CLUSTERING

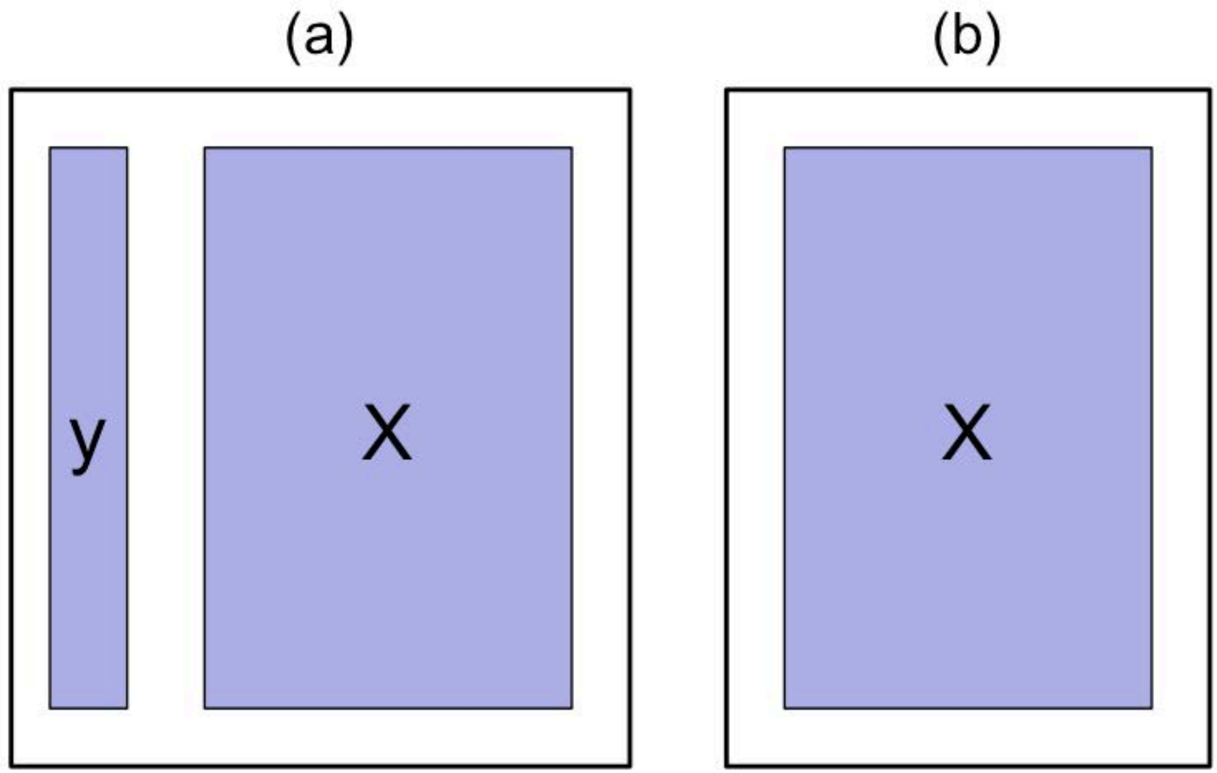
Outline

- What is Clustering?
- Hierarchical Clustering
- K-Means Clustering

WHAT IS CLUSTERING?

