UNIVERSITY *of* STIRLING

Artificial Intelligence (CSC9YE)
Machine Learning: Clustering[1]

Gabriela Ochoa
goc@cs.stir.ac.uk

[1] Based on a Lecture by Dr. Nadarajen Veerapen

Overview

Clustering

Unsupervised Learning

Supervised vs Unsupervised Learning

- In the last couple of lectures, you looked at machine learning in general and learnt about supervised learning in particular: regression and classification. In that context, we are given a set of features about each object as well as an outcome variable. The objective is to predict the outcome based on the features.
- Today, we focus on unsupervised learning where we only observe the features but we are not provided with any outcome variable. We'll look at K-means and hierarchical clustering.

Unsupervised Learning

The Goals of Unsupervised Learning

- We want to find interesting things about a set of data. Is there an informative way to visualize the data? Can we discover subgroups among the variables or among the observations?
- This means grouping and separating data points at the same time.
- We need a way to measure how (dis)similar the data points are: distance.

Applications of Clustering

- Market Segmentation
  - Suppose we have access to a large number of measurements (e.g. median household income, occupation, distance from nearest urban area, and so forth) for a large number of people.
  - Our goal is to perform market segmentation by identifying subgroups of people who might be more receptive to a particular form of advertising, or more likely to purchase a particular product.
  - The task of performing market segmentation amounts to clustering the people in the data set.
- Internet and the Web
  - Document classification
  - Cluster Weblog data to discover groups of similar access patterns
  - Pattern recognition
- Image processing
  - Astronomy – aggregation of stars, galaxies, or super-galaxies
  - Medicine – separating healthy from diseased tissue

Distance

Euclidean Distance

- In the 2D plane, the Euclidean distance between $p_1 = (x_1, y_1)$ and $p_2 = (x_2, y_2)$ is given by the Pythagoras theorem:

$$d(p_1, p_2) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

- In 3D, the Euclidean distance between (x_1, y_1, z_1) and (x_2, y_2, z_2) is given by the Pythagoras theorem:

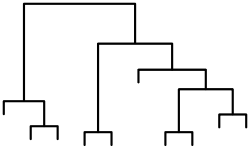
$$d(p_1, p_2) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2}$$

- In general, the distance between points $\boldsymbol{x}$ and $\boldsymbol{y}$ in $\mathbb{R}^n$ (n dimensions):

$$d(\boldsymbol{x}, \boldsymbol{y}) = |\boldsymbol{x} - \boldsymbol{y}| = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

Two Clustering Methods

- In K-means clustering, we seek to partition the observations into a pre-specified number of clusters.
- In hierarchical clustering, we do not know in advance how many clusters we want; in fact, we end up with a tree-like visual representation of the observations, called a dendrogram, that allows us to view at once the clusterings obtained for each possible number of clusters, from 1 to n.



K-means: An Optimisation Problem

- The idea behind K-means clustering is that a good clustering is one for which the within-cluster variation is as small as possible.
- The within-cluster variation for cluster C_k is a measure $WCV(C_k)$ of the amount by which the observations within a cluster differ from each other.
- Hence we want to solve the problem

$$\text{minimize}_{C_1,\ldots,C_K} \left(\sum_{k=1}^{K} WCV(C_k) \right) \tag{1}$$

- In words, this formula says that we want to partition the observations into K clusters such that the total within-cluster variation, summed over all K clusters, is as small as possible.

K-means: An Optimisation Problem

- Typically we use the Euclidean distance

$$WCV(C_k) = \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 \qquad (2)$$

  where $|C_k|$ denotes the number of observations in the k^{th} cluster.

- Combining (1) and (2) gives the optimization problem that defines K-means clustering,

$$\text{minimize}_{C_1,\ldots,C_K} \left(\sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 \right) \qquad (3)$$

K-means Clustering

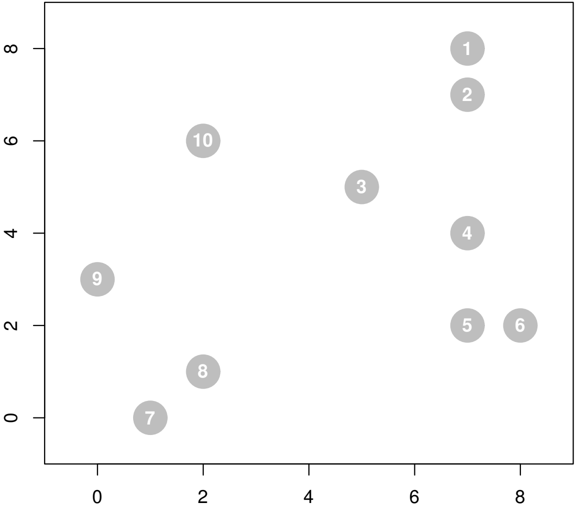
1. Randomly select k points. These serve as initial cluster centroids for the observations.
2. Assign each observation to the cluster whose centroid is closest.
3. Iterate until the cluster assignments stop changing:
    3.1 For each of the k clusters, compute the cluster centroid. The k^{th} cluster centroid is the vector of the p feature means for the observations in the k^{th} cluster.
    3.2 Assign each observation to the cluster whose centroid is closest.

