NOTES 8

Clustering

Acknowledgement: some of the contents are borrowed with or without modification from *An Introduction to Statistical Learning, with applications in R* (Springer, 2013) with permission from the authors: G. James, D. Witten, T. Hastie and R.Tibshirani.

Unsupervised Learning

- Unsupervised Learning: observe only the features X_1, X_2, . . . , X_p. We are not interested in prediction, because we do not have an associated response variable Y .

- The goal is to discover interesting things about the measurements: Can we discover subgroups among the variables or among the observations?

Challenge of Unsupervised Learning

- Unsupervised learning is more subjective than supervised learning, as there is no simple goal for the analysis, such as prediction of a response.

- But techniques for unsupervised learning are of growing importance in a number of fields:
  - subgroups of breast cancer patients grouped by their gene expression measurements,
  - groups of shoppers characterized by their browsing and purchase histories,
  - movies grouped by the ratings assigned by movie

Clustering

- Clustering refers to a very broad set of techniques for finding subgroups, or clusters, in a data set
- Clustering looks for homogeneous subgroups among the observations.
- We seek a partition of the data into distinct groups so that the observations within each group are quite similar to each other
- Thus, we must define what it means for two or more observations to be similar or different.
  - Euclidean distance
  - Many more measures

Correlation-based distance (Pearson Correlation)

Two Clustering Methods

- K-means clustering: partition the observations into a pre-specified number of clusters.

- Hierarchical clustering: 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.

Technique Characteristics

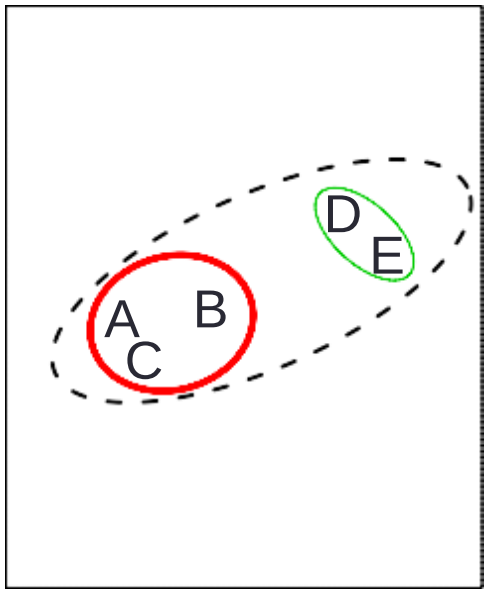
- Agglomerative vs Divisive
  - Agglomerative: each instance is its own cluster and the algorithm merges clusters
  - Divisive: begins with all instances in one cluster and divides it up
- Hierarchical clustering
  - Agglomerative
  - Builds a hierarchy in a "bottom-up" fashion...

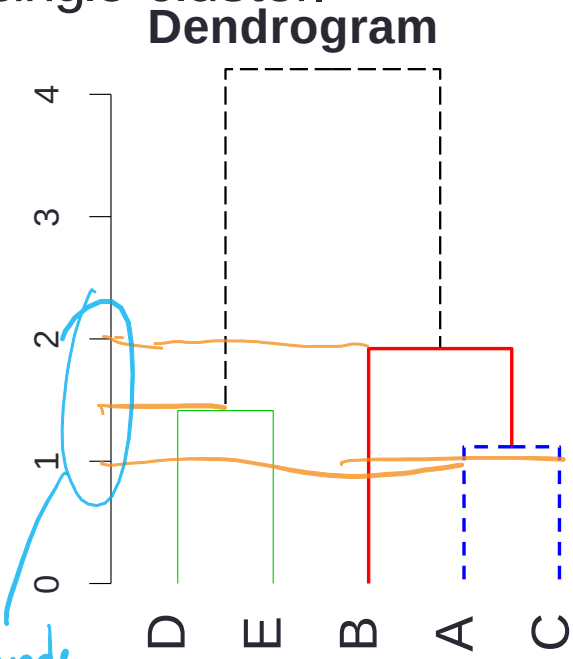
from bottom to the top

Hierarchical clustering algorithm

The approach in words:

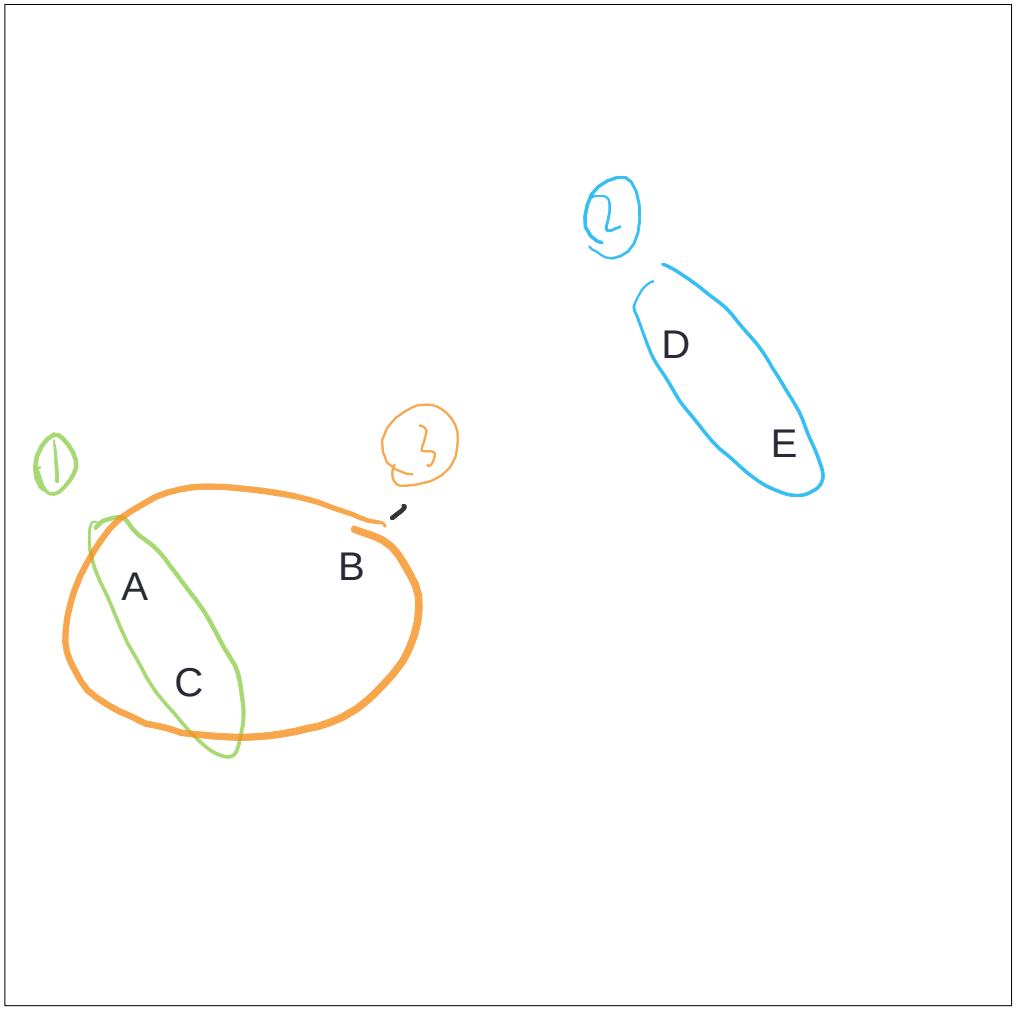
- Start with each point in its own cluster.
- Identify the closest two clusters and merge them.
- Repeat.
- Ends when all points are in a single cluster.



