Digital Adventure Ride to the Future

7 – 18 January, 2024

ENRICHMENT FOR ALL

Utrecht University

1

Unsupervised Learning – Clustering

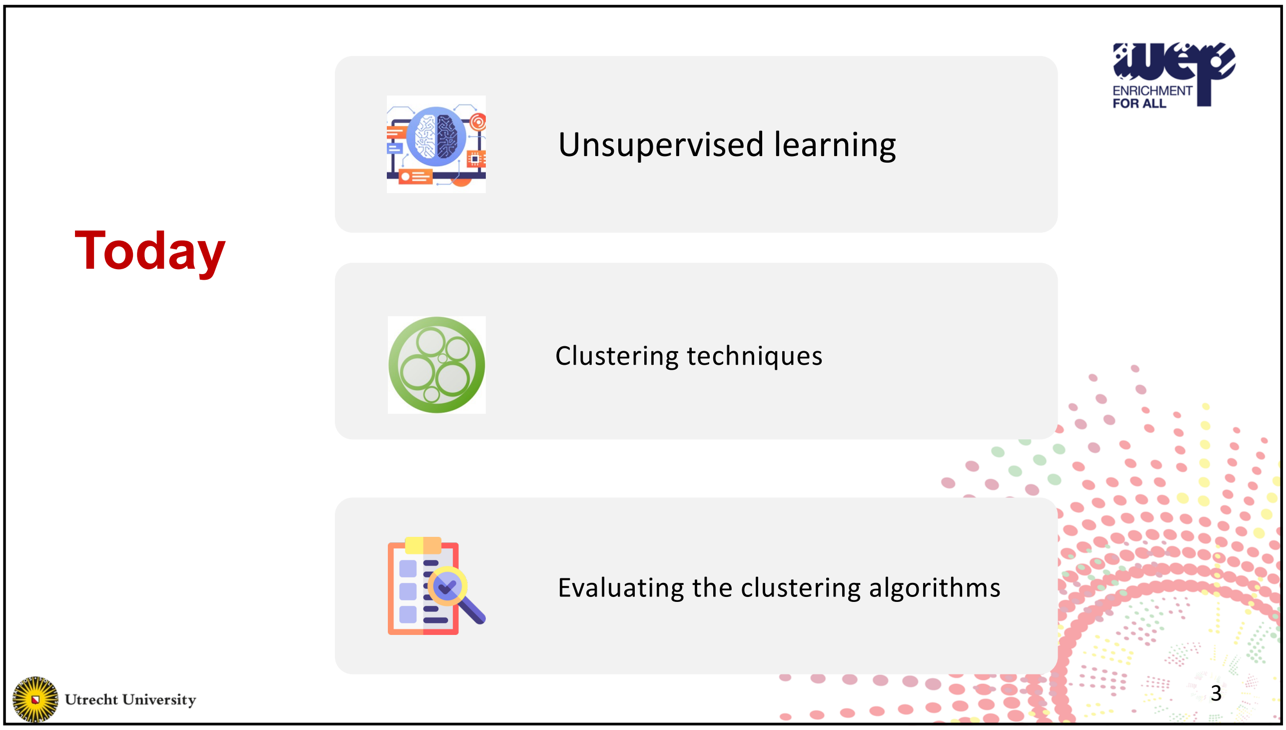
Hakim Qahtan

Department of Information and Computing Sciences
Utrecht University

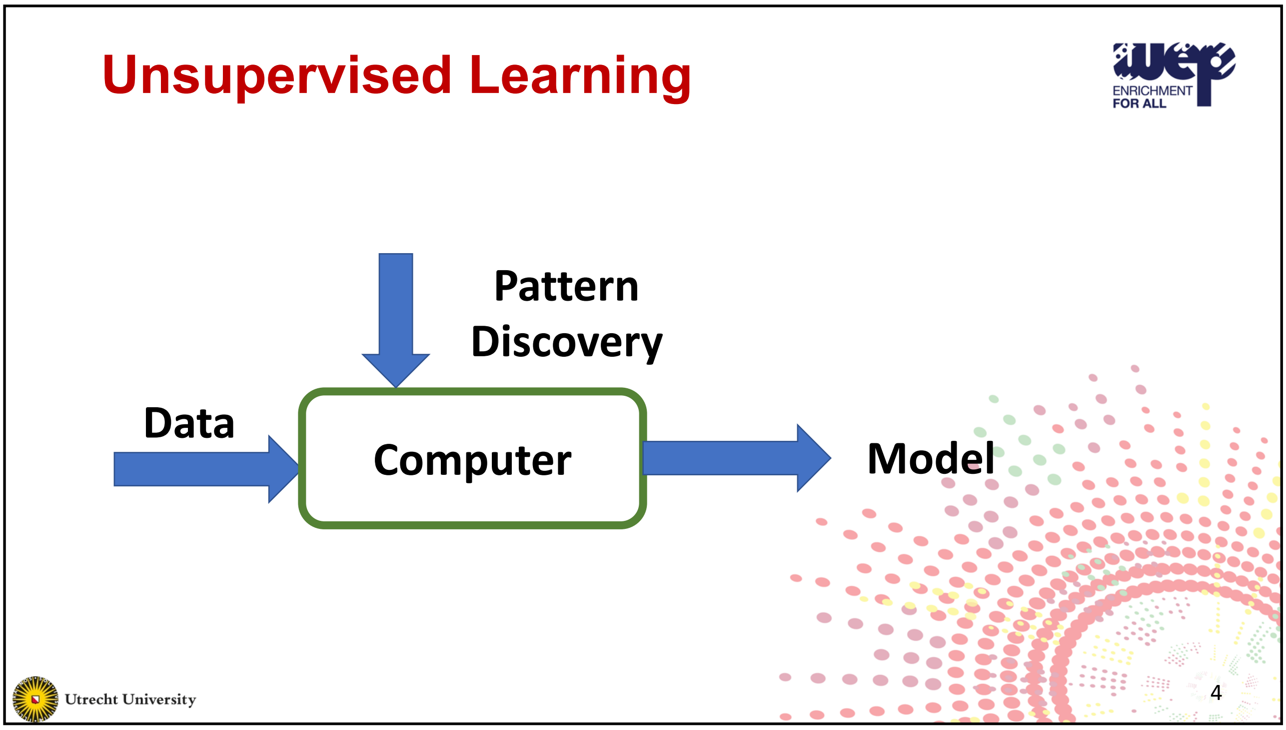
ENRICHMENT FOR ALL

Utrecht University

2


Today
Unsupervised learning
Clustering techniques
Evaluating the clustering algorithms