Notes:

- The notion of *closest* is usually defined using the Euclidean distance.
- Any ties in the assignment of an observation to a cluster should be broken deterministically to avoid looping. Example: assign to the cluster with lowest index.

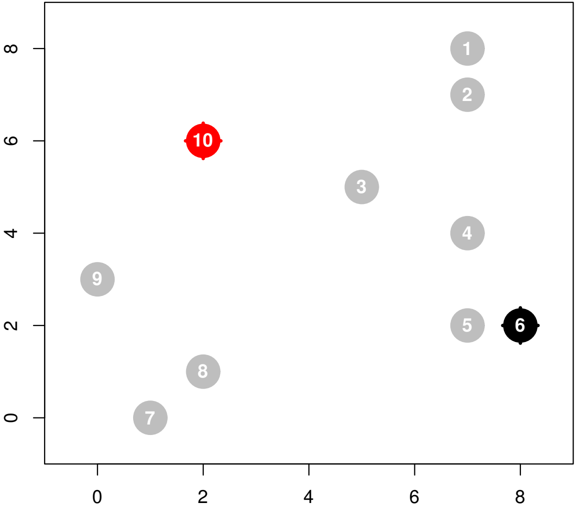
K-means Algorithm

K-means with k=2



K-means Algorithm

Randomly choose centroids

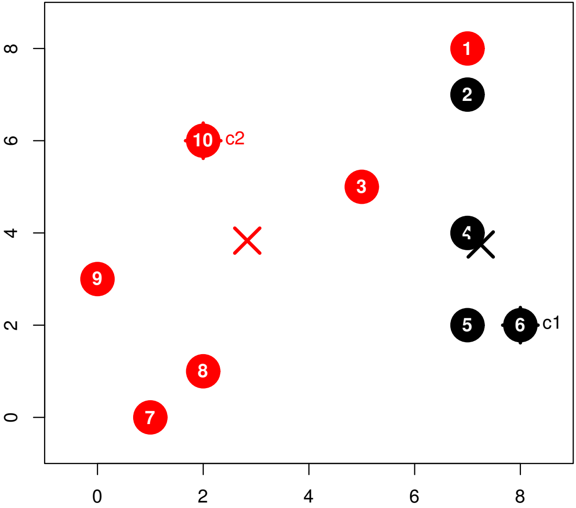


Distances

	c1	c2
1	6.08	5.39
2	5.10	5.10
3	4.24	3.16
4	2.24	5.39
5	1.00	6.40
6	0.00	7.21
7	7.28	6.08
8	6.08	5.00
9	8.06	3.61
10	7.21	0.00

K-means Algorithm

Assign points to clusters

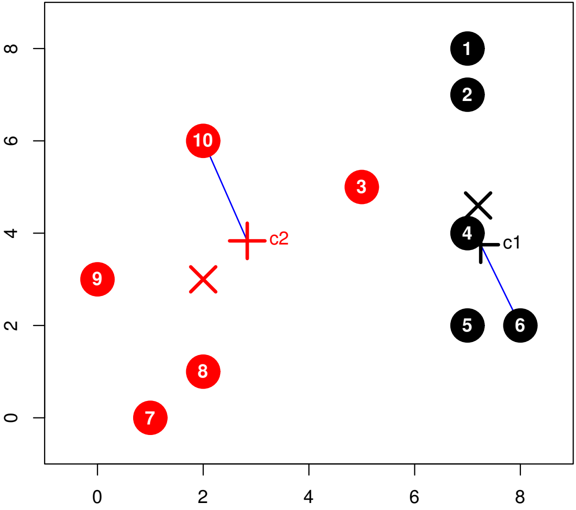


Distances

	c1	c2
1	6.08	$\boxed{5.39}$
2	$\boxed{5.10}$	5.10
3	4.24	$\boxed{3.16}$
4	$\boxed{2.24}$	5.39
5	$\boxed{1.00}$	6.40
6	$\boxed{0.00}$	7.21
7	7.28	$\boxed{6.08}$
8	6.08	$\boxed{5.00}$
9	8.06	$\boxed{3.61}$
10	7.21	$\boxed{0.00}$

