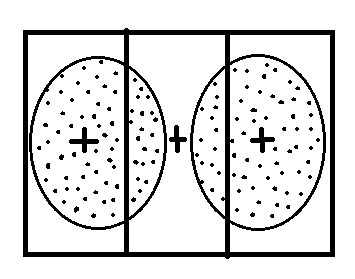
****

1. If we begin with initial centroids that are real points, we would necessarily get particular solution because of the restriction that the circles are more than 1 radius apart from. The bisector could have any angle, as above and it could be the other circle that is split. All these solutions have the same globally minimal error.

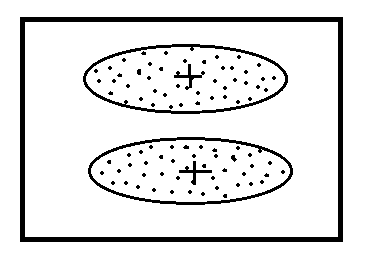
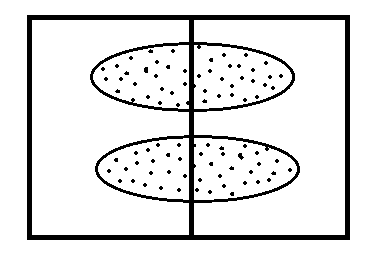




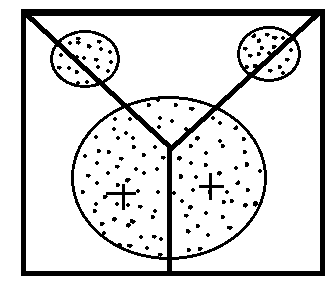
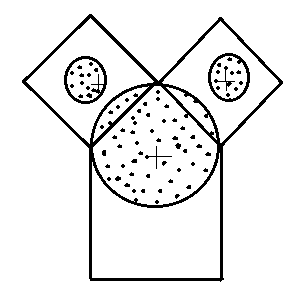
1. 3 boxes show 3 cluster that will result in the realistic case that the initial centroid are actual data points.



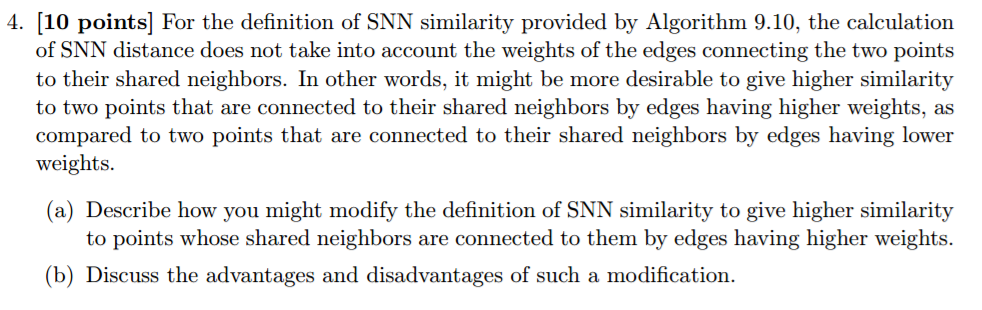
1. In both case, the rectangles show the cluster. In 1st case, the 2 clusters are only a local minimum while in the second case the clusters represent a globally minimal solution.



1. For the solution shown in the below figure the 2 top cluster are enclosed in two boxes, while the 3rd cluster is enclosed by the regions defined by a triangle and a rectangle.( the two smaller clusters in the diagram are supposed to be symmetrical



**Question No: 4**

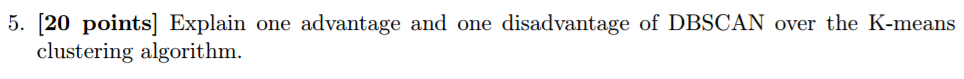


**Answer:**

1. This can be done by taking into account the weights of the edges into the similarity computation. We can add a function like mean or min of the weighs of the edges that connect the two points to this common neighbor. Adding up the value of this function for each common neighbor will give an edge weight-incorporated SNN similarity measure between the two points.
2. **Disadvantage**: The method is very complex and very difficult to determine. It is also computationally expensive to include the weights of the edges.

