2. **Describe the goal of all clustering methods.**

   The primary objective of Clustering is to divide the entire data set into relatively homogeneous subgroups or clusters such that similarity of observations within a cluster is maximized and at the same time similarity of observations between clusters is minimized.

3. **Suppose that we have the following data (one variable). Use single linkage to identify the clusters. Data: 0 0 1 3 3 6 7 9 10 10**

   The agglomerative clustering method that starts by placing each observation into its own sub-cluster, and then progressively agglomerates each observation based on the desired linkage rule. The ***single-linkage*** approach will agglomerate the observation or sub-cluster in question to the sub-cluster having the shortest distance to ***any*** observation in another sub-cluster.

   The steps are described below as follows:

   - Step 1 – All ten observations are placed in their own clusters.

   - Steps 2, 3, and 4 – The algorithm searches the distance matrix for the closest two clusters to merge. There are actually three candidates {0, 0}, {3, 3}, and {10, 10} all having a distance of zero. The algorithm selects them in some arbitrary order. At the end of Step 4, the clusters are then formed as **{0, 0}, {1}, {3, 3}, {6}, {7}, {9},** and **{10, 10}**.

   - Steps 5, 6, and 7 – The algorithm searches the distance matrix for the closest two clusters to merge and there are actually three candidates {0, 0, 1}, {6, 7}, and {9, 10, 10} all having a distance of 1. The algorithm selects them in some arbitrary order. At the end of Step 4, the clusters are then formed as **{0, 0, 1}, {3, 3}, {6, 7},** and **{9, 10, 10}.**

   - Steps 8 and 9 – The algorithm searches the distance matrix for the closest two clusters to merge. There are two candidates {0,0,1,3,3} and {6,7,9,10,10} each having a distance of 2. The algorithm selects them in some arbitrary order. At the end of Step 9, the clusters are then formed as **{0,0,1,3,3}** and **{6,7,9,10,10}.**

Discovering Knowledge In Data: An Introduction to Data Mining, By Daniel T. Larose & Chantal D. Larose
Copyright © 2014 John Wiley & Sons, Inc.

- Step 10 – The algorithm searches the distance matrix for the closest two clusters to merge. Since there are only two clusters, the algorithm merges them into a single cluster **{0, 0, 1, 3, 3, 6, 7, 9, 10, 10}** and then ends.

4. Suppose that we have the following data (one variable). Use complete linkage to identify the clusters. Data: 0 0 1 3 3 6 7 9 10 10

The agglomerative clustering method that starts by placing each observation into its own cluster, and then progressively agglomerates each observation based on the desired linkage rule. The ***complete-linkage*** approach will agglomerate the observation or sub-cluster in question to the sub-cluster having the shortest distance to the ***most distant*** observation in another sub-cluster.

The steps are described below as follows:

- Step 1 - All ten observations are placed in their own clusters.

- Steps 2, 3, and 4 – The algorithm searches the distance matrix for the closest two clusters to merge. There are actually three candidates {0, 0}, {3, 3}, and {10, 10} all having a distance of zero. The algorithm selects them in some arbitrary order. At the end of Step 4, the clusters are then formed as **{0, 0}, {1}, {3, 3}, {6}, {7}, {9},** and **{10, 10}.**

- Steps 5, 6, and 7 – The algorithm searches the distance matrix for the closest two clusters to merge based on the most distant points in each cluster and there are actually three candidates {0, 0, 1}, {6, 7}, and {9, 10, 10} all having a distance of 1. The algorithm selects them in some arbitrary order. At the end of Step 4, the clusters are then formed as **{0, 0, 1}, {3, 3}, {6, 7},** and **{9, 10, 10}**.

- Steps 8 and 9 – The algorithm searches the distance matrix for the closest two clusters to merge. There are two candidates {0,0,1,3,3} and {6,7,9,10,10} each having a distance of 3. The algorithm selects them in some arbitrary order. At the end of Step 9, the clusters are then formed as **{0,0,1,3,3}** and **{6,7,9,10,10}**. Note that although the result is identical to the single-linkage approach, the distance measure calculated using single-linkage was 2 whereas it is 3 using complete linkage.

- Step 10 - The algorithm searches the distance matrix for the closest two clusters to merge. Since there are only two clusters, the algorithm merges them into a single cluster **{0, 0, 1, 3, 3, 6, 7, 9, 10, 10}** and then ends.

Discovering Knowledge In Data: An Introduction to Data Mining, By Daniel T. Larose & Chantal D. Larose
Copyright © 2014 John Wiley & Sons, Inc.

5. What is an intuitive idea for the meaning of the centroid of a cluster?

The centroid of a cluster can be envisioned as its geometric center. In other words, we can envision a cluster of observations say in R^3 as a point cloud, and the center of that cloud would be the cluster centroid.

6. Suppose that we have the following data:

a	b	c	d	e	f	g	h	i	j
(2,0)	(1,2)	(2,2)	(3,2)	(2,3)	(3,3)	(2,4)	(3,4)	(4,4)	(3,5)

Identify the cluster by applying the k-means algorithm, with k = 2. Try using initial cluster centers as far apart as possible.

Using points a and j as the initial cluster centers, the figure below illustrates how the K-means algorithm clusters the data points in two passes.

K-Means K = 2	Pass 1		
	Distance (2,0)	Distance (3,5)	Assignment
a (2,0)	0	5.099	C1
b(1,2)	2.2361	3.6056	C1
c(2,2)	2	3	C1
d(3,2)	2.3607	3	C1
e(2,3)	3	2.23607	C2
f(3,3)	3.1623	2	C2
g(2,4)	4	1.4142	C2
h(3,4)	4.1231	1	C2
i(4,4)	4.4721	1.4142	C2
j(3.5)	5.099	0	C2

K-Means K = 2	Pass 2		
	Distance (2,1.5)	Distance (2.833,3.833)	Assignment
a (2,0)	1.5	3.9229	C1
b(1,2)	1.118	2.5927	C1
c(2,2)	0.5	2.0138	C1
d(3,2)	1.118	1.8409	C1
e(2,3)	1.5	1.1785	C2
f(3,3)	1.8028	0.8498	C2
g(2,4)	2.5	0.8498	C2
h(3,4)	2.6926	0.2357	C2
i(4,4)	3.2012	1.1785	C2
j(3.5)	3.7401	1.1785	C2

Figure 6.1. K-means clustering in two passes

From the figure above, we observe that the algorithm completes in two passes. The algorithm for this exercise is initialized to form two clusters using points a at (2,0) and j at (3,5) as the initial cluster centers. For each point in the data set, the algorithm calculates the Euclidean distance to each initial cluster center, and then assigns that point to the cluster with the shortest distance. The cluster assignments for Pass 1 define Cluster 1 as containing points **{a, b, c, d}** and Cluster 2 as containing points **{e, f, g, h, i, j}.**

The centroids are adjusted using these new cluster assignments. In Pass 2, for each point in the data set, the algorithm calculates the Euclidean distance to each adjusted centroid and then reassigns that point to the cluster with the shortest distance. However, in Pass 2 no new cluster assignments are made. Therefore, the centroids do not change and execution of the algorithm completes.

Discovering Knowledge In Data: An Introduction to Data Mining, By Daniel T. Larose & Chantal D. Larose
Copyright © 2014 John Wiley & Sons, Inc.