Supervised vs. Unsupervised Learning

- Supervised Learning: both X and Y are known
- Unsupervised Learning: only X



Supervised Learning

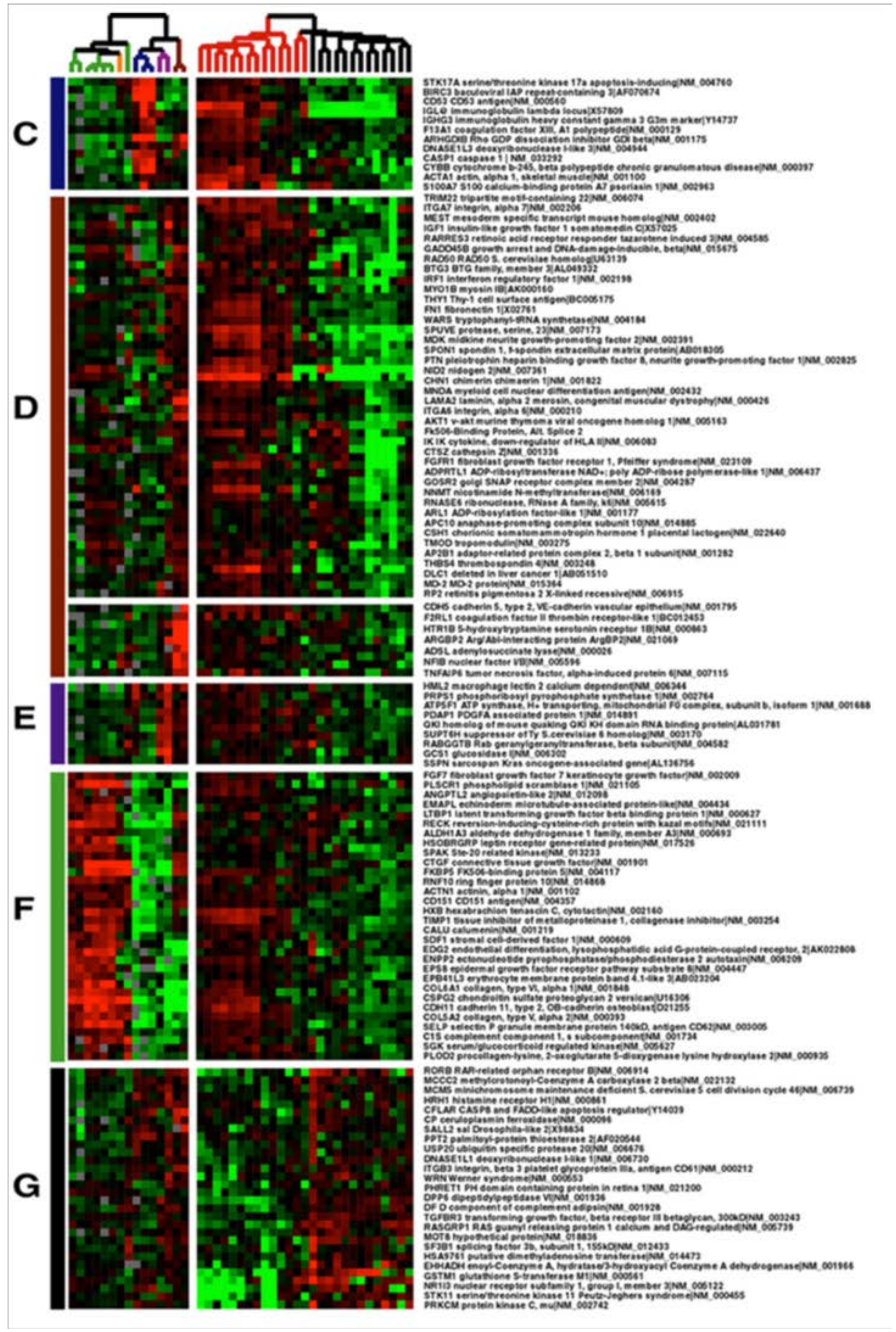
Unsupervised Learning

Clustering

- Clustering refers to a set of techniques for finding subgroups, or clusters, in a data set.
- A good clustering is one when the observations within a group are similar but between groups are very different

An example

- We collect measurements *p* genes on each of *n* breast cancer patients.
    - Different unknown types of cancer could be discovered by clustering (grouping patients).
    - Expression values of genes may change together as a group, these patterns may discovered by clustering (grouping genes).



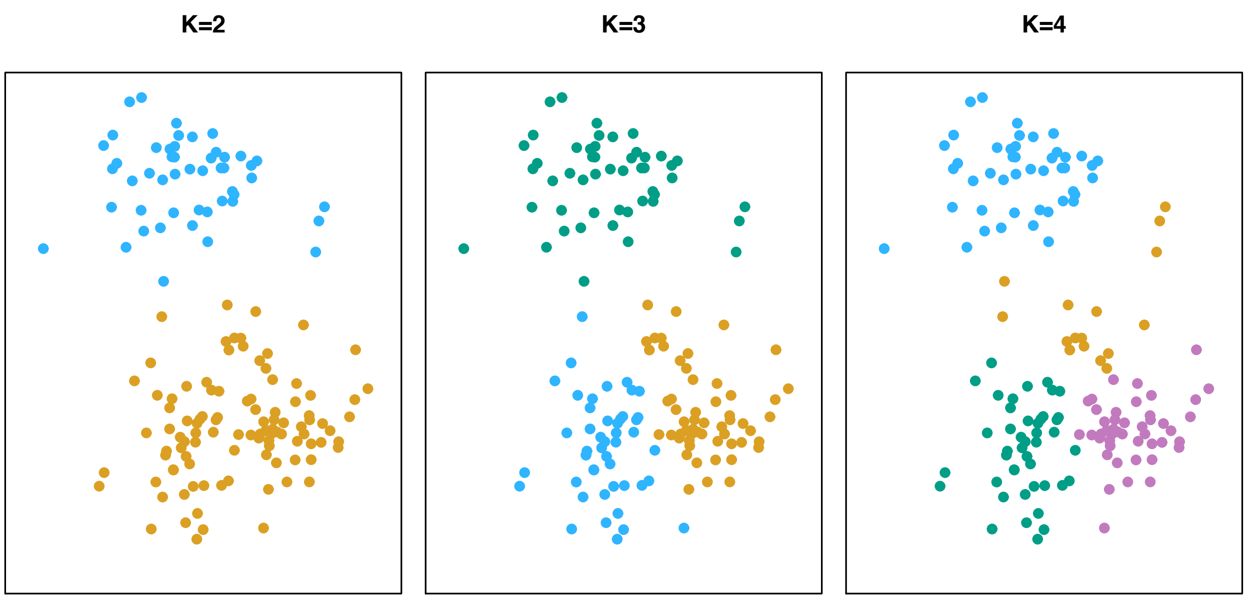
Different Clustering Methods

- There are many different types of clustering methods
- We will concentrate on two of the most commonly used approaches
  - K-Means Clustering
  - Hierarchical Clustering (Similar to grow a decision tree)
  - Model-based Clustering will be introduced later if we have time (soft-clustering, simultaneous clustering, etc)

K-MEANS CLUSTERING

K-Means Clustering

- To perform K-means clustering, one must first specify the desired number of clusters K
- Then the K-means algorithm will assign each observation to exactly one of the K clusters



How does K-Means work?

- We would like to partition that data set into K clusters

$$C_1, \ldots, C_K$$

  - Each observation belong to at least one of the K clusters
  - The clusters are non-overlapping, i.e. no observation belongs to more than one cluster
- The objective is to have a minimal “within-cluster-variation”, i.e. the elements within a cluster should be as similar as possible
- One way of achieving this is to minimize the sum of all the pair-wise squared Euclidean distances between the observations in each cluster.

$$\underset{C_1,\ldots,C_K}{\text{minimize}} \left\{ \sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 \right\} \quad (4)$$

$\Leftrightarrow$ Minimize $J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \|\mathbf{x}_n - \boldsymbol{\mu}_k\|^2$

K-Means Algorithm (special case of EM)

- Initial Step: Randomly assign each observation to one of K clusters

- Iterate until the cluster assignments stop changing:
  - For each of the K clusters, compute the cluster centroid. The k^th cluster centroid if the mean of the observations assigned to the k^th cluster

$$\boldsymbol{\mu}_k = \frac{\sum_n r_{nk}\mathbf{x}_n}{\sum_n r_{nk}}$$

solving $\frac{\partial J}{\partial \mu_k} = 0$

  - Assign each observation to the cluster whose centroid is closest (where "closest" is defined using Euclidean distance.

$$r_{nk} = \begin{cases} 1 & \text{if } k = \arg\min_j \|\mathbf{x}_n - \boldsymbol{\mu}_j\|^2 \\ 0 & \text{otherwise.} \end{cases}$$