K-means Algorithm

Iteration 1

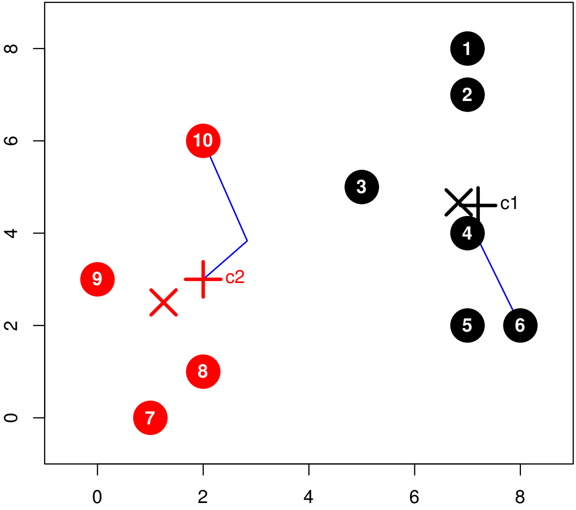


Distances

	c1	c2
1	$\boxed{4.26}$	5.89
2	$\boxed{3.26}$	5.23
3	2.57	$\boxed{2.46}$
4	$\boxed{0.35}$	4.17
5	$\boxed{1.77}$	4.55
6	$\boxed{1.90}$	5.48
7	7.29	$\boxed{4.25}$
8	5.93	$\boxed{2.95}$
9	7.29	$\boxed{2.95}$
10	5.71	$\boxed{2.32}$

K-means Algorithm

Iteration 2

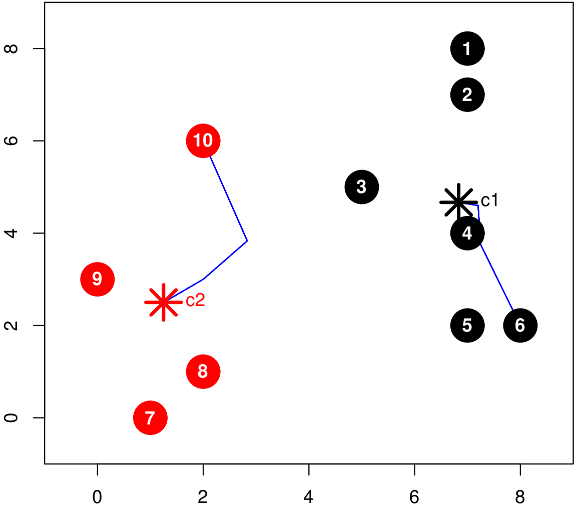


Distances

	c1	c2
1	$\boxed{3.41}$	7.07
2	$\boxed{2.41}$	6.40
3	$\boxed{2.24}$	3.61
4	$\boxed{0.63}$	5.10
5	$\boxed{2.61}$	5.10
6	$\boxed{2.72}$	6.08
7	7.72	$\boxed{3.16}$
8	6.32	$\boxed{2.00}$
9	7.38	$\boxed{2.00}$
10	5.39	$\boxed{3.00}$

K-means Algorithm

Iteration 3: no change in centroids



Distances

	c1	c2
1	$\boxed{3.34}$	7.96
2	$\boxed{2.34}$	7.30
3	$\boxed{1.86}$	4.51
4	$\boxed{0.69}$	5.94
5	$\boxed{2.67}$	5.77
6	$\boxed{2.91}$	6.77
7	7.47	$\boxed{2.51}$
8	6.07	$\boxed{1.68}$
9	7.03	$\boxed{1.35}$
10	5.01	$\boxed{3.58}$

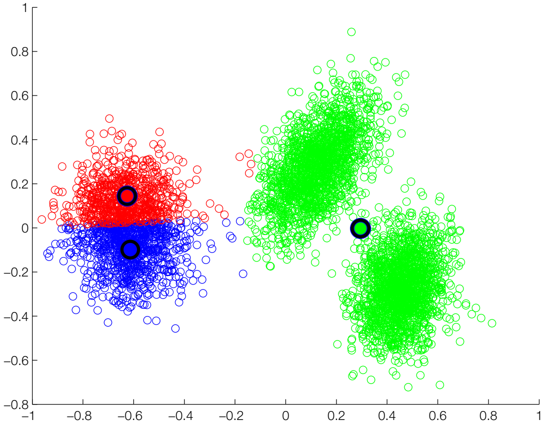
Properties of the Algorithm

- This algorithm is guaranteed to decrease the value of the objective (3).
- However it is not guaranteed to give the global minimum.

Properties of the Algorithm

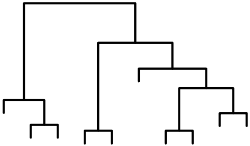
- This algorithm is guaranteed to decrease the value of the objective (3).
- However it is not guaranteed to give the global minimum.
- The algorithm may get stuck in a local optimum.

