Week 5: Clustering

Data Analysis

Leo Hofste & Martin Wesselink

01-03-2020

SAXION UNIVERSITY OF APPLIED SCIENCES

Agenda

1. **What is Cluster Analysis?**
   - Clustering application examples

2. **Cluster Analysis: Basic Concepts**

3. **Cluster methods in general**
   - Partitioning Methods
   - Hierarchical Methods
   - Other Methods

4. **Evaluation of Clustering**

5. **Summary**

6. **Exercises with Orange**

WEEK 5: CLUSTERING

SAXION UNIVERSITY OF APPLIED SCIENCES

CRISP-DM

Modeling

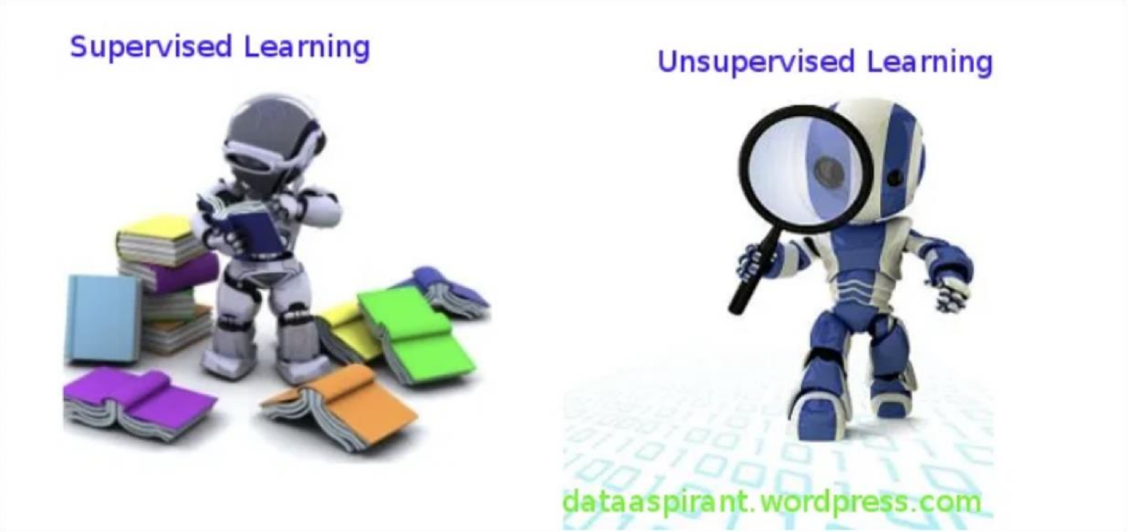
1. **Recap of week 3**
2. **What is Cluster Analysis?**
   Clustering application examples
3. **Cluster Analysis: Basic Concepts**
4. **Cluster methods in general**
   Partitioning Methods
   Hierarchical Methods
   Other Methods
5. **Evaluation of Clustering**
6. **Summary**
7. **Exercises with Orange**

Unsupervised learning with Clustering

Machine learning paradigms

Supervised learning
- Classification
- Regression (sometimes called ‘prediction’)
- Sequence prediction (including time series forecasting)

Unsupervised learning
- Clustering (or ‘cluster analysis’)
- Association



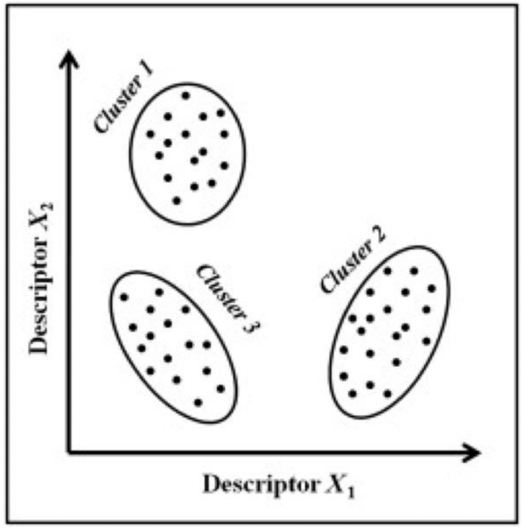
CRISP-DM

Modeling

1. **Recap of week 3**
2. **What is Cluster Analysis?**
   - Clustering application examples
3. **Cluster Analysis: Basic Concepts**
4. **Cluster methods in general**
   - Partitioning Methods
   - Hierarchical Methods
   - Other Methods
5. **Evaluation of Clustering**
6. **Summary**
7. **Exercises with Orange**

Unsupervised learning with Clustering

Unsupervised learning *is a branch of machine learning that learns from test data that has not been labeled, classified or categorized*



(https://towardsdatascience.com/unsupervised-machine-learning-9329c97d6d9f)

CRISP-DM

Modeling

1. **Recap of week 3**
2. **What is Cluster Analysis?**
   - Clustering application examples
3. **Cluster Analysis: Basic Concepts**
4. **Cluster methods in general**
   - Partitioning Methods
   - Hierarchical Methods
   - Other Methods
5. **Evaluation of Clustering**
6. **Summary**
7. **Exercises with Orange**

SAXION UNIVERSITY OF APPLIED SCIENCES

Unsupervised learning with Clustering

What is Cluster Analysis?

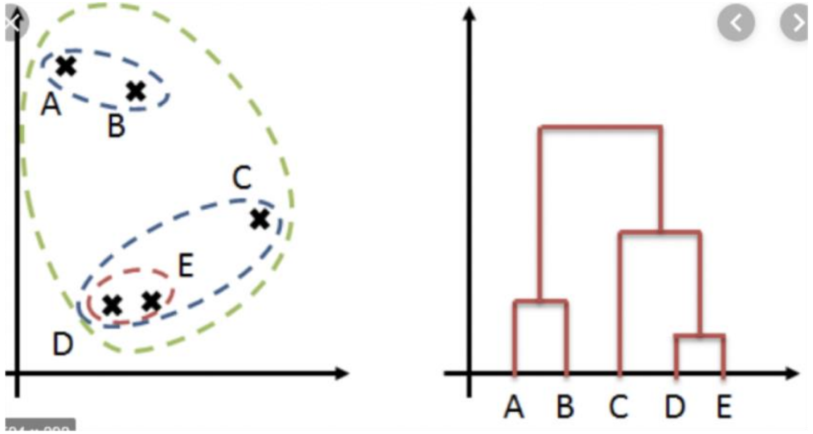
Cluster: A collection of data objects

- similar (or related) to one another within the same group
- dissimilar (or unrelated) to the objects in other groups

Cluster analysis (or Clustering)

- Finding similarities between data according to the characteristics found in the data and grouping similar data objects into clusters

Unsupervised learning: no predefined classes



CRISP-DM

Modeling

1. **Recap of week 3**
2. **What is Cluster Analysis?**
   Clustering application examples
3. **Cluster Analysis: Basic Concepts**
4. **Cluster methods in general**
   Partitioning Methods
   Hierarchical Methods
   Other Methods
5. **Evaluation of Clustering**
6. **Summary**
7. **Exercises with Orange**

Unsupervised learning with Clustering

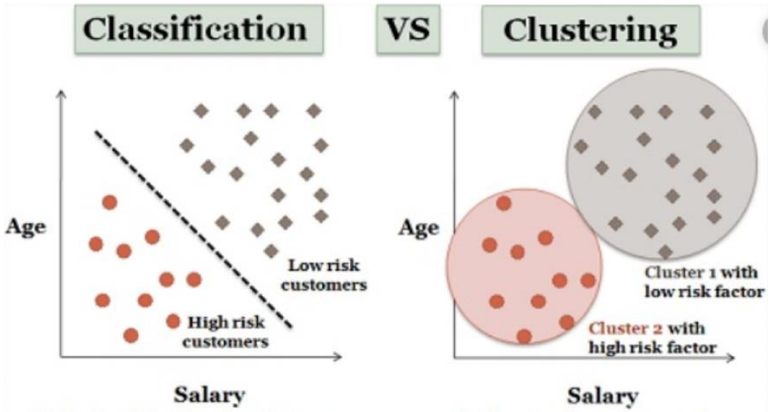
Clustering ≠ Classification

Clustering

- Does our population (f.i. customers.) naturally fall into different groups?
- Find these groups and the characteristics of these groups (f.i. using exploratory data analysis)
- There is no prediction of the group to which new observations belong(**!**)

Classification

- Groups (f.i. customers) that differ with respect to a particular target characteristic (f.i. risk to churn)
- Find a model to predict (and/or explain) the target characteristic



Risk classification for the loan payees on the basis of customer salary

Unsupervised learning with Clustering

CRISP-DM

Modeling

1. Recap of week 3
2. What is Cluster Analysis?
   - Clustering application examples
3. Cluster Analysis: Basic Concepts
4. Cluster methods in general
   - Partitioning Methods
   - Hierarchical Methods
   - Other Methods
5. Evaluation of Clustering
6. Summary
7. Exercises with Orange

What is Cluster Analysis?

