Prof. Dr. Alfred Benedikt Brendel
Chair of Business Information Systems, esp. Intelligent Systems and Services

Data Science: Advanced Analytics
Basics of Segmentation

Dresden // 26.04.2023
Sommersemester 2023

TECHNISCHE UNIVERSITÄT DRESDEN
CHAIR OF BUSINESS INFORMATION SYSTEMS
ESP. INTELLIGENT SYSTEMS AND SERVICES
DRESDEN concept

Prof. Dr. Alfred Benedikt Brendel
Chair of Business Information Systems, esp. Intelligent Systems and Services

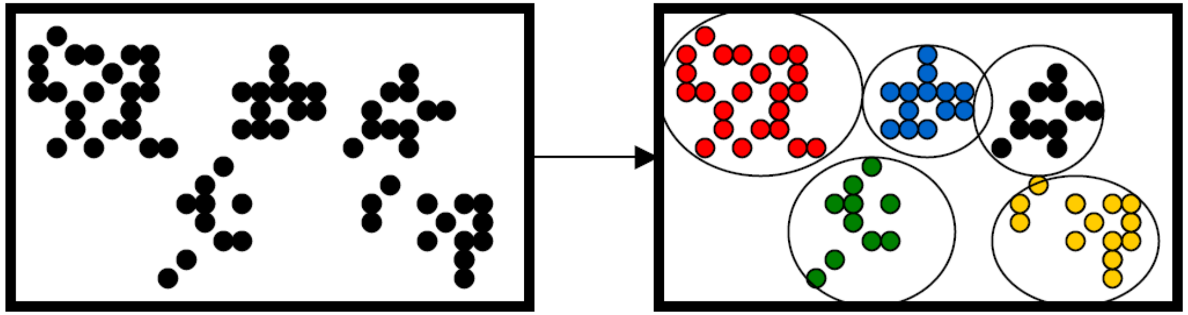
Basics of Segmentation

Introduction

Aims of the cluster analysis

- **Segmentation** or **cluster analysis** is used to group objects and/or characteristics into classes or groups, so that
  - between the elements of the same classes as **similar as possible,**
  - between the elements of different classes the **greatest possible dissimilarity**

  is achieved.
- The clusters are not known beforehand.
- So-called **unsupervised learning methods** are used, for example.



Prof. Dr. Alfred Benedikt Brendel
Chair of Business Information Systems, esp. Intelligent Systems and Services

Basics of segmentation

Cluster types

Subdivision of cluster approaches
Disjunctive and exhaustive methods

- First, a distinction is made between disjoint and non-disjoint segmentations, where in the first case a given element may be assigned to only one class, in the second case to several classes.
- A segmentation K = {K1, K2, … } of the set N is thus called
  - **disjoint** if holds: $K_i, K_j \boldsymbol{K},\ i\ \neq j\ K\ L\ =\ \emptyset$
  - **non-disjoint** if holds: $K_i, K_j \boldsymbol{K},\ \ \exists\ E\ \boldsymbol{K}: E K_i \wedge E K_j, i\ \neq j$
- In addition, a distinction is also made between exhaustive methods, in which each element is assigned to at least one class, and non-exhaustive methods, which allow unclassified elements.
- A segmentation K = {K1, K2, … } of the set N is thus called
  - **exhaustive** if applies:
  - **non-exhaustive** if applies:

$$\bigcup_{K_i \in \boldsymbol{K}} K_i = N$$

$$\bigcup_{K_i \in \boldsymbol{K}} K_i \subset N$$

Subdivision of cluster approaches
Agglomerative and single-modal methods

Furthermore, one still distinguishes:

- Agglomerative and divisive methods. The former start from one-element classes and combine elements into classes step by step. Divisive methods, on the other hand, start from an initial decomposition which is refined (i.e., divided) step by step.
- Single-modal methods that group objects or features into classes, and dual-modal methods that simultaneously classify objects and features of a data matrix.