Local Optimum



Hierarchical Clustering

- K-means clustering requires us to pre-specify the number of clusters K. This can be a disadvantage.
- Hierarchical clustering is an alternative approach which does not require that we commit to a particular choice of K.
- Here, we describe bottom-up or agglomerative clustering. This is the most common type of hierarchical clustering, and refers to the fact that a dendrogram (a tree) is built starting from the leaves and combining clusters up to the trunk.

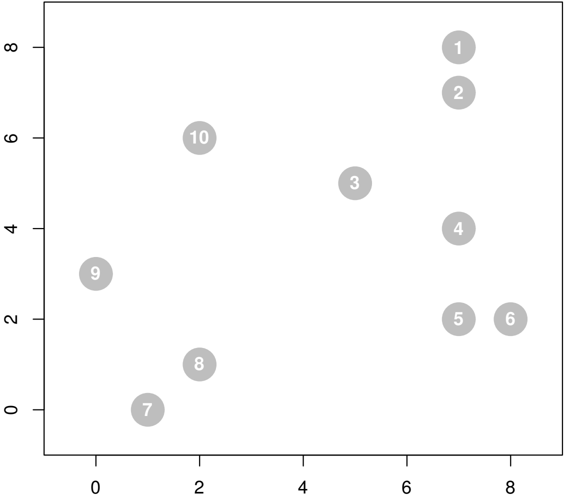


Hierarchical Clustering Algorithm

- Start with each point in its own cluster.
- Identify the closest two clusters and merge them.
- Repeat.
- Ends when all points are in a single cluster.

Hierarchical Clustering

Example using Single Linkage: minimal inter-cluster difference



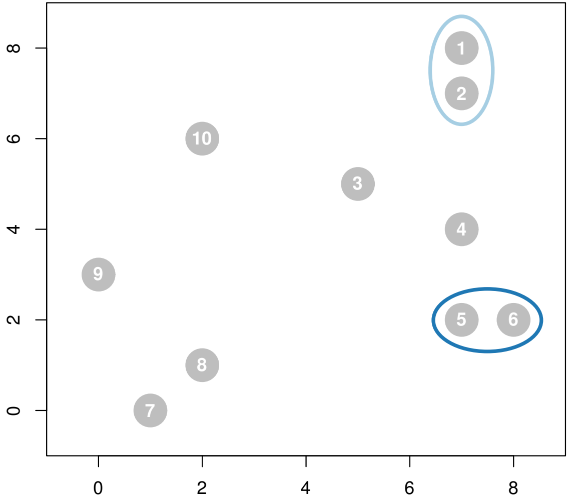
Hierarchical Clustering

Distance Matrix

	1	2	3	4	5	6	7	8	9	10
1	0.0									
2	1.0	0.0								
3	3.6	2.8	0.0							
4	4.0	3.0	2.2	0.0						
5	6.0	5.0	3.6	2.0	0.0					
6	6.1	5.1	4.2	2.2	1.0	0.0				
7	10.0	9.2	6.4	7.2	6.3	7.3	0.0			
8	8.6	7.8	5.0	5.8	5.1	6.1	1.4	0.0		
9	8.6	8.1	5.4	7.1	7.1	8.1	3.2	2.8	0.0	
10	5.4	5.1	3.2	5.4	6.4	7.2	6.1	5.0	3.6	0.0

Clusters: {1}, {2}, {3}, {4}, {5}, {6}, {7}, {8}, {9}, {10}

Hierarchical Clustering



Hierarchical Clustering

d	k	Clusters	Comment
0.0	10	$\{1\}$, $\{2\}$, $\{3\}$, $\{4\}$, $\{5\}$, $\{6\}$, $\{7\}$, $\{8\}$, $\{9\}$, $\{10\}$	Start with each observation as one cluster.
1.0	8	$\{1, 2\}$, $\{3\}$, $\{4\}$, $\{5, 6\}$, $\{7\}$, $\{8\}$, $\{9\}$, $\{10\}$	Merge $\{1\}$ and $\{2\}$ as well as $\{5\}$ and $\{6\}$ since they are the closest: d(1,2)=1 and d(5,6)=1