Clustering Application Examples

- **Biology**: taxonomy of living things: kingdom, phylum, class, order, family, genus and species

- **Information retrieval**: document clustering
- Land use: Identification of areas of similar land use in an earth observation database

- **Marketing**: Help marketers discover distinct groups in their customer bases, and then use this knowledge to develop targeted marketing programs

- **City-planning**: Identifying groups of houses according to their house type, value, and geographical location

- **Climate**: understanding earth climate, find patterns of atmospheric and ocean

CRISP-DM

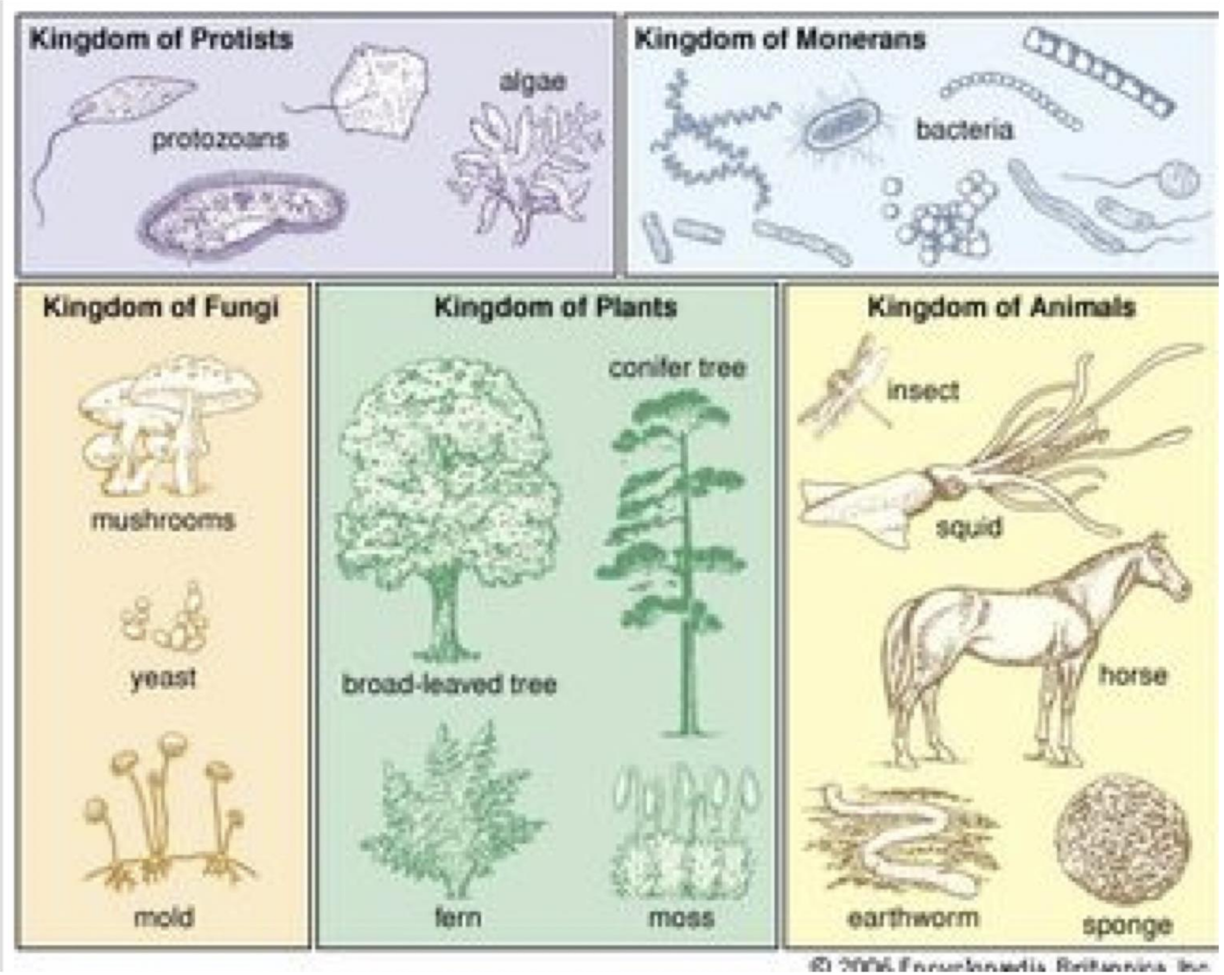
Modeling

1. **Recap of week 3**
2. **What is Cluster Analysis?**
   - Clustering application examples
3. **Cluster Analysis: Basic Concepts**
4. **Cluster methods in general**
   - Partitioning Methods
   - Hierarchical Methods
   - Other Methods
5. **Evaluation of Clustering**
6. **Summary**
7. **Exercises with Orange**

Unsupervised learning with Clustering

What is Cluster Analysis?

Clustering Application Example: Biology



© 2006 Encyclopædia Britannica, Inc.

CRISP-DM

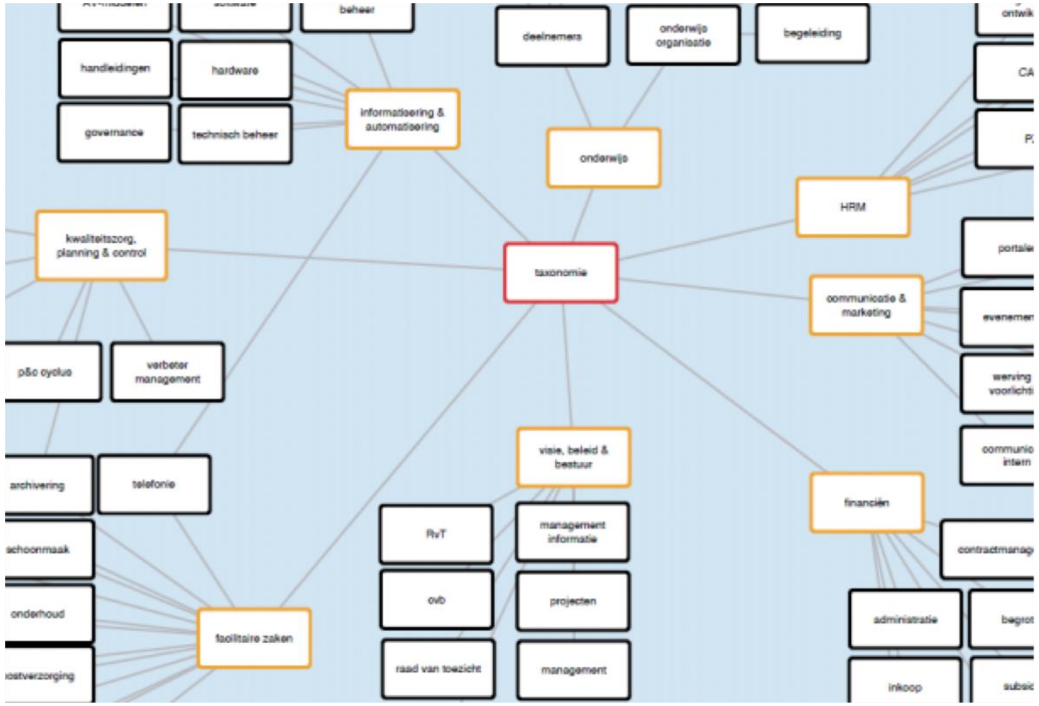
Modeling

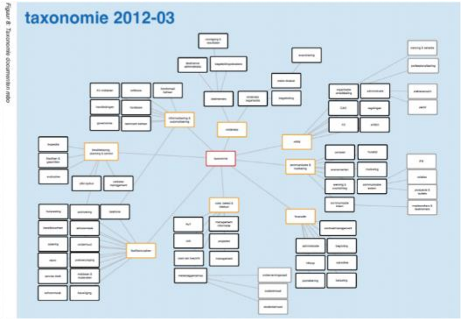
1. Recap of week 3
2. What is Cluster Analysis?
   - Clustering application examples
3. Cluster Analysis: Basic Concepts
4. Cluster methods in general
   - Partitioning Methods
   - Hierarchical Methods
   - Other Methods
5. Evaluation of Clustering
6. Summary
7. Exercises with Orange

Unsupervised learning with Clustering

What is Cluster Analysis?