Utrecht University
3

3


Unsupervised Learning
Pattern
Discovery
Data
Computer
Model
Utrecht University
4

4

Unsupervised Learning Models

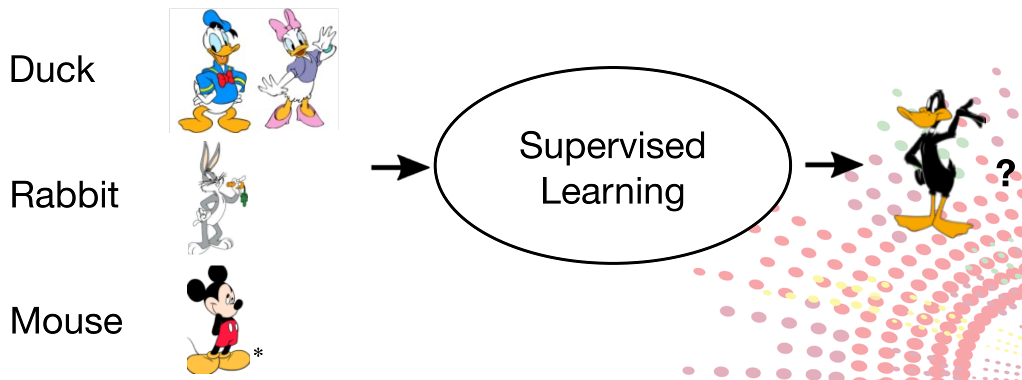
- Unsupervised learning models include:
  - Clustering (the most common unsupervised task)
  - Dimensionality reduction
  - Association rules
  - Outlier detection
  - Novelty detection
  - …
- Today, we are focusing on – Clustering

