Unsupervised Learning. Hierarchical Clustering.

W. Wang[1]

[1]Department of Mathematics
University of Houston

MATH 4323

Hierarchical Clustering.

One potential disadvantage of K-means approach: we have to pre-specify the number of clusters K.

Hierarchical clustering is an alternative approach which has a few advantages:

- doesn't require to know the # of clusters in advance,
- it results in an attractive tree-based representation of the observations, called a dendrogram.

We will discuss the most common type of hierarchical clustering - bottom-up or agglomerative clustering:

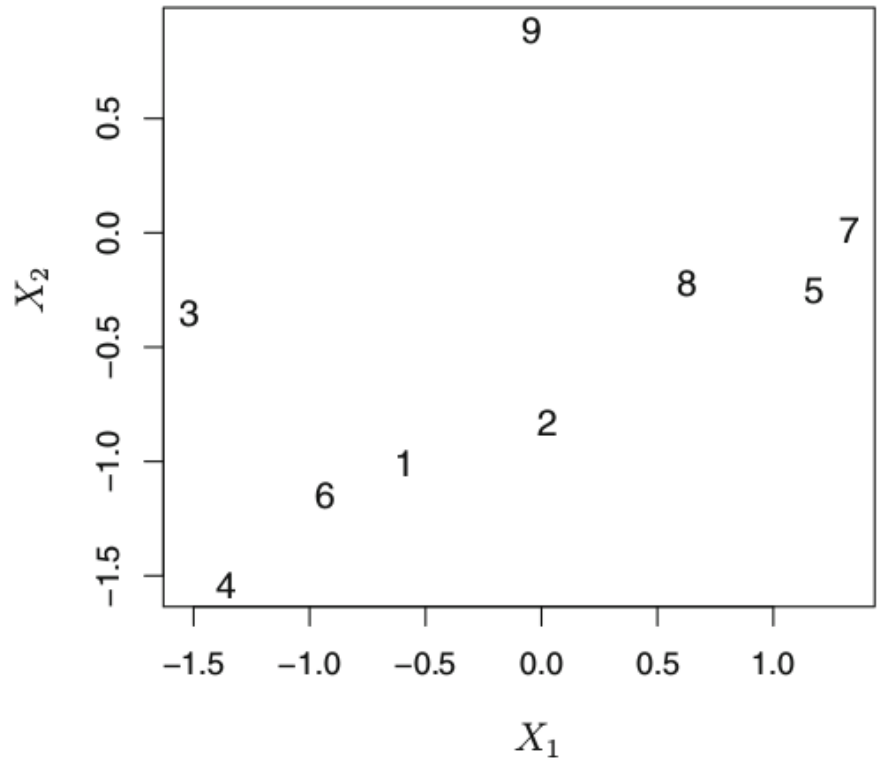
1. Treat each of n observations as it's own cluster ($\Longrightarrow$ n clusters).
2. Merge the **most similar** (or **least dissimilar**) clusters into one cluster, until all observations end up in a single huge cluster. Classic measure of **dissimilarity** is Eucledian distance.

Hierarchical Clustering Algorithem:

1. Start with n observations and a measure of all the pairwise dissimilarities. Treat each observation as its own cluster.
2. Identify the pair of clusters that are least dissimilar and fuse them. The dissimilarity between these two clusters indicates the height in the dendrogram at which the fusion should be placed.
3. Repeat until all observations are in a single cluster.

Hierarchical Clustering: Toy Example.

Example. Look at the toy example with $n = 9$ observations, described by $p = 2$ features. Each observation i is it's own "cluster" $\{i\}$.

