COMP30120 Tutorial

Clustering

Q1

(a) The data set below contains 10 items represented by 4 numeric features.

Item	Feature1	Feature2	Feature3	Feature4
x1	5.1	3.8	1.6	0.2
x2	4.6	3.2	1.4	0.2
x3	5.3	3.7	1.5	0.2
x4	5	3.3	1.4	0.2
x5	7	3.2	4.7	1.4
x6	6.4	3.2	4.5	1.5
x7	6.9	3.1	4.9	1.5
x8	5.5	2.3	4	1.3
x9	6.5	2.8	4.6	1.5
x10	5.7	2.8	4.5	1.3

These items have been randomly assigned to two clusters in order to initialise the *k*-Means algorithm. The assignments are as follows:

$$C1 = \{ x1, x3, x7, x8 \} \quad C2 = \{ x2, x4, x5, x6, x9, x10 \}$$

Based on the data and cluster assignments, calculate the centroid vector for each cluster.

(b) Based on the centroids calculated above, which clusters will the items *x1* and *x10* next be assigned to? Calculate distances using the Euclidean distance measure.

Q2

(a) Describe the difference between the *single-linkage, complete* and *average linkages*, which are used as cluster metrics in Agglomerative Hierarchical Clustering.

(b) Calculate the distances between *x2* and *C1* using *single*, *complete* and *average linkage* for the data below, if the cluster *C1 = {x1, x3}.* Assume that distances between items are calculated using Euclidean distance.

Item	Feature1	Feature2
x1	1.3	1.5
x2	0.5	2.4
x3	0.0	3.0

Q3

The following table depicts a pairwise distance matrix for 5 items:

	x1	x2	x3	x4	x5
x1	0				
x2	2	0			
x3	6	5	0		
x4	10	9	4	0	
x5	9	8	5	3	0

Calculate the dendrogram representing the agglomerative hierarchical clustering of these items based on the single-linkage method. The answer should illustrate the distance matrices originating from each clustering step.

Q4

In Weka, apply *k*-Means with Euclidean distance to the *Iris* ARFF dataset provided on the course Moodle page.

Report the *Within cluster sum of squared errors* (SSE) for runs with different numbers of clusters: *k=2*, *k=3* and *k=4*.