Subdivision of cluster approaches

Hierarchical and sharp methods

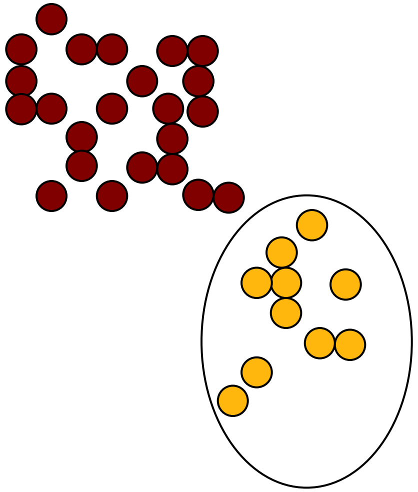
- Hierarchical segmentation methods are characterized by the fact that a cluster located at a higher fusion level completely contains the corresponding clusters located at a lower level.
- Non-hierarchical methods are based on optimization methods that estimate class membership or attempt to iteratively improve classification quality by swapping elements between classes.
- Furthermore, a distinction is made between sharp and fuzzy segmentations. In the sharp approach, each object - if assigned to at least one class - is uniquely assigned to the corresponding class. In the fuzzy approach, there is no clear assignment to a class. Instead, share values are assigned that determine the degree to which an object belongs to a class.

Disjoint segmentation
Example

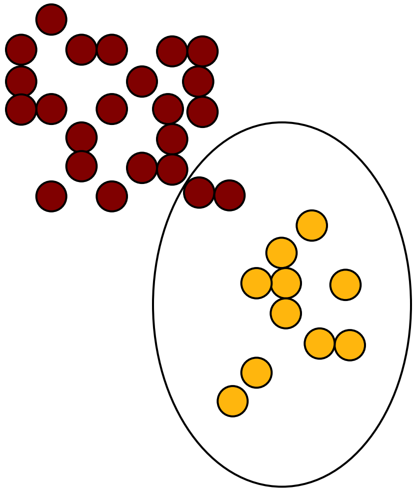
Two classes may contain common elements, but a subset relationship is excluded.