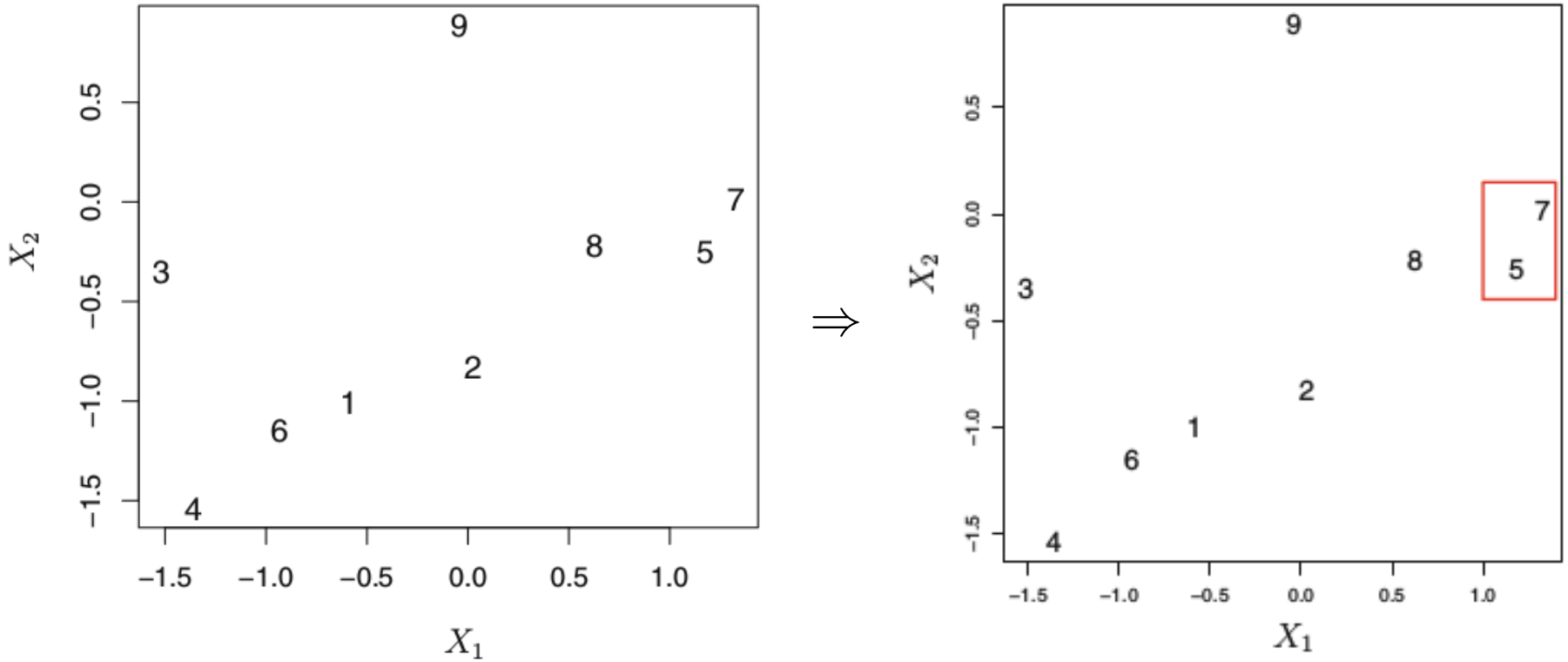


Question: which pairs of "clusters" will be merged first?

Answer: $\{5\}$ and $\{7\}$; $\{1\}$ and $\{6\}$ - very close via Eucledian distance.

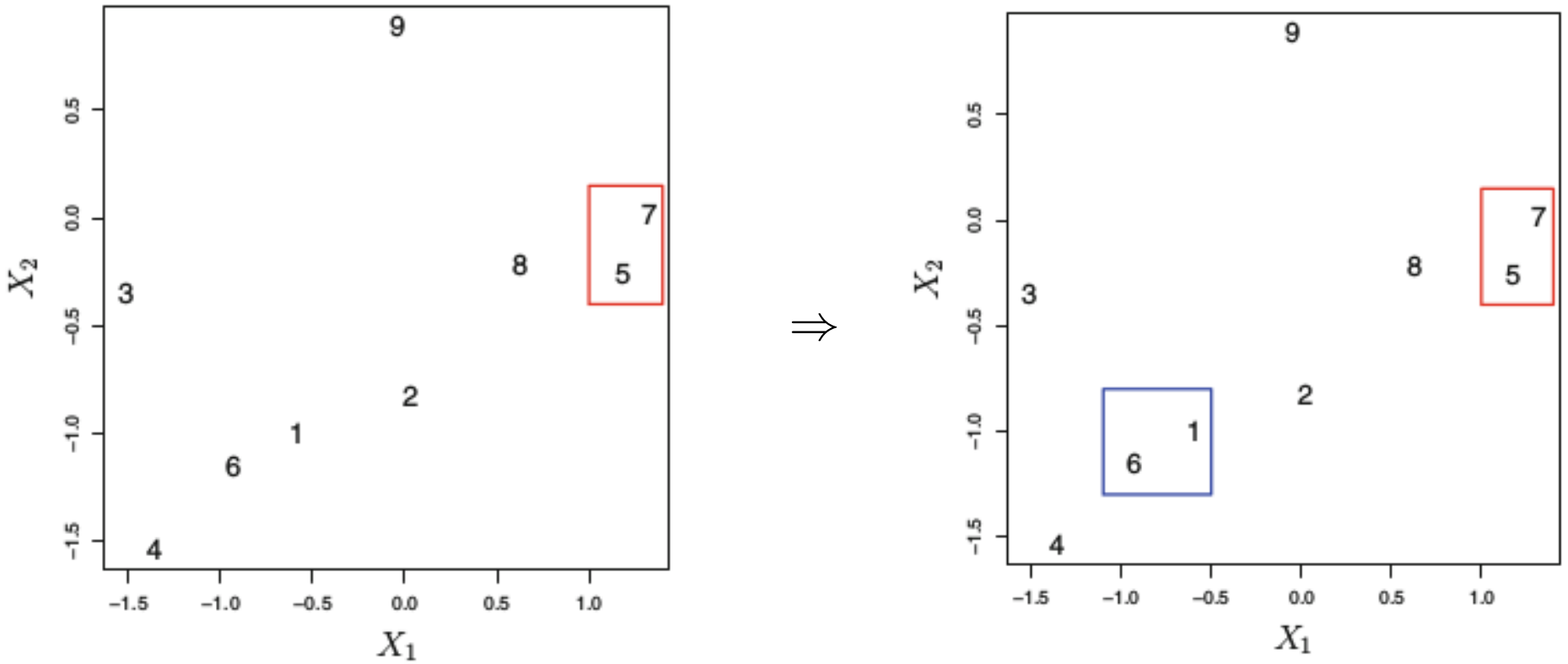
Hierarchical Clustering: Toy Example.

Example (con't). Starting from a $K = n$ cluster solution (each observation is its own cluster), merging most similar "one-observation clusters" $\{5\}$ and $\{7\}$ leads us to a $K = n - 1$ cluster solution.



