Christian-Albrechts-Universität zu Kiel

Lecture
“Intelligent Systems“

Chapter 7: Clustering of Time-Series

Prof. Dr.-Ing. habil. Sven Tomforde / Intelligent Systems

Winter term 2020/2021

About this chapter

Contents

- Basics
- Offline techniques
- Online techniques
- Conclusion and further readings

Goals

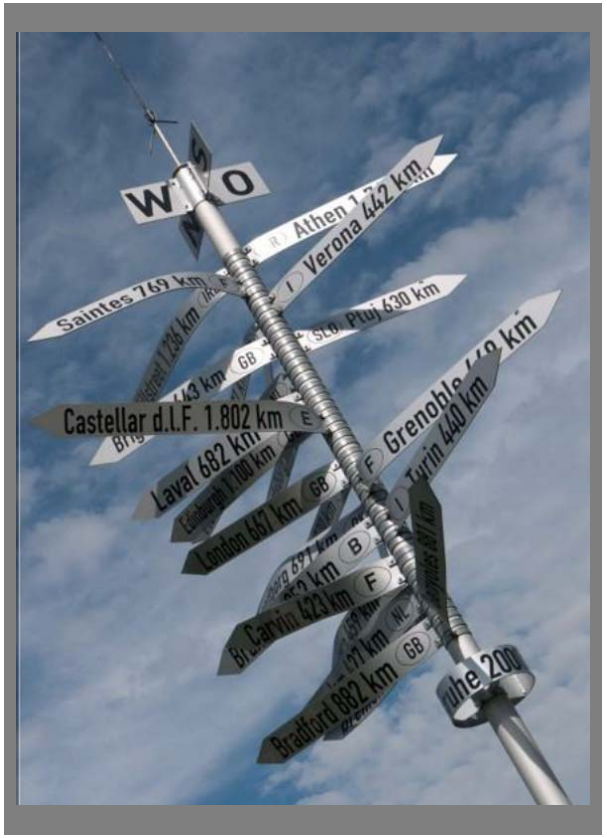
Students should be able to:

- define what segmentation of time series is and why it is necessary.
- explain the differences between online and offline approaches.
- compare bottom-up and top-down concepts.
- compare sliding and growing window as well as SWAB approaches.

Agenda

Christian-Albrechts-Universität zu Kiel

- Basics
- Hierarchical Clustering
- c-Means
- Further techniques
- Cluster evaluation
- Conclusion and further readings



Clustering

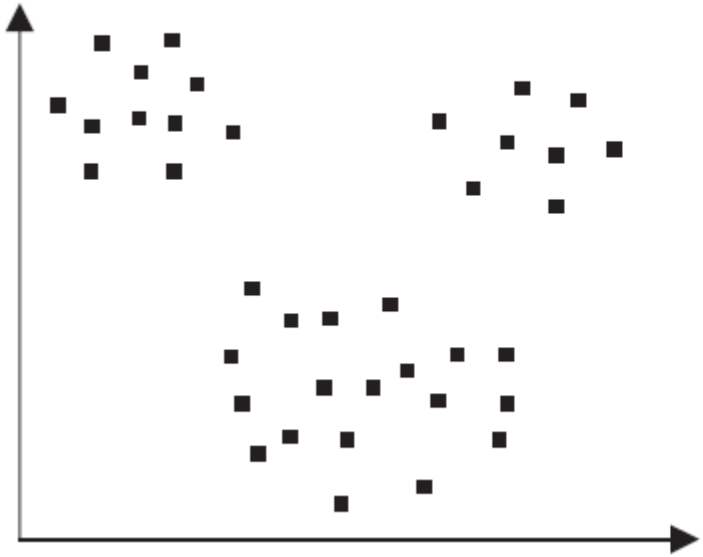
- Subdivision of a set of given patterns into groups (clusters)
- I.e. a task of unsupervised learning
- In contrast to classification, which is used to predict class affiliation (with the help of examples).
- Typically, the result of the process is to summarise similar patterns:
  - Homogeneity within clusters:
    → Patterns belonging to the same cluster should be as similar as possible.
  - Heterogeneity between clusters:
    → Patterns assigned to different clusters should be as different as possible.

Clustering (continued)

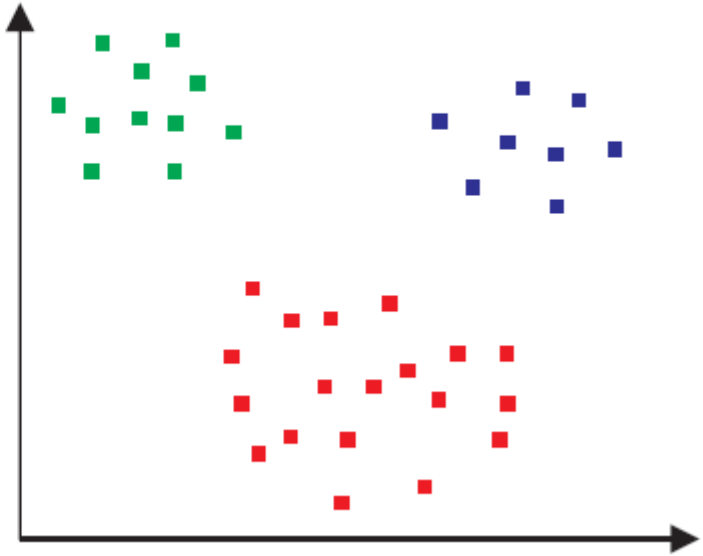
- There are no specifications as to which patterns belong in which clusters.
- The most important tool in clustering is the distance (or similarity) between two patterns
- For time-series: clustering using a similarity measure on time series (or using a suitable representation/model)

Example:

Data set:

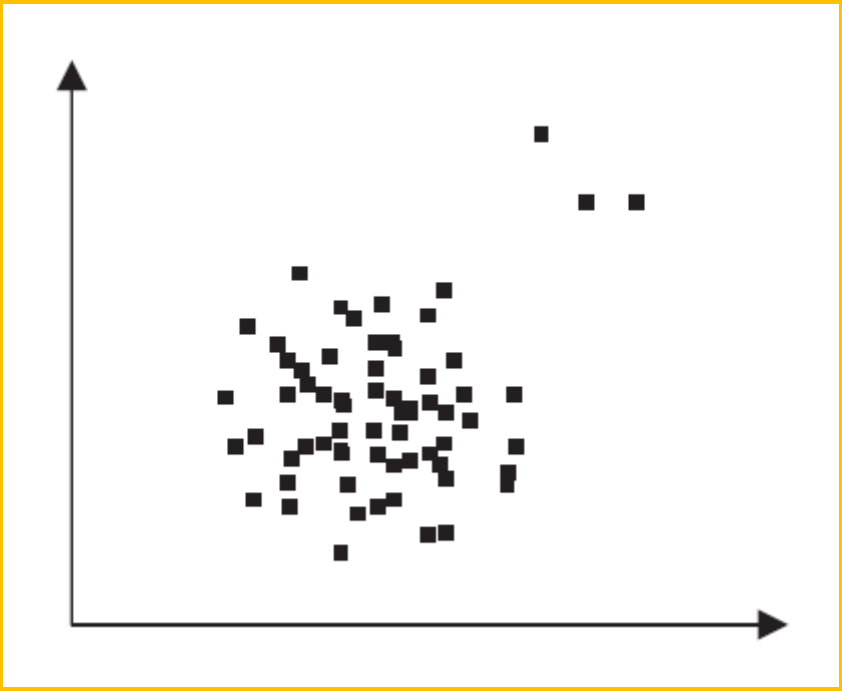


Clustering result:



Question:

