

Clustering Methods

CS 530
Chapman
Spring 2021

1

Take 🏠 Message for rest of course: Clustering

- What is unsupervised learning?
- What is clustering?
 - The k -Means algorithm
 - Hierarchical clustering

2

TABLE OF CONTENTS

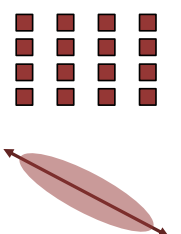
- Introduction to Unsupervised Learning
- k -Means Clustering
 - The Clustering Algorithm
 - Uses of k -Means
 - How to choose k ?
- k -Medoids Clustering, briefly
- Hierarchical Clustering, briefly

3

Unsupervised Learning

Unsupervised learning strives to find structure in unlabeled data, typically changing its representation. Unlabeled data are a collection of explanatory variables (x_1, \dots, x_p) without a response (label) variable y .

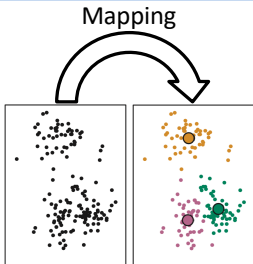
- Clustering Methods
 - k -Means Clustering: Group samples based on some measure of proximity
 - Hierarchical Clustering: Group samples based on similarity in a proximity-based tree
- Dimensionality reduction
 - Principal Component Analysis (PCA): Represent variables using combinations with the greatest variance.



4

Clustering Methods

Clustering methods map unlabeled data into clusters according to some inner structure. This can help explain data that might otherwise appear unrelated and to have a more-concise representation of the data. Clustering thus searches for homogeneous subgroups among the observations.



[Image Source](#)

5

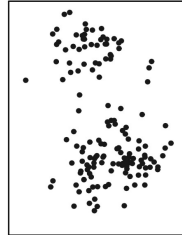
THE K-MEANS ALGORITHM

6

k-Means Clustering

Given a set of unlabeled samples that we want to cluster into subgroups, how do we decide on the mapping of samples to clusters? Which samples belong together?

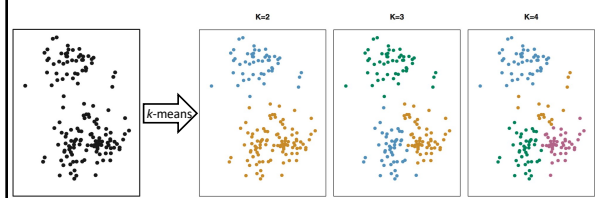
The k -means clustering algorithm needs the number of clusters, k , as input. It then starts with a random clustering of the data into k clusters and iteratively rearranges the clusters until they converge (stop changing with iterations).



[Image Source](#)

7

k-Means Clustering



[Image Source](#)

Above are result of applying k -means clustering—with $k = 2, 3$, and 4 —to the same data set. Which one is the “correct” clustering?

8

k-Means Clustering

Formally, say we have n samples in m -dimensional space; so $\vec{x}_i \in \mathbb{R}^m, i = 1, \dots, n$

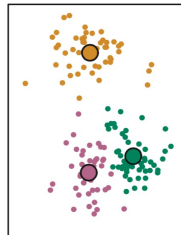
$$X = \{\vec{x}_1, \vec{x}_2, \dots, \vec{x}_n\},$$

The k -means algorithm partitions X into a set of k clusters

$$\{C_1, \dots, C_k\}$$

such that

- Each sample belongs to a cluster
 $C_1 \cup C_2 \cup \dots \cup C_k = X$
- Clusters do not intersect
 $C_i \cap C_j = \emptyset \quad \forall i, j$



[Image Source](#)

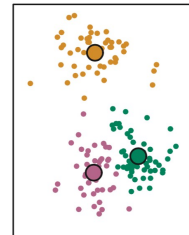
9

k-Means Clustering

The (rather intuitive) idea of k -means is that—for a given k —we want to minimize the variation within each cluster, $W(C_p)$, where

$$W(C_p) = \frac{1}{|C_p|} \sum_{i,j \in C_p} (\vec{x}_i - \vec{x}_j)^2.$$

↑ Intra-cluster variation
 L2 norm



[Image Source](#)

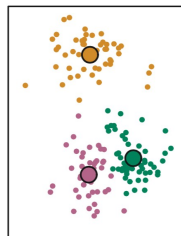
10

k-Means Clustering

The (rather intuitive) idea of k -means is that—for a given k —we want to minimize the variation within each cluster, $W(C_p)$, where

$$W(C_p) = \frac{1}{|C_p|} \sum_{i,j \in C_p} (\vec{x}_i - \vec{x}_j)^2.$$

We want to minimize the squared Euclidean distance within each cluster.
 $|C_p|$ denotes the number of elements in cluster p —or the cardinality of cluster p . We normalize the within-cluster distance by this cardinality to balance different cluster sizes.