**Advantage**: Can obtain more accurate result in calculation of similarity measure because along with the number, even the weights will be taken into account. Also, due to the presence of weight reduces the dependence of the result on the number of neighbor.

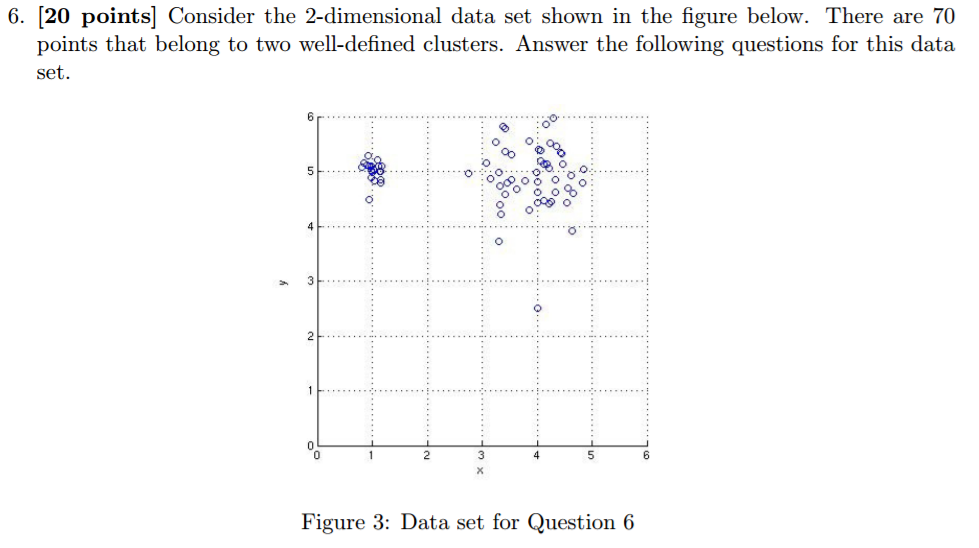
**Question No: 5**

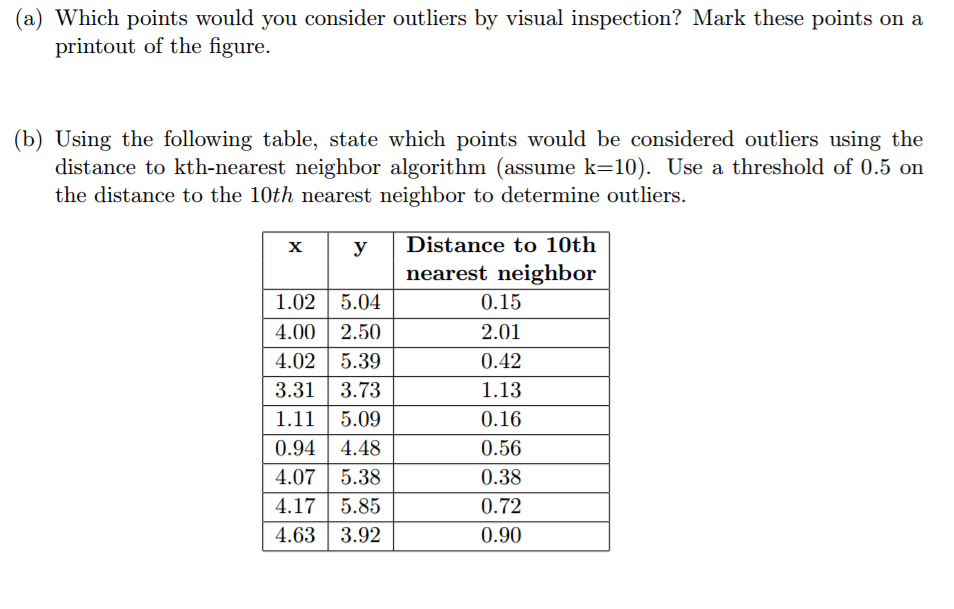


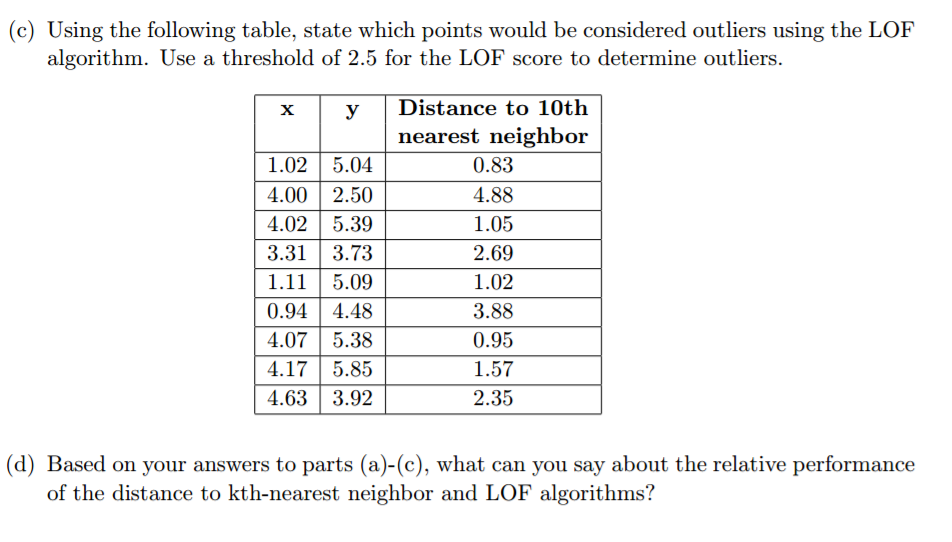
**Answer:**

**Advantage**: DBSCAN handles cluster of arbitrary shapes and sizes much better than k-means because it is based on the density. It also handles noise in an efficient manner. Also, it is not necessary to specify the number of cluster in advance.

**Disadvantage**: Cannot handle cluster of widely varying densities. Also due to this reason, the computational efficiency than k means. Also, not very easy to setup the Eps and Min parameters.

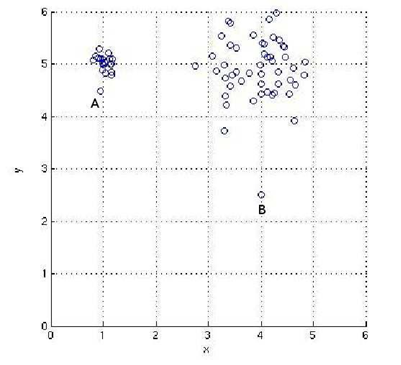
**Question No: 6**





**Answer:**

1. The points A and B is shown in the figure below are the outliers of the data set

****

1. The points (4.00, 2.50), (3.31, 3.73), (0.94, 4.48), (4.17, 5.85) and (4.63, 3.92) are the outliers determined by the kth nearest neighbor algorithm, as their distance to the 10th neighbor is higher than the threshold.
2. The points (4.00, 2.50), (3.31, 3.73) and (0.94, 4.48) are the outliers determined by the LOF algorithm since their LOF score is higher than the threshold.
3. LOF, which is based on local density, made only one false discovery as against three made by the distance-based algorithm. The outlier detection algorithm is based on the distance to the kth-nearest neighbor. As the distance between A and its kth nearest neighbor is less than the distance between some pairs of the points in the larger cluster, the outlier detection algorithm marked more non-outlier points as outliers.