Clustering Application Example: Information retrieval





CRISP-DM

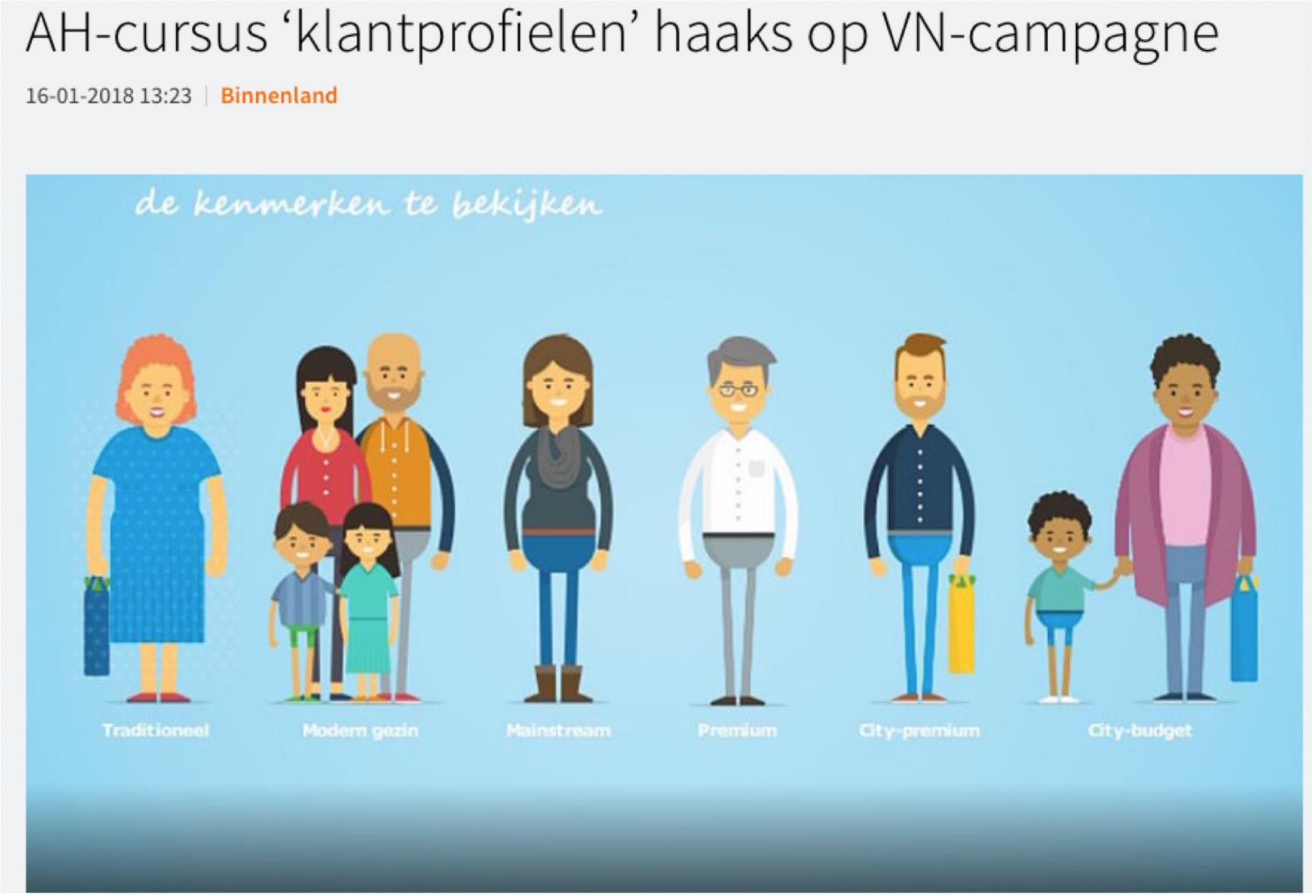
Modeling

1. **Recap of week 3**
2. **What is Cluster Analysis?**
   - Clustering application examples
3. **Cluster Analysis: Basic Concepts**
4. **Cluster methods in general**
   - Partitioning Methods
   - Hierarchical Methods
   - Other Methods
5. **Evaluation of Clustering**
6. **Summary**
7. **Exercises with Orange**

Unsupervised learning with Clustering

What is Cluster Analysis?

Clustering Application Example: Marketing – Customer Profiles



AH-cursus ‘klantprofielen’ haaks op VN-campagne

16-01-2018 13:23 | Binnenland

CRISP-DM

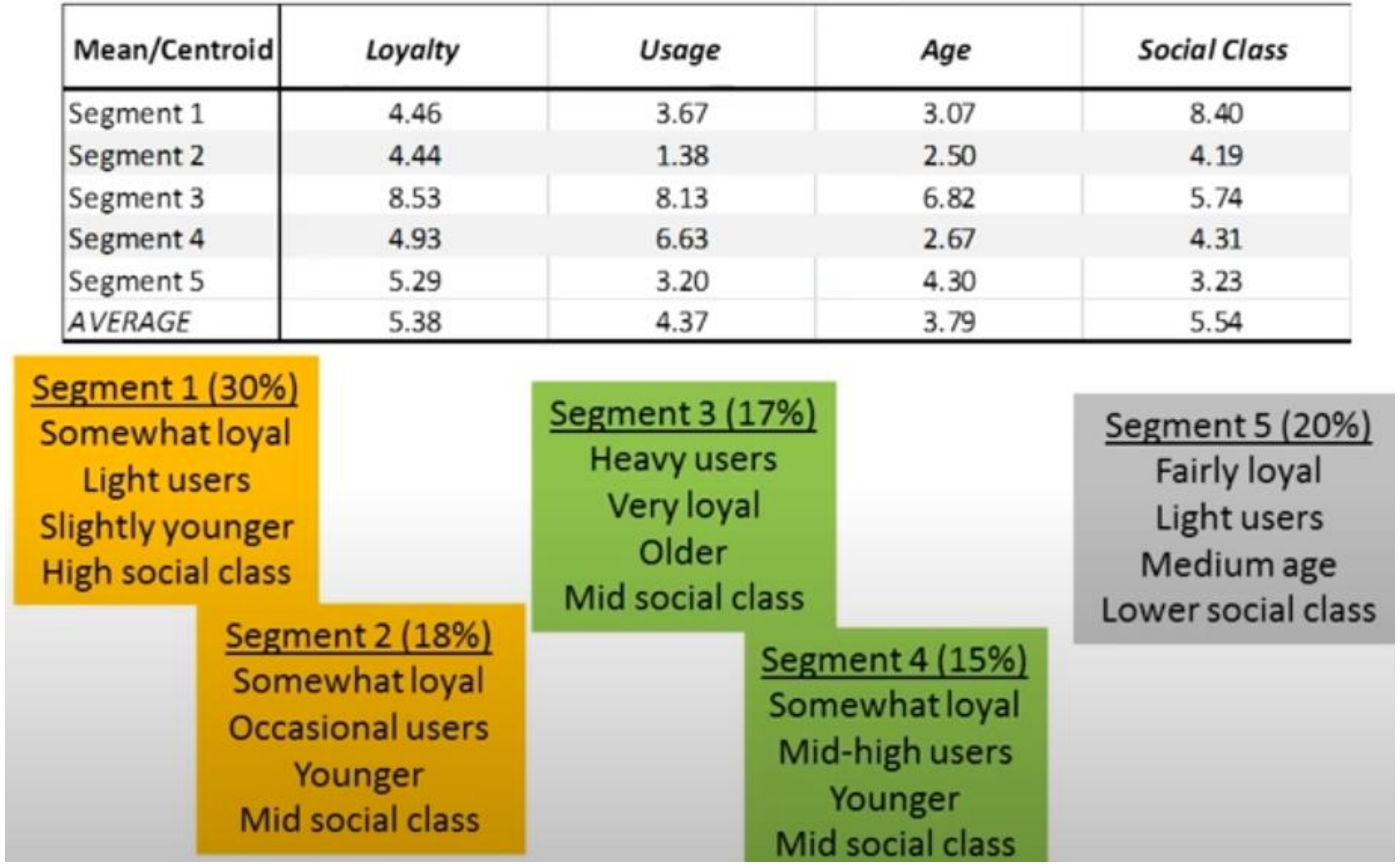
Modeling

1. **Recap of week 3**
2. **What is Cluster Analysis?**
   - Clustering application examples
3. **Cluster Analysis: Basic Concepts**
4. **Cluster methods in general**
   - Partitioning Methods
   - Hierarchical Methods
   - Other Methods
5. **Evaluation of Clustering**
6. **Summary**
7. **Exercises with Orange**

Unsupervised learning with Clustering

What is Cluster Analysis?

Clustering Application Example: Marketing – Customer Profiles



Mean/Centroid	*Loyalty*	*Usage*	*Age*	*Social Class*
Segment 1	4.46	3.67	3.07	8.40
Segment 2	4.44	1.38	2.50	4.19
Segment 3	8.53	8.13	6.82	5.74
Segment 4	4.93	6.63	2.67	4.31
Segment 5	5.29	3.20	4.30	3.23
AVERAGE	5.38	4.37	3.79	5.54