[Image Source](#)

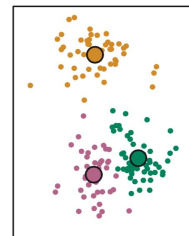
11

k-Means Clustering

Mathematically stated, we want to solve the following optimization problem:

$$\min_{C_1, \dots, C_k} \left\{ \sum_{p=1}^k W(C_p) \right\} = \min_{C_1, \dots, C_k} \left\{ \sum_{p=1}^k \frac{1}{|C_p|} \sum_{i,j \in C_p} (\vec{x}_i - \vec{x}_j)^2 \right\}.$$

In English, we want to find the division into k clusters that minimizes the overall within-cluster distances across all the clusters.



[Image Source](#)

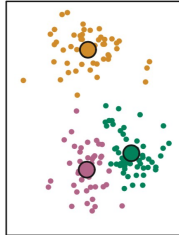
12

k-Means Clustering

Mathematically stated, we want to solve the following optimization problem:

$$\min_{C_1, \dots, C_k} \left\{ \sum_{p=1}^k W(C_p) \right\} = \min_{C_1, \dots, C_k} \left\{ \sum_{p=1}^k \frac{1}{|C_p|} \sum_{i,j \in C_p} (\vec{x}_i - \vec{x}_j)^2 \right\}.$$

Can you suggest an algorithm guaranteed to find the global minimum of $\{\sum_{p=1}^k W(C_p)\}$ over C_1, \dots, C_k ?



[Image Source](#)

13

Group exercise



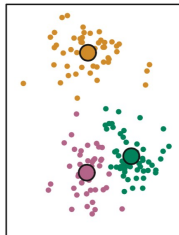
In breakout rooms, we will now work on a google doc

14

k-Means Clustering

There are almost k^n ways to partition n observations into k clusters. So, brute-force testing all partitions to find the absolute minimum of $\{\sum_{p=1}^k \frac{1}{|C_p|} \sum_{i,j \in C_p} (\vec{x}_i - \vec{x}_j)^2\}$ is intractable. But we can find a *local minimum* with the algorithm:

1. Pick k , the total number of clusters.
2. Randomly assign each sample to a cluster.
3. Repeat the below until the clusters stabilize:
 - I. Compute each cluster's centroid:
 - II. $\vec{x}_{C_p} = \frac{1}{|C_p|} \sum_{i \in C_p} \vec{x}_i$.
 - III. Assign each sample to the cluster to whose centroid it is closest (typically using Euclidean distance).



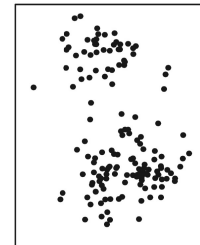
[Image Source](#)

15

k-Means Clustering Algorithm

The k -means clustering algorithm in operation:

We decide to cluster our unlabeled data into $k = 3$ clusters.

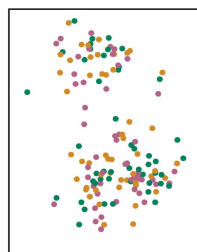


[Image Source](#)

17

k-Means Clustering Algorithm

The algorithm randomly assigns each sample to one of the three clusters (depicted as green, orange, or purple).

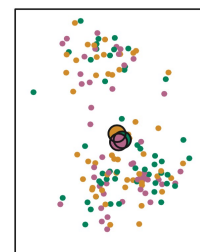


[Image Source](#)

18

k-Means Clustering Algorithm

The centroids for each cluster are then calculated. They are depicted as large, colored circles.

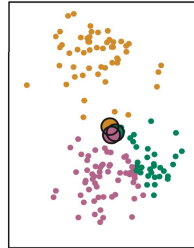


[Image Source](#)

19

k-Means Clustering Algorithm

The cluster labels for each sample are reassigned to the cluster with the closest centroid.