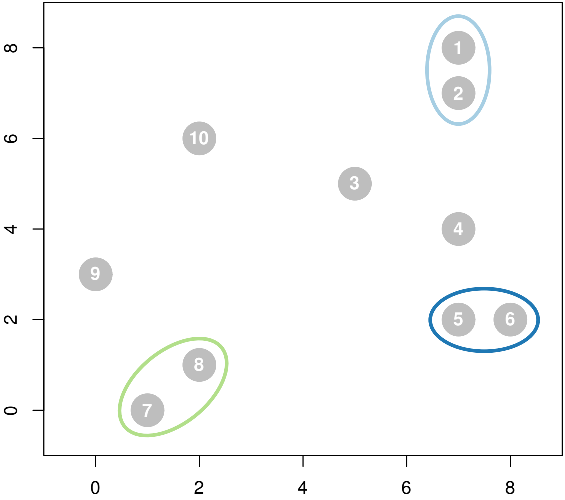
Hierarchical Clustering

Distance Matrix

	1	2	3	4	5	6	7	8	9	10
1	0.0									
2	1.0	0.0								
3	3.6	2.8	0.0							
4	4.0	3.0	2.2	0.0						
5	6.0	5.0	3.6	2.0	0.0					
6	6.1	5.1	4.2	2.2	1.0	0.0				
7	10.0	9.2	6.4	7.2	6.3	7.3	0.0			
8	8.6	7.8	5.0	5.8	5.1	6.1	1.4	0.0		
9	8.6	8.1	5.4	7.1	7.1	8.1	3.2	2.8	0.0	
10	5.4	5.1	3.2	5.4	6.4	7.2	6.1	5.0	3.6	0.0

Clusters: {1, 2}, {3}, {4}, {5, 6}, {7}, {8}, {9}, {10}

Hierarchical Clustering



Hierarchical Clustering

d	k	Clusters	Comment
0.0	10	{1}, {2}, {3}, {4}, {5}, {6}, {7}, {8}, {9}, {10}	Start with each observation as one cluster.
1.0	8	{1, 2}, {3}, {4}, {5, 6}, {7}, {8}, {9}, {10}	Merge {1} and {2} as well as {5} and {6} since they are the closest: d(1,2)=1 and d(5,6)=1
1.4	7	{1, 2}, {3}, {4}, {5, 6}, {7, 8}, {9}, {10}	Merge {7} and {8} since they are the closest: d(7,8)=1.4

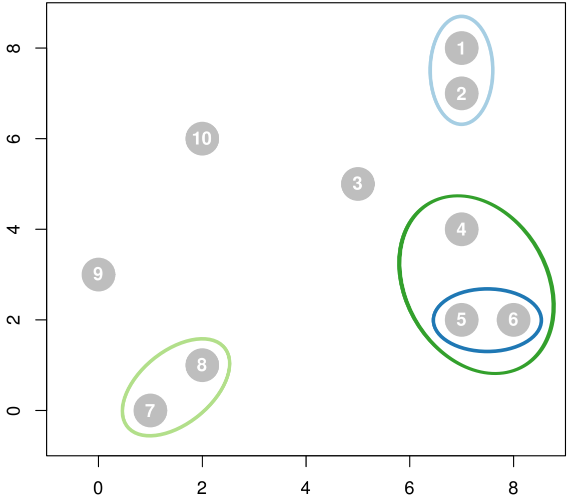
Hierarchical Clustering

Distance Matrix

	1	2	3	4	5	6	7	8	9	10
1	0.0									
2	1.0	0.0								
3	3.6	2.8	0.0							
4	4.0	3.0	2.2	0.0						
5	6.0	5.0	3.6	2.0	0.0					
6	6.1	5.1	4.2	2.2	1.0	0.0				
7	10.0	9.2	6.4	7.2	6.3	7.3	0.0			
8	8.6	7.8	5.0	5.8	5.1	6.1	1.4	0.0		
9	8.6	8.1	5.4	7.1	7.1	8.1	3.2	2.8	0.0	
10	5.4	5.1	3.2	5.4	6.4	7.2	6.1	5.0	3.6	0.0

Clusters: {1, 2}, {3}, {4}, {5, 6}, {7, 8}, {9}, {10}

Hierarchical Clustering



Hierarchical Clustering

d	k	Clusters	Comment
0.0	10	{1}, {2}, {3}, {4}, {5}, {6}, {7}, {8}, {9}, {10}	Start with each observation as one cluster.
1.0	8	{1, 2}, {3}, {4}, {5, 6}, {7}, {8}, {9}, {10}	Merge {1} and {2} as well as {5} and {6} since they are the closest: d(1,2)=1 and d(5,6)=1
1.4	7	{1, 2}, {3}, {4}, {5, 6}, {7, 8}, {9}, {10}	Merge {7} and {8} since they are the closest: d(7,8)=1.4
2.0	6	{1, 2}, {3}, {4, 5, 6}, {7, 8}, {9}, {10}	Merge {4} and {5, 6} since 4 and 5 are the closest: d(4,5)=2.0