Utrecht University

5

Supervised VS Unsupervised Learning

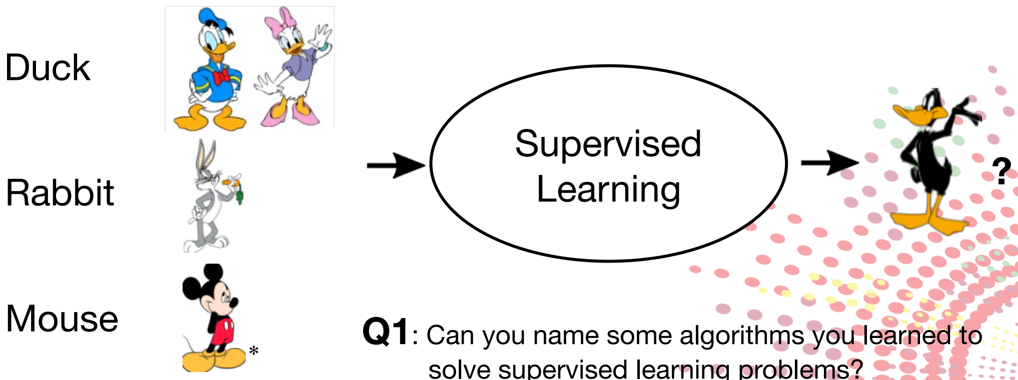
Learn **with** supervision: train a model with labelled data



Utrecht University

Supervised VS Unsupervised Learning

Learn **with** supervision: train a model with labelled data



Q1: Can you name some algorithms you learned to solve supervised learning problems?

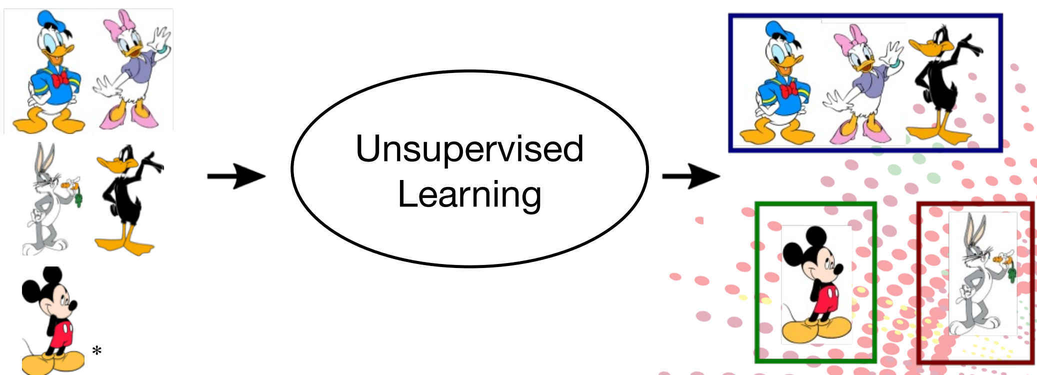
Utrecht University

7

7

Supervised VS Unsupervised Learning

Learn **without** supervision: discover hidden patterns in data



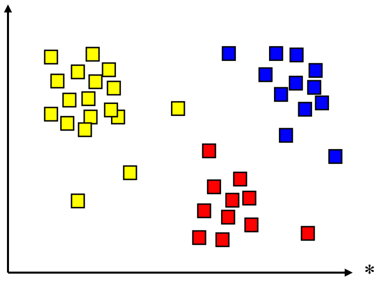
Utrecht University

8

8

Unsupervised Learning - Clustering

- Grouping similar objects together
- Organize unlabeled data into similar groups (called clusters)
  - Similarity can be defined in different ways.
  - Widely used: healthcare, text mining, computer graphics, etc.



Close in the dimensional space

Share similar features

Clustering - A historic application

- John Snow, a British physician plotted the location of cholera deaths on a map during the London cholera epidemic in 1854 [1].
- The map was used to indicate the specific clusters, certain intersections, where serious infections were reported.

[1] Brody, Howard, et al. "Map-making and myth-making in Broad Street: the London cholera epidemic, 1854." The Lancet (2000).