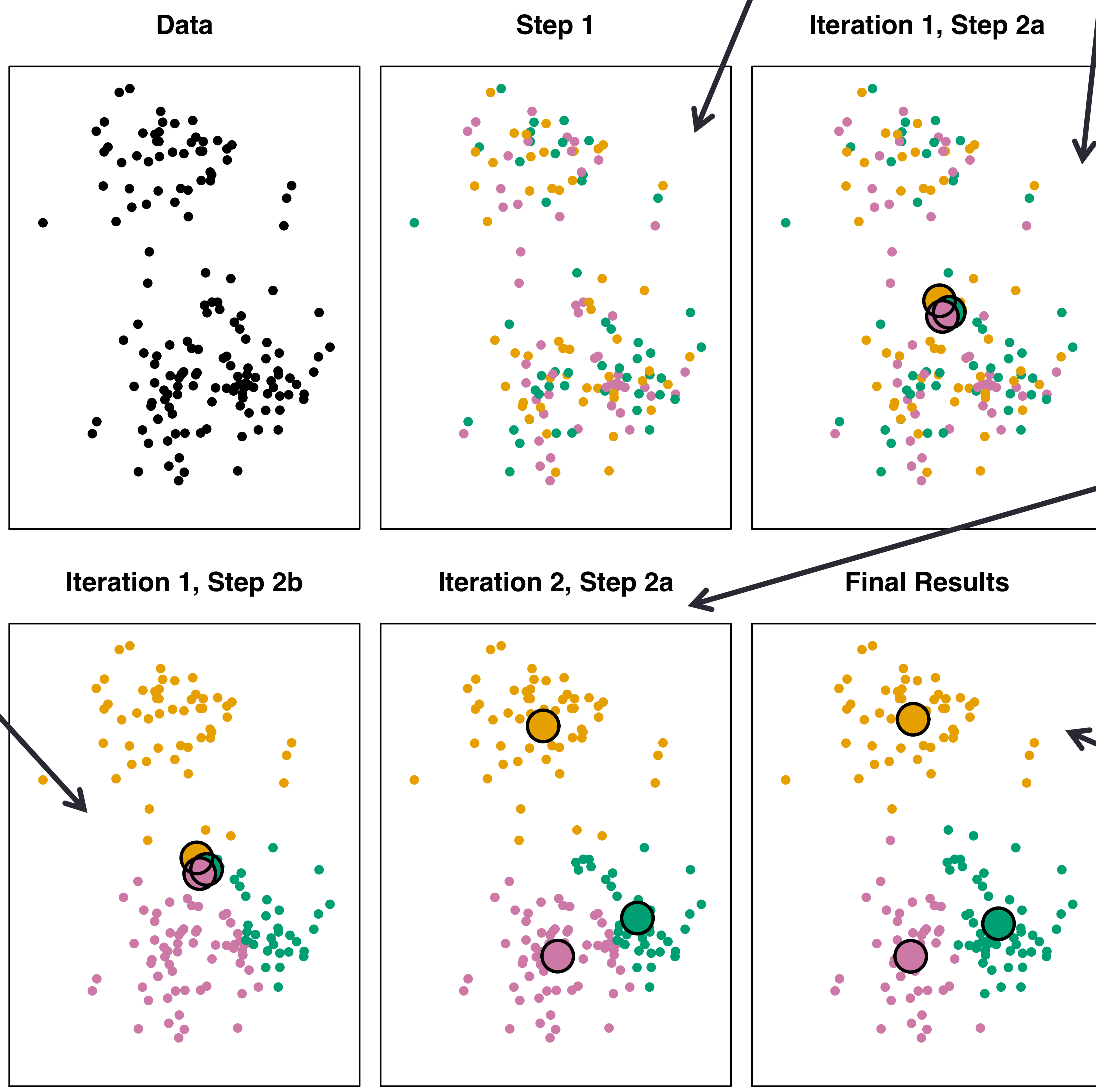
⬅️ Hard
Soft ➡️

$$z_{ij} = \frac{e^{-\beta\left\|x_j - \mu_i\right\|^2}}{\sum_{l=1}^{k} e^{-\beta\left\|x_j - \mu_l\right\|^2}}$$