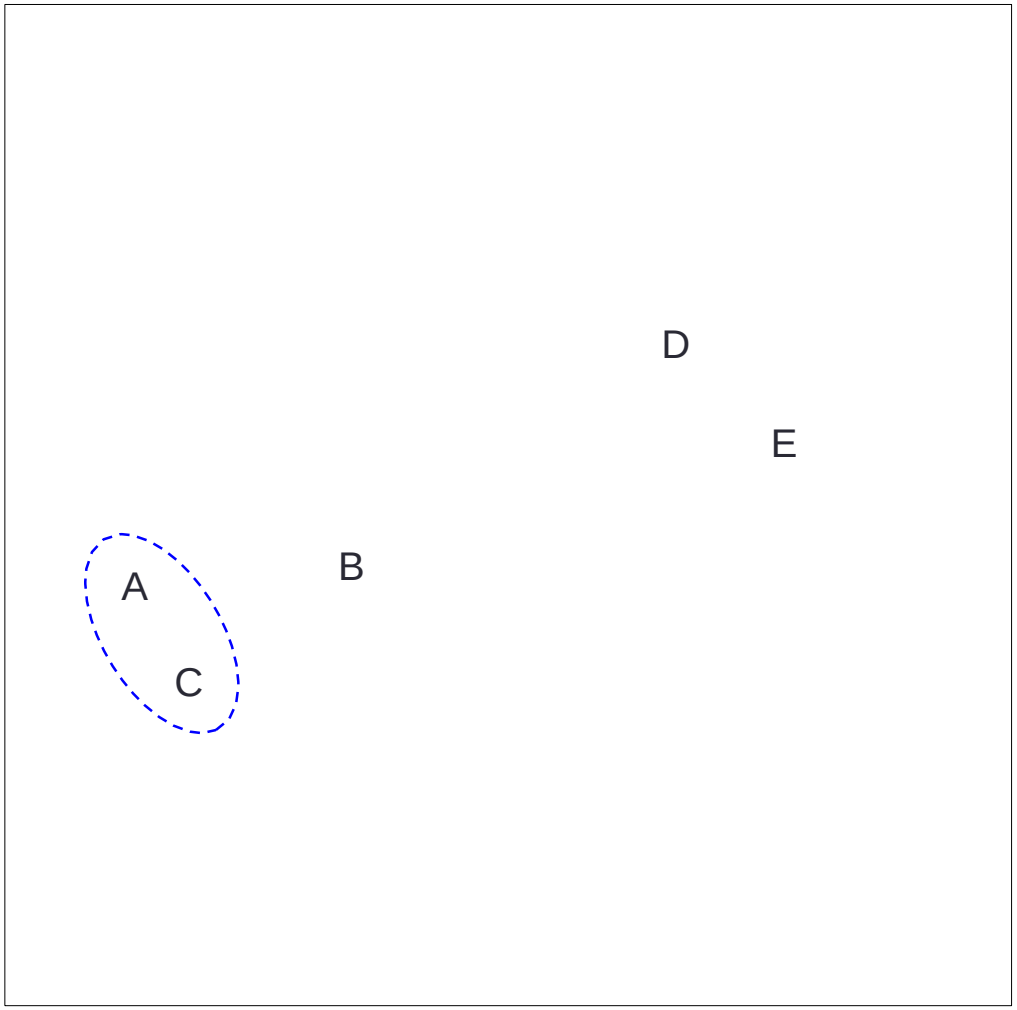


Depends on which methods u use (Complete; Single; Average)