Figure 1: **Snow's map of cholera deaths in the Broad Street area**

Clustering – Applications

Image compression

- k-means clustering applied to image processing
  Goal: image compression -- less storage!
  - Cluster blocks of 4 pixels, then replace blocks by their cluster centroid

Original (1MB) K = 200 (0.2375 MB) K = 4 (0.0625 MB)

Utrecht University

11

11

Cluster Analysis

- A Cluster: A collection of data objects where
  - similar (or related) objects fall within the same group
  - dissimilar (or unrelated) objects fall in different groups
- Cluster analysis
  - Finding similarities between data according to the characteristics found in the data and grouping similar data objects into clusters
  - Increase intraclass similarity
  - Decrease interclass similarity
- Unsupervised learning: no predefined classes (data labels are not required)

Utrecht University

12

12

Measuring the Similarity between Objects

- A cluster is a collection of similar objects
- Dissimilarity/Similarity metric
  - Similarity between objects in a cluster is expressed in terms of a distance function $d(a, b)$
    - Simple way to model similarity is Gaussian kernel $s(a, b) = e^{-\lambda d(a,b)}$
    - Takes values in the interval [0,1] (1 means $a = b$).
    - Most common distance metric is Minkowski distance $d(a, b) = \left(\sum_{i=1}^{d} |a_i - b_i|^p\right)^{1/p}$
      - $a, b \in R^d$, d is the data dimensionality
      - $p = 1$, Manhattan distance or Taxi-cap distance
      - $p = 2$, Euclidean distance
      - $p = \infty$, the distance is the max difference between the a_i, b_i, $i = 1, \ldots, d$ ($d_\infty([1,2,3,4], [5,2,3,6]) = |1 - 5| = 4$)

Utrecht University

13

13

Clustering Techniques

Utrecht University

14

Major Clustering Approaches

- Partitioning approach:
  - Construct various partitions and then evaluate them by some criterion, e.g., minimizing the sum of square errors
  - Typical methods: k-means, k-medoids, CLARANS
- Hierarchical approach:
  - Create a hierarchical decomposition of the set of data (or objects) using some criterion
  - Typical methods: Divisive (Diana), Agglomerative (Agnes), BIRCH, CAMELEON
- Density-based approach:
  - Based on connectivity and density functions
  - Typical methods: DBSACN, OPTICS, DenClue

Utrecht University

15

15

K-means Clustering

Utrecht University

16

K-means Clustering

- Minimizes the distance between the objects in each cluster to the mean of that cluster (also called the centroid)
- Solves the following optimization problem

$$\arg\min_{c} \sum_{j=1}^{k} \sum_{x \in c_j} d(x, \mu_i) = \arg\min_{c} \sum_{j=1}^{k} \sum_{x \in c_j} ||x - \mu_i||_2^2$$

- k is the number of clusters and should be provided by the user
- c_i represent the set of objects in cluster (i) and μ_i is the mean (centroid) of cluster (i)
- $||x - \mu_i||_2^2$ is the square of the Euclidean distance

Utrecht University 17

17

K-means Clustering

- Input: the dataset and the number of clusters k
- Each cluster is determined by its average point (centroid)
- Each point in the dataset is assigned to the cluster with the closest centroid
- Algorithm:

1. Select k points as initial centroids
2. **repeat**
3. Form k clusters by assigning all clusters to the closest centroid
4. Recompute the centroid of each cluster
5. **until** the centroids do not change

Utrecht University 18

18

K-means Clustering – Pros and Cons

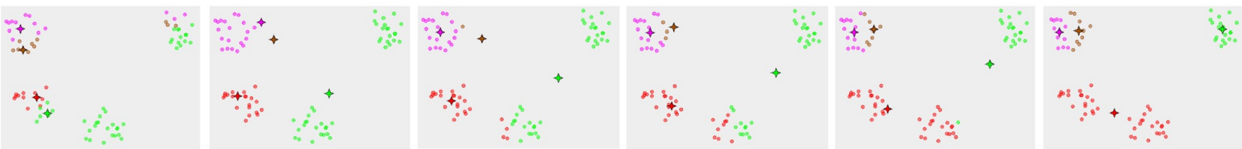
- Pros:
  - Simple, scale well for large number of samples
  - Computationally efficient
  - Tighter clusters than the hierarchical clustering
- Cons
  - Difficult to guess the value of *k*
  - Cannot handle clusters with arbitrary shapes
  - Clustering accuracy depends heavily on the initial selection of the centroids
  - Based on the use of squared Euclidean distance as the measure of dissimilarity
  - Cluster means are not robust to outliers
  - May not ensure the convergence to the global minimum