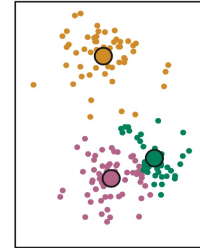
[Image Source](#)

20

k-Means Clustering Algorithm

The centroids are recalculated based on current reassignment into clusters.

After this *single iteration*, the clusters have already taken shape. The centroids and cluster members are already close to their final positions.



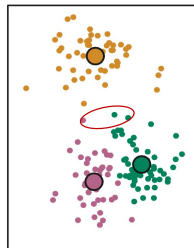
[Image Source](#)

21

k-Means Clustering Algorithm

After a few more iterations, the centroids reach a stable equilibrium. The *k*-means algorithm then stops.

Note, very few samples change clusters in later iterations.

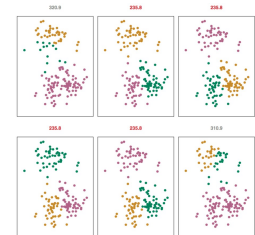


[Image Source](#)

22

k-Means Clustering

Running *k*-means 6 times on the same data set, with $k = 3$. The overall within-cluster variance, $\left\{ \sum_{p=1}^k \frac{1}{|C_p|} \sum_{i,j \in C_p} (\vec{x}_i - \vec{x}_j)^2 \right\}$, is given above each run.



[Image Source](#)

The *k*-means algorithm is therefore typically run several times, each time with a different random initial clustering. This often yields different results. The clustering resulting in the minimal within-cluster variability over all runs is then selected. It is closest (or equal) to the global minimum.

23

How to choose k ?

How to determine which value of k to use when using *k*-means clustering?

The higher k is, the lower the inter-cluster variance. When p (number of clusters) = n (number of observations, samples):

$$\left\{ \sum_{p=1}^k \frac{1}{|C_p|} \sum_{i,j \in C_p} (\vec{x}_i - \vec{x}_j)^2 \right\} = 0$$

But the clustering is trivial and achieves nothing.

27

How to choose k ?

How to determine which value of k to use when using *k*-means clustering? Sometimes more art than science.

- AIC, BIC.
- Elbow Method with percent of variance explained.
- Gap Statistic
- Silhouette Coefficient

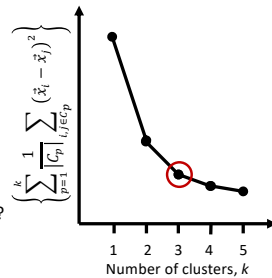
None of these methods can be used blindly. You must make sure that your results make sense.

28

How to choose k ? Elbow Method

The elbow method consists of scrutinizing the plot of the minimization formula for each value of k , and then choosing the best value for k based on the “elbow” of the plot. The elbow is where the decrease in the plot begins to diminish.

What the problem with this method?

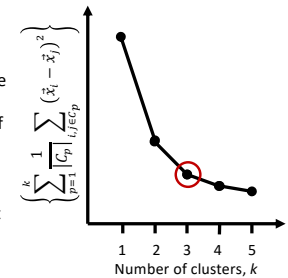


29

How to choose k ? Elbow Method

The elbow method consists of scrutinizing the plot of the minimization formula for each value of k , and then choosing the best value for k based on the “elbow” of the plot. The elbow is where the decrease in the plot begins to diminish.

It is a subjective measure. Different people may select different k 's.



30

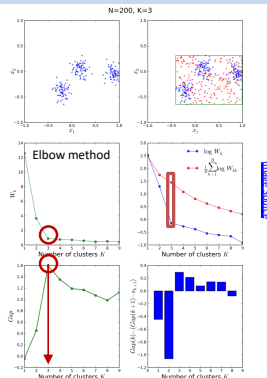
How to choose k ? Gap Method

The gap statistic improves on the elbow method with a more objective method to pick an “ideal” number of clusters.

Clustering into C_1, \dots, C_k , we denote

$$W_k = \sum_{p=1}^k \frac{1}{|C_p|} \sum_{i,j \in C_p} (\vec{x}_i - \vec{x}_j)^2$$

for the original dataset and W_{kb} for each of the reference datasets, computed by sampling uniformly B times from the original dataset's bounding box.