Segment 1 (30%)
Somewhat loyal
Light users
Slightly younger
High social class

Segment 2 (18%)
Somewhat loyal
Occasional users
Younger
Mid social class

Segment 3 (17%)
Heavy users
Very loyal
Older
Mid social class

Segment 4 (15%)
Somewhat loyal
Mid-high users
Younger
Mid social class

Segment 5 (20%)
Fairly loyal
Light users
Medium age
Lower social class

CRISP-DM

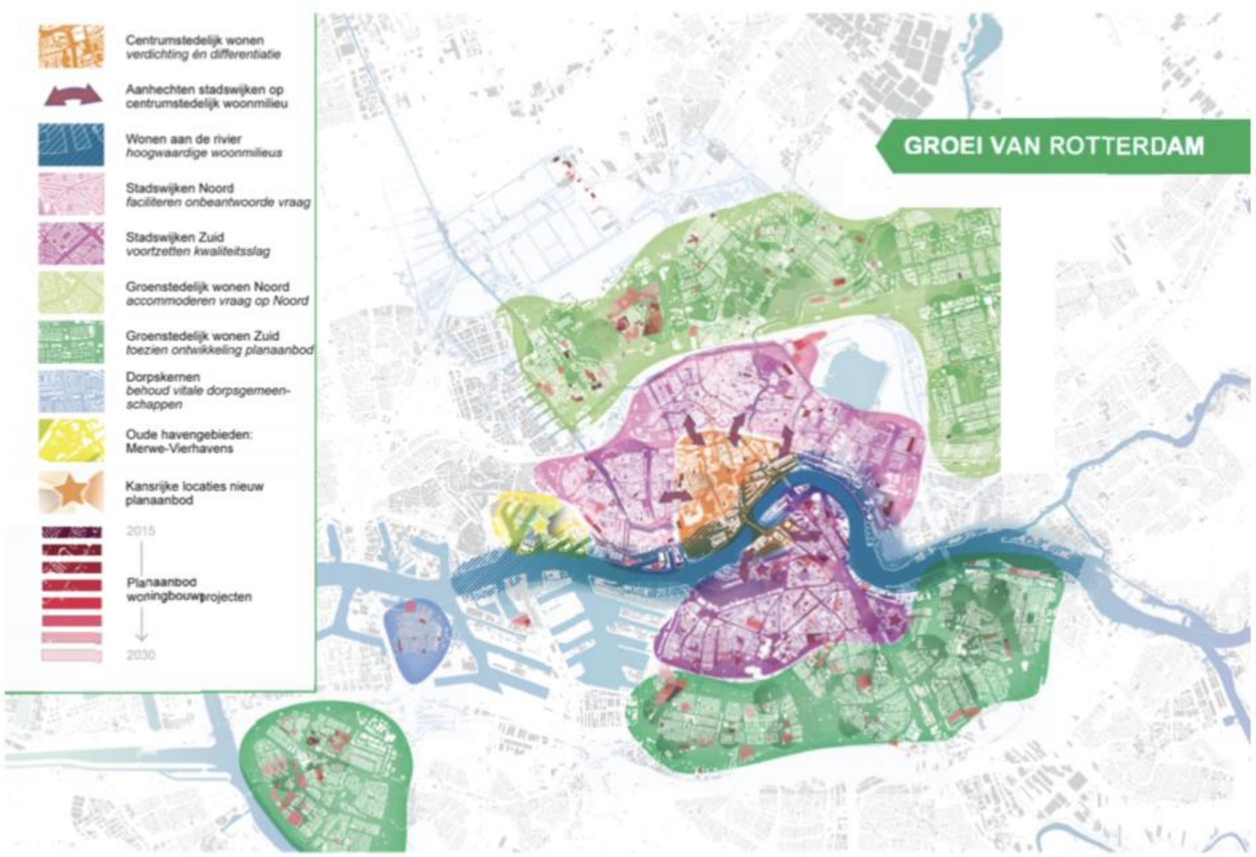
Modeling

1. **Recap of week 3**
2. **What is Cluster Analysis?**
   Clustering application examples
3. **Cluster Analysis: Basic Concepts**
4. **Cluster methods in general**
   Partitioning Methods
   Hierarchical Methods
   Other Methods
5. **Evaluation of Clustering**
6. **Summary**
7. **Exercises with Orange**

Unsupervised learning with Clustering

What is Cluster Analysis?

Clustering Application Example: City planning - Rotterdam



GROEI VAN ROTTERDAM
Centrumstedelijk wonen
verdichting én differentiatie
Aanhechten stadswijken op centrumstedelijk woonmilieu
Wonen aan de rivier
hoogwaardige woonmilieus
Stadswijken Noord
faciliteren onbeantwoorde vraag
Stadswijken Zuid
voortzetten kwaliteitsslag
Groenstedelijk wonen Noord
accommoderen vraag op Noord
Groenstedelijk wonen Zuid
toezien ontwikkeling planaanbod
Dorpskernen
behoud vitale dorpsgemeenschappen
Oude havengebieden: Merwe-Vierhavens
Kansrijke locaties nieuw planaanbod
2015
Planaanbod woningbouwprojecten
2030

CRISP-DM

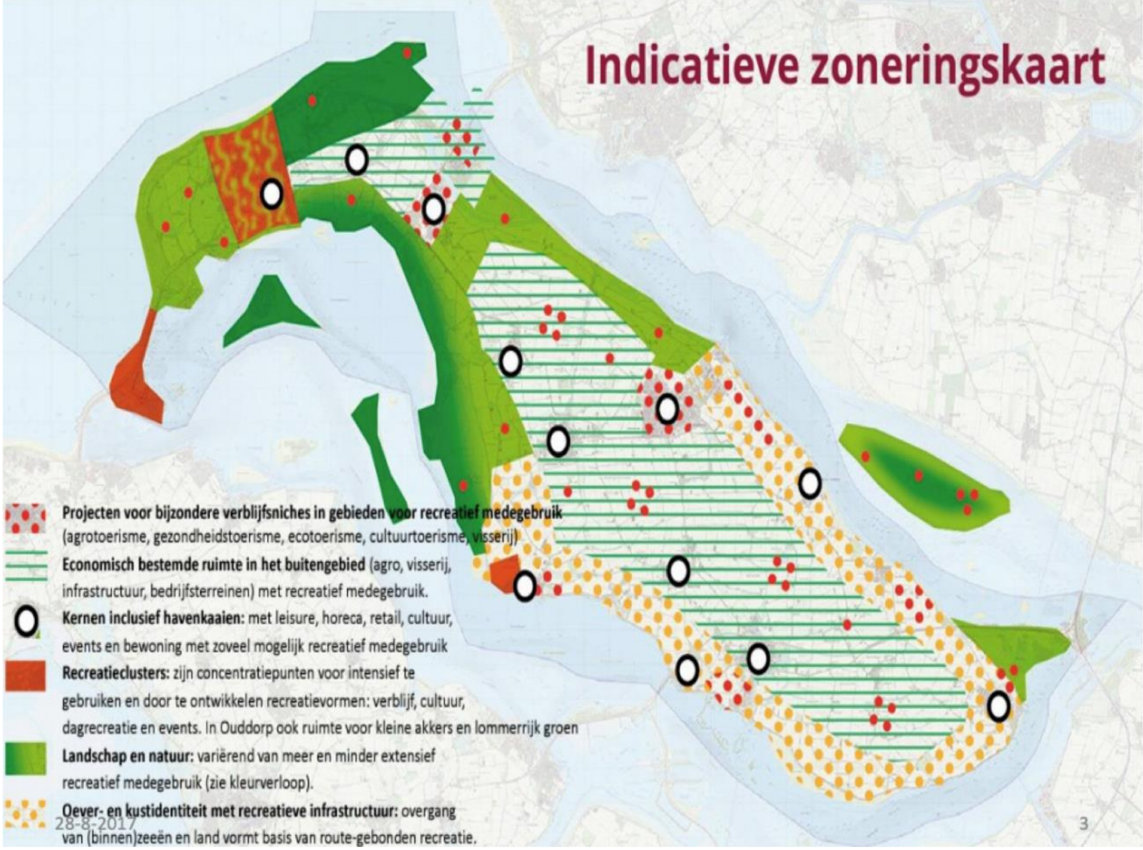
Modeling

1. **Recap of week 3**
2. **What is Cluster Analysis?**
   Clustering application examples
3. **Cluster Analysis: Basic Concepts**
4. **Cluster methods in general**
   Partitioning Methods
   Hierarchical Methods
   Other Methods
5. **Evaluation of Clustering**
6. **Summary**
7. **Exercises with Orange**

Unsupervised learning with Clustering

What is Cluster Analysis?

Clustering Application Example: Land use


Indicatieve zoneringskaart
Projecten voor bijzondere verblijfsniches in gebieden voor recreatief medegebruik (agrotoerisme, gezondheidstoerisme, ecotoerisme, cultuurtoerisme, visserij)
Economisch bestemde ruimte in het buitengebied (agro, visserij, infrastructuur, bedrijfsterreinen) met recreatief medegebruik.
Kernen inclusief havenkaaien: met leisure, horeca, retail, cultuur, events en bewoning met zoveel mogelijk recreatief medegebruik
Recreatieclusters: zijn concentratiepunten voor intensief te gebruiken en door te ontwikkelen recreatievormen: verblijf, cultuur, dagrecreatie en events. In Ouddorp ook ruimte voor kleine akkers en lommerrijk groen
Landschap en natuur: variërend van meer en minder extensief recreatief medegebruik (zie kleurverloop).
Oever- en kustidentiteit met recreatieve infrastructuur: overgang van (binnen)zeeën en land vormt basis van route-gebonden recreatie.
28-8-2017
3

CRISP-DM

Modeling

1. **Recap of week 3**
2. **What is Cluster Analysis?**
   - Clustering application examples
3. **Cluster Analysis: Basic Concepts**
4. **Cluster methods in general**
   - Partitioning Methods
   - Hierarchical Methods
   - Other Methods
5. **Evaluation of Clustering**
6. **Summary**
7. **Exercises with Orange**

Unsupervised learning with Clustering

What is Cluster Analysis?

Clustering Application Example: Land use / city planning

Middelduinen
Detailkaart z.o.z.
Ouddorp
Nieuwland
Kleistee
Polder Oudeland
Westhoofd
Klingendaal
Volgerland
Westduinen
Westeinde
Springert-duinen
Springert-polder
Poldersbos
Preekhilpolder
Zondags rustgebied

Voorstel Natuur-lijk Ouddorp voor Zondagsrust gebied

CRISP-DM

Modeling

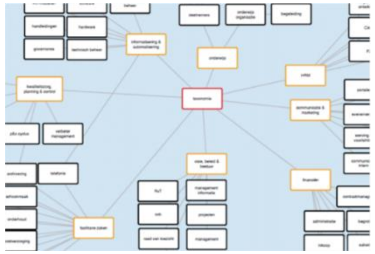
1. **Recap of week 3**
2. **What is Cluster Analysis?**
   Clustering application examples
3. **Cluster Analysis: Basic Concepts**
4. **Cluster methods in general**
   Partitioning Methods
   Hierarchical Methods
   Other Methods
5. **Evaluation of Clustering**
6. **Summary**
7. **Exercises with Orange**

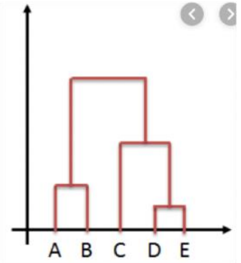
SAXION UNIVERSITY OF APPLIED SCIENCES

Unsupervised learning with Clustering

Basic Concepts

Considerations for Cluster Analysis



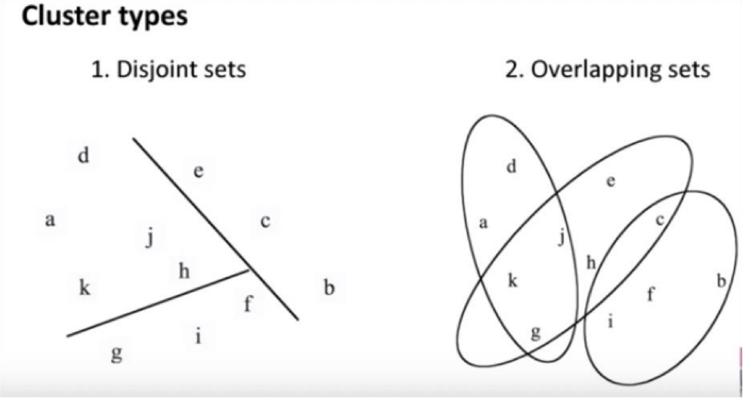


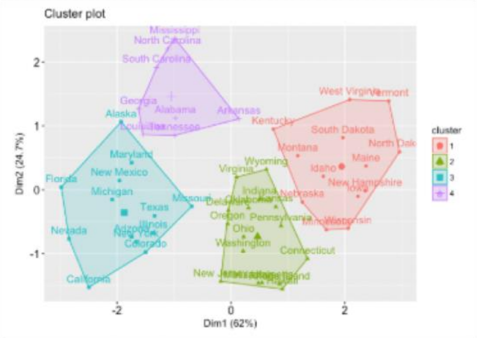
Partitioning criteria

- Single level vs. hierarchical partitioning (often, multi-level hierarchical partitioning is desirable)