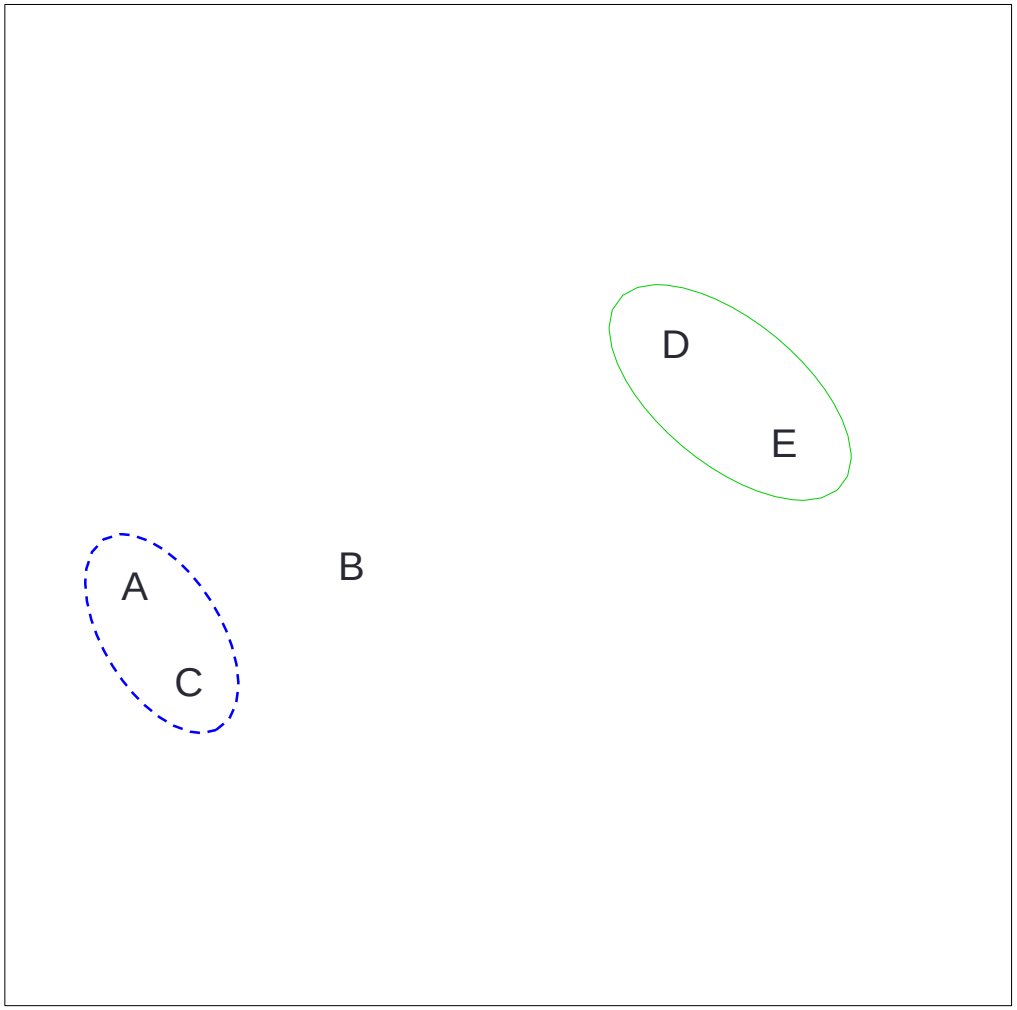
Hierarchical clustering: the idea



Hierarchical clustering



Hierarchical clustering



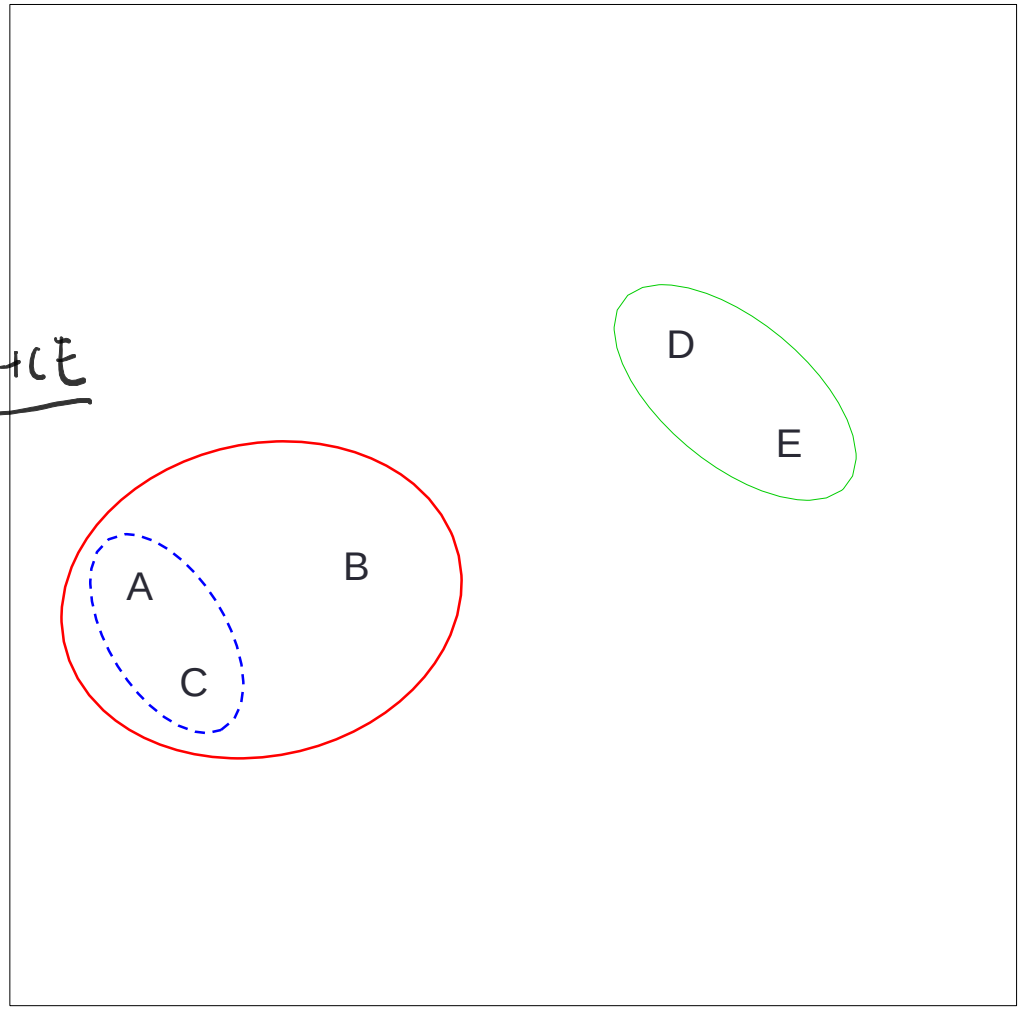
Hierarchical clustering

Average:

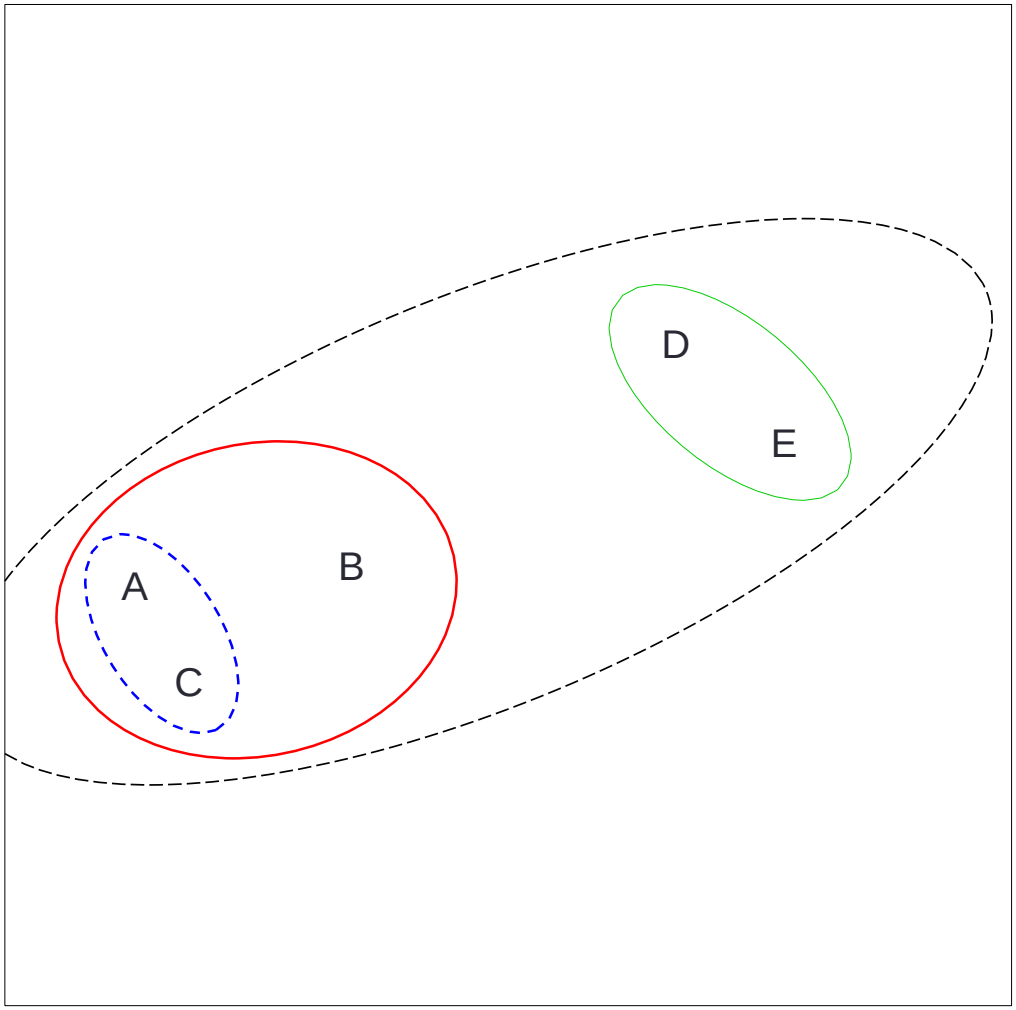
$$\frac{BD+BE+AD+AE+CD+CE}{6}$$

Single:
BD.