N = {1,2,3,4,5}

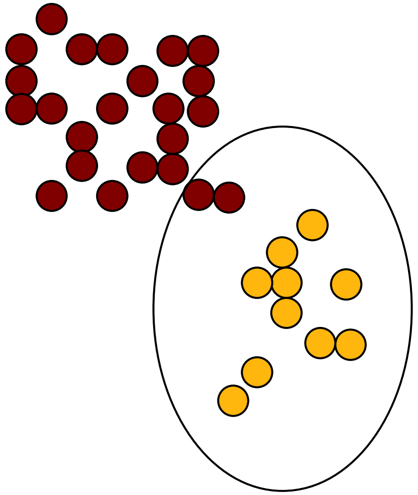
disjunct



non-disjunct



not allowed

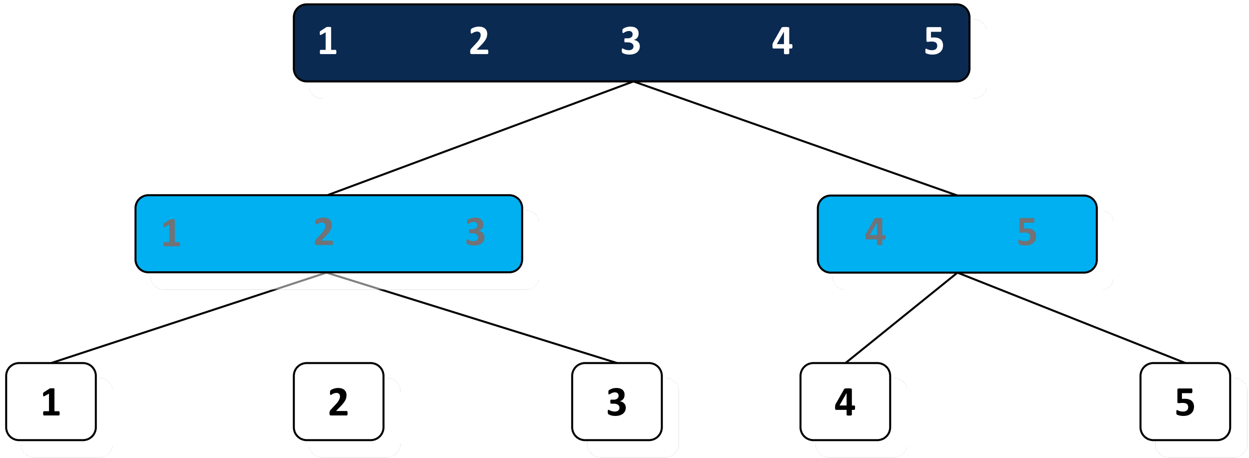


Hierarchical segmentation
Example

A hierarchy is a union of disjoint segmentations, i.e., a sequence of disjoint segmentations for one class, two classes, ..., n classes. An overlap of the classes is excluded.

$\boldsymbol{K}$ = { {1}, ..., {5}, {1,2,3}, {4,5}, {1,2,3,4,5} }

"Union of decompositions"

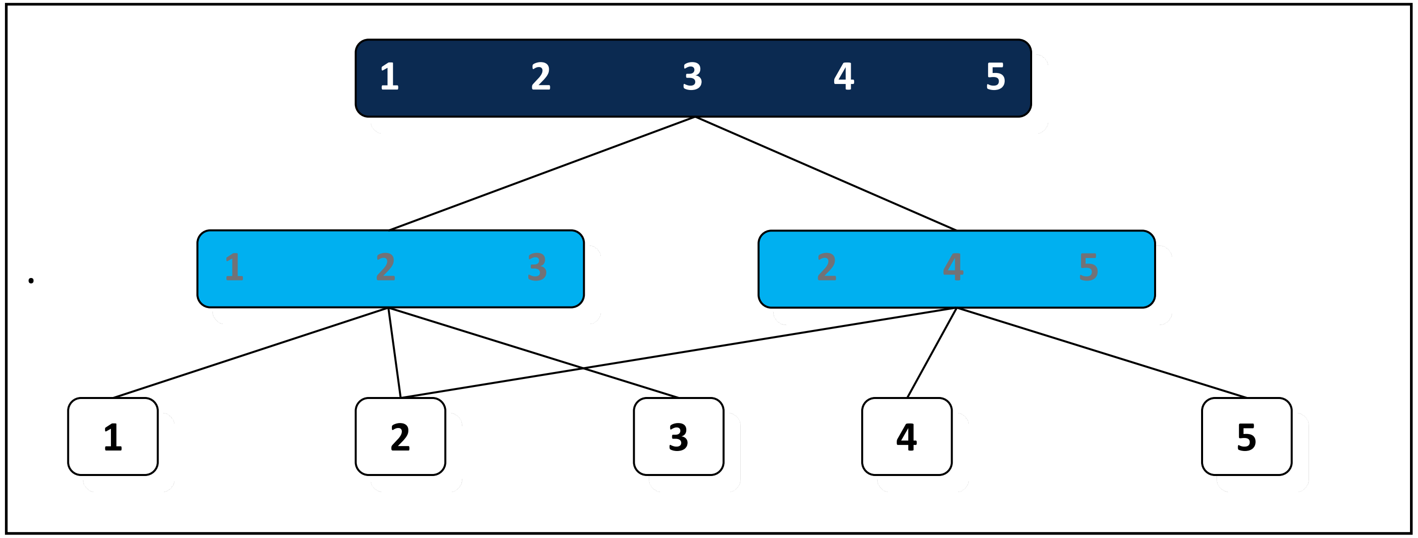


Quasi-hierarchical segmentation

Example

A quasi-hierarchy is a union of non-disjoint segmentations, i.e., a sequence of non-disjoint segmentations for one class, two classes, ..., n classes. An overlap of classes is not excluded. Furthermore, it holds that the union of all real subsets of a class K yields just K again.