31

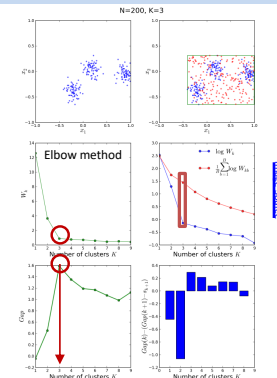
How to choose k ? Gap Method

The gap statistic is then defined as

$$\text{Gap}_n(k) = \frac{1}{B} \sum_{b=1}^B \log W_{kb} - \log W_k$$

$$= E_n^*[\log W_k] - \log W_k$$

The first k for which $\text{Gap}_n(k) - \text{Gap}_n(k+1) > 0$ is the one we select.



32

How to choose k ? Silhouette Coefficient

The silhouette coefficient, $s(i)$, for sample i is defined as

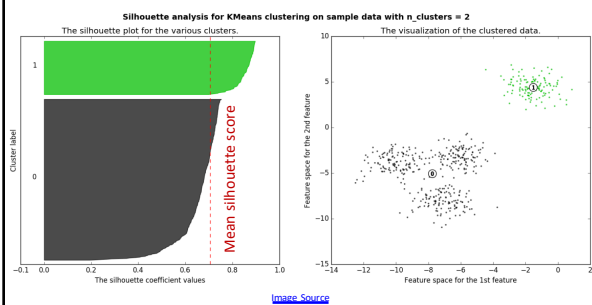
$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

Here, $a(i)$ is the average dissimilarity between sample i and all other samples in i 's own cluster. And $b(i)$ is the lowest average dissimilarity between i and any other cluster—i.e., i 's dissimilarity with the neighboring cluster. By definition, $s(i) \in [-1, 1]$, and we would like $s(i)$ to be close to 1.

The average silhouette coefficient over the dataset is a measure of how tightly grouped the data are in their clusters. A greater silhouette coefficient corresponds to higher quality clustering. $s(i) < 0$ suggest a problem.

How to choose k ? Silhouette Coefficient

Silhouette plots for $k = 2$ for a dataset

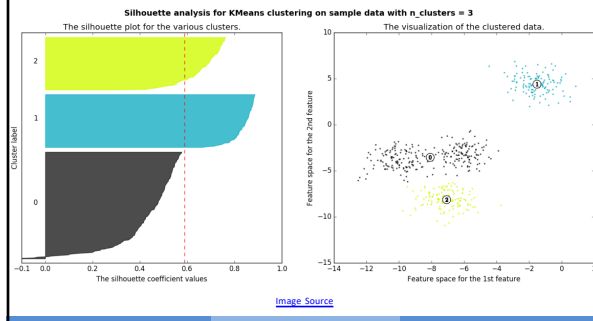


33

34

How to choose k ? Silhouette Coefficient

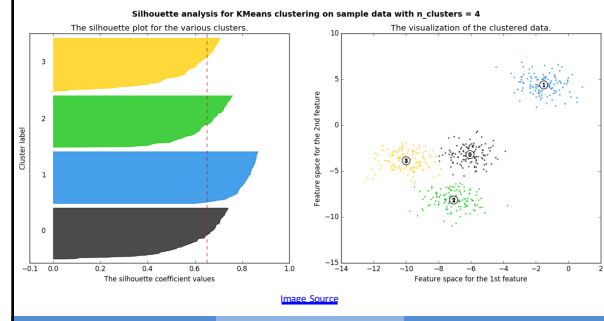
Silhouette plots for $k = 3$ for a dataset



35

How to choose k ? Silhouette Coefficient

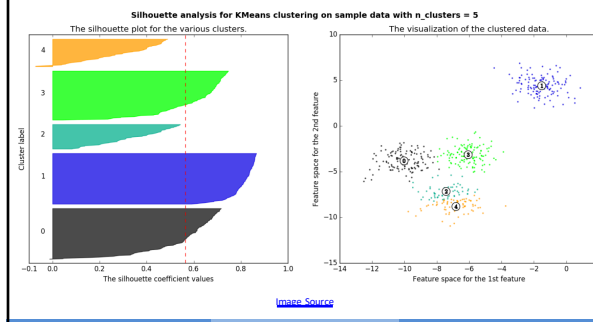
Silhouette plots for $k = 4$ for a dataset