Complete:
AE

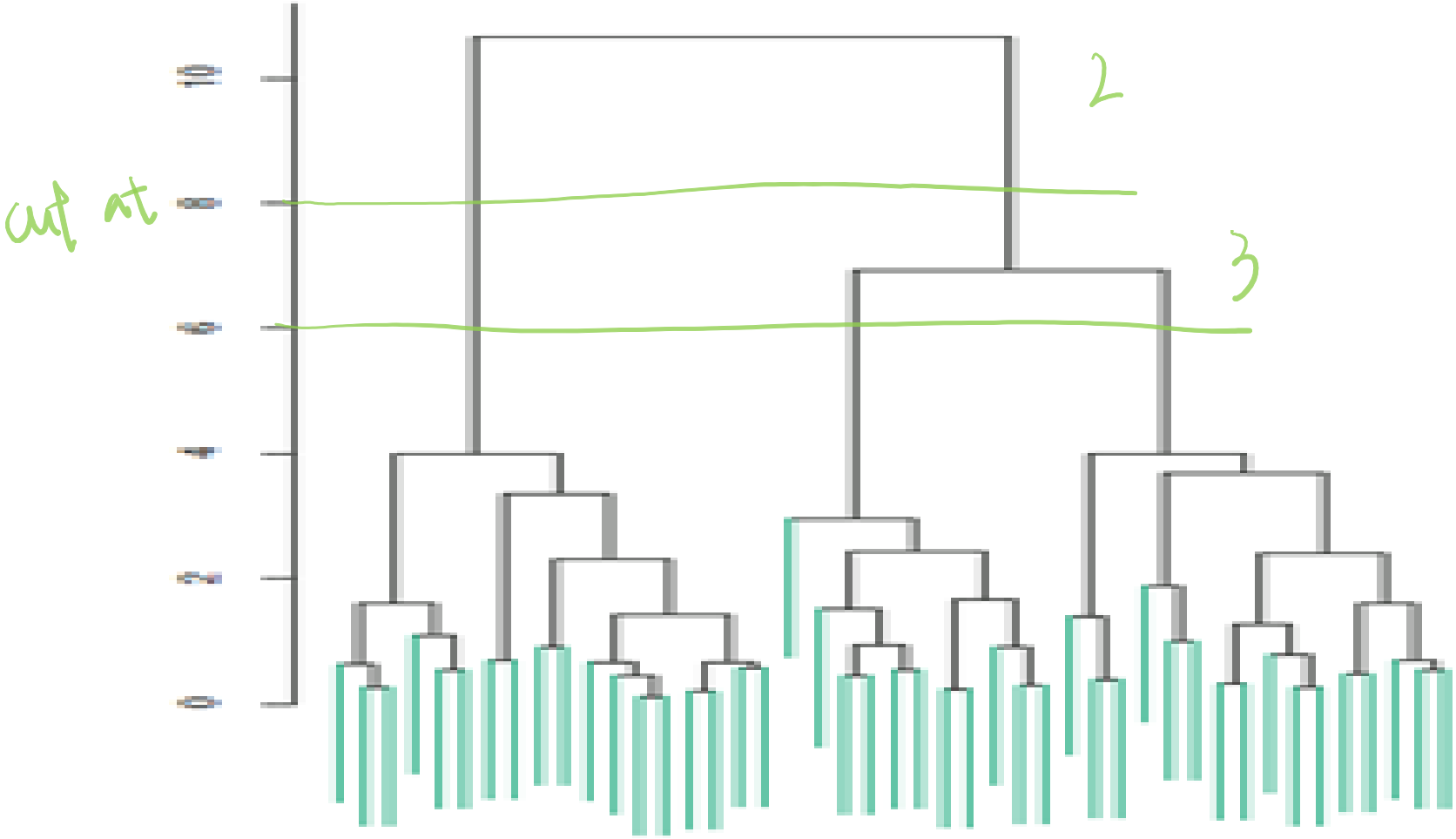


Hierarchical clustering

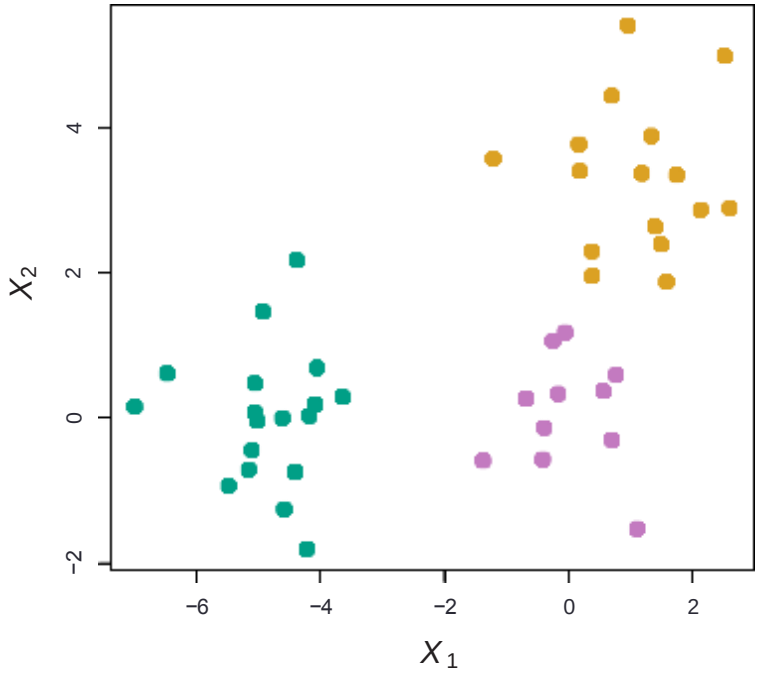


Dendrogram

Subjective



An Example

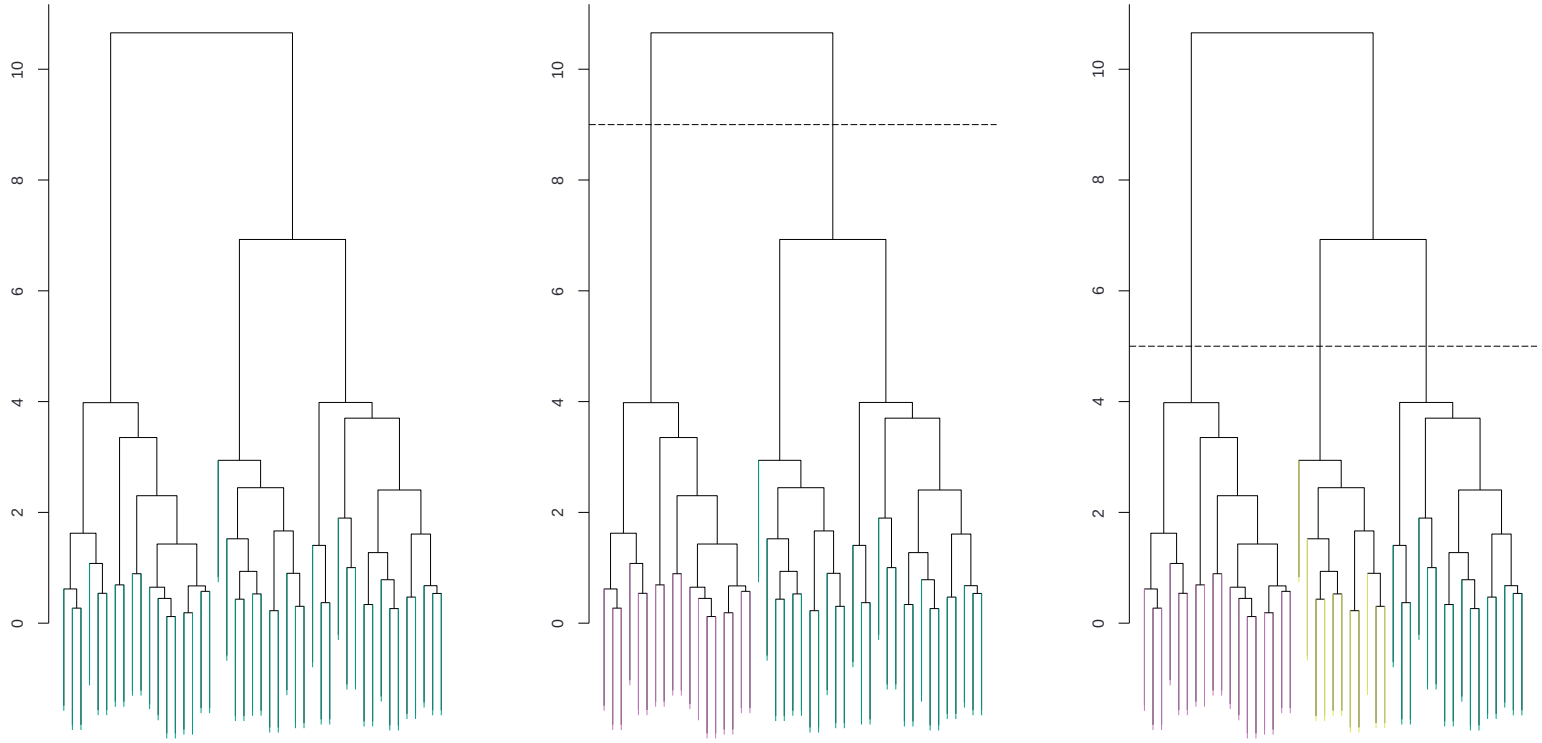


Suppose there's no class in advance

45 observations generated in 2-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.