An Illustration of the K-Means Algorithm

Random Assignment of points

Compute cluster centers from initial assignments



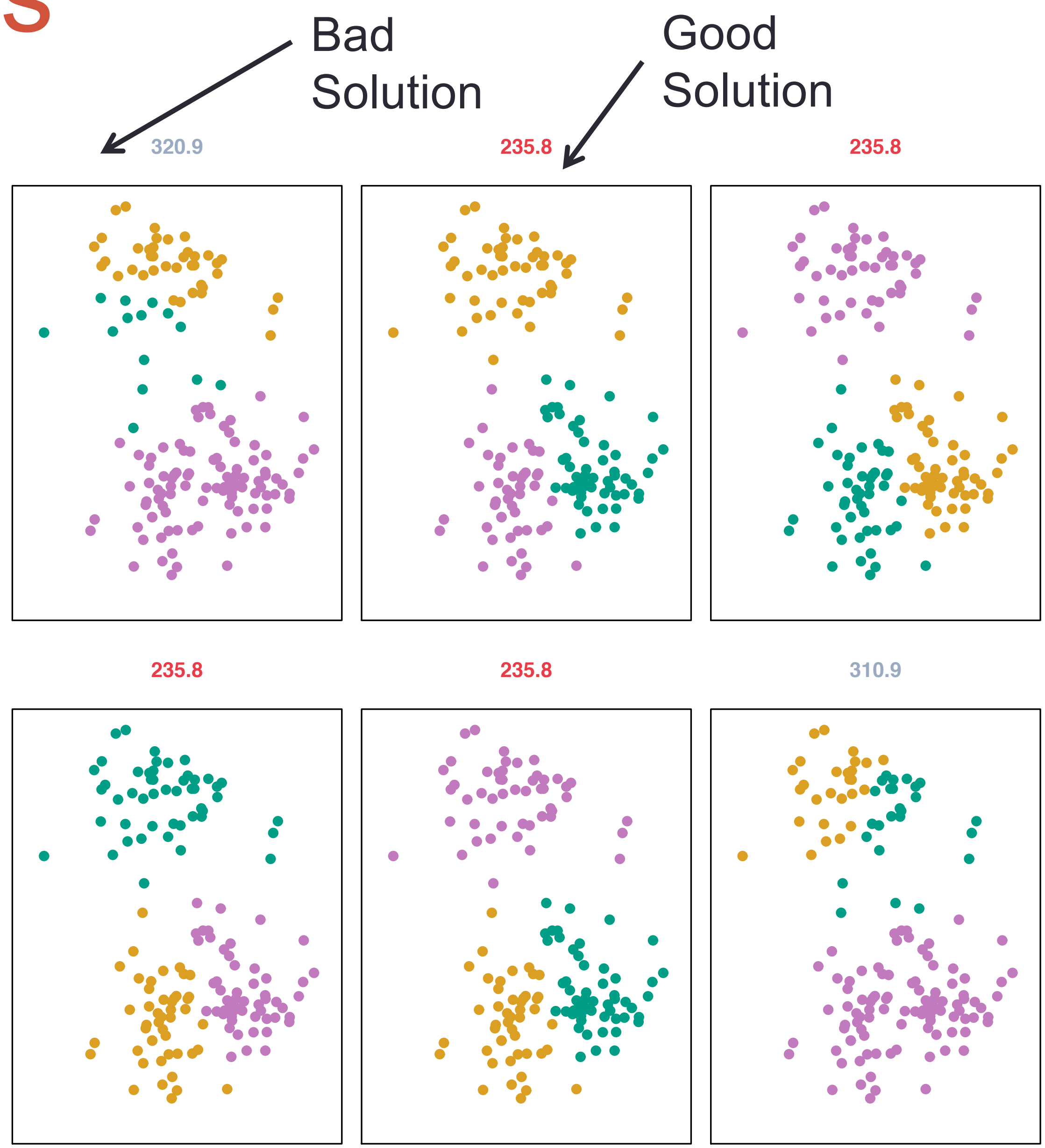
Computer new cluster centers

Assign points to closest cluster center

Now there is no further change so stop

Local Optimums

- The K-means algorithm can get stuck in "local optimums" and not find the best solution
- Hence, it is important to run the algorithm multiple times with random starting points to find a good solution



Property of K-Means

- This algorithm is guaranteed to decrease the value of the objective (4) at each step. *Why?* Note that

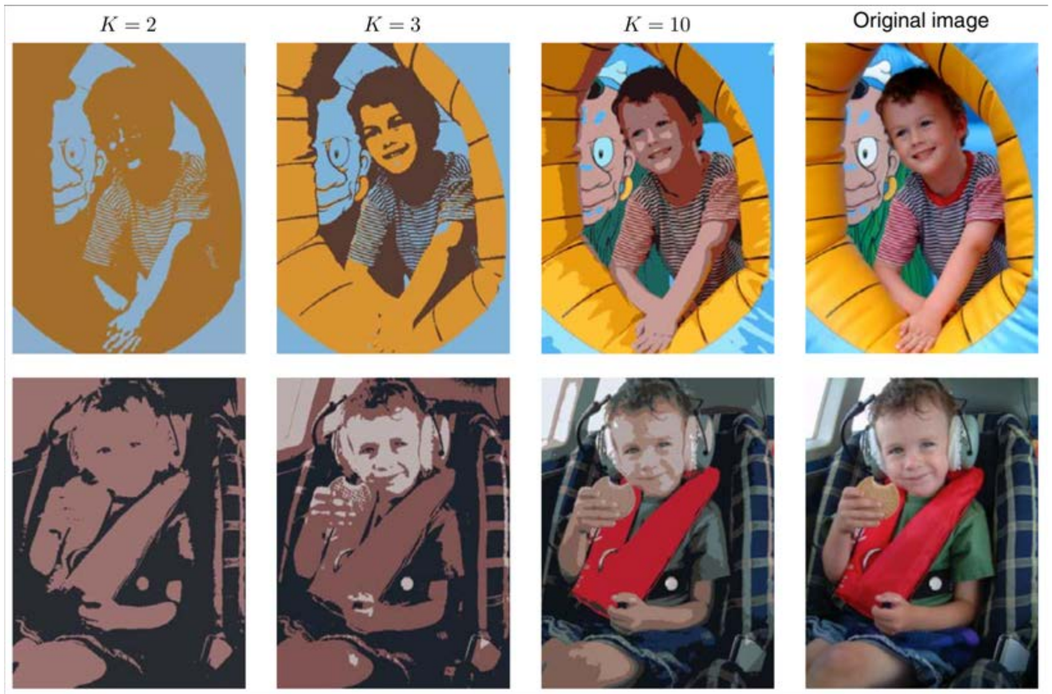
$$\frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 = 2 \sum_{i \in C_k} \sum_{j=1}^{p} (x_{ij} - \bar{x}_{kj})^2,$$

  where $\bar{x}_{kj} = \frac{1}{|C_k|} \sum_{i \in C_k} x_{ij}$ is the mean for feature j in cluster C_k.

- however it is not guaranteed to give the global minimum. *Why not?*

Application: Image segmentation and compression

- Treats each pixel in the image as a separate data point.
- 4K resolution 3840 x 2160 -> 8.3M data points
- Each data point contains 3 variables, comprising the intensities of the red, blue, and green channels.
- Original data: 8.3M x 3 matrix, Values: double precision (64bit/number)
- Apply K-Means to group pixels
- Compress image to cluster membership only, i.e. 8.3M X 1 vector, value: integer or log2(K) bit binary.