36

How to choose k ? Silhouette Coefficient

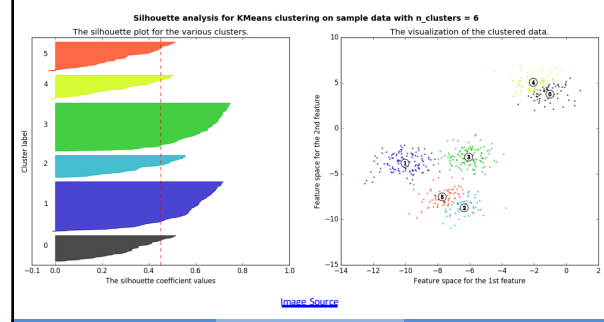
Silhouette plots for $k = 5$ for a dataset



37

How to choose k ? Silhouette Coefficient

Silhouette plots for $k = 6$ for a dataset

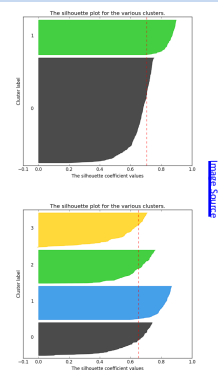


38

How to choose k ? Silhouette Coefficient

The silhouette plots suggest that $k = 3, 5, 6$ might be bad because of the presence of clusters with below average silhouette scores and due to the wide fluctuations in the size of the silhouette plots (assuming we think the sizes should be similar). It is harder to decide between Silhouette analysis is more ambivalent in deciding between $k = 2, 4$.

Naturally, k is easy to choose for 2-dimensional problems. These methods are intended for high-dimensional ones, where visualization is difficult.



39

K-MEDOIDS

42

k-Medoids Clustering

The k -means clustering uses centroids to delineate clusters. The formula for calculating the centroid C of a set of m points is

$$C = \frac{\sum_{j=1}^m \vec{x}_j}{m}$$

This formula is not robust to outliers. An alternative method is k -medoids clustering, where instead of centroids, we use elements from our data set as *median* elements, or *medoids*; typically one element per cluster.

- k -medoids is more flexible than k -means with respect to the similarity measures it can use (e.g., Euclidean distance, absolute Pearson correlation, ...), ones on which k -means might not converge.
- k -medoids is more robust to outliers than k -means.
- But k -means is $O(kn)$ while k -medoids might be $O(kn^2)$.

43

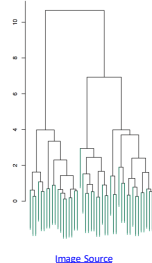
HIERARCHICAL CLUSTERING

44

Clustering Methods: Hierarchical Clustering

Hierarchical clustering also involves joining data points based on similarity or proximity. But it represents this using a tree structure, called a dendrogram. Decisions about how to classify points depend on how many branches of the dendrogram we want to include.

Hierarchical clustering is a form of unsupervised decision trees (which we will discuss later).



[Image Source](#)

45

Intro to Hierarchical Clustering

With k -means and k -medoids clustering, we must define the number of clusters k before applying the method. What if we want a method that doesn't require a pre-defined number of clusters up front?

We can use hierarchical clustering. With hierarchical clustering, elements group together in a tree diagram, and we can choose the end number of clusters *after* we create the tree. We can use hierarchical clustering in two different ways:

- Agglomerative clustering (bottom-up approach): Each data point is a cluster, and we group clusters based on similarity.
- Divisive clustering (top-down approach): The entire data set is a cluster that is split recursively until only singleton clusters of individual data points remain.

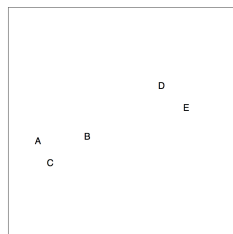
46

Hierarchical Clustering

This is agglomerative clustering with a small toy example. Suppose we have data points

$\{A, B, C, D, E\}$

organized as seen on the right. We will show how these points are combined into larger and larger clusters, and the resulting hierarchy of clusters defines our dendrogram, or tree diagram.

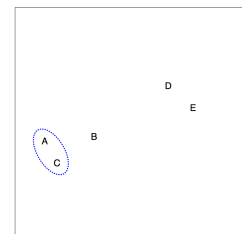


[Image Source](#)

47

Hierarchical Clustering

First, A and C are combined because they are the two closest elements.