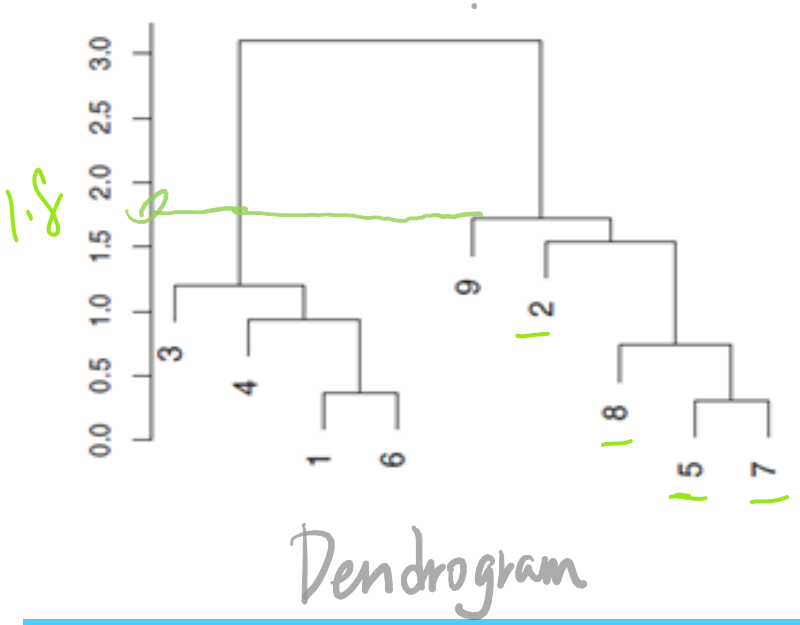
Unsupervised Learning

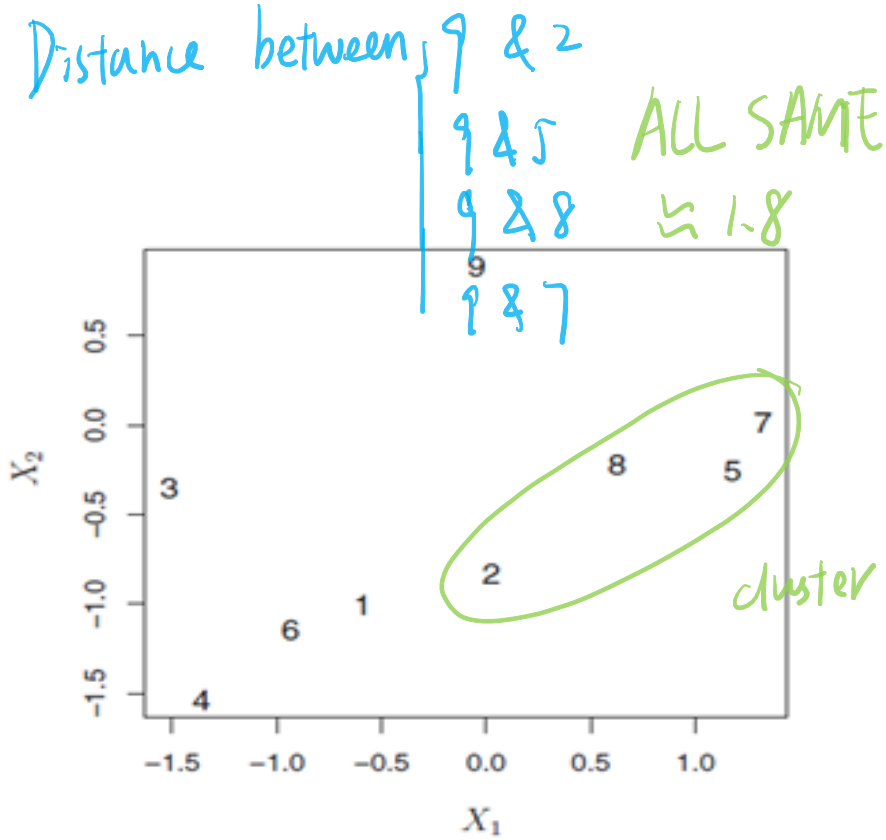
Interpretation of Dendrogram



Cutoff $\Rightarrow$ limits size of hierarchy

Another Example





- 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.