Utrecht University

19

19

K-means Clustering – Demo

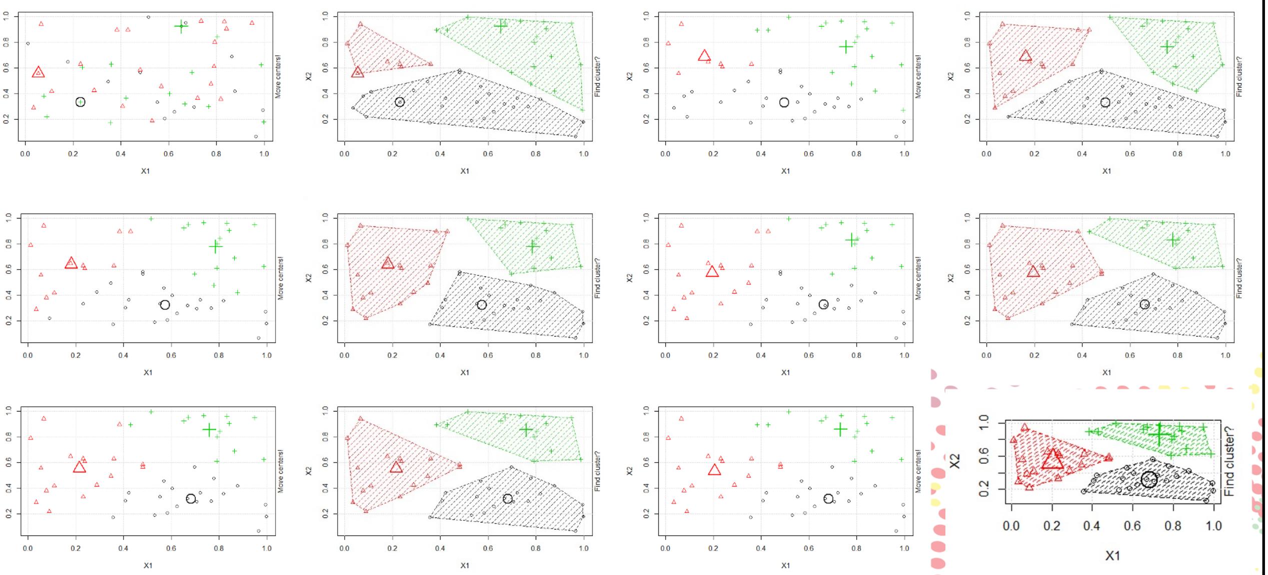


Utrecht University

20

20

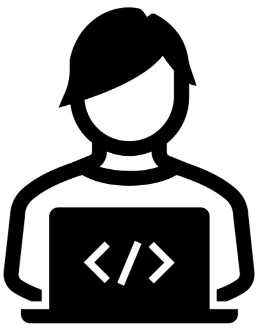
K-means Clustering – Demo



Utrecht University

21

21



Coding Time

Utrecht University

22

ENRICHMENT FOR ALL

Coffee Break

Utrecht University

23

ENRICHMENT FOR ALL

DBSCAN Clustering

Utrecht University

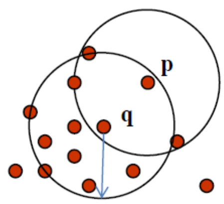
24

Density Based Clustering Methods

- Basic idea: points within density connected region belong to the same cluster
- Major Features:
  - Discover clusters of arbitrary shapes
  - Handle noise
  - Single scan over the data
- Requirements:
  - Parameters as terminating conditions
- You can look at DBSCAN: Ester, et al. (SIGKDD'96)

DBSCAN

- Two parameters:
  - *Eps* ϵ: maximum radius of the neighborhood
  - MinPts: minimum number of points in the Eps-neighborhood of each point to be considered in the cluster
- Eps-neighborhood of q:
  - $N_{Eps}(q) = \{x \in D \mid dist(q, x) \le \epsilon\}$
- Directly density-reachable: a point p is *directly density-reachable* from a point q w.r.t. Eps and MinPts if:
  - q is core point i.e. $card(N_{Eps}(q)) \ge MinPts$
  - p belongs to $N_{Eps}(q)$

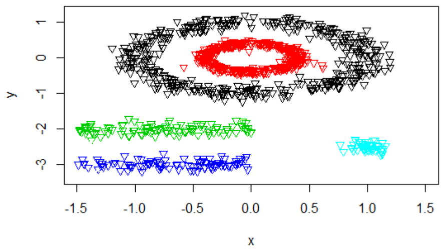


Eps = 1cm
MinPts = 5

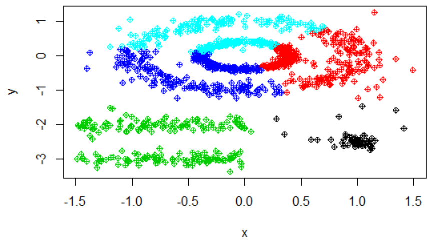
DBSCAN

- DBSCAN is based on:
  - Density: number of points within a specified radius (Eps)
  - Core points: those points which has more than the number of MinPts within Eps-neighborhood
    - These are the points in the interior of the clusters
  - Border points: has fewer number of points within Eps-neighborhood but they are in the neighborhood of core points
  - Noise points: all the points that are neither core points nor border points.