K = { {1}, ..., {5}, {1,2,3} {2,4,5}, {1,2,3,4,5} }



Choice of segmentation type

Further requirements for the classification type arise

— from the problem definition,

— of the required number of classes,

— Upper or lower limits for the number of objects in the classes

— etc.

The segmentation method is then also derived from the desired classification type.

Prof. Dr. Alfred Benedikt Brendel
Chair of Business Information Systems, esp. Intelligent Systems and Services

Basics of segmentation

Distance measures

Similarity measures of the objects

—Since cluster analysis is used to group objects and/or features into classes or groups according to their similarity, one needs a measure that quantifies the similarity of two objects described by any features.

- Similarity measure AM: The larger a value, the more similar two objects are.
- Consequence: What does a value AM = 0 mean? or How great is the similarity of two identical objects?

⇒ Transition to a difference measure or distance measure

! The main problem of this measure is not the determination of the dissimilarity of two objects on the basis of one characteristic, but the dissimilarity in the presence of several characteristics (aggregation problem).

Aggregation of nominal characteristics

— Objects are generally characterized on the basis of nominal features by the possession of a certain property.

—Considerin binary features, two objects (i,j) show a high similarity with respect to several features, if frequently - i.e. for many features - the same expression can be observed.

- Yes / Yes i.e. 1/1
- No / No i.e. 0/0

Common combinations like

- Yes / No or No / Yes

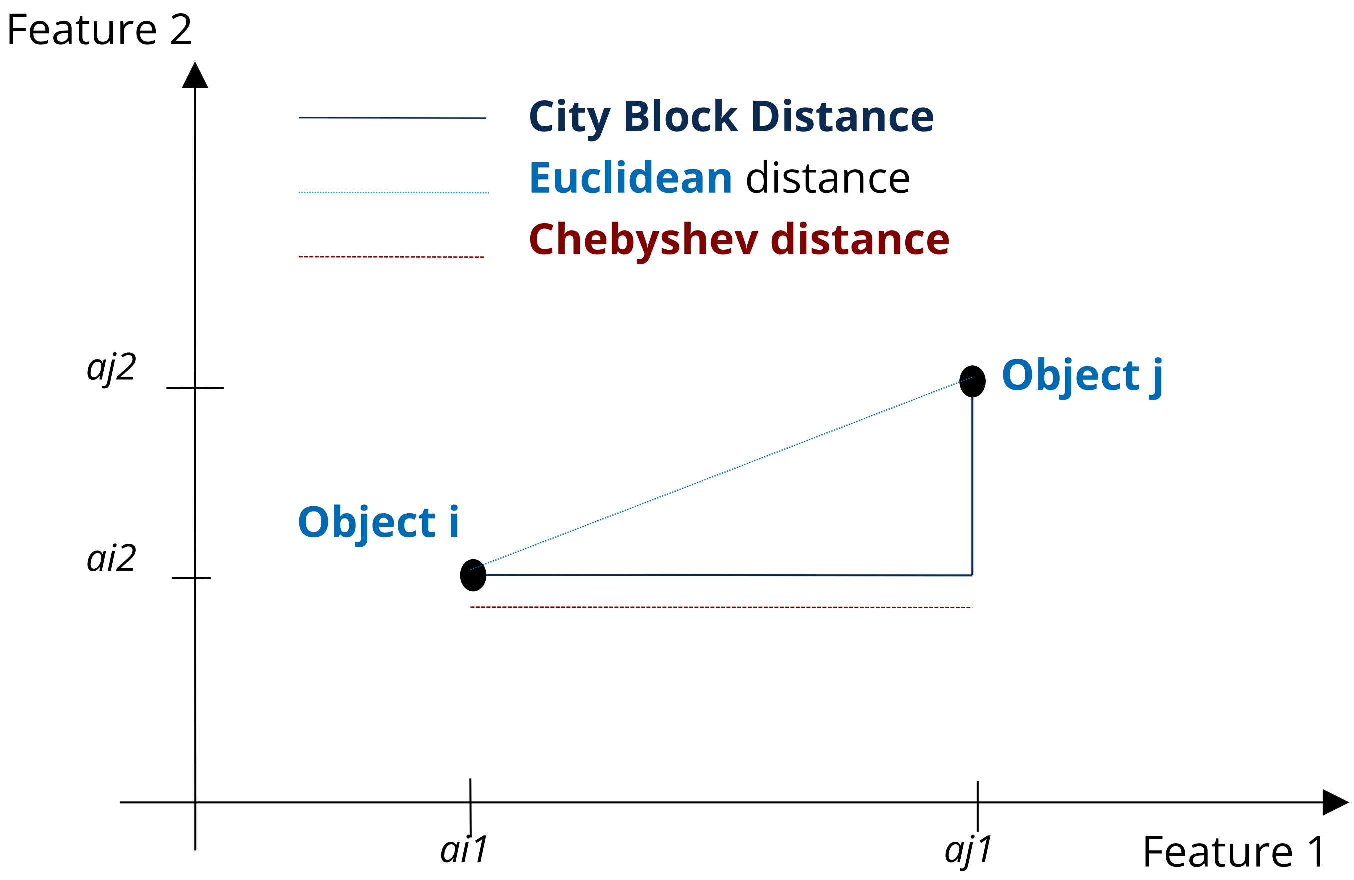
indicate a high level of dissimilarity.