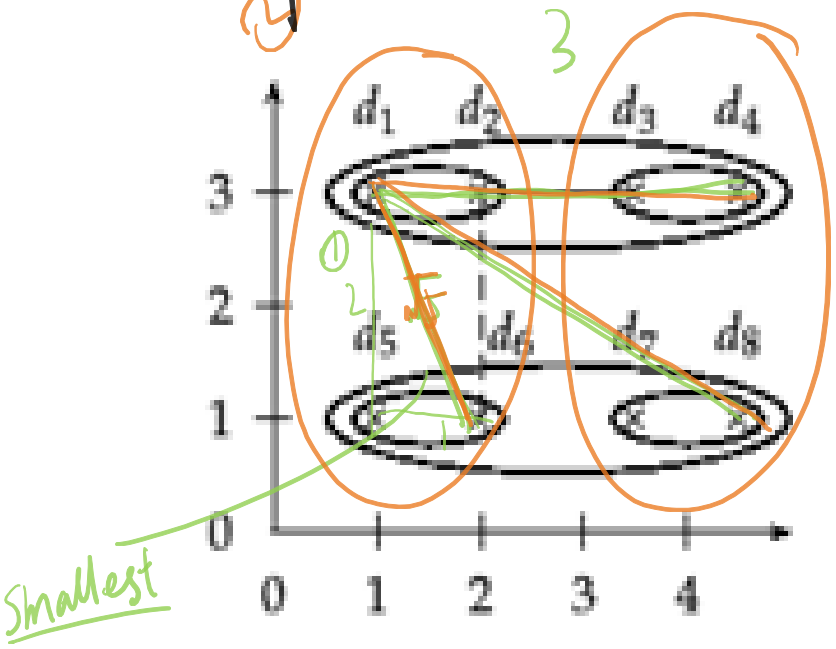
Linkage Type

Define the distance

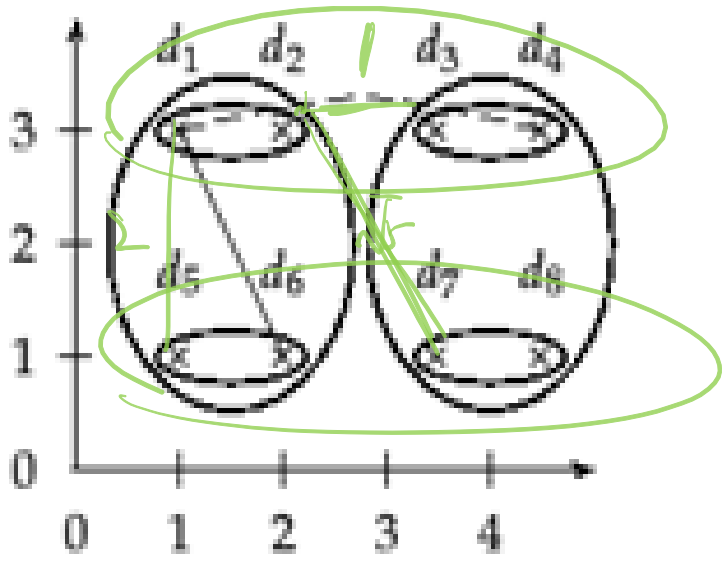
Linkage	Description
Complete	Maximal inter-cluster dissimilarity. Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, and record the largest of these dissimilarities. as the distance of these two cluster
Single	Minimal inter-cluster dissimilarity. Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, and record the smallest of these dissimilarities.
Average	Mean inter-cluster 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.

An Example

Use Complete Method:



Use Single Method



Source: Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, *Introduction to Information Retrieval*, Cambridge University Press. 2008.

Partitional Clustering

- Output a single partition of the data into clusters
- Good for large data sets
- Determining the number of clusters is a major challenge
- K-means clustering: partition the observations into a pre-specified number of clusters.

Details of K-means clustering

Let $C_1, \ldots, C_K$ denote sets containing the indices of the observations in each cluster. These sets satisfy two properties:

1. $C_1 \cup C_2 \cup \ldots \cup C_K = \{1, \ldots, n\}$. In other words, each observation belongs to at least one of the K clusters.
2. $C_k \cap C_{k'} = \emptyset$ for all $k \neq k'$. In other words, the clusters are non-overlapping: no observation belongs to more than one cluster.

No overlap

For instance, if the ith observation is in the kth cluster, then $i \in C_k$.

Details of K-means clustering

- 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 $\mathrm{WCV}(C_k)$ of the amount by which the observations within a cluster differ from each other.
- Hence we want to solve the problem

minimize the total WCV

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

- 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.

How to define within-cluster variation?

- Typically we use Euclidean distance

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

  - where $|C_k|$ denotes the number of observations in the kth cluster.
  - Combining the above equations gives the optimization problem that defines K-means clustering,

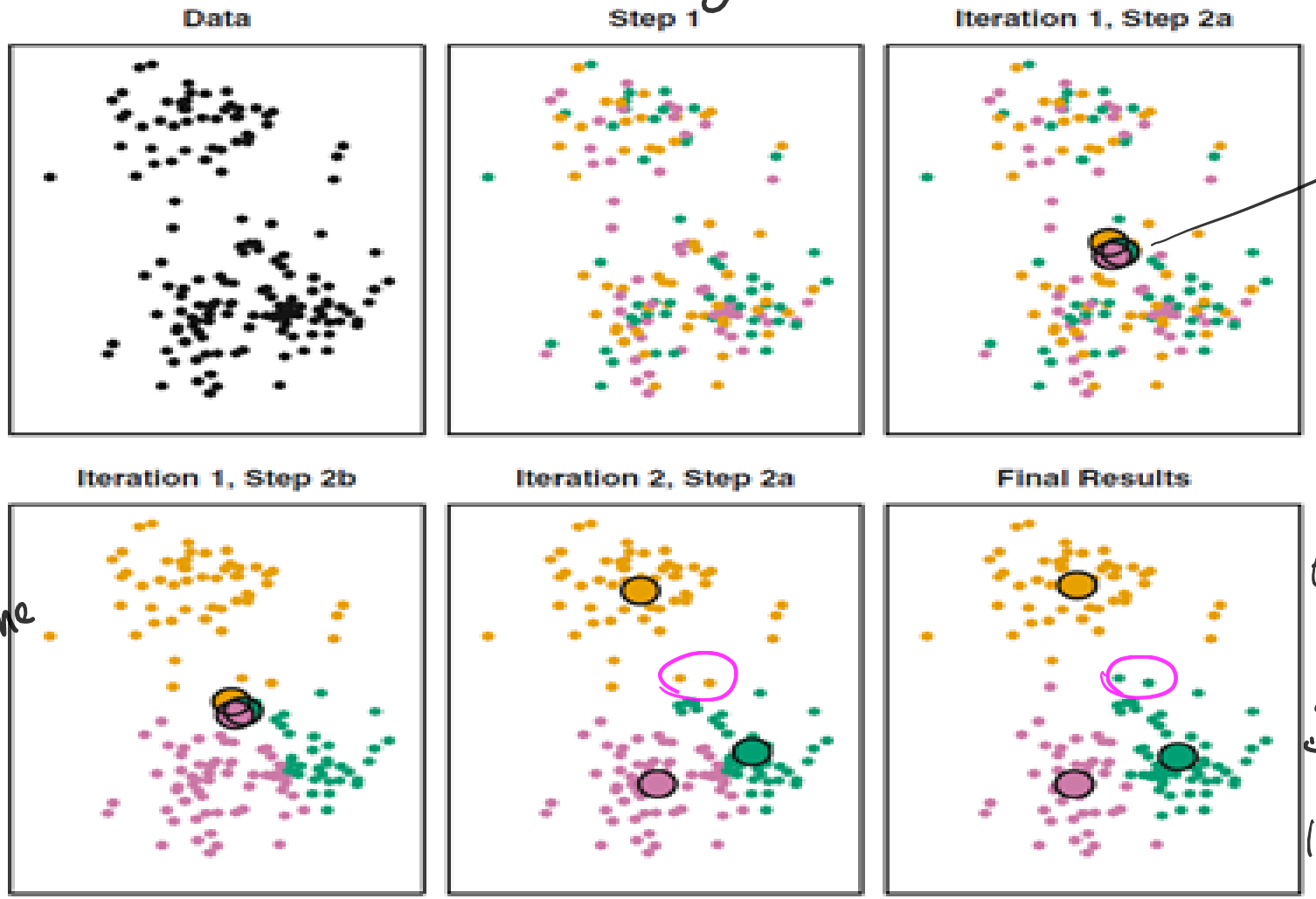
$$\underset{C_1,\ldots,C_K}{\text{minimize}} \left\{ \sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 \right\}$$

K-Means Clustering Algorithm

- Randomly assign a number, from 1 to K, to each of the observations. These serve as initial cluster assignments for the observations.