Comparing K-means and DBSCAN

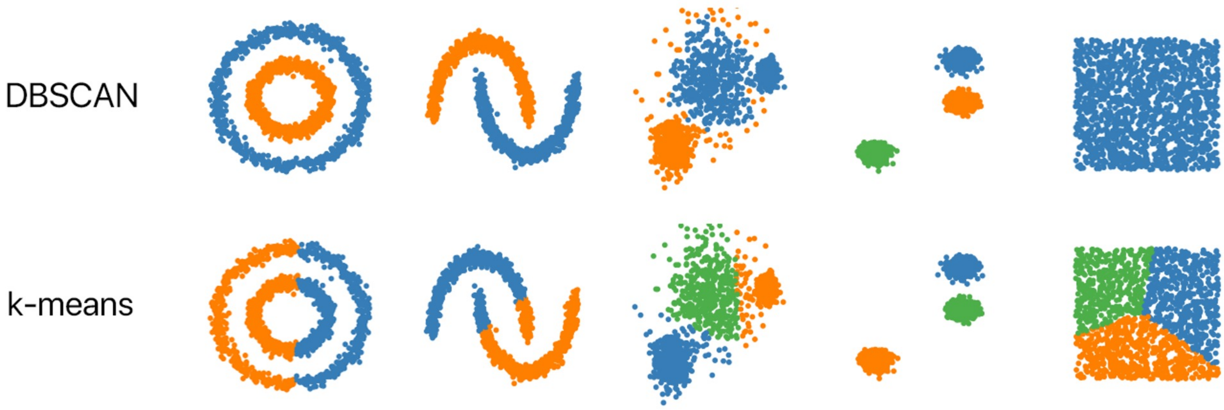


DBSCAN

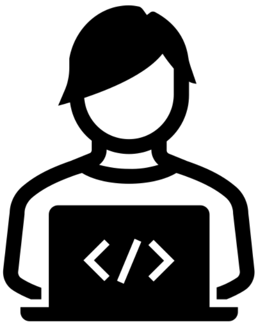


K-means

Comparing K-means and DBSCAN



29



Coding Time

30

Hierarchical Clustering

Utrecht University

31

Hierarchical Clustering

- Construct nested clusters by successive merging (agglomerative/bottom up) or splitting (divisive/top down)
- The hierarchy of the clusters is represented as a tree which is called dendrogram.
- The root represents a single cluster for all the data
- The leaves represent clusters with single object