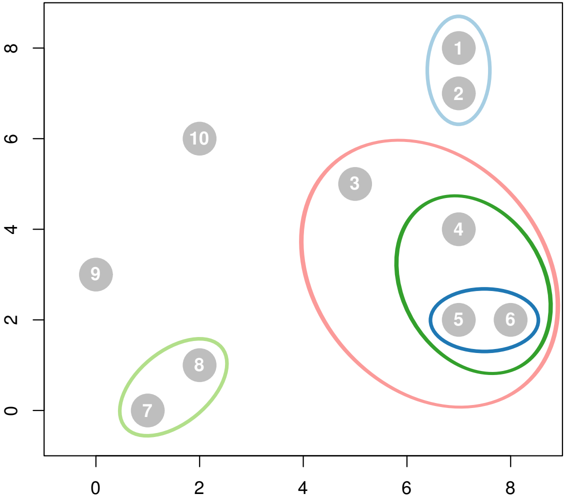
Hierarchical Clustering

Distance Matrix

	1	2	3	4	5	6	7	8	9	10
1	0.0									
2	1.0	0.0								
3	3.6	2.8	0.0							
4	4.0	3.0	2.2	0.0						
5	6.0	5.0	3.6	2.0	0.0					
6	6.1	5.1	4.2	2.2	1.0	0.0				
7	10.0	9.2	6.4	7.2	6.3	7.3	0.0			
8	8.6	7.8	5.0	5.8	5.1	6.1	1.4	0.0		
9	8.6	8.1	5.4	7.1	7.1	8.1	3.2	2.8	0.0	
10	5.4	5.1	3.2	5.4	6.4	7.2	6.1	5.0	3.6	0.0

Clusters: $\{1, 2\}, \{3\}, \{4, 5, 6\}, \{7, 8\}, \{9\}, \{10\}$

Hierarchical Clustering



Hierarchical Clustering

d	k	Clusters	Comment
0.0	10	{1}, {2}, {3}, {4}, {5}, {6}, {7}, {8}, {9}, {10}	Start with each observation as one cluster.
1.0	8	{1, 2}, {3}, {4}, {5, 6}, {7}, {8}, {9}, {10}	Merge {1} and {2} as well as {5} and {6} since they are the closest: d(1,2)=1 and d(5,6)=1
1.4	7	{1, 2}, {3}, {4}, {5, 6}, {7, 8}, {9}, {10}	Merge {7} and {8} since they are the closest: d(7,8)=1.4
2.0	6	{1, 2}, {3}, {4, 5, 6}, {7, 8}, {9}, {10}	Merge {4} and {5, 6} since 4 and 5 are the closest: d(4,5)=2.0
2.2	5	{1, 2}, {3, 4, 5, 6}, {7, 8}, {9}, {10}	Merge {3} and {4, 5, 6} since 3 and 4 are the closest: d(3,4)=2.2

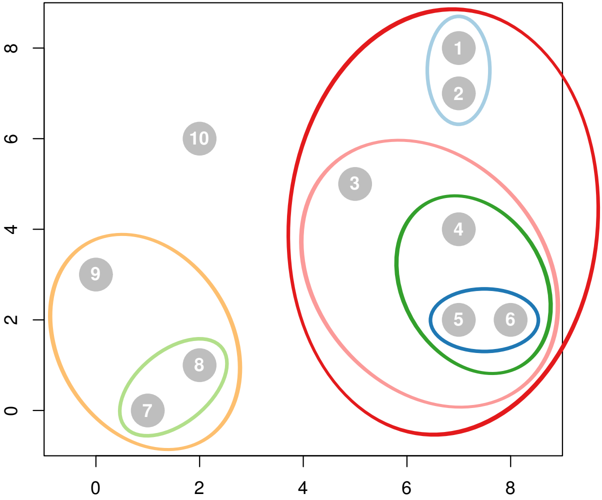
Hierarchical Clustering

Distance Matrix

	1	2	3	4	5	6	7	8	9	10
1	0.0									
2	1.0	0.0								
3	3.6	2.8	0.0							
4	4.0	3.0	2.2	0.0						
5	6.0	5.0	3.6	2.0	0.0					
6	6.1	5.1	4.2	2.2	1.0	0.0				
7	10.0	9.2	6.4	7.2	6.3	7.3	0.0			
8	8.6	7.8	5.0	5.8	5.1	6.1	1.4	0.0		
9	8.6	8.1	5.4	7.1	7.1	8.1	3.2	2.8	0.0	
10	5.4	5.1	3.2	5.4	6.4	7.2	6.1	5.0	3.6	0.0

Clusters: {1, 2}, {3, 4, 5, 6}, {7, 8}, {9}, {10}

Hierarchical Clustering



Hierarchical Clustering

d	k	Clusters	Comment
0.0	10	{1}, {2}, {3}, {4}, {5}, {6}, {7}, {8}, {9}, {10}	Start with each observation as one cluster.
1.0	8	{1, 2}, {3}, {4}, {5, 6}, {7}, {8}, {9}, {10}	Merge {1} and {2} as well as {5} and {6} since they are the closest: d(1,2)=1 and d(5,6)=1
1.4	7	{1, 2}, {3}, {4}, {5, 6}, {7, 8}, {9}, {10}	Merge {7} and {8} since they are the closest: d(7,8)=1.4
2.0	6	{1, 2}, {3}, {4, 5, 6}, {7, 8}, {9}, {10}	Merge {4} and {5, 6} since 4 and 5 are the closest: d(4,5)=2.0
2.2	5	{1, 2}, {3, 4, 5, 6}, {7, 8}, {9}, {10}	Merge {3} and {4, 5, 6} since 3 and 4 are the closest: d(3,4)=2.2
2.8	3	{1, 2, 3, 4, 5, 6}, {7, 8, 9}, {10}	Merge {1, 2} and {3, 4, 5, 6} as well as {7, 8} and {9} since 2 and 3 as well as 8 and 9 are the closest: d(2,3)=2.8 and d(8,9)=2.8

