Practice Exercise 6 – Clustering (solution)

Question 3: Use min/single link to perform agglomerative clustering by showing the dendrogram for the data described by the distance matrix below:

	A	B	C	D
A	0	1	4	5
B		0	2	6
C			0	3
D				0

Note: the height of each "junction" in the dendrogram represents the distance between the pair of clusters.

A,B,C,D

@1: (A,B), C,D

@2: ((A,B),C),D

@3: (((A,B),C)),D)

@4: (((A,B),C)),D)

@5: (((A,B),C)),D)

@6: (((A,B),C)),D)

The dendrogram will be drawn accordingly.

Question 4: Repeat the above with max/complete link to perform agglomerative clustering.

A,B,C,D

@1: (A,B), C,D

@2: (A,B),C,D

@3: (A,B),(C,D)

@4: (A,B),(C,D)

@5: (A,B),(C,D)

@6: ((A,B),(C,D))

The dendrogram will be drawn accordingly.

Question 7: What is the goal of clustering, and how it differs from classification?

Clustering aims to distribute data samples into a number of groups such that the labels of these groups are unknown in advance. On the other hand, classification aims to assign each data sample into one of the groups such that the labels of these groups are known in advance.

For doing these, both of these may rely on similar metrics (e.g. distance metrics) or could employ entirely different metrics (e.g. as in the case of SVM).

Question 8: Describe in what situation the conventional k-means algorithm would fail to cluster the data. Can you suggest a modification to overcome the problem?

K-means expects to determine means of data groups and then assign each data sample to its nearest cluster on the basis of distance of this sample to the cluster mean. The algorithm fails to work in cases, where for example, the means of different groups are lying very close to each other (or overlapping). In such cases, kernel k-means algorithm may be used which implicitly transforms the data such that means may not be overlapping.