Separation of clusters

- Exclusive (e.g., one customer belongs to only one region) vs. non-exclusive(e.g., one document may belong to more than one class)



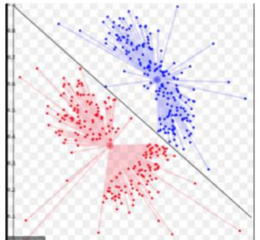


Similarity measure

- Distance-based (e.g., Euclidian) vs. connectivity-based (e.g., density)

Clustering space

- Full space (often when low dimensional) vs. subspaces (often in high-dimensional clustering)



CRISP-DM

Modeling

1. Recap of week 3
2. What is Cluster Analysis?
   - Clustering application examples
3. **Cluster Analysis: Basic Concepts**
4. Cluster methods in general
   - Partitioning Methods
   - Hierarchical Methods
   - Other Methods
5. Evaluation of Clustering
6. Summary
7. Exercises with Orange

Unsupervised learning with Clustering

Basic Concepts

Major Clustering Approaches

Partitioning approach

- Construct various partitions and then evaluate them by some criterion
- 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: Diana, Agnes, BIRCH, CAMELEON

Density-based approach

- Based on connectivity and density functions
- Typical methods: DBSACN, OPTICS, DenClue

Grid-based approach

- based on a multiple-level granularity structure
- Typical methods: STING, WaveCluster, CLIQUE

Model-based approach

Unsupervised learning with Clustering

CRISP-DM

Modeling

1. **Recap of week 3**
2. **What is Cluster Analysis?**
   - Clustering application examples
3. **Cluster Analysis: Basic Concepts**
4. **Cluster methods in general**
   - Partitioning Methods
   - Hierarchical Methods
   - Other Methods
5. **Evaluation of Clustering**
6. **Summary**
7. **Exercises with Orange**

Cluster methods in general

Partitioning Methods

Partitioning method:
Partitioning a dataset of ***n*** objects into a set of ***k*** clusters, such that the sum of squared distances is minimized (where c_i is the centroid or medoid of cluster C_i)

Global optimal:
exhaustively enumerate all partitions
(Volledig opsommen van alle partities)

Heuristic methods:
k-means and *k-medoids* algorithms (distance-based clustering)

1. **Recap of week 3**
2. **What is Cluster Analysis?**
   Clustering application examples
3. **Cluster Analysis: Basic Concepts**
4. **Cluster methods in general**
   Partitioning Methods
   Hierarchical Methods
   Other Methods
5. **Evaluation of Clustering**
6. **Summary**
7. **Exercises with Orange**

Unsupervised learning with Clustering

Cluster methods in general

Partitioning Methods: The K-Means Clustering Method

Given *k*, the *k-means* algorithm is implemented in four steps:

1. Partition objects into *k* nonempty clusters

2. Compute the centroids of the clusters of the current partitioning (the centroid is the center, i.e., *mean point*, of the cluster)
   Update cluster centroids

3. Assign each object to the cluster with the nearest centroid
   Reassign objects to clusters

4. Go back to Step 2, stop when the assignment does not change

CRISP-DM

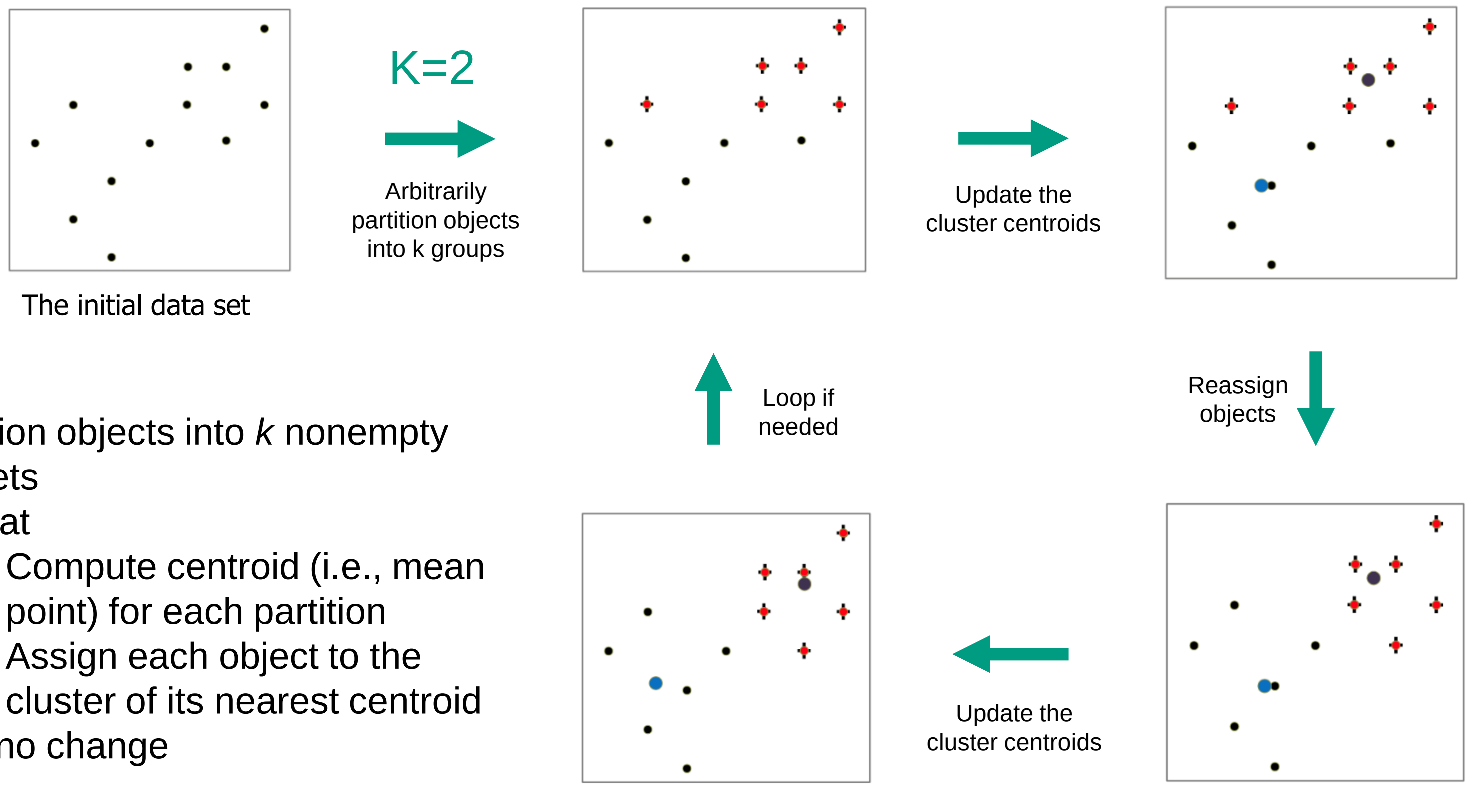
Modeling

1. **Recap of week 3**
2. **What is Cluster Analysis?**
   Clustering application examples
3. **Cluster Analysis: Basic Concepts**
4. **Cluster methods in general**
   Partitioning Methods
   Hierarchical Methods
   Other Methods
5. **Evaluation of Clustering**
6. **Summary**
7. **Exercises with Orange**

Unsupervised learning with Clustering

Cluster methods in general

Partitioning Methods: The K-Means Clustering Method



The initial data set

K=2

Arbitrarily partition objects into k groups

Update the cluster centroids

Reassign objects

Update the cluster centroids

Loop if needed

- Partition objects into *k* nonempty subsets
- Repeat
  - Compute centroid (i.e., mean
  - point) for each partition
  - Assign each object to the
  - cluster of its nearest centroid
- Until no change