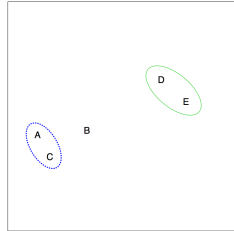


[Image Source](#)

48

Hierarchical Clustering

Then, D and E are combined into a cluster, again due to similarity.

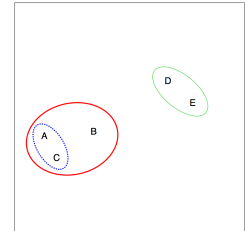


[Image Source](#)

49

Hierarchical Clustering

B is closer to the $\{A, C\}$ cluster than the $\{D, E\}$ cluster, so the red ellipse defines the next grouping.

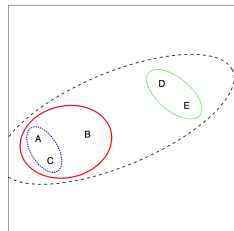


[Image Source](#)

50

Hierarchical Clustering

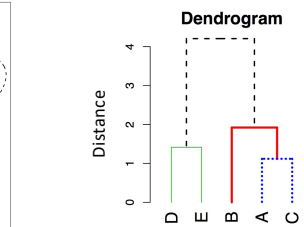
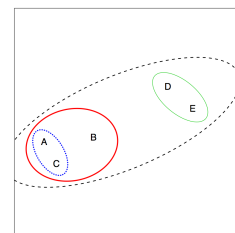
Finally, all clusters are combined to form one large group containing all points.



[Image Source](#)

51

Dendrograms

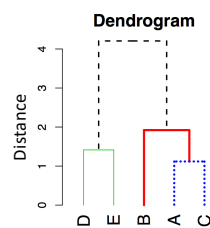


[Image Source](#)

52

Dendrograms

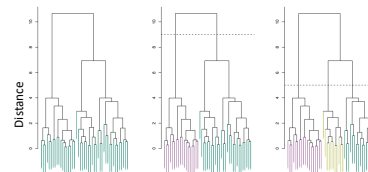
The ordering of clusters determines the dendrogram for hierarchical clustering. The dendrogram gives us an idea about how related different data points are. And the height on the dendrogram, where two groups are linked, gives us the linkage order. Links higher up on the dendrogram imply greater dissimilarity between the groups.



[Image Source](#)

53

Dendrograms



[Image Source](#)

We can determine the number of clusters resulting from hierarchical clustering by making a "cut" in the dendrogram according to the maximal distance we allow. We then use the branches at the cut to:

1. determine the number of cluster, and
2. determine the membership in the clusters.

54

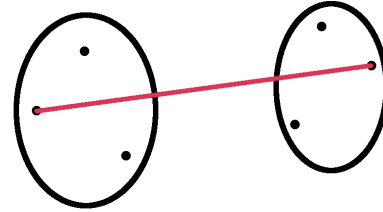
Linkage

How to decide which clusters to join together in hierarchical clustering? At least 4 dissimilarity measures exist:

- Complete (Maximal) Linkage: Dissimilarity between two clusters, A and B , is determined by the furthest distance between any element of A and any element of B .
- Single (Minimal) Linkage: Dissimilarity between A and B is determined by shortest distance between any element of A and any element of B .
- Average Linkage: Dissimilarity between A and B is determined by average distance between any element of A and any element of B .
- Centroid Linkage: Dissimilarity between A and B is determined by the distance between the centroid of A and the centroid of B . (This method may lead to inversions.)

The top 3 methods are computationally expensive, $O(|A| \cdot |B|)$ per linkage computation. The bottom method suffers from *inversion*.

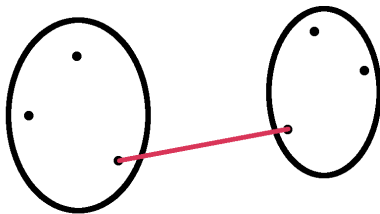
Complete (Maximal) Linkage



55

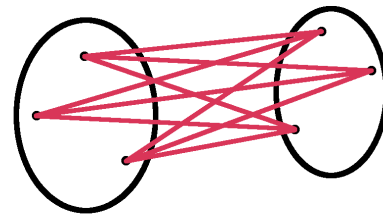
56

Single (Minimal) Linkage



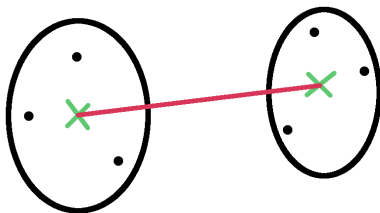
57

Average Linkage



58

Centroid Linkage



59

Group exercise



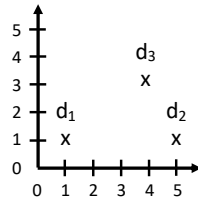
In breakout rooms, we will now work on a google doc