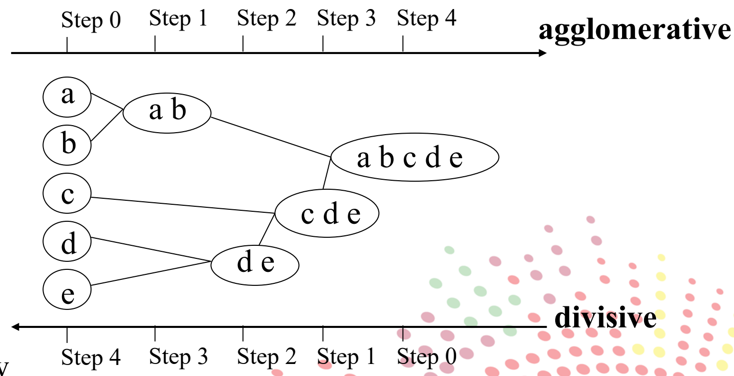
Utrecht University

32

32

Hierarchical Clustering

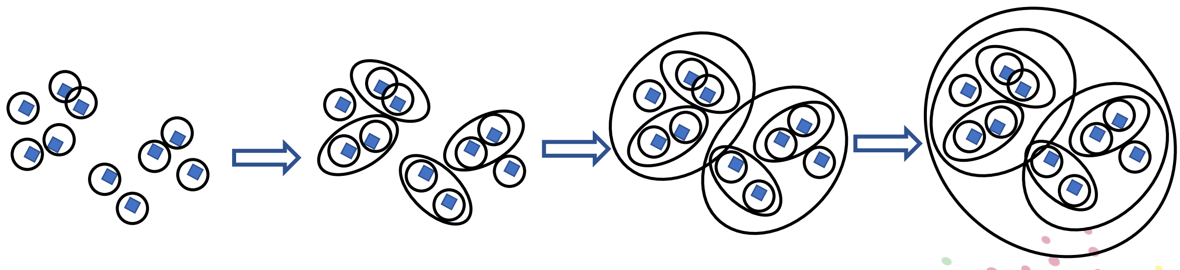
- Two approaches
  - Agglomerative (bottom-up)
  - Divisive (top-down)
- Requires stopping condition
  - Number of clusters
  - Agglomerative: similarity between merged clusters is low
  - Divisive: maximum distance between all possible partition divisions is small



Utrecht University 33

33

Hierarchical Clustering (Agglomerative)

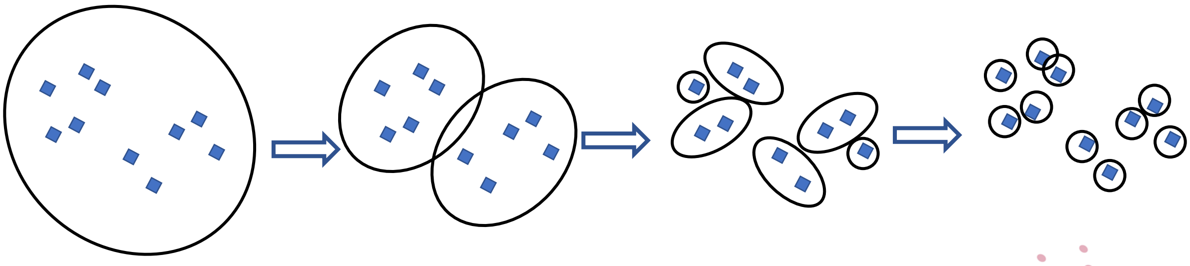


- Also known as bottom-up approach
- Start with each object as a cluster in its own
- Repeatedly merge nodes that have the least dissimilarity
- Go on in a non-descending fashion
- At every step, check the stopping condition

Utrecht University 34

34

Hierarchical Clustering (Divisive)

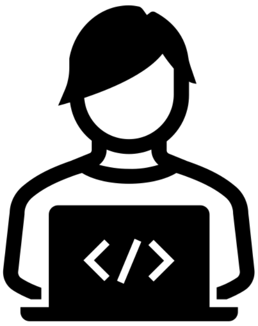


- Also known as top-down approach
- Start with assuming all the data belong to the same cluster
- Check all possible ways to divide the clusters
- Choose the best division (that will reduce intraclass distance and increase interclass distance)
- Repeat the process until stopping condition is met

Utrecht University

35

35



Coding Time

Utrecht University

36

Coffee Break

Utrecht University

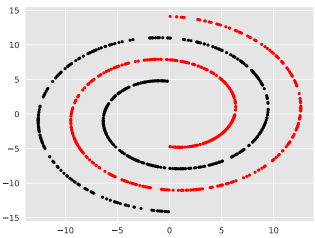
37

Spectral Clustering

Utrecht University

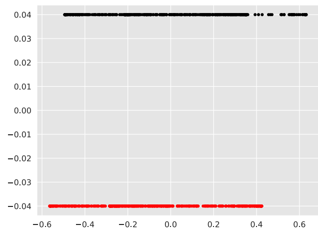
38

Spectral Clustering



- Dataset exhibits complex cluster shapes
- K-means performs very poorly in this space due to bias toward dense spherical clusters.

In the embedded space given by two leading eigenvectors, clusters are trivial to separate.