Special distance indices
Interpretation of distances

- The **idea of distance aggregation** can be understood as follows:

  *Each **oject vector ai** can be represented as a point in m-dimensional space.*

- So this **geometric view** suggests to measure the distance between i and j by **Euclid's distance** (or a generalization of it).



Aggregation of quantitative characteristics

— Let $A = (a_{ik})_{nxm}$ be a quantitative data matrix. Then d(i,j) is

$$d(i,j) = \left(\sum_{k=1}^{m} \gamma_k \left| a_{ik} - a_{jk} \right|^p \right)^{\frac{1}{p}}, \qquad \gamma_k > 0, \quad p \in \mathbb{N}$$

weighted L_p distance of i and j.

Special distance indices
Definition of selected distances

Specifically, one speaks of a

— City block distance for p = 1:

$$d(i,j) = \sum_{k=1}^{m} \gamma_k \left| a_{ik} - a_{jk} \right|$$

— Euclidean distance for p = 2:

$$d(i,j) = \sqrt{\sum_{k=1}^{m} \gamma_k \left| a_{ik} - a_{jk} \right|^2}$$

— Chebyshev distance for $p \to \infty$:

$$d(i,j) = \max_k \ \gamma_k \left| a_{ik} - a_{jk} \right|$$

Problematic properties of the Lp distance

— If two features k1 and k2 are highly correlated, then both features provide approximately the same information regarding the similarity of the objects.

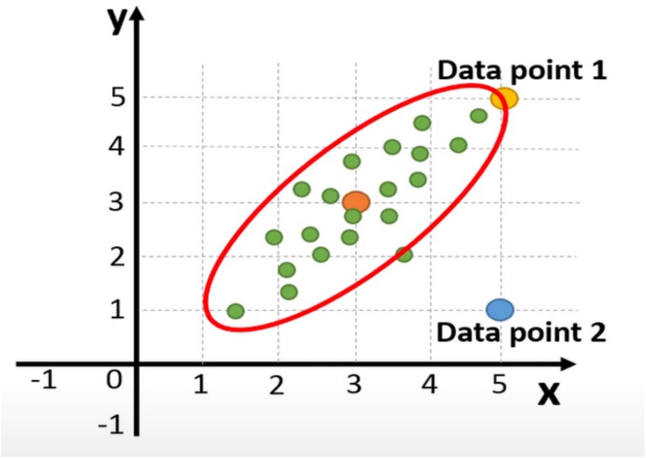
The same information is considered "multiple".