- Iterate until the cluster assignments stop changing:
  - For each of the K clusters, compute the cluster centroid. The kth cluster centroid is the vector of the p feature means for the observations in the kth cluster.
  - Assign each observation to the cluster whose centroid is closest (where closest is defined using Euclidean distance).

An Example: Clustering Results with K=3



Details of Previous Figure

- The progress of the K-means algorithm with K=3.
  - Top left: The observations are shown.
  - Top center: In Step 1 of the algorithm, each observation is randomly assigned to a cluster.
  - Top right: In Step 2(a), the cluster centroids are computed.

  These are shown as large colored disks. Initially the centroids are almost completely overlapping because the initial cluster assignments were chosen at random.
  - Bottom left: In Step 2(b), each observation is assigned to the nearest centroid.
  - Bottom center: Step 2(a) is once again performed, leading to new cluster centroids.
  - Bottom right: The results obtained after 10 iterations.

An illustration of K-means, with K=2

Data Point	X1	X2
P1	1	1
P2	4	2
P3	2	0
P4	3	5
P5	1	5
P6	3	3
P7	4	10
p8	2	6

Step 1

Randomly assigns cluster to each observation

Cluster	Data Point	X1	X2
1	P1	1	1
2	P2	4	2
1	P3	2	0
2	P4	3	5
1	P5	1	5
2	P6	3	3
1	P7	4	10
2	P8	2	6

Step 2: Iteration 1

Cluster	Data Point	X1	X2
1	P1	1	1
1	P3	2	0
1	P5	1	5
1	P7	4	10

Centroid (2 , 4)

Cluster	Data Point	X1	X2
2	P2	4	2
2	P4	3	5
2	P6	3	3
2	P8	2	6

Centroid (3 , 4)

Step 2: Iteration 2

Cluster	Data Point	X1	X2
1	P1	1	1
1	P3	2	0
1	P5	1	5
1	P8	2	6

Centroid (1.5 , 3)

1 + 2 + 1 + 2 = 6 ÷ 4 = 1.5

1 + 0 + 5 + 6 = 12 ÷ 4 = 3

Cluster	Data Point	X1	X2
2	P2	4	2
2	P4	3	5
2	P6	3	3
2	P7	4	10

Centroid (3.5 , 5)

Step 2: Iteration 3

Cluster	Data Point	X1	X2
1	P1	1	1
1	P3	2	0
1	P5	1	5
1	P6	3	3
1	P2	4	2

Cluster	Data Point	X1	X2
2	P4	3	5
2	P7	4	10
2	P8	2	6

Step 2: Iteration 4

Cluster	Data Point	X1	X2
1	P1	1	1
1	P3	2	0
1	P6	3	3
1	P2	4	2

Cluster	Data Point	X1	X2
2	P4	3	5
2	P7	4	10
2	P8	2	6
2	P5	1	5

K-Means with different random assignment of the observations

Details of Previous Figure

- K-means clustering performed six times on the data from previous figure with $K = 3$, each time with a different random assignment of the observations in Step 1 of the K-means algorithm.
- Above each plot is the value of the objective (4).
- Three different local optima were obtained, one of which resulted in a smaller value of the objective and provides better separation between the clusters.
- Those labeled in red all achieved the same best solution, with an objective value of 235.8

Practical issues

- 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?
  - How many clusters to choose?

Choice of Dissimilarity Measure

- So far have used Euclidean distance.

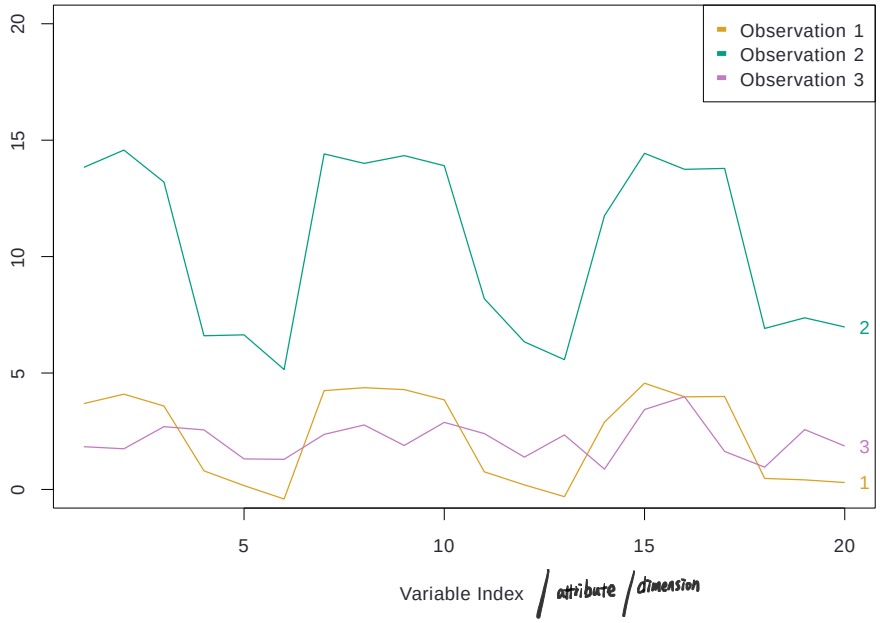
- Other distance measures:
  - Manhattan distance : $d(x, y) = \sum_i^n (|x_i - y_i|)$
  - Chebyshev distance: $d(x, y) = max_i(|x_i - y_i|)$

- An alternative is correlation-based distance which considers two observations to be similar if their features are highly correlated.

Correlation-Based Similarity

- Pearson Correlation
- $r(x,y) = \frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\overline{y})^2}}$

if there're change in the similar way



C_1, C_2 change in similar way

Binary Customer-Item Matrix

Customer	Item A	Item b	Item c
1	1	0	0
2	1	1	0
3	0	0	1
4	1	1	1
5	0	1	1
6	0	1	0
7	0	0	0
8	0	0	0
9	0	0	0

Which type of dissimilarity measure is a better choice?