Online K-Means algorithm (MacQueen, 1967)

- Why do we need online K-Means?
  - Data comes in sequential. We need to keep updating clustering results.
  - Data is too huge to process all in one time. (alternative solution is down-sampling)
- For new data point x_n

$$\boldsymbol{\mu}_k^{\text{new}} = \boldsymbol{\mu}_k^{\text{old}} + \eta_n(\mathbf{x}_n - \boldsymbol{\mu}_k^{\text{old}})$$

- η_n is learning rate parameter which decrease with n.

K-medoids algorithm

- K-means algorithm is based on the use of squared Euclidean distance (L2 norm)
- The determination of the cluster means non-robust to outliers
- K-*medoids* algorithm use any choice of dissimilarity dissimilarity measure V(x,x')
- Solution found by minimize

$$\widetilde{J} = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \mathcal{V}(\mathbf{x}_n, \boldsymbol{\mu}_k)$$

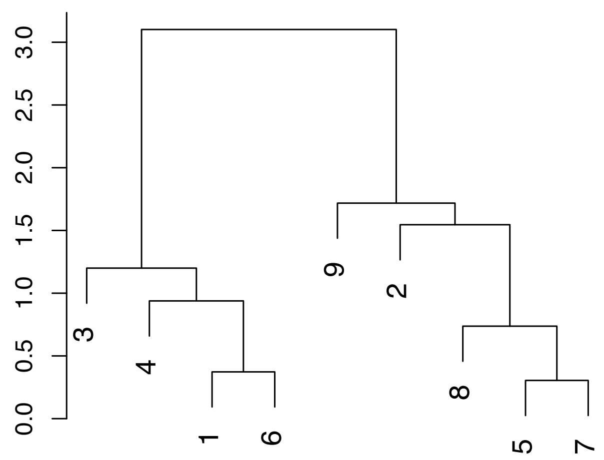
HIERARCHICAL CLUSTERING

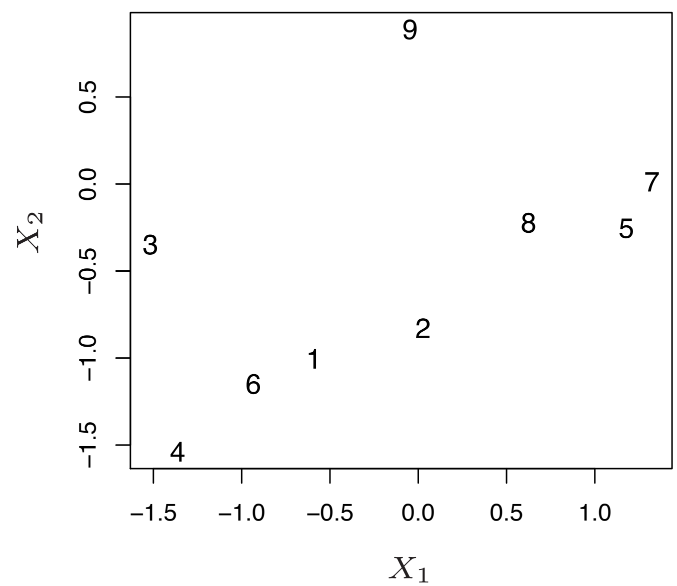
Hierarchical Clustering

- K-Means clustering requires choosing the number of clusters.
- If we don’t want to do that, an alternative is to use Hierarchical Clustering
- Hierarchical Clustering has an added advantage that it produces a tree based representation of the observations, called a Dendogram (looks similar to decision tree)

Dendograms

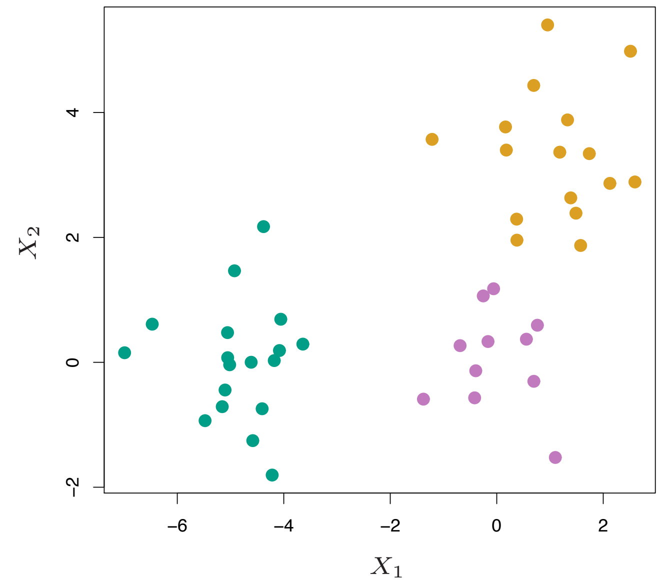
- 9 samples, 2 variables, hierarchical clustering
- First join closest points (5 and 7)
- Height of fusing/merging (on vertical axis) indicates how similar the points are
- After the points are fused they are treated as a single observation and the algorithm continues