— Characteristics with large dispersion (i.e. variance) have a higher weight in aggregation

— The general remedy here is the **Mahalanobis distance**

cluster has centroid & distance to centroid describes if they are similiar to the whole cluster

→



Prof. Dr. Alfred Benedikt Brendel
Chair of Business Information Systems, esp. Intelligent Systems and Services

Basics of segmentation

Evaluation criteria

Evaluation criteria

If one has two different segmentations K_1 und K_2, the question arises, which is more suitable.

Criteria for the evaluation of segmentations

A distinction is made between measures for evaluating the

— Heterogeneity of **a class** (intra-class dissimilarity) - heterogeneity indices

— Difference of **two classes** (interclass difference) - dissimilarity indices

— Quality of **a segmentation** (quality index) - quality indices

Examples of heterogeneity indices

— Evaluation of the **maximum distances between** two objects:

$$h(K) = \max_{i,j} d(i,j)$$

— Evaluation of the (weighted) **sum of all distances** between objects:

$$h(K) = \frac{1}{c} \sum_{\substack{i<j \\ i,j \in K}} d(i,j) \quad \text{mit} \quad \begin{array}{l} c = 1 \\ c = \frac{1}{2} |K| \cdot (|K| - 1) \end{array}$$

Examples of dissimilarity indices

— **Single Linkage:** Evaluation of the minimum distances of two objects from the different classes:

$$v(K,L) = \min_{i \in K, j \in L} d(i,j)$$

— **Complete Linkage:** Evaluation of the maximum distances of two objects from the classes:

$$v(K,L) = \max_{i \in K, j \in L} d(i,j)$$

— Evaluation of the (weighted) **sum of all distances** between objects:

$$v(K,L) = \frac{1}{c} \sum_{i \in K} \sum_{j \in L} d(i,j)$$

Examples of quality indices
Part 1

— Evaluation of the classification on the **basis of heterogeneity**

$$b(\boldsymbol{K}) = \frac{1}{c} \sum_{K \in \boldsymbol{K}} h(K)$$

- c = 1 -> Sum of heterogeneity indices
- c = |K| ->Mean class heterogeneity

or

$$b(\boldsymbol{K}) = \max_{K \in \boldsymbol{K}} h(K)$$

Examples of quality indices

Part 2

— Evaluation of the classification on the **basis of disimilarity**

$$b(\boldsymbol{K}) = c \cdot \left(\sum_{\substack{K,L \in \boldsymbol{K} \\ K \neq L}} v(K,L) \right)^{-1}$$

— Evaluation of the classification on the **basis of heterogeneity** and on the **basis of disimilarity.**

$$b(\boldsymbol{K}) = \left(\sum_{K \in \boldsymbol{K}} h(K) \right) \cdot \left(\sum_{\substack{K,L \in \boldsymbol{K} \\ K \neq L}} v(K,L) \right)^{-1}$$

The problem of quality indices

— In general, the quality index falls as the number of classes increases.