Correlation

Distance doesn't tell us anything

Numeric Customer-Item Matrix

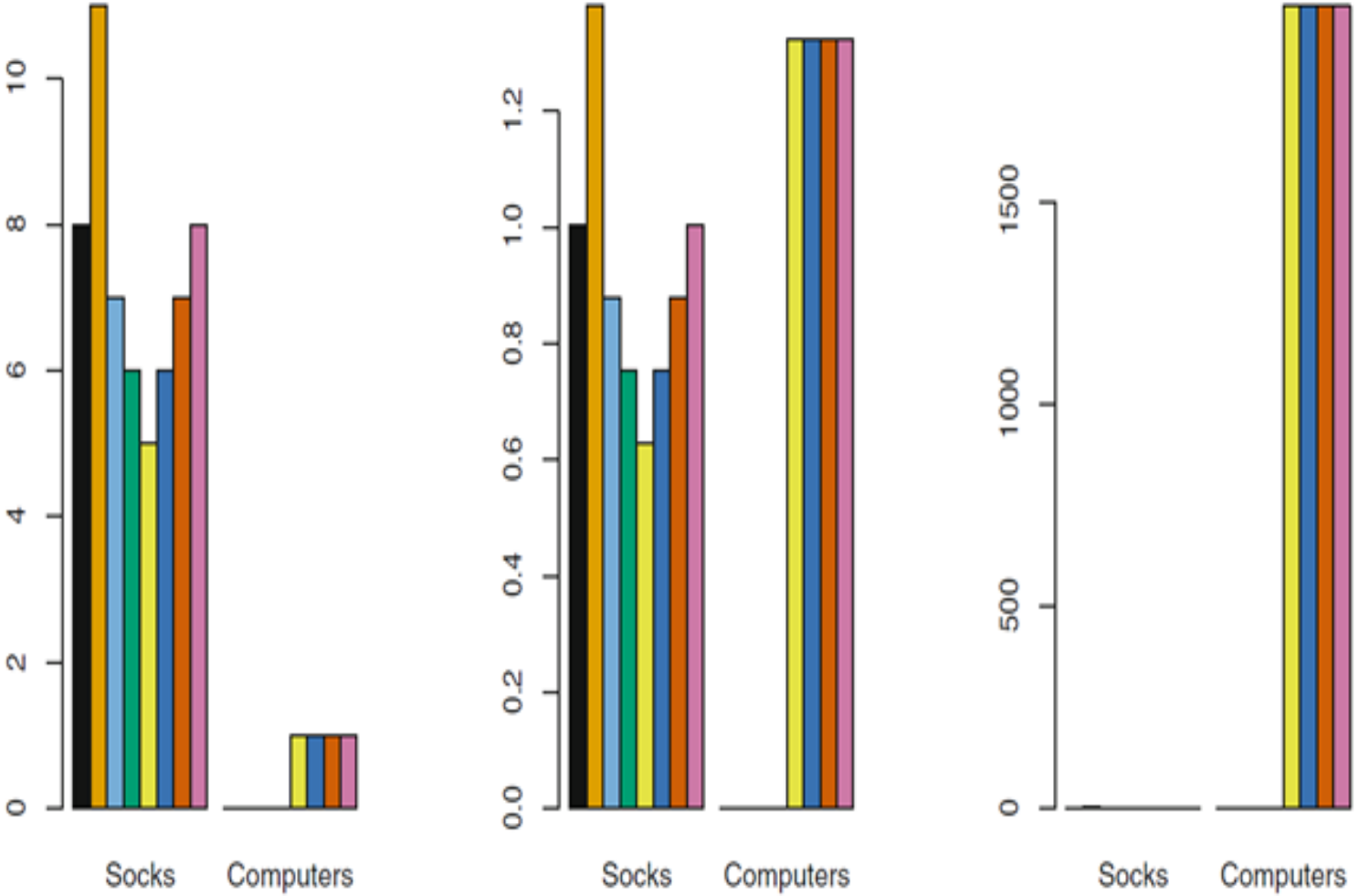
Customer	Socks	Computers
1	8	0
2	12	0
3	7	0
4	6	0
5	4	1
6	5	1
7	7	1
8	8	1

Which type of dissimilarity measure is a better choice?

Neither ??

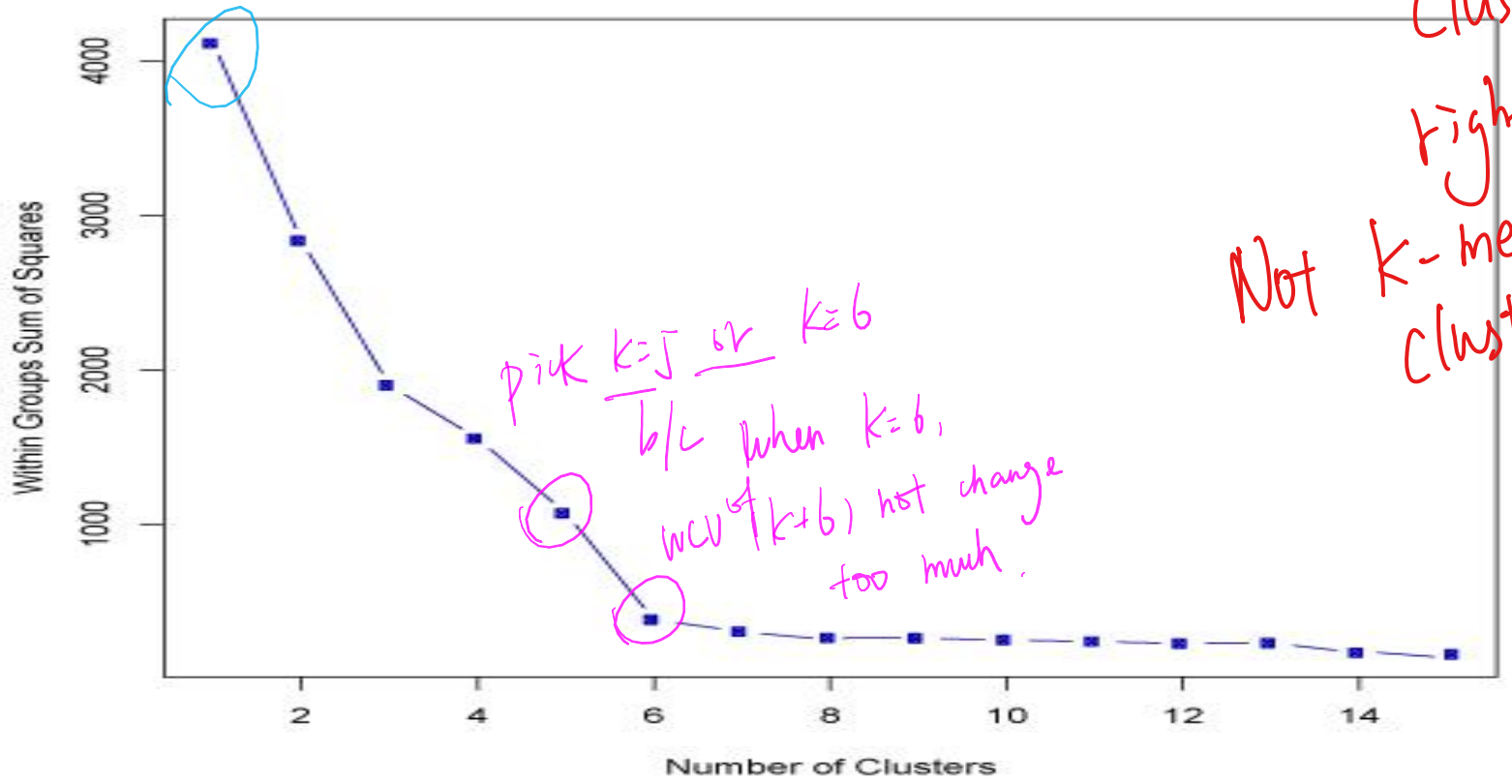
Euclidean Distance after standardization?

Scaling of variables



If intend to standardize the variables to have mean zero and standard deviation one, use the scale() function in R have huge difference between price

Select the number of Cluster → Only hierarchical clustering right ? Not K-means clustering



- *Source: Peeples, Matthew A. 2011 R Script for K-Means Cluster Analysis. Electronic document, http://www.mattpeeples.net/kmeans.html*