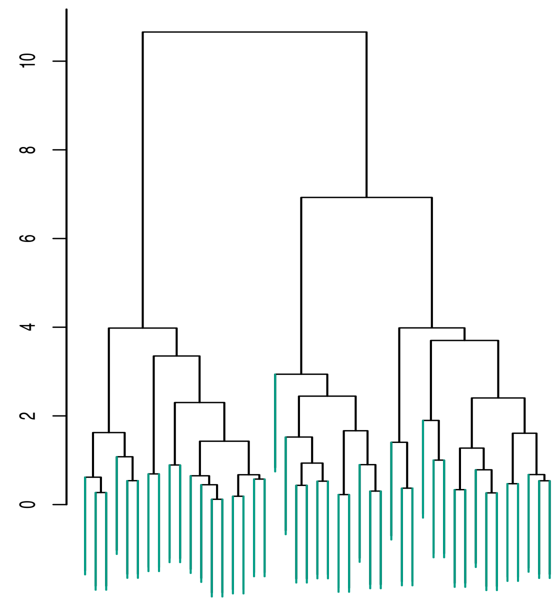




Interpretation

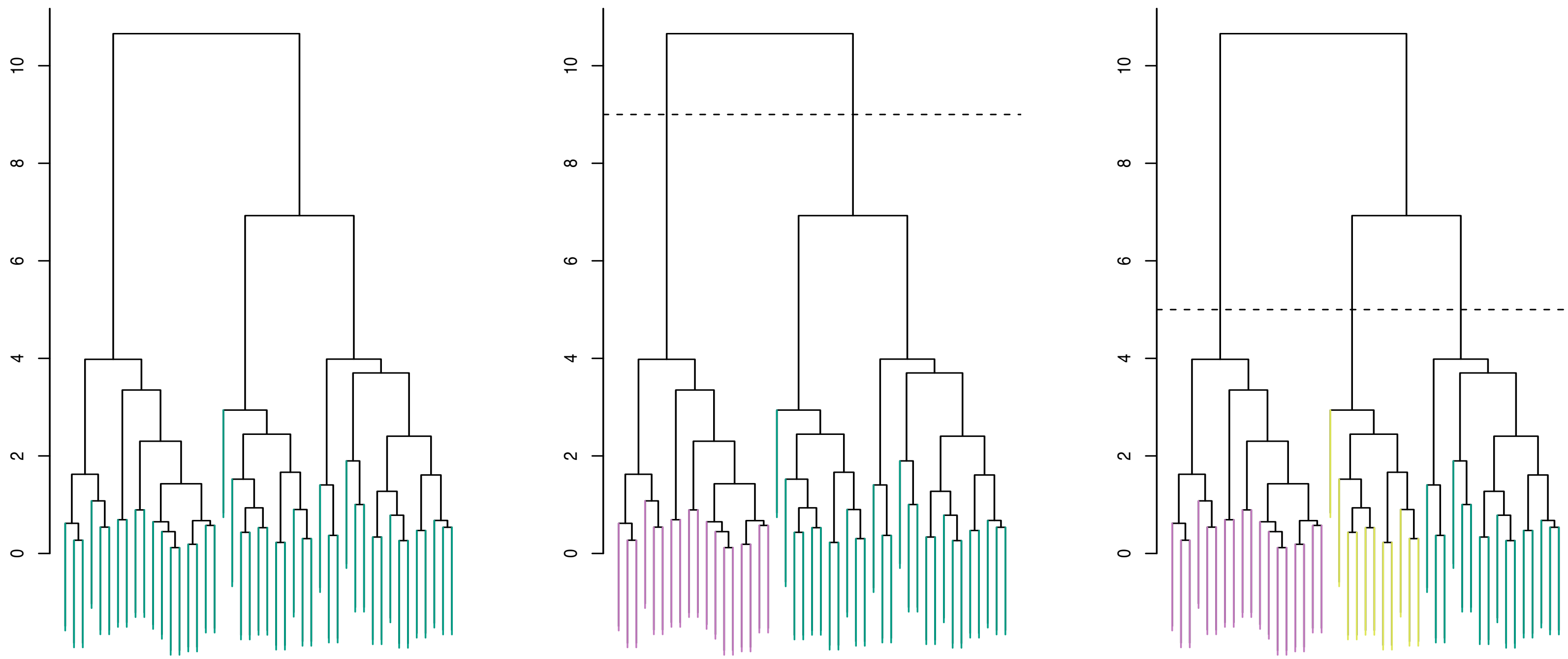
- Each "leaf" of the dendogram represents one of the 45 observations
- At the bottom of the dendogram, each observation is a distinct leaf. However, as we move up the tree, some leaves begin to fuse. These correspond to observations that are similar to each other.
- As we move higher up the tree, an increasing number of observations have fused. The earlier (lower in the tree) two observations fuse, the more similar they are to each other.
- Observations that fuse later are usually but not always quite different
- Greedy algorithm!





Choosing Clusters

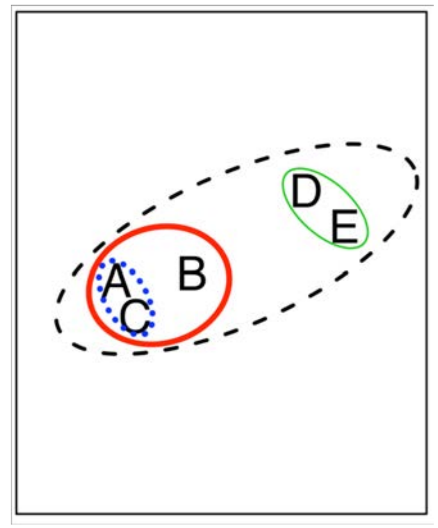
- To choose clusters we draw lines across the dendogram
- We can form any number of clusters depending on where we draw the break point.
- No agreed criteria to trim these trees

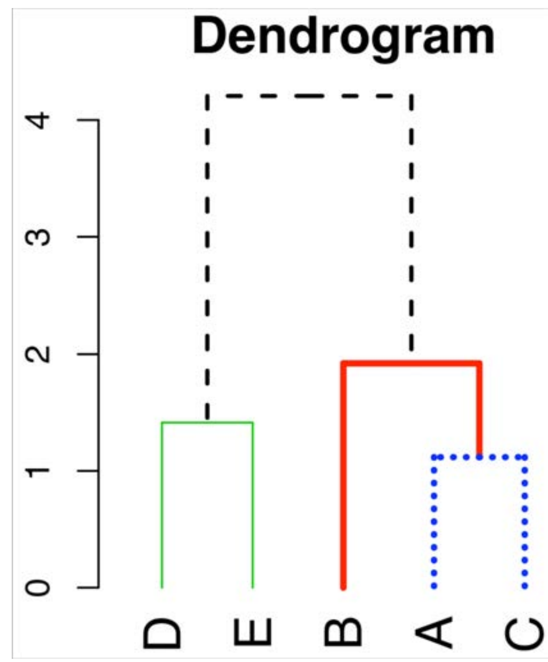


One Cluster Two Clusters Three Clusters

Algorithm (Agglomerative Approach)