CRISP-DM

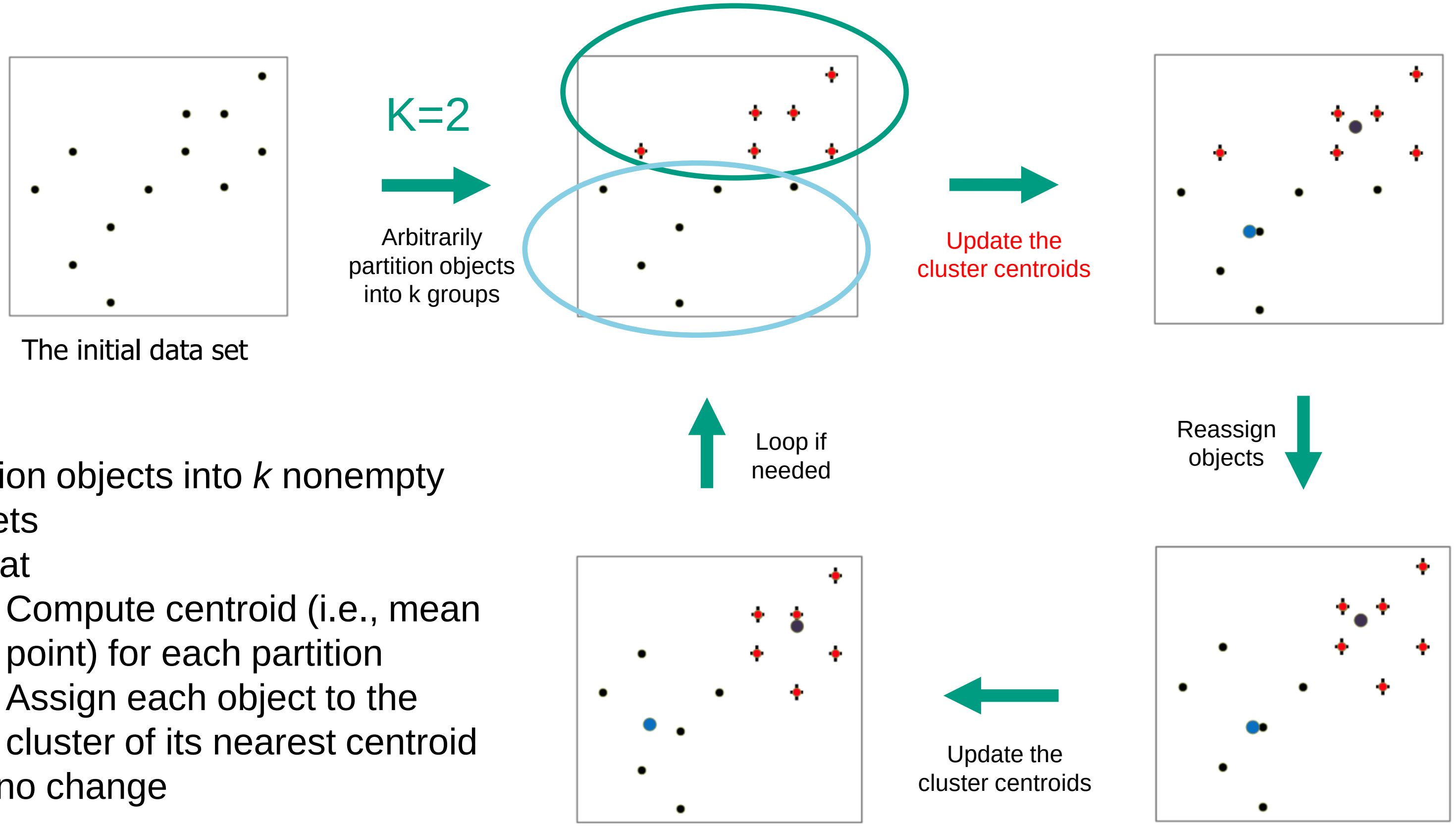
Modeling

1. **Recap of week 3**
2. **What is Cluster Analysis?**
   - Clustering application examples
3. **Cluster Analysis: Basic Concepts**
4. **Cluster methods in general**
   - Partitioning Methods
   - Hierarchical Methods
   - Other Methods
5. **Evaluation of Clustering**
6. **Summary**
7. **Exercises with Orange**

Unsupervised learning with Clustering

Cluster methods in general

Partitioning Methods: The K-Means Clustering Method



K=2
Arbitrarily partition objects into k groups
The initial data set
Update the cluster centroids
Reassign objects
Loop if needed
Update the cluster centroids

- Partition objects into *k* nonempty subsets
- Repeat
  - Compute centroid (i.e., mean
  - point) for each partition
  - Assign each object to the
  - cluster of its nearest centroid
- Until no change

CRISP-DM

Modeling

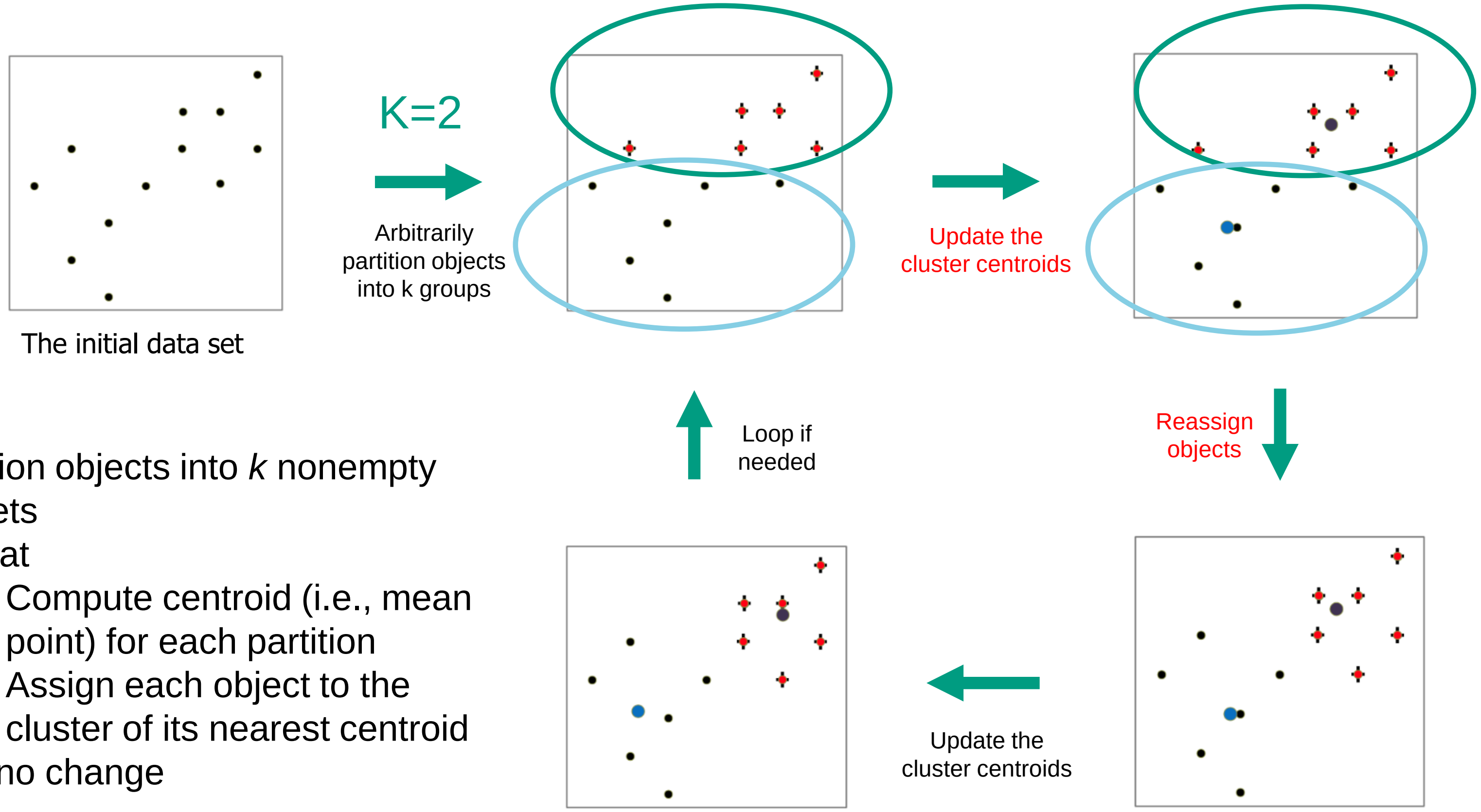
1. **Recap of week 3**
2. **What is Cluster Analysis?**
   Clustering application examples
3. **Cluster Analysis: Basic Concepts**
4. **Cluster methods in general**
   Partitioning Methods
   Hierarchical Methods
   Other Methods
5. **Evaluation of Clustering**
6. **Summary**
7. **Exercises with Orange**

SAXION UNIVERSITY OF APPLIED SCIENCES

Unsupervised learning with Clustering

Cluster methods in general

Partitioning Methods: The K-Means Clustering Method



The initial data set
K=2
Arbitrarily partition objects into k groups
Update the cluster centroids
Reassign objects
Update the cluster centroids
Loop if needed

- Partition objects into *k* nonempty subsets
- Repeat
  - Compute centroid (i.e., mean
  - point) for each partition
  - Assign each object to the
  - cluster of its nearest centroid
- Until no change