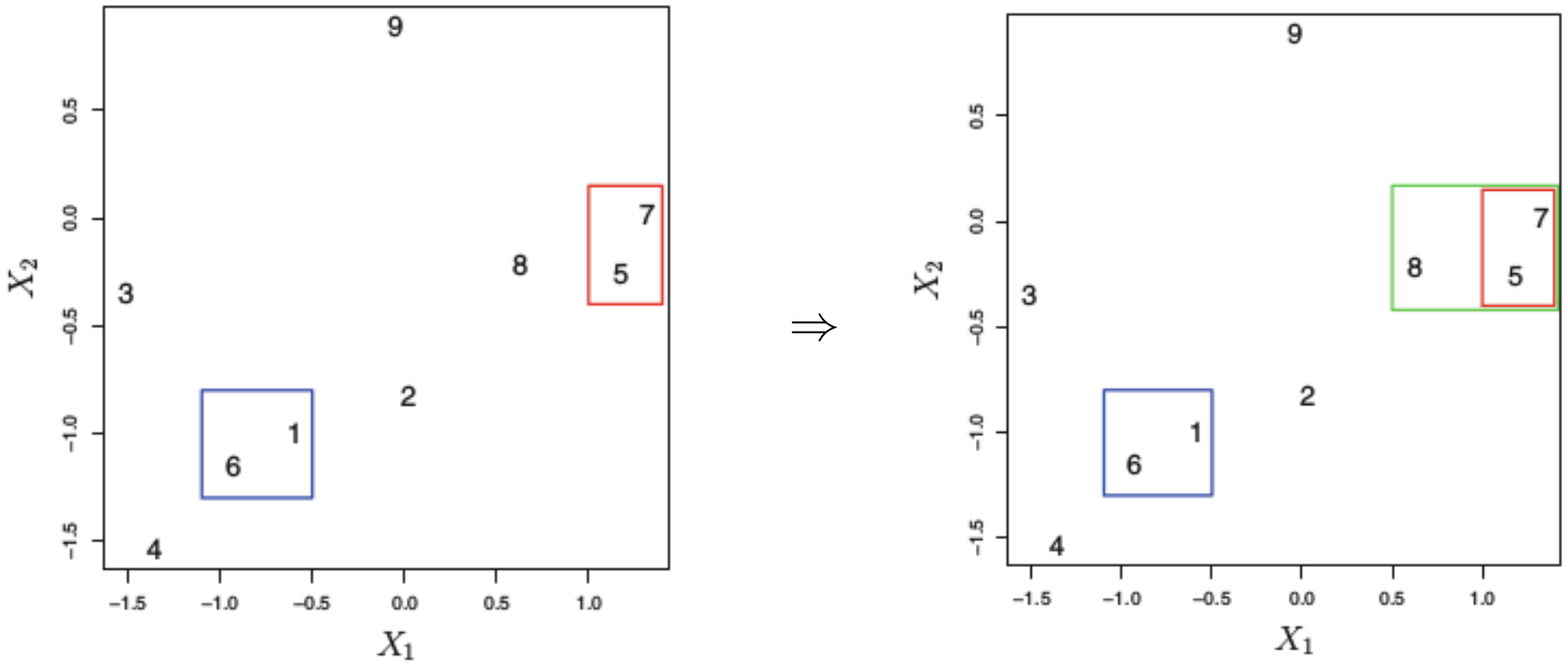
Hierarchical Clustering: Toy Example.

Example (con't). Further, merging most similar clusters $\{1\}$ and $\{6\}$ we go from $K = n - 1$ to a $K = n - 2$ clustering solution:



Hierarchical Clustering: Toy Example.

Example (con't). Next, merging most similar clusters $\{5, 7\}$ and $\{8\}$ leads us from $K = n - 2$ to a $K = n - 3$ clustering solution:



Question: How did we decide on merging $\{5, 7\}$ and $\{8\}$?

Hierarchical Clustering: Linkage.

The notion of linkage defines the dissimilarity between two groups of observations.

Linkage	*Description*
Complete	Maximal intercluster dissimilarity. Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, and record the *largest* of these dissimilarities.
Single	Minimal intercluster dissimilarity. Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, and record the *smallest* of these dissimilarities. Single linkage can result in extended, trailing clusters in which single observations are fused one-at-a-time.
Average	Mean intercluster dissimilarity. Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, and record the *average* of these dissimilarities.
Centroid	Dissimilarity between the centroid for cluster A (a mean vector of length p) and the centroid for cluster B. Centroid linkage can result in undesirable *inversions*.

Hierarchical Clustering: Dissimilarity Matrix.

Example (cont'): Below is the dissimilarity matrix of the initial $n = 9$ "clusters" according to Eucledian distance as the dissimilarity measure (see the lecture code for more detail).

```
> round(dist(x),2)
     1    2    3    4    5    6    7    8
2 0.66
3 1.08 1.70
4 0.97 1.52 1.15
5 2.02 1.48 2.80 2.98
6 0.36 0.96 1.00 0.61 2.38
7 2.24 1.74 2.93 3.20 0.32 2.59
8 1.46 1.04 2.16 2.43 0.65 1.82 0.79
9 1.96 2.01 1.91 2.77 1.84 2.25 1.75 1.37
```

Question: Which pair of "clusters" yields the smallest dissimilarity?

Answer: $\{5\}$ & $\{7\}$ (0.32), hence we merge them into one cluster.

Dissimilarity Matrix, Complete Linkage.

Example (cont'): Now, let's use complete linkage to compute dissimilarities of the resulting $\{5, 7\}$ cluster with other 7 "clusters":

- $dist(\{1\}, \{5, 7\}) = max(dist(\{1\}, \{5\}), dist(\{1\}, \{7\})) = max(2.02, 2.24) = 2.24$,
- $dist(\{2\}, \{5, 7\}) = max(dist(\{2\}, \{5\}), dist(\{2\}, \{7\})) = max(1.48, 1.74) = 1.74$,
- ....
- $dist(\{6\}, \{5, 7\}) = max(dist(\{6\}, \{5\}), dist(\{6\}, \{7\})) = max(2.38, 2.59) = 2.59$,
- $dist(\{8\}, \{5, 7\}) = max(dist(\{8\}, \{5\}), dist(\{8\}, \{7\})) = max(0.65, 0.79) = 0.79$,
- $dist(\{9\}, \{5, 7\}) = max(dist(\{9\}, \{5\}), dist(\{9\}, \{7\})) = max(1.84, 1.75) = 1.84$,

Dissimilarity Matrix, Complete Linkage.