- The dendogram is produced as follows:
  - Start with each point as a separate cluster (n clusters)
  - Calculate a measure of dissimilarity between all points/ clusters
  - Fuse two clusters that are most similar so that there are now n-1 clusters
  - Fuse next two most similar clusters so there are now n-2 clusters
  - Continue until there is only 1 cluster



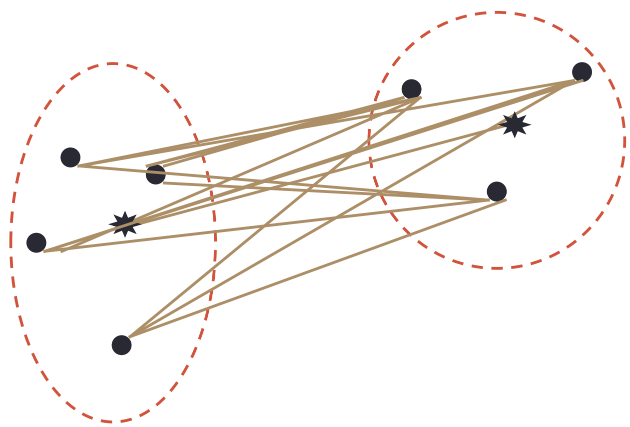


How do we define dissimilarity?

- Implementing hierarchical clustering involves one obvious issue
- How do we define the dissimilarity, or linkage, between the fused (A,B) cluster and C?
- There are four options:
  - Complete Linkage
  - Single Linkage
  - Average Linkage
  - Centriod Linkage

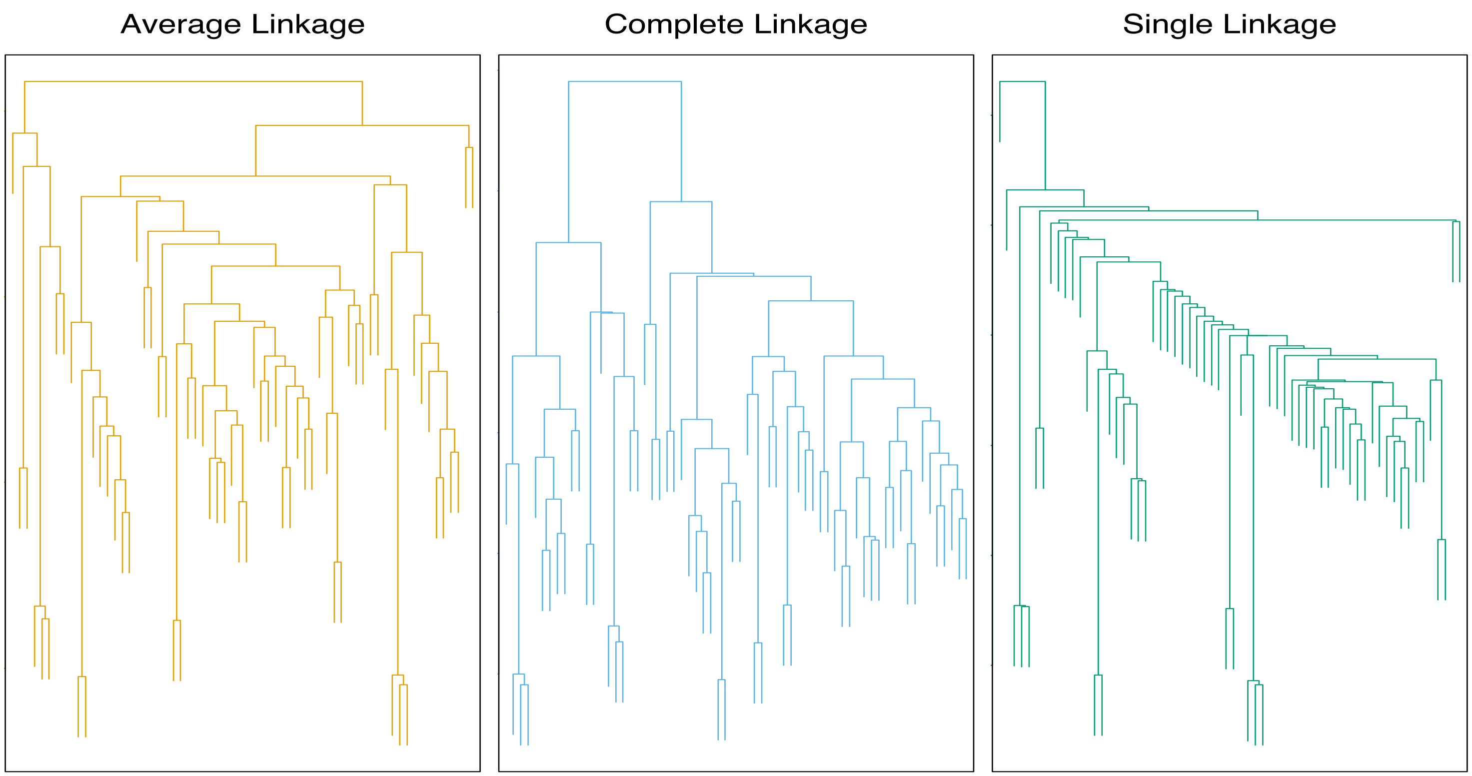
Linkage Methods: Distance Between Clusters

- **Complete Linkage:** Largest distance between observations
- **Single Linkage:** Smallest distance between observations
- **Average Linkage**: Average distance between observations
- **Centroid:** distance between centroids of the observations



Linkage Can be Important

- Here we have three clustering results for the same data
- The only difference is the linkage method but the results are very different
- Complete and average linkage tend to yield evenly sized clusters whereas single linkage tends to yield extended clusters to which single leaves are fused one by one.