CRISP-DM

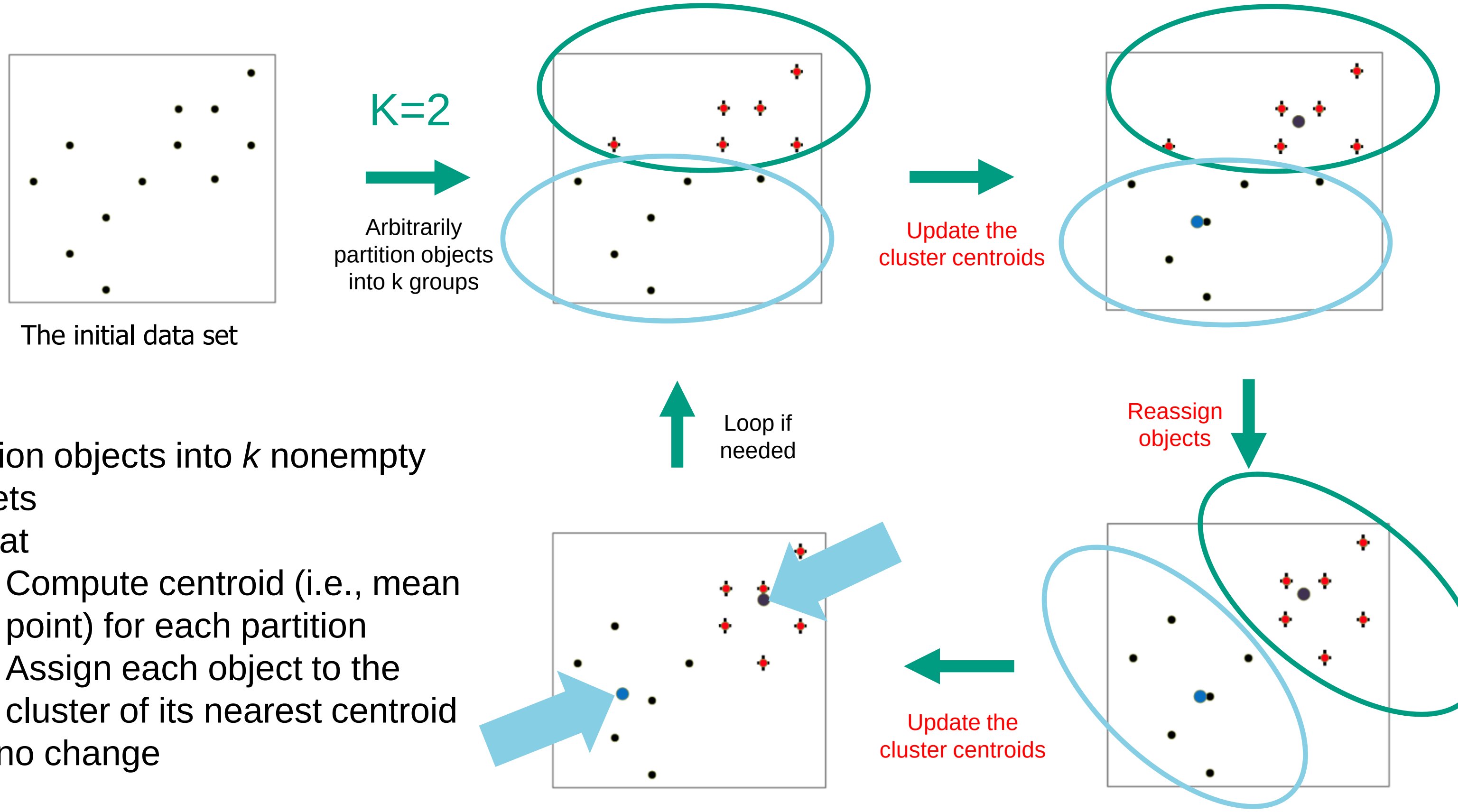
Modeling

1. **Recap of week 3**
2. **What is Cluster Analysis?**
   Clustering application examples
3. **Cluster Analysis: Basic Concepts**
4. **Cluster methods in general**
   Partitioning Methods
   Hierarchical Methods
   Other Methods
5. **Evaluation of Clustering**
6. **Summary**
7. **Exercises with Orange**

Unsupervised learning with Clustering

Cluster methods in general

Partitioning Methods: The K-Means Clustering Method



The initial data set
K=2
Arbitrarily partition objects into k groups
Update the cluster centroids
Reassign objects
Update the cluster centroids
Loop if needed

- Partition objects into *k* nonempty subsets
- Repeat
  - Compute centroid (i.e., mean
  - point) for each partition
  - Assign each object to the cluster of its nearest centroid
- Until no change

Unsupervised learning with Clustering

CRISP-DM

Modeling

1. **Recap of week 3**
2. **What is Cluster Analysis?**
   - Clustering application examples
3. **Cluster Analysis: Basic Concepts**
4. **Cluster methods in general**
   - Partitioning Methods
   - Hierarchical Methods
   - Other Methods
5. **Evaluation of Clustering**
6. **Summary**
7. **Exercises with Orange**

Cluster methods in general

Partitioning Methods: The K-Means Clustering Method

What Is the Problem of the K-Means Method?

The k-means algorithm is sensitive to outliers!
Since an object with an extremely large value may substantially distort the distribution of the data

K-Medoids: Instead of taking the **mean** value of the object in a cluster as a reference point, **medoids** can be used, which is the **most centrally located** object in a cluster

We want two or more clusters with a very small cluster sum of squared errors

CRISP-DM

Modeling

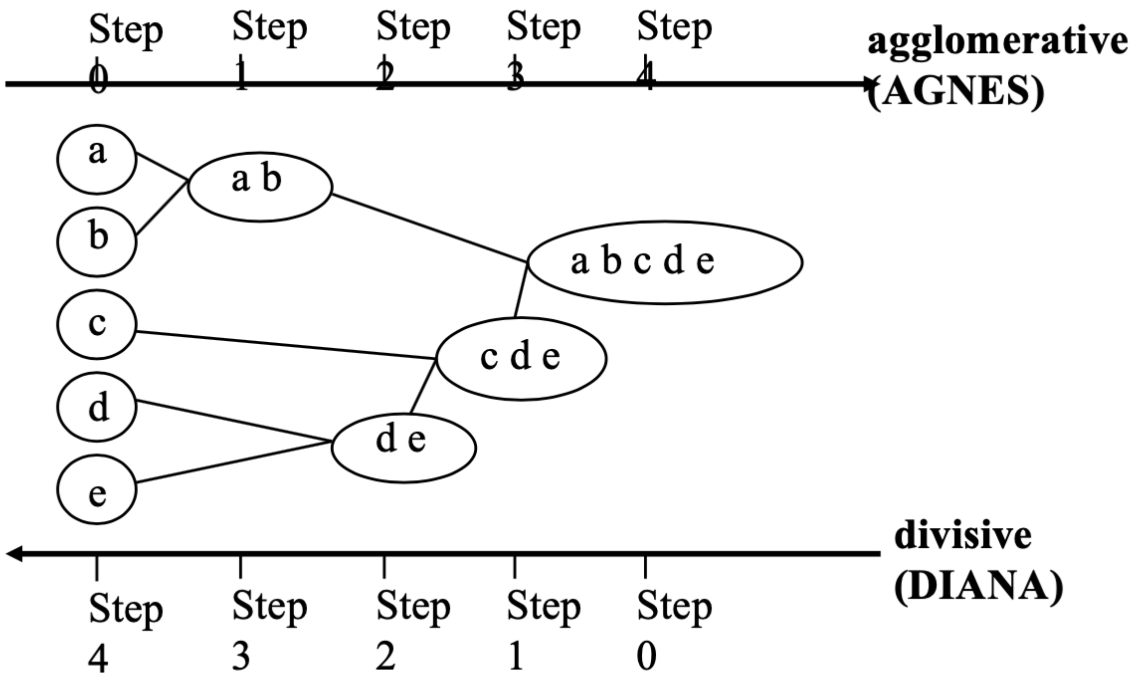
1. **Recap of week 3**
2. **What is Cluster Analysis?**
   Clustering application examples
3. **Cluster Analysis: Basic Concepts**
4. **Cluster methods in general**
   Partitioning Methods
   Hierarchical Methods
   Other Methods
5. **Evaluation of Clustering**
6. **Summary**
7. **Exercises with Orange**

Unsupervised learning with Clustering

Cluster methods in general

Hierarchical Clustering

Use distance matrix as clustering criteria. This method starts with all data points assigned to a cluster of their own, merges the two nearest clusters until there is only a single cluster left



CRISP-DM

Modeling

1. **Recap of week 3**
2. **What is Cluster Analysis?**
   - Clustering application examples
3. **Cluster Analysis: Basic Concepts**
4. **Cluster methods in general**
   - Partitioning Methods
   - Hierarchical Methods
   - Other Methods
5. **Evaluation of Clustering**
6. **Summary**
7. **Exercises with Orange**

Unsupervised learning with Clustering

Cluster methods in general

Determine the Number of Clusters

Empirical method:
# of clusters: k ≈√(n/2) for a dataset of n points, e.g., n = 200, k = 10

Other methods:

- Silhouette method: average silhouette score of the dataset
- Elbow method (k-means)
- Dendrogram heuristic (hierarchical clustering)

Silhouette score *s*(*i*) of point *i*:

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

- *a*(*i*): mean distance between *i* and all other data points in same cluster
- *b*(*i*): smallest mean distance of *i* to all points in any other cluster, of which *i* is not a member

SAXION UNIVERSITY OF APPLIED SCIENCES

CRISP-DM

Modeling

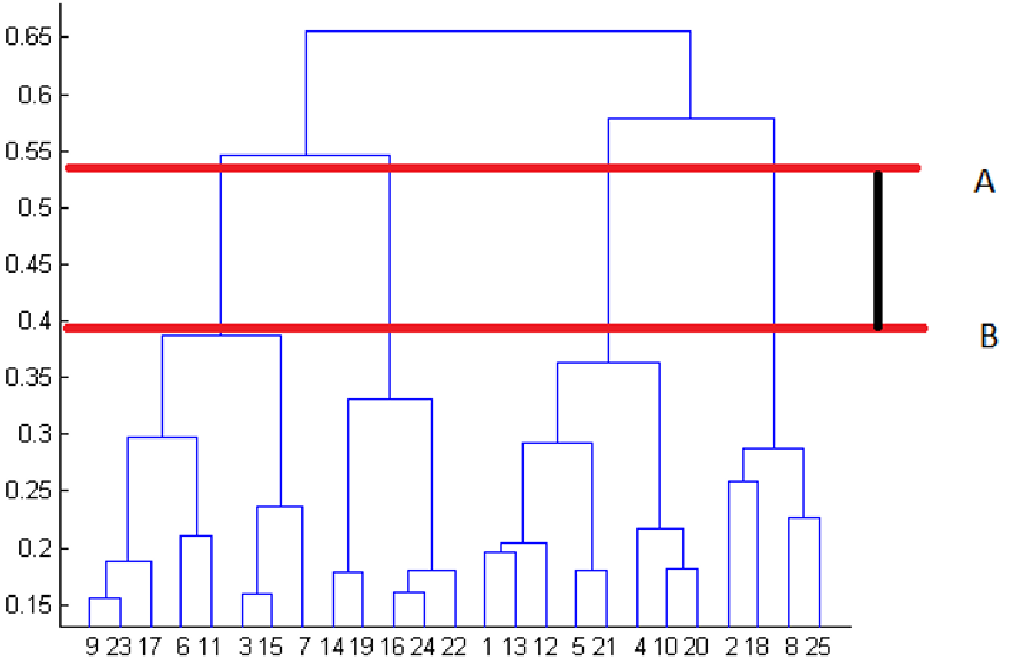
1. **Recap of week 3**
2. **What is Cluster Analysis?**
   - Clustering application examples
3. **Cluster Analysis: Basic Concepts**
4. **Cluster methods in general**
   - Partitioning Methods
   - Hierarchical Methods
   - Other Methods
5. **Evaluation of Clustering**
6. **Summary**
7. **Exercises with Orange**