— **A conflict of objectives** arises between

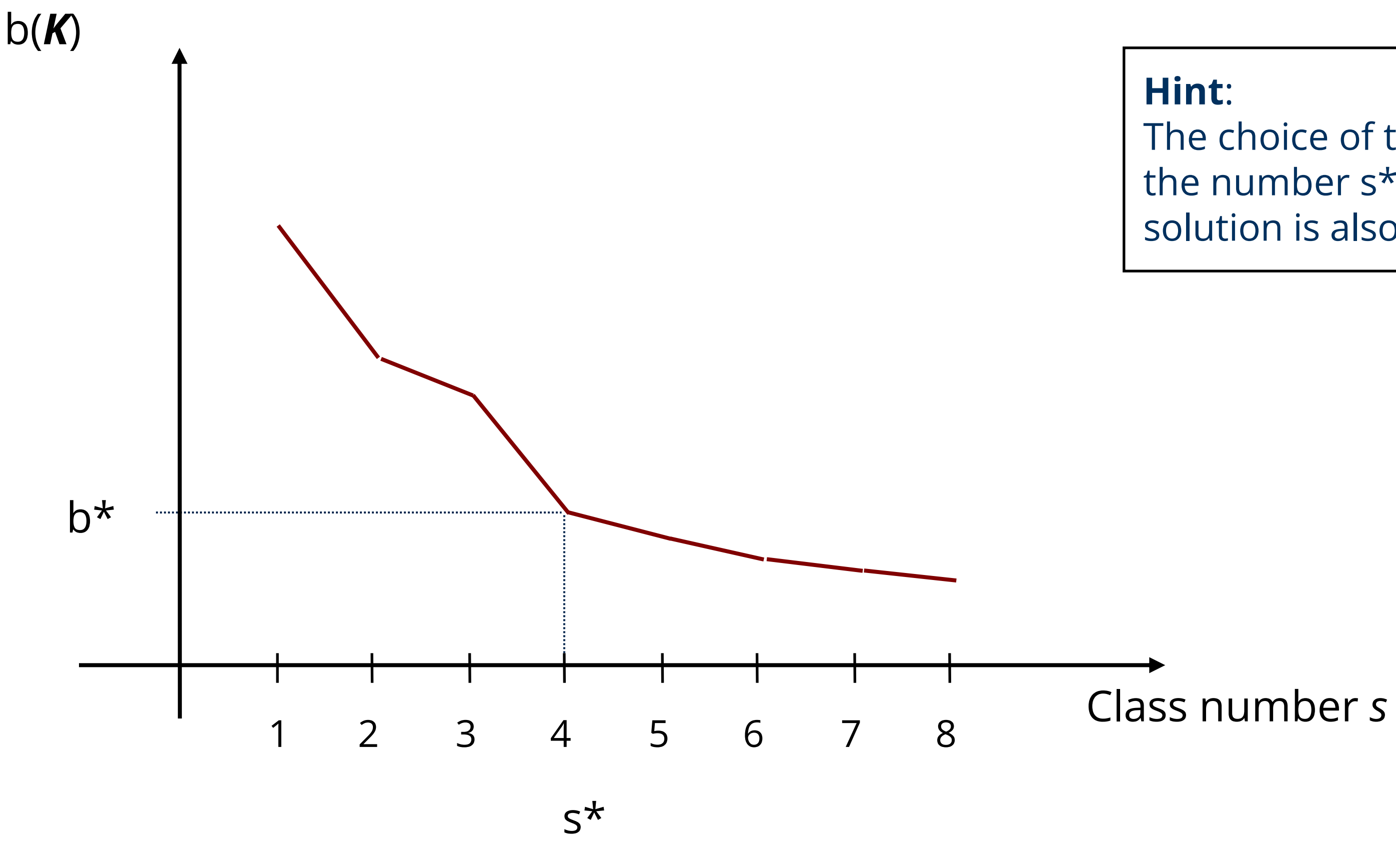
- as small a number of classes as possible and
- quality index as small as possible.

Decision support: **elbow criterion**

With the help of this so-called criterion, the "optimal" choice of the number of classes can be made. The decision is made in favor of the number of classes s* with the quality b*, which is the same with

- a **reduction in the** number of classes **s** leads to a **sharp increase in b**
- an **increase in the** number of classes **s** leads only to a **slight improvement in b.**

Elbow diagram



Hint:
The choice of the number of classes falls in favor of the number s*=4. The choice of the two-class solution is also justifiable.