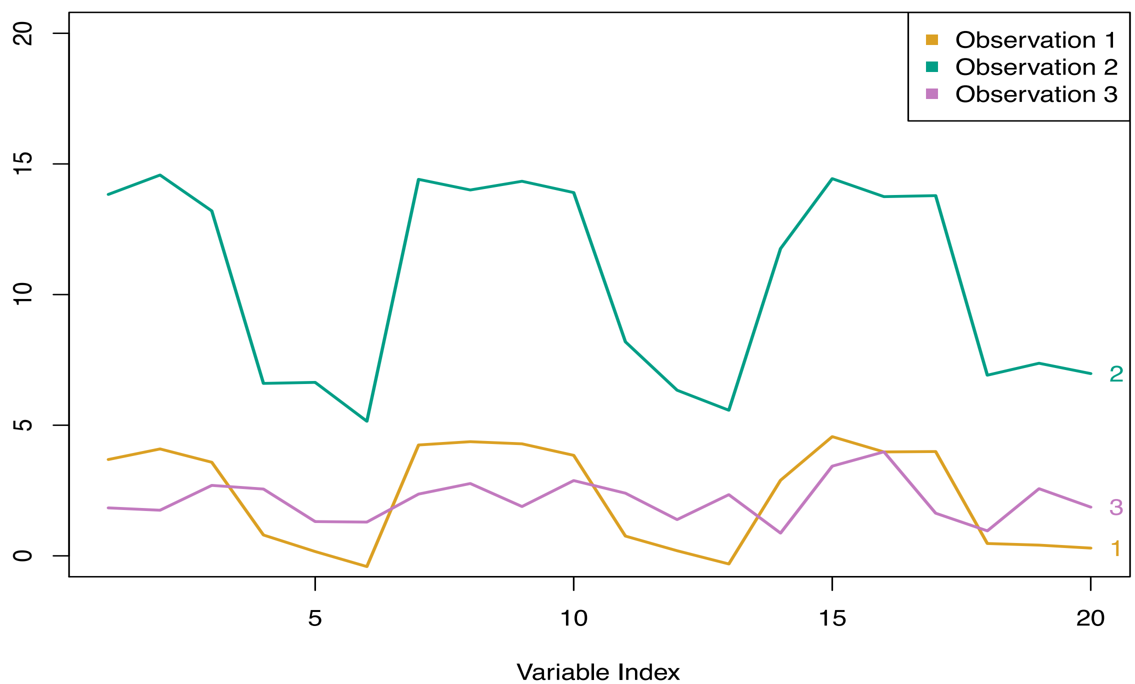


Choice of Dissimilarity Measure

- So far, we have considered using Euclidean distance as the dissimilarity measure
- However, an alternative measure that could make sense in some cases is the correlation based distance

Comparing Dissimilarity Measures

- In this example, we have 3 observations and p = 20 variables
- In terms of Euclidean distance obs. 1 and 3 are similar
- However, obs. 1 and 2 are highly correlated so would be considered similar in terms of correlation measure

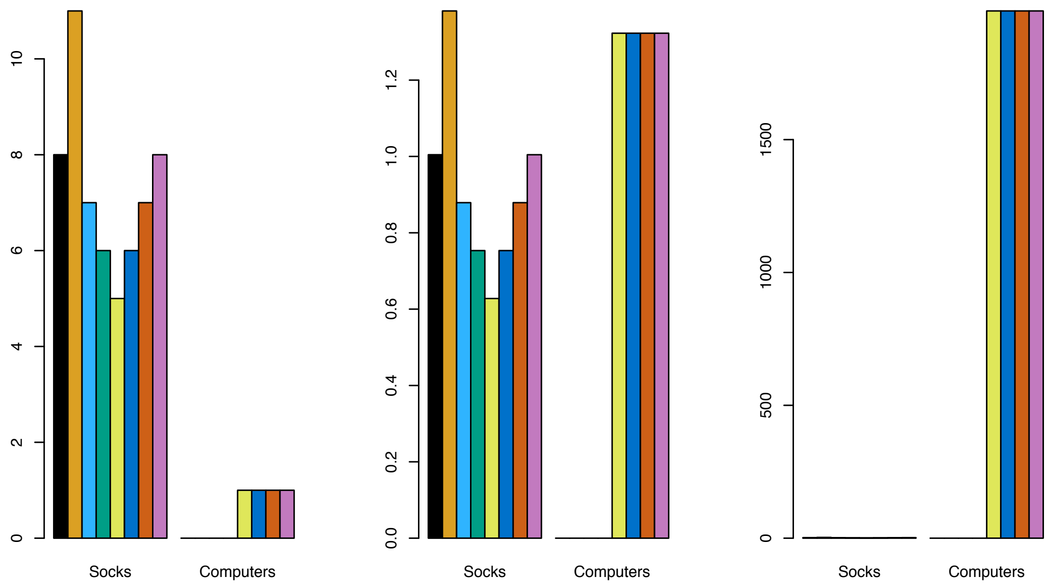


Online Shopping Example

- Suppose we record the number of purchases of each item (columns) for each customer (rows)
- Using Euclidean distance, customers who have purchases very little will be clustered together
- Using correlation measure, customers who tend to purchase the same types of products will be clustered together even if the magnitude of their purchase may be quite different

Standardizing the Variables

- Consider an online shop that sells two items: socks and computers
  - Left: In terms of quantity, socks have higher weight
  - Center: After standardizing, socks and computers have equal weight
  - Right: In terms of dollar sales, computers have higher weight



FINAL THOUGHTS

Practical Issues in Clustering

- In order to perform clustering, some decisions must be made:
  - Should the features first be standardized? i.e. Have the variables centered to have a mean of zero and standard deviation of one.
  - In case of hierarchical clustering:
    - What dissimilarity measure should be used? (*Kmeans only use Euclidean, since "centroid" is from Euclidean geometry.*)
    - What type of linkage should be used?
    - Where should we cut the dendogram in order to obtain clusters?
  - In case of K-means clustering:
    - How many clusters should we look for the data?
- In practice, we try several different choices, and look for the one with the most useful or interpretable solution. There is no single right answer!

Final Thoughts

- Most importantly, one must be careful about how the results of a clustering analysis are reported
- These results should **not** be taken as the absolute truth about a data set
- Rather, they should constitute a starting point for the developments of a scientific hypothesis and further study, preferably on independent data