Unsupervised learning with Clustering

Dendrogram

- Shows all clusters (bottom up)
- The height at which two clusters are merged represents the distance between the clusters (agglomerative)

Heuristic: The number of clusters is equal to the number of vertical lines cut by a horizontal line that can transverse the maximum distance without intersecting a cluster



CRISP-DM

Modeling

1. **Recap of week 3**
2. **What is Cluster Analysis?**
   Clustering application examples
3. **Cluster Analysis: Basic Concepts**
4. **Cluster methods in general**
   Partitioning Methods
   Hierarchical Methods
   Other Methods
5. **Evaluation of Clustering**
6. **Summary**
7. **Exercises with Orange**

SAXION UNIVERSITY OF APPLIED SCIENCES

Unsupervised learning with Clustering

Evaluation of Clustering

Measuring Clustering Quality

- 3 kinds of measures: External, internal and relative

- External: supervised, employ criteria not inherent to the dataset
  - Compare a clustering against prior or expert-specified knowledge using certain clustering quality measure

- Internal: unsupervised, criteria derived from data itself
  - Evaluate the goodness of a clustering by considering how well the clusters are separated, and how compact the clusters are, e.g., Silhouette coefficient

- Relative: directly compare different clusterings, usually those obtained via different parameter settings for the same algorithm

CRISP-DM

Modeling

1. **Recap of week 3**
2. **What is Cluster Analysis?**
   - Clustering application examples
3. **Cluster Analysis: Basic Concepts**
4. **Cluster methods in general**
   - Partitioning Methods
   - Hierarchical Methods
   - Other Methods
5. **Evaluation of Clustering**
6. **Summary**
7. **Exercises with Orange**

SAXION UNIVERSITY OF APPLIED SCIENCES

Unsupervised learning with Clustering

Summary

- Cluster analysis groups objects based on their similarity and has wide applications
- Clustering algorithms can be categorized into partitioning methods, hierarchical methods, density-based methods, grid-based methods, and model-based methods
- K-means and K-medoids algorithms are popular partitioning-based clustering algorithms
- Birch and Chameleon are interesting hierarchical clustering algorithms, and there are also probabilistic hierarchical clustering algorithms
- DBSCAN, OPTICS, and DENCLU are interesting density-based algorithms
- STING and CLIQUE are grid-based methods, where CLIQUE is also a subspace clustering algorithm
- Quality of clustering results can be evaluated in various ways

EXERCISES

Let’s start with the exercises in Orange

Exercise 1
Do the following e-learnings to get familiar with clustering.

1. Getting Started with Orange 11: k-Means
   https://www.youtube.com/watch?v=vgmL808eSw4

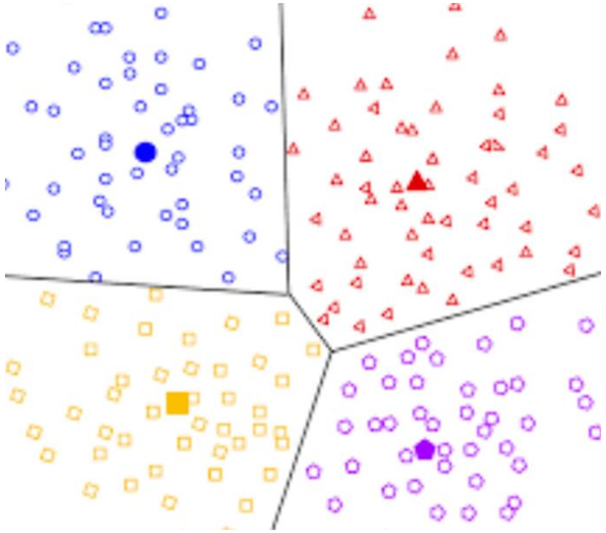
2. Getting Started with Orange 12: k-Means Explained
   https://www.youtube.com/watch?v=I0e0Qyev8Ac

3. Getting Started with Orange 13: Silhouette
   https://www.youtube.com/watch?v=5TPldC_dC0s

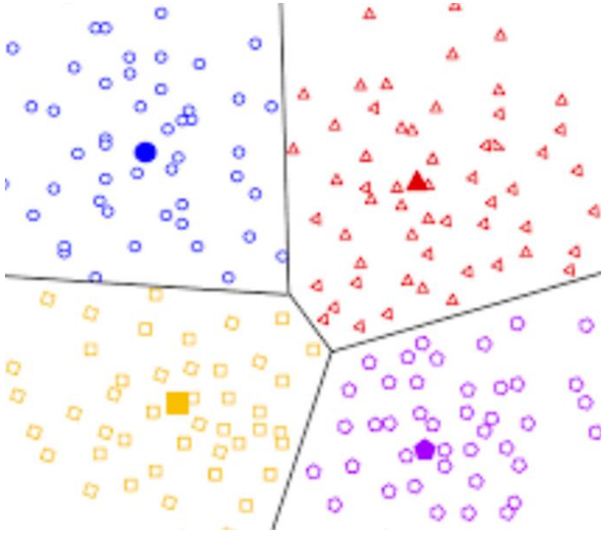
4. Text clustering:
   https://www.youtube.com/watch?v=rH_vQxQL6oM

5. Image clustering:
   https://www.youtube.com/watch?v=Iu8g2Twjn9U

6. Image clustering predictions:
   https://www.youtube.com/watch?v=lvgx62a8XQk



Let’s start with the exercises in Orange



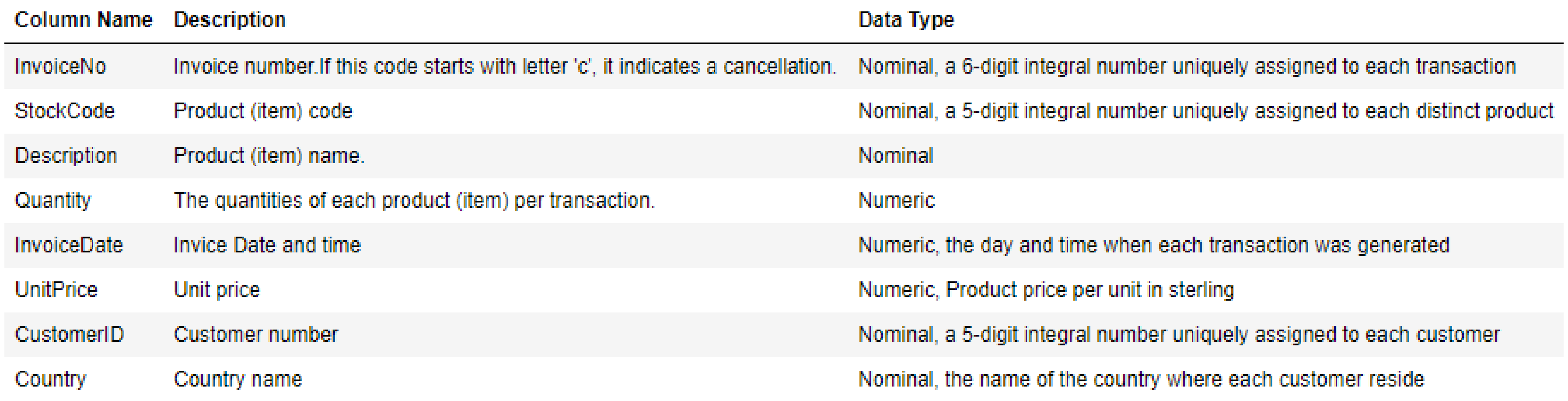
Exercise 2

Cleaning and clustering exercise:

1. Download the file ‘Online Retail small’ from Blackboard (week 4).
2. Use CRISP-DM (exploratory data analysis and preprocessing)
3. Method for this file;
   - We want to make clusters for this file, with Orange;
   - Investigate some interesting clusters for this file;
   - Cluster using the best scoring clustering

Let’s start with the exercises in Orange

Column Descriptions



Column Name	Description	Data Type
InvoiceNo	Invoice number.If this code starts with letter 'c', it indicates a cancellation.	Nominal, a 6-digit integral number uniquely assigned to each transaction
StockCode	Product (item) code	Nominal, a 5-digit integral number uniquely assigned to each distinct product
Description	Product (item) name.	Nominal
Quantity	The quantities of each product (item) per transaction.	Numeric
InvoiceDate	Invice Date and time	Numeric, the day and time when each transaction was generated
UnitPrice	Unit price	Numeric, Product price per unit in sterling
CustomerID	Customer number	Nominal, a 5-digit integral number uniquely assigned to each customer
Country	Country name	Nominal, the name of the country where each customer reside