TECHNISCHE UNIVERSITÄT DRESDEN

CHAIR OF BUSINESS INFORMATION SYSTEMS
ESP. INTELLIGENT SYSTEMS AND SERVICES

DRESDEN concept

Prof. Dr. Alfred Benedikt Brendel
Chair of Business Information Systems, esp. Intelligent Systems and Services

Basics of segmentation

Launch heuristics

Launch heuristics

are segmentation methods that

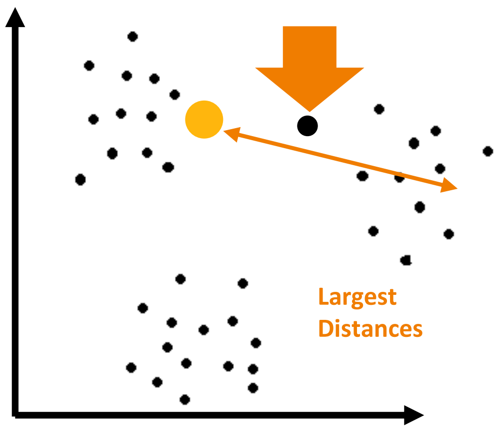
- using **simple algorithmic approaches,**
- without the claim to **optimality,**
- without high (computing) **effort** and
- i.A. on the basis of a **distance matrix**

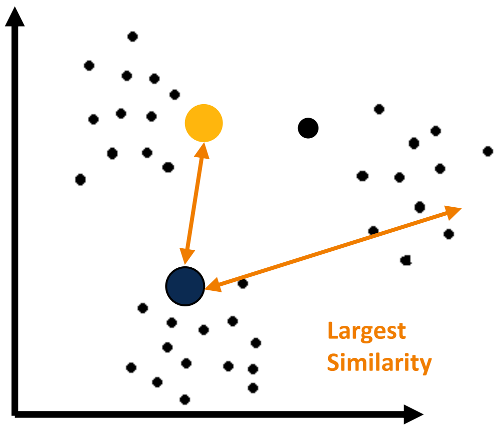
divide a set of objects into

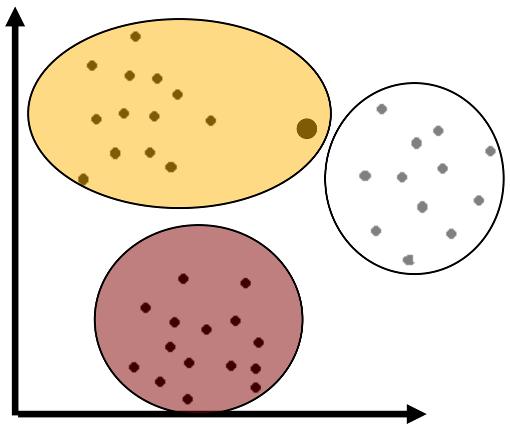
- a **decomposition** or
- an **overlap**.

Sequence of heuristics for a decomposition

Random selection The starting point	Formation of the Cluster representatives	Formation of the Cluster







Procedure and application

1) First, s different class centers are determined if possible.

2) After that, the remaining elements are assigned to the nearest class center.

Application of heuristics:

— Determination of initial segmentations that can be iteratively improved using other methods.

— Determination of segmentations when very large data sets preclude other methods for computation time reasons.

Heuristic for a decomposition

Object set N, distance matrix D, class number s

1) Choose 1st class center (cc) $i_1 \in N$ randomly

2) Select 2nd cc $i_2 \in N$ with:

$$\max_j d(i_1,j) = d(i_1,i_2)$$

3) Choose for t=3, …,s cc $i_t \in N$ with

$$\max_j \min_{\tau=1,\dots,t-1} d(i_\tau,j) = d(i_\tau,i_t)$$

4) Form classes around the centers i_1, …, i_s according to

$$\{K_1,\dots,K_s\} \qquad K_\sigma = \left\{ j \in N : \min_s d(i_s,j) = d(i_\sigma,j) \right\}$$

TECHNISCHE UNIVERSITÄT DRESDEN

CHAIR OF BUSINESS INFORMATION SYSTEMS
ESP. INTELLIGENT SYSTEMS AND SERVICES

DRESDEN concept

Prof. Dr. Alfred Benedikt Brendel
Chair of Business Information Systems, esp. Intelligent Systems and Services

Thank you for your attention