

COMP47460 Tutorial

Clustering

1.

- (a) The dataset below contains 10 examples, each represented by 4 numeric features.

Item	f1	f2	f3	f4
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 examples 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 in part (a), which clusters will the examples *x1* and *x10* next be assigned to? Calculate distances using the Euclidean distance measure.

2.

The table below shows three examples represented by 2 numeric features.

Example	f1	f2
x1	1.3	1.5
x2	0.5	2.4
x3	0.0	3.0

If the cluster $C1 = \{x1, x3\}$, use the Euclidean distance measure to calculate the distances between the example $x2$ and cluster $C1$ based on *single*, *complete*, and *average linkage*.

3.

The following table depicts a symmetric distance matrix for 5 examples:

	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 examples based on the single-linkage method. The answer should illustrate the distance matrices originating from each clustering step.

4.

In Weka, apply k -Means with Euclidean distance to the *Iris* ARFF dataset.

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

Repeat the above process again, but change the random seed parameter for k -Means. Are the SSE scores identical? If not, explain why not.