60

Centroid Linkage: Try it out

Plot the dendrogram for the data below:

$$\begin{aligned} d_1 &= (1, 1) \\ d_2 &= (5, 1) \\ d_3 &= (4, 3) \end{aligned}$$



[Image Source](#)

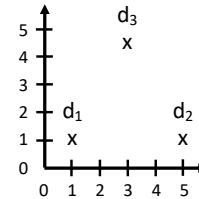
61

Centroid Linkage: Try it out

Plot the dendrogram for the data below:

$$\begin{aligned} d_1 &= (1 + \varepsilon, 1) \\ d_2 &= (5, 1) \\ d_3 &= (3, 2\sqrt{3}) \end{aligned}$$

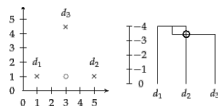
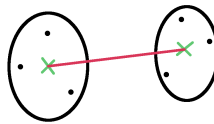
$(\varepsilon \ll 1)$



[Image Source](#)

62

Centroid Linkage: Inversion Example

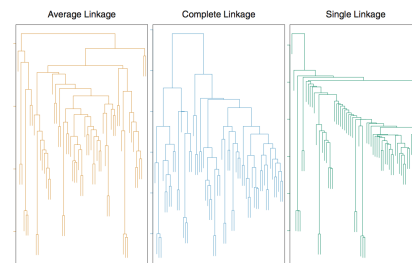


► Figure 17.9 Centroid clustering is not monotonic. The documents d_1 at $(1 + \varepsilon, 1)$, d_2 at $(5, 1)$, and d_3 at $(3, 1 + 2\sqrt{3})$ are almost equidistant, with d_1 and d_2 closer to each other than to d_3 . The non-monotonic inversion in the hierarchical clustering of the three points appears as an intersecting merge line in the dendrogram. The intersection is circled.

[Image Source](#)

63

Linkages: Visual Comparison



[Image Source](#)

64

Dissimilarity Measures

Cluster linkage in hierarchical clustering depends on the type of linkage and on the type of dissimilarity measure used:

- Euclidean distance (L_2) or other Minkowski distances (Chebyshev distance (L_∞), Manhattan distance (L_1), etc.)
- Hamming distance (mismatches between words)
- Tempo/rhythm similarity (for songs)
- Shared keywords (for webpages)

65

Pros/Cons of Hierarchical Clustering

Pros:

- Hierarchical clustering is an instance-based method. It makes no assumptions about the form of clusters
- It generates the entire hierarchy at once
- The clusters need not be centrally dense (unlike k-means or k-medoids)
- Sometimes maps nicely onto human intuition

Cons:

- Suffers from the Curse of Dimensionality. In higher dimensions, data points are spread further apart, making the formation of clusters difficult
- The hierarchical clustering algorithm does not scale well for larger data sets. It is computationally expensive, at least $O(n^2)$
- Only guaranteed to reach local minimum in whichever distance
- Dendrogram interpretation is often very subjective

66

Summary

- Unsupervised learning strives to find structure in unlabeled data
- *k*-Means clustering is an iterative procedure that, given the number of clusters, finds clusters that locally minimize the within cluster variability
 - Choosing “best” *k* is not trivial. There are several methods, but—like everything else in machine learning—they cannot be used blindly
- *k*-Medoids finds “median-sample clusters”
 - Can use additional clustering measures, more robust to outliers, but is computationally more expensive
- Hierarchical clustering
 - Generates entire hierarchy at once as dendrogram irrespective of number of clusters, but does not work well for large number of points (computationally expensive) or dimensions

69

Take 🏠 Message for rest of course: Clustering

- Unsupervised learning finds structure in unlabeled data based on a measure of distance, variance, etc.
- Clustering methods (*k*-means, *k*-medoids, hierarchical) cluster samples based on a given distance measure. They thus offer a compressed representation of the data.
 - Clustering is sometimes carried out as a preprocessing step before regression or classification
 - Clustering can also be used to better visualize large and/or high-dimensional datasets

70