Example (cont'): Having calculated the dissimilarities between new cluster $\{5, 7\}$ (denoted as 5&7) and other $n - 2 = 7$ "clusters", the resulting dissimilarity matrix is:

	1	2	3	4	6	8	9
2	0.66						
3	1.08	1.70					
4	0.97	1.52	1.15				
6	0.36	0.96	1.00	0.61			
8	1.46	1.04	2.16	2.43	1.82		
9	1.96	2.01	1.91	2.77	2.25	1.37	
5&7	2.24	1.74	2.93	3.20	2.59	0.79	1.84

Next, we apply the same approach to this new dissimilarity matrix:

- Find pair of clusters with the smallest dissimilarity:

$$\{1\} \text{ and } \{6\},\ dist(\{1\}, \{6\}) = 0.36$$

- Re-calculate the dissimilarity matrix when substituting clusters $\{1\}$ and $\{6\}$ for the merged cluster $\{1, 6\}$, etc.

Complete Linkage.

Example. So, back to question on how we decided to merge $\{5, 7\}$ and $\{8\}$: according to complete linkage, the dissimilarity measure

$$dist(\{5, 7\}, \{8\}) = max(dist(\{5\}, \{8\}), dist(\{7\}, \{8\}))$$

yielded the smallest value across all other $\binom{n-2}{2}$ pair-wise dissimilarities between $n - 2 = 7$ clusters. So, as a result, we went from a $n - 2 = 7$ cluster solution:

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

to a $n - 3 = 6$ cluster solution:

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

Hierarchical Clustering: Algorithm.

Algorithm 10.2 *Hierarchical Clustering*

1. Begin with n observations and a measure (such as Euclidean distance) of all the $\binom{n}{2} = n(n-1)/2$ pairwise dissimilarities. Treat each observation as its own cluster.

2. For $i = n, n-1, \ldots, 2$:

   (a) Examine all pairwise inter-cluster dissimilarities among the i clusters and identify the pair of clusters that are least dissimilar (that is, most similar). Fuse these two clusters. The dissimilarity between these two clusters indicates the height in the dendrogram at which the fusion should be placed.

   (b) Compute the new pairwise inter-cluster dissimilarities among the $i-1$ remaining clusters.

Hierarchical Clustering: Algorithm.

Some **advantages** of hierarchical clustering:

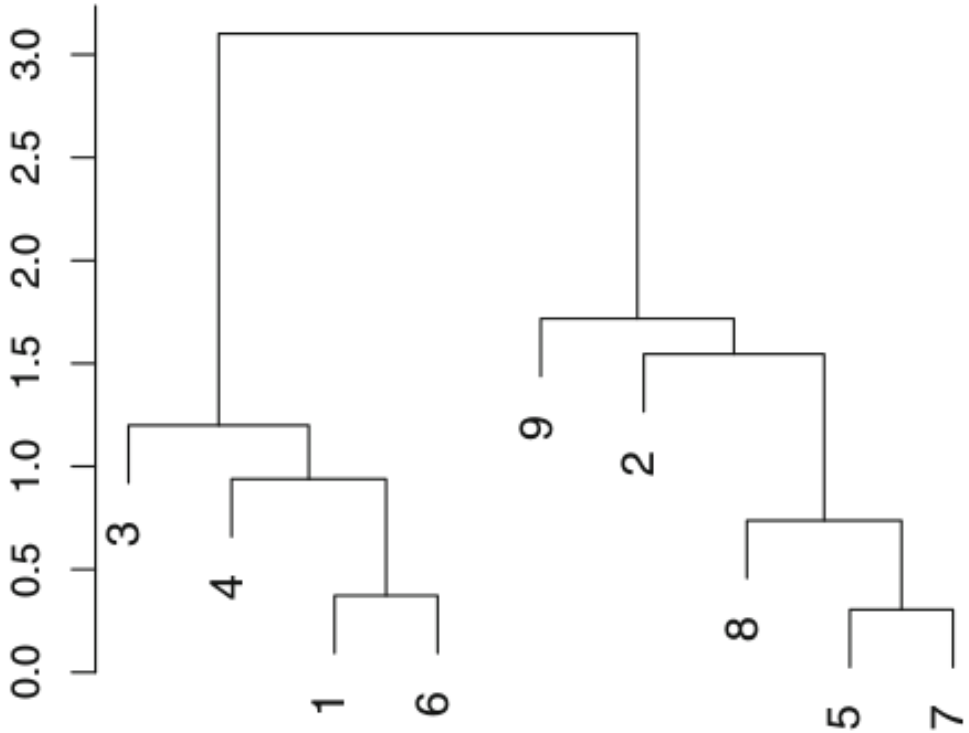
- In order to run, hierarchical clustering doesn't need a pre-specified value K for the # of clusters.
- There's no random initialization, hence it always leads to the same clustering solution.
- Gives a pretty tree-based representation, called a dendrogram, of the similarity between observations.

Some **disadvantages** of hierarchical clustering:

- Might be very slow for large number n of observations, because it builds a long hierarchy of clustering solutions, and has to calculate tons of dissimilarity matrices.
- Results might heavily depend on choice of linkage.

Hierarchical Clustering: Dendrogram.

Example (con't). Below is the **dendrogram** for our toy example:



Interpretation:

- Each "leaf" is one of $n = 9$ observations.
- As we move up, leaves fused into "branches" ($\leftrightarrow$ similar observations fuse into clusters).

Hierarchical Clustering: Dendrogram.

Interpretation (cont'd):

- The earlier (lower in the tree) fusions occur, the more similar the groups of observations are to each other.
- Observations that fuse later (higher in the tree) can be quite different.