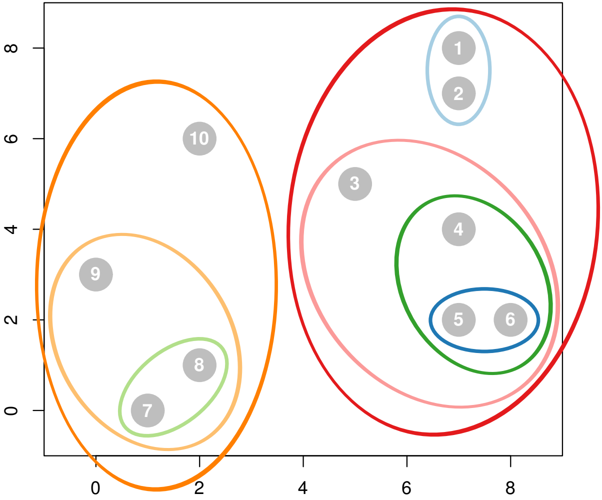
Hierarchical Clustering

Distance Matrix

	1	2	3	4	5	6	7	8	9	10
1	0.0									
2	1.0	0.0								
3	3.6	2.8	0.0							
4	4.0	3.0	2.2	0.0						
5	6.0	5.0	3.6	2.0	0.0					
6	6.1	5.1	4.2	2.2	1.0	0.0				
7	10.0	9.2	6.4	7.2	6.3	7.3	0.0			
8	8.6	7.8	5.0	5.8	5.1	6.1	1.4	0.0		
9	8.6	8.1	5.4	7.1	7.1	8.1	3.2	2.8	0.0	
10	5.4	5.1	3.2	5.4	6.4	7.2	6.1	5.0	3.6	0.0

Clusters: $\{1, 2, 3, 4, 5, 6\}$, $\{7, 8, 9\}$, $\{10\}$

Hierarchical Clustering



Hierarchical Clustering

d	k	Clusters	Comment
0.0	10	{1}, {2}, {3}, {4}, {5}, {6}, {7}, {8}, {9}, {10}	Start with each observation as one cluster.
1.0	8	{1, 2}, {3}, {4}, {5, 6}, {7}, {8}, {9}, {10}	Merge {1} and {2} as well as {5} and {6} since they are the closest: d(1,2)=1 and d(5,6)=1
1.4	7	{1, 2}, {3}, {4}, {5, 6}, {7, 8}, {9}, {10}	Merge {7} and {8} since they are the closest: d(7,8)=1.4
2.0	6	{1, 2}, {3}, {4, 5, 6}, {7, 8}, {9}, {10}	Merge {4} and {5, 6} since 4 and 5 are the closest: d(4,5)=2.0
2.2	5	{1, 2}, {3, 4, 5, 6}, {7, 8}, {9}, {10}	Merge {3} and {4, 5, 6} since 3 and 4 are the closest: d(3,4)=2.2
2.8	3	{1, 2, 3, 4, 5, 6}, {7, 8, 9}, {10}	Merge {1, 2} and {3, 4, 5, 6} as well as {7, 8} and {9} since 2 and 3 as well as 8 and 9 are the closest: d(2,3)=2.8 and d(8,9)=2.8
3.2	2	{1, 2, 3, 4, 5, 6, 10}, {7, 8, 9}	Merge {1, 2, 3, 4, 5, 6} and {10} since 3 and 10 are the closest: d(3,10)=3.2