Spectral Clustering

- Many datasets can be transformed into a graph representation (similarity graph).
- Given set of data points
  - Compute the similarity matrix $S = [S_{ij}],\ i,j = 1,\dots,n,\ S_{ij} = s(x_i, x_j)$
    - s is a similarity measure.
- Construct graph:
  - Data points are vertices
  - Connect close points
  - Intuition: graph captures local neighborhood
- Clustering is partitioning the graph into connected components

Measuring the Clustering Quality

Utrecht University

41

Measuring the Clustering Quality

- Two methods: intrinsic vs. extrinsic
- Intrinsic: unsupervised, i.e., the ground truth is unavailable
  - Evaluate the goodness of a clustering by considering how well the clusters are separated, and how compact the clusters are
  - Ex. Rand Index and Silhouette coefficient
- Extrinsic: supervised, i.e., the ground truth is available (for evaluation only)
  - Compare a clustering against the ground truth using certain clustering quality measure
  - Ex. Precision and recall metrics

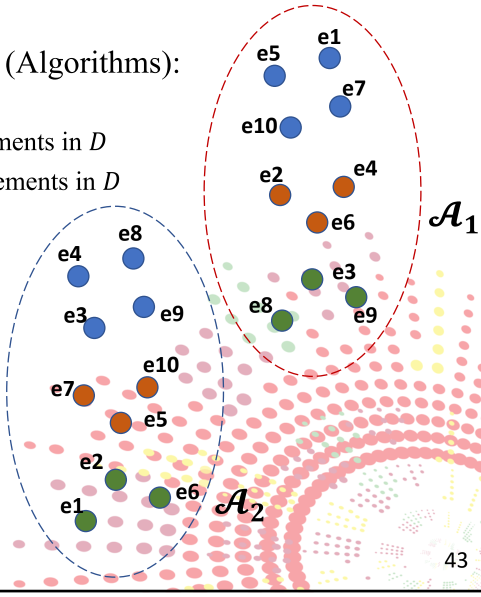
Utrecht University

42

42

Measuring the Clustering Quality – Rand Index

- Is used for comparing two clustering approaches (Algorithms):
- $D = \{e_1, e_2, \dots, e_n\}$
- $\boldsymbol{\mathcal{A}_1} = \{A_{11}, A_{12}, \dots, A_{1r}\}$ r-clusters created from the elements in D
- $\boldsymbol{\mathcal{A}_2} = \{A_{21}, A_{22}, \dots, A_{2s}\}$ s-clusters created from the elements in D
- Rand index R is defined as $R = \frac{a+b}{a+b+c+d} \in [0,1]$
  - a = set of pairs that are in the same subset of $\boldsymbol{\mathcal{A}_1}, \boldsymbol{\mathcal{A}_2}$
  - b = set of elements in same subset in $\boldsymbol{\mathcal{A}_1}$ ONLY
  - c = set of elements in same subset in $\boldsymbol{\mathcal{A}_2}$ ONLY
  - d = set of pairs in different subsets
- If the dataset is labeled and $\boldsymbol{\mathcal{A}_2}$ is the actual labels
  - R is analogous to the accuracy in classification



Utrecht University

43

43

Measuring the Clustering Quality – Silhouette Score

- Intrinsic approach:
- Compares the final shape of the clusters
- Let $v_i \in C_r,\ \ r = 1, \dots, k$ (k clusters)
  - $a(v_i) = \frac{1}{|C_r|-1} \sum_{v_j \in C_r, v_j \neq v_i} d(v_i, v_j)$ (the average distance between v_i and all other objects in the same cluster)
  - $b(v_i) = \min_{C_t \neq C_r} \frac{1}{|C_t|} \sum_{v_t \in C_t} d(v_i, v_t)$ (the average distance between v_i and all other objects in the nearest cluster to its cluster)
  - $sil(v_i) = \frac{b(v_i) - a(v_i)}{\max\{a(v_i), b(v_i)\}} \in [-1,1]$

Utrecht University

44

44

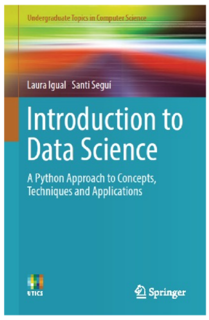
Reading Material for Interested Students

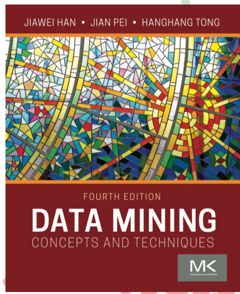
- Introduction to Data Science, Ch 7.

  Unsupervised Learning

- Data Mining: Concepts-and-Techniques

  Ch 8. Cluster Analysis





Acknowledgement: parts of the material were prepared by Mel Chekol and Shihan Wang

Figures with * are from the Internet

Utrecht University

45

45

Thank You

Utrecht University

46

47