Question 2

Candidate Number: 59069

Data Science Algorithms & Tools - CS3DS19

2. (a) Briefly discuss Cluster Analysis in general and, in particular, two types of clustering: partitional and hierarchical.

(4 marks)

Clustering: Clustering is the operation of grouping a set of data points, in such a way that those grouped together share properties or similarities with their own group, more than any other group.

Partitional: Partitional Clustering is a method of clustering in which the data points are split or “partitioned” into a predefined number of groups of data, before being evaluated against specific criterion.

Hierarchical: Hierarchical Clustering conversely creates a hierarchical decomposition of the dataset through the use of specific criteria. Produces a set of nested clusters as a hierarchical tree.

(b) Compare and contrast one algorithm for partitional clustering and one for hierarchical clustering in terms of advantages (at least two) and disadvantages (at least two), including their computational complexity.

(6 marks)

Chosen Algorithms:

K-means - Partitioning Clustering
AGNES - Hierarchical Clustering

One determining factor between using Hierarchical Clustering or Partition Clustering is the necessity for flexibility or scalability. Hierarchical clustering is *flexible*, but futile when considering large amounts of data. Whereas a Partitioning clustering algorithm such as K-means is *scalable*, but extremely ineffective with flexible data.

K-means clustering also requires a prerequisite knowledge of how many clusters the data should be grouped into, whereas hierarchical methods such as AGNES do not require the number of clusters to be specified prior. Hierarchical clustering is extremely easy to implement.

Due to the nature of K-means clustering, the clusters produced are circular or spherical, depending on the dataset, this can be positive or negative, however it is seldom beneficial.

K-Means Clustering often starts with random cluster centroids, resulting in varying results which are lacking in consistency and repeatability whereas hierarchical clustering will always yield the same clustering results.

K-means also yields a time complexity of $O(I \cdot n \cdot k \cdot d)$ where Irepresents the number of iterations, $n$is the number of data points ,k represents the number of clusters and $d$is the number of attributes or dimensions of the data. Overall this is recognised as a linear time complexity, whereas most hierarchical clustering algorithms including AGNES have an $O(N^2)$time complexity. The result of this is that K-means is much more appropriate for large amounts of data compared to clustering. Whereas hierarchical often performs better on smaller datasets.

(c) Consider the set of 6 data points in 2 dimensions (x,y) in the table in Figure Q2-1. Apply one iteration of the k-means algorithm (for k=2) to find the cluster allocation (C0 or C1) of each data point and the values of the centroids at the end of the iteration (it-1).

Which of the two alternative initialisations of the centroids (c_0 and c_1 at it-0) given below produced the best clustering according to the cost function (SSE) optimised by k-means?

Provide the results in the following page as well as a worked solution (formulas and your arithmetic calculations) to compute the values of the centroids at the end of the iteration (it-1), the cluster allocations and the values of the cost function before and after the iteration (it-0 and it-1).

Find cluster allocation C0/C1. Find Centroids at end of Iteration 1

	Case 1	Case 2
Data_ID	Cluster ID (0/1) after it-1	Cluster ID (0/1) after it-1
0	0	0
1	0	0
2	1	0
3	1	1
4	1	1
5	1	1

This was calculated by calculating the euclidean distance if each data point from the nearest cluster centroid. $\sqrt{(\overline{X} - x)^2 + (\overline{Y} - y)^2}$ where x,y are the coordinates of the datapoint and $\overline{X}, \overline{Y}$ are the coordinates of each cluster centroid respectively.

Once the distance from each centroid had been computed, the datapoint was assigned to the closest cluster.

Which of two cases provide best clustering according to SSE optimised by Kmeans

Centroids

Centroid	Iteration0 X	Iteration0 Y	Iteration1 X	Iteration1 Y
Case 1 Cluster 0	0.2	1	0.2	0.4
Case 1 Cluster 1	0.5	1	0.575	0.4
Case 2 Cluster 0	0.3	1	0.26666666	0.4
Case 2 Cluster 1	0.6	1	0.63333333	0.4

The initial (it0) coordinates were given. The it1 coordinates were calculated as an average (mean) of all the data points assigned to that cluster.

Sum Squared Errors

	Iteration 0	Iteration 1
Case 1 SSE	2.265	0.0825
Case 2 SSE	2.235	0.068333333

The sum of squared errors, is the sum of all the squared euclidean distances of each point to its respective cluster. The smaller this value, the more accurate the clustering is.

According to the cost function (SSE) **Case 2** produced the best initialisation of the centroids, when optimised by k-means.

Best results = Case 2

FOR FURTHER WORKINGS SEE:

https://docs.google.com/spreadsheets/d/1HUa-pfdsb09MAIW3MT7oYBHjmfMI2sTh5eU3xK9iGxU/edit?usp=sharing

data ID	x	y
p_0	0.2	0.3
p_1	0.2	0.5
p_2	0.4	0.4
p_3	0.6	0.4
p_4	0.65	0.3
p_5	0.65	0.5

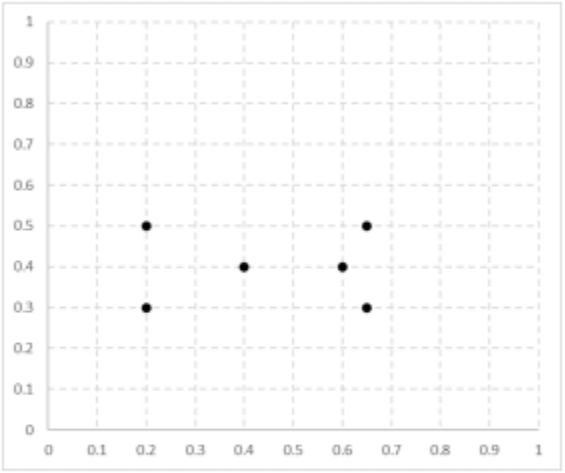

Figure Q2-1. The input data points

Case 1:

Centroid c_0 for Cluster 0:

c_0	x	y
it-0	0.2	1
it-1		

Centroid c_1 for Cluster 1:

c_1	x	y
it-0	0.5	1
it-1		

Cluster allocations:

data ID	Cluster ID (0\1) after it-1
p_0	
p_1	
p_2	
p_3	
p_4	
p_5	

	cost function (SSE) value
it-0	
it-1	

Case 2:

Centroid c_0 for Cluster 0:

c_0	x	y
it-0	0.3	1
it-1		

Centroid c_1 for Cluster 1:

c_1	x	y
it-0	0.6	1
it-1		

Cluster allocations:

data ID	*Cluster ID (0\1) after it-1*
p_0	
p_1	
p_2	
p_3	
p_4	
p_5	

	cost function (SSE) value
it-0	
it-1	

The best clustering solution is given by:

- ☐ Case 1
- ☐ Case 2

(10 marks)