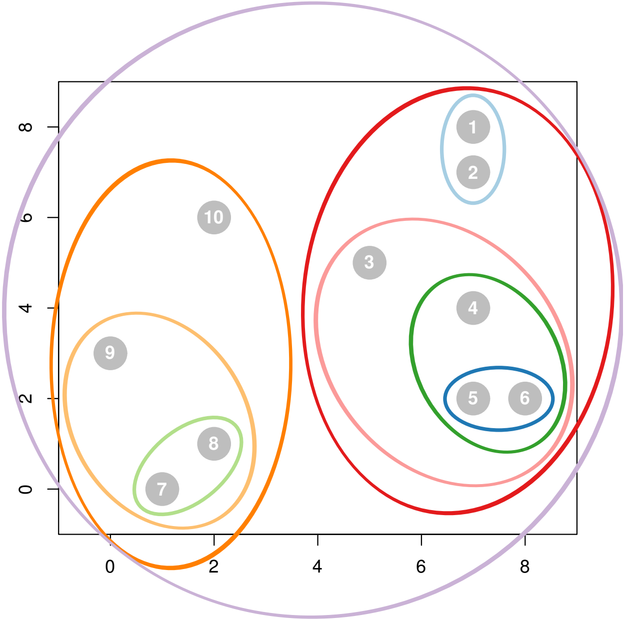
Hierarchical Clustering

Distance Matrix

	1	2	3	4	5	6	7	8	9	10
1	0.0									
2	1.0	0.0								
3	3.6	2.8	0.0							
4	4.0	3.0	2.2	0.0						
5	6.0	5.0	3.6	2.0	0.0					
6	6.1	5.1	4.2	2.2	1.0	0.0				
7	10.0	9.2	6.4	7.2	6.3	7.3	0.0			
8	8.6	7.8	5.0	5.8	5.1	6.1	1.4	0.0		
9	8.6	8.1	5.4	7.1	7.1	8.1	3.2	2.8	0.0	
10	5.4	5.1	3.2	5.4	6.4	7.2	6.1	5.0	3.6	0.0

Clusters: {1, 2, 3, 4, 5, 6, 10}, {7, 8, 9}

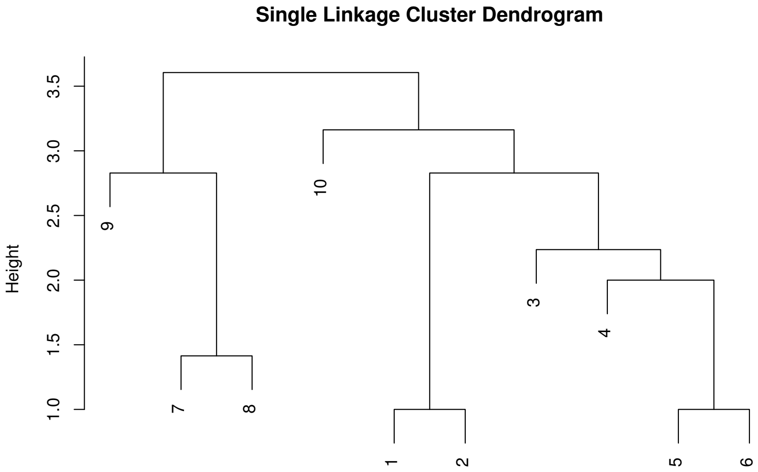
Hierarchical Clustering



Hierarchical Clustering

d	k	Clusters	Comment
0.0	10	{1}, {2}, {3}, {4}, {5}, {6}, {7}, {8}, {9}, {10}	Start with each observation as one cluster.
1.0	8	{1, 2}, {3}, {4}, {5, 6}, {7}, {8}, {9}, {10}	Merge {1} and {2} as well as {5} and {6} since they are the closest: d(1,2)=1 and d(5,6)=1
1.4	7	{1, 2}, {3}, {4}, {5, 6}, {7, 8}, {9}, {10}	Merge {7} and {8} since they are the closest: d(7,8)=1.4
2.0	6	{1, 2}, {3}, {4, 5, 6}, {7, 8}, {9}, {10}	Merge {4} and {5, 6} since 4 and 5 are the closest: d(4,5)=2.0
2.2	5	{1, 2}, {3, 4, 5, 6}, {7, 8}, {9}, {10}	Merge {3} and {4, 5, 6} since 3 and 4 are the closest: d(3,4)=2.2
2.8	3	{1, 2, 3, 4, 5, 6}, {7, 8, 9}, {10}	Merge {1, 2} and {3, 4, 5, 6} as well as {7, 8} and {9} since 2 and 3 as well as 8 and 9 are the closest: d(2,3)=2.8 and d(8,9)=2.8
3.2	2	{1, 2, 3, 4, 5, 6, 10}, {7, 8, 9}	Merge {1, 2, 3, 4, 5, 6} and {10} since 3 and 10 are the closest: d(3,10)=3.2
3.6	1	{1, 2, 3, 4, 5, 6, 7, 8, 9, 10}	Merge remaining two clusters, d(9,10)=3.6

Hierarchical Clustering



Conclusions

- Unsupervised learning is important for understanding the variation and grouping structure of a set of unlabeled data, and can be a useful pre-processor for supervised learning.
- It is intrinsically more difficult than supervised learning because there is no gold standard (like an outcome variable) and no single objective (like test set accuracy).

Reference

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013).
An Introduction to Statistical Learning: with Applications in R.
Springer.