To sum it up:

For any two observations, the **height** of their **fusion**, as measured on the vertical axis, indicates how different the two observations are. Thus, observations that fuse at the very bottom of the tree are quite similar to each other, whereas observations that fuse high in the tree will tend to be quite different.

Hierarchical Clustering: Dendrogram.

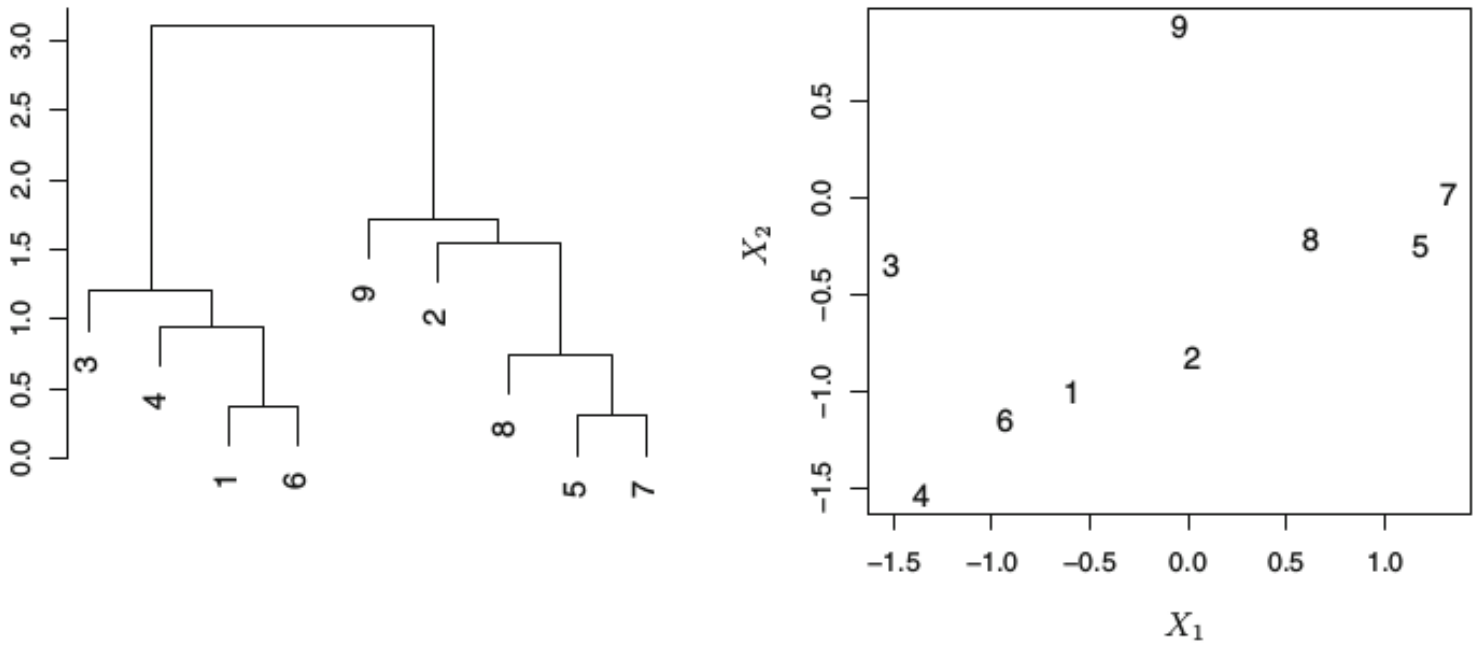


FIGURE 10.10. *An illustration of how to properly interpret a dendrogram with nine observations in two-dimensional space.* Left: *a dendrogram generated using Euclidean distance and complete linkage. Observations* 5 *and* 7 *are quite similar to each other, as are observations* 1 *and* 6. *However, observation* 9 *is* no more similar to *observation* 2 *than it is to observations* 8, 5, *and* 7, *even though observations* 9 *and* 2 *are close together in terms of horizontal distance. This is because observations* 2, 8, 5, *and* 7 *all fuse with observation* 9 *at the same height, approximately* 1.8. Right: *the raw data used to generate the dendrogram can be used to confirm that indeed, observation* 9 *is no more similar to observation* 2 *than it is to observations* 8, 5, *and* 7.

Dendrogram: Object Similarity.

Example (cont'd). Given the fact that observations 9 and 2 appear close to each other on dendrogram (**left**), while not being similar in the actual $2D$ data visualization (**right**) , this example goes on to show that

- We cannot draw conclusions about the similarity of two observations based on their proximity along the **horizontal** axis.
- Rather, we draw conclusions about the similarity of two observations based on the location on the **vertical** axis where branches containing those two observations first are fused.

Hierarchical Clustering: Example.

Example. We will work with simulated data on $n = 45$ observations, described by $p = 2$ features. See more info below.

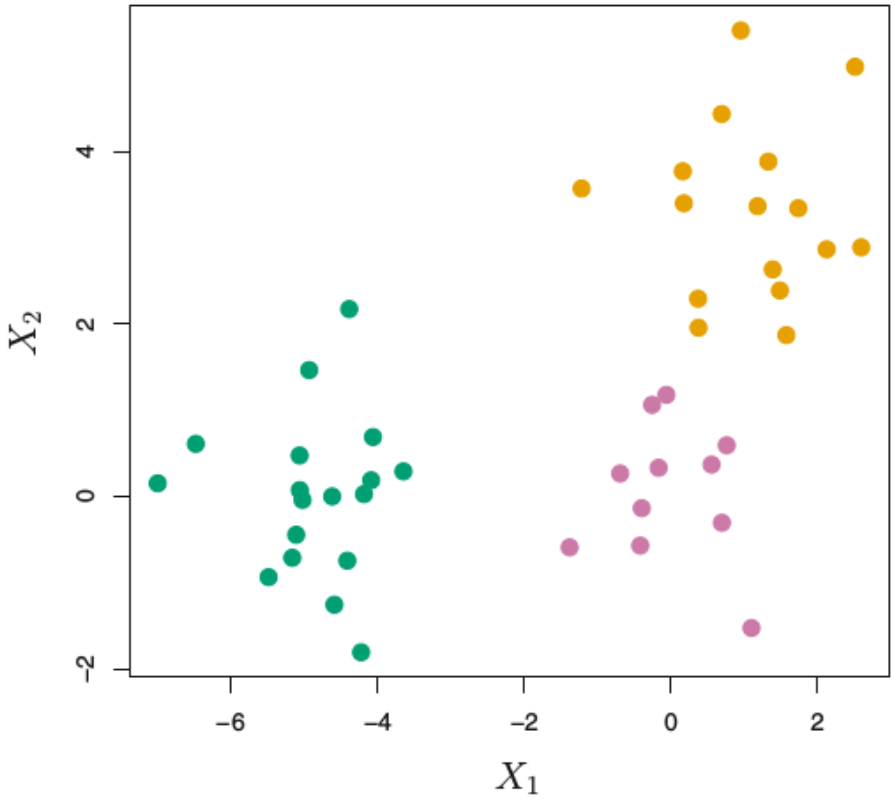


FIGURE 10.8. *Forty-five observations generated in two-dimensional space. In reality there are three distinct classes, shown in separate colors. However, we will treat these class labels as unknown and will seek to cluster the observations in order to discover the classes from the data.*

Hierarchical Clustering: Dendrogram.

Example (cont'd). Hierarchical clustering yields the following result (right).

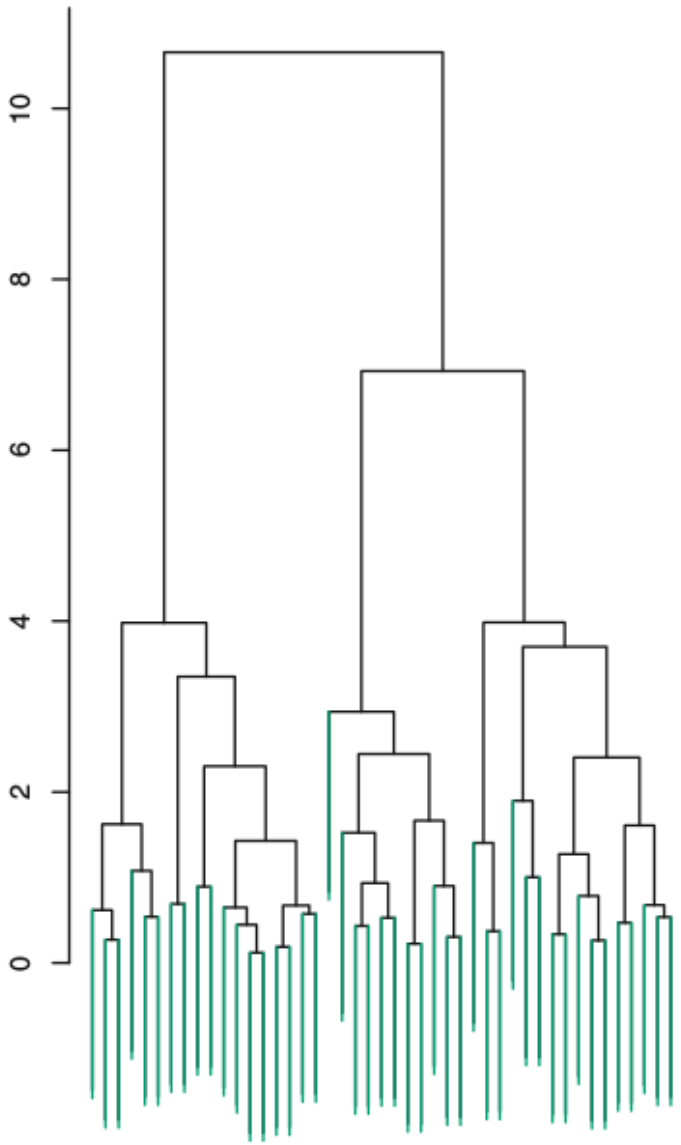
Each **leaf** (at the bottom) is one of $n = 45$ observations.

Moving up the tree, leaves of **similar** observations begin to fuse into **branches**.

Moving even higher, **branches themselves** fuse with other leaves or branches.

The earlier (lower in the tree) fusions occur, the more similar the groups of observations are to each other.

Marked in **green** are the initial **"single-observation clusters"**.



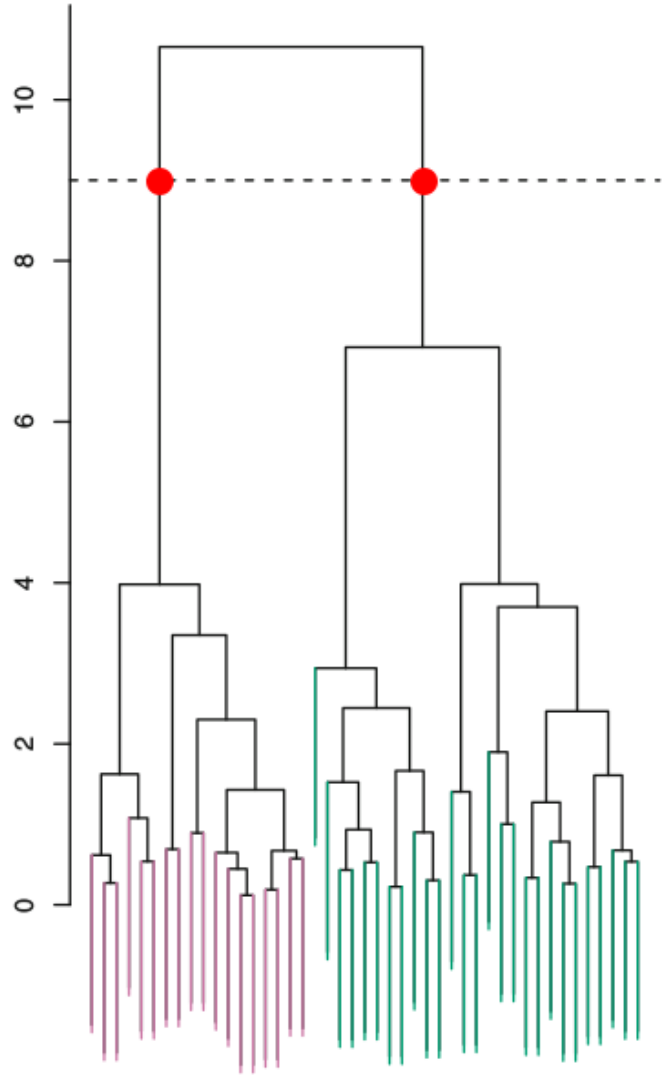
Dendrogram: Identifying Clusters.

Having understood how to interpret the object similarity from a dendrogram, we can move on to the issue of identifying clusters:

Example (cont'd). In order to identify clusters, we make a **horizontal** cut across the dendrogram (see picture on the right).

This particular cut at at a height of 9 sets **"roots"** (marked in red on the plot) for **two** distinct sets of observations, which can be interpreted as **clusters**:

- **Cluster #1** shows in purple color,
- **Cluster #2** - in green.



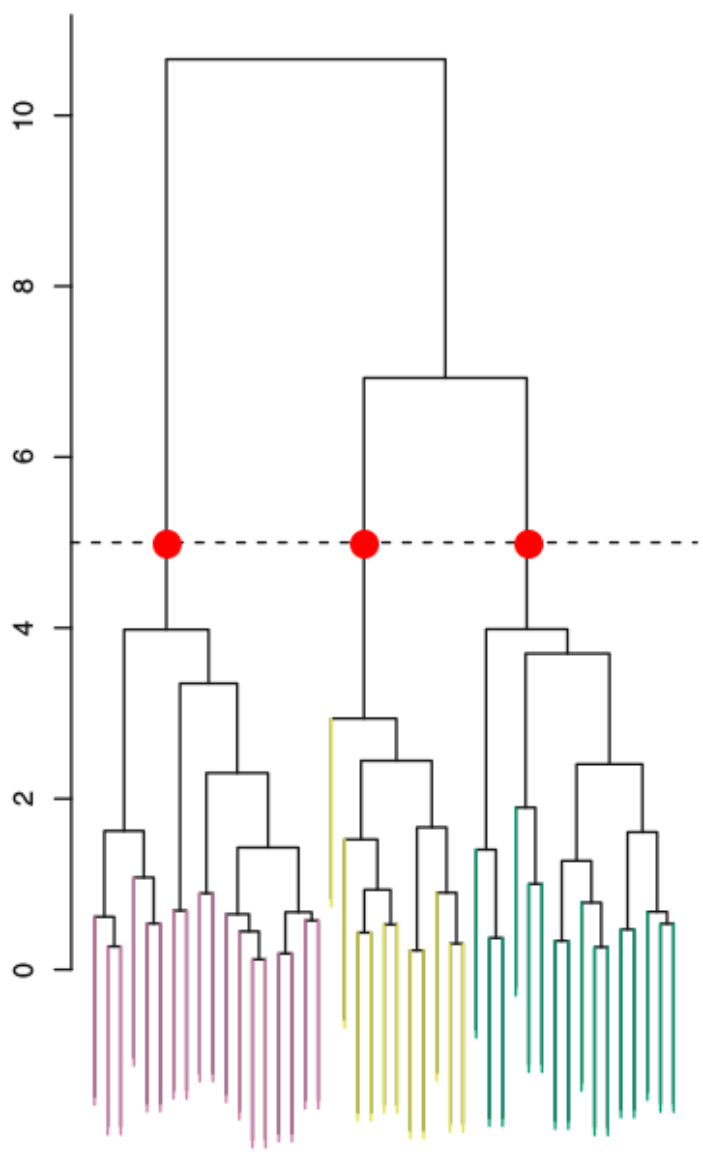
Dendrogram: Identifying Clusters.

Example (cont'd). We can also make a horizontal cut - this time at height 5 - resulting into **three** clusters (see their roots as red dots on the plot).

- **Cluster #1** shows in purple color,
- **Cluster #2** shows in orange color,
- **Cluster #3** - in green.

Further cuts can be made as one descends the dendrogram in order to obtain any number of clusters, between

- 1 ($\equiv$ no cut, a single cluster filled with all observations), and
- n ($\equiv$ cut at $height = 0$, each observation is in its own cluster).

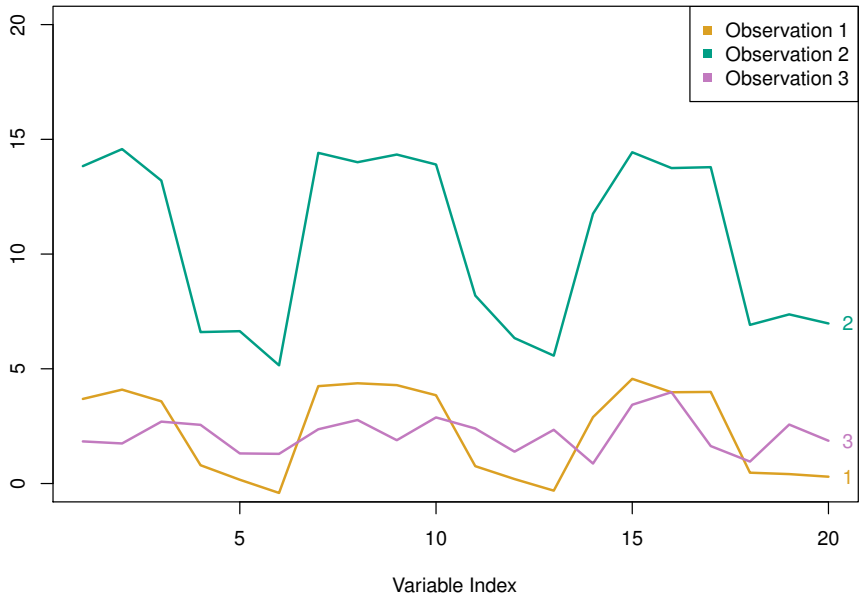


Choice of Dissimilarity Measure

- Euclidean distance is commonly-used for the dissimilarity between two observations.
- An alternative is correlation-based distance which considers two observations to be similar if their features are highly correlated (even though the observed values may be far apart in terms of Euclidean distance).
- This is an unusual use of correlation, which is normally computed between variables; here it is computed between the observation profiles for each pair of observations.

Choice of Dissimilarity Measure

- Correlation-based distance focuses on the shapes of the observation profiles rather than their magnitudes.



  - Three observations with measurements on 20 variables are shown.
  - Observation 1 and 3: similar in values for each variables, thus small Euclidean distance. But weakly correlated.
  - Observation 1 and 2: very different values for each variable (large Euclidean distance), but highly correlated (small correlation-based distance between them).

Hierarchical vs K-Means: Similarities & Differences.

Similarity between **hierarchical** and ***K*-Means** clustering:

the **height of the cut** to the dendrogram in hierarchical clustering

serves the **same role** as the

K in K-means clustering

which is

to control the **# of clusters** obtained.

Differences:

- Hierarchical clustering does **not** need a pre-specified value of clusters in order to compute a clustering solution.
- Hierarchical clustering has one single dendrogram that can be used to obtain any number of clusters.

Practical issues with Clustering

- Should the observations or features first be standardized in some way? For instance, maybe the variables should be centered to have mean zero and scaled to have standard deviation one.
- In the case of hierarchical clustering,
  - What dissimilarity measure should be used?
  - What type of linkage should be used?
  - Where should we cut the dendrogram in order to obtain clusters?
- In the case of K-means clustering, how many clusters should we look for in the data?

Each of these decisions can have a strong impact on the results obtained. With these methods, there is no single right answer—any solution that exposes some interesting aspects of the data should be considered.

Clustering with mixed data

- So far, we only discussed examples of clustering with numerical data.
- In practice, most data sets contain a mixture of numeric, categorical, and ordinal variables.
- When performing clustering analysis, whether an observation is similar to another observation should depend on these data type attributes.
- How to perform clustering analysis with mixed data types?
  We can convert any categorical variables to numeric using one-hot encoding.

  in R, *to_categorical*(...).

Choices of linkage

The linkage method helps to measure the dissimilarity between two clusters of observations.

Linkage	*Description*
Complete	Maximal intercluster dissimilarity. Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, and record the *largest* of these dissimilarities.
Single	Minimal intercluster dissimilarity. Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, and record the *smallest* of these dissimilarities. Single linkage can result in extended, trailing clusters in which single observations are fused one-at-a-time.
Average	Mean intercluster dissimilarity. Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, and record the *average* of these dissimilarities.
Centroid	Dissimilarity between the centroid for cluster A (a mean vector of length p) and the centroid for cluster B. Centroid linkage can result in undesirable *inversions*.

TABLE 10.2. *A summary of the four most commonly-used types of linkage in hierarchical clustering.*

Another commonly used linkage:

- Ward's minimum variance method which minimizes the total within-cluster variance. At each step the pair of clusters with the smallest between-cluster distance are merged. This kind of linkage tends to generate more compact clusters.

Choice of Distance Measure

How do we measure the dissimilarity between two clusters of observations? Distance metrics to compute the pairwise differences between observations.

- The most common distance measure is the Euclidean distance.

$$\sqrt{\sum_{i=1}^{P}(x_{1j} - x_{2j})^2}$$

- Other common options:
  - Weighted Euclidean distance: $\sqrt{\sum_{i=1}^{P} w_i(x_{1j} - x_{2j})^2}$
  - Manhattan distance: point-to-point measurement and is commonly used for binary predictors. $\sum_{i=1}^{P} |x_{1j} - x_{2j}|$
  - Minkowski distance: a generalization of the Euclidean and Manhattan distances.$(\sum_{i=1}^{P} |x_{1j} - x_{2j}|^q)^{1/q}$
  - Other options, check ?*dist*() in R.

Association Rule

Association rule analysis has emerged as a popular tool for mining commercial data bases.

- The goal is to find joint values of the variables $X = (X_1, X_2, \cdots, X_p)$ that appear most frequently in the data base.

- It is most often applied to binary-valued data $X_j \in \{0, 1\}$, where it is referred to as "market basket" analysis.

Marketing

In this context,

- The observations are sales transactions, such as those occurring at the checkout counter of a store.
- The variables represent all of the items sold in the store.
- For observation i, each variable X_j is binary; $X_{ij} = 1$ if the jth item is purchased as part of the transaction, whereras $X_{ij} = 0$ if it was not purchased.
- Those variables that frequently have joint values of one represent items that are frequently purchased together.

This information can be quite useful for stocking shelves, cross-marketing in sales promotions, catalog design, and consumer segmentation on buying patterns.

Information

For interested students, refer to "The Elements of Statistical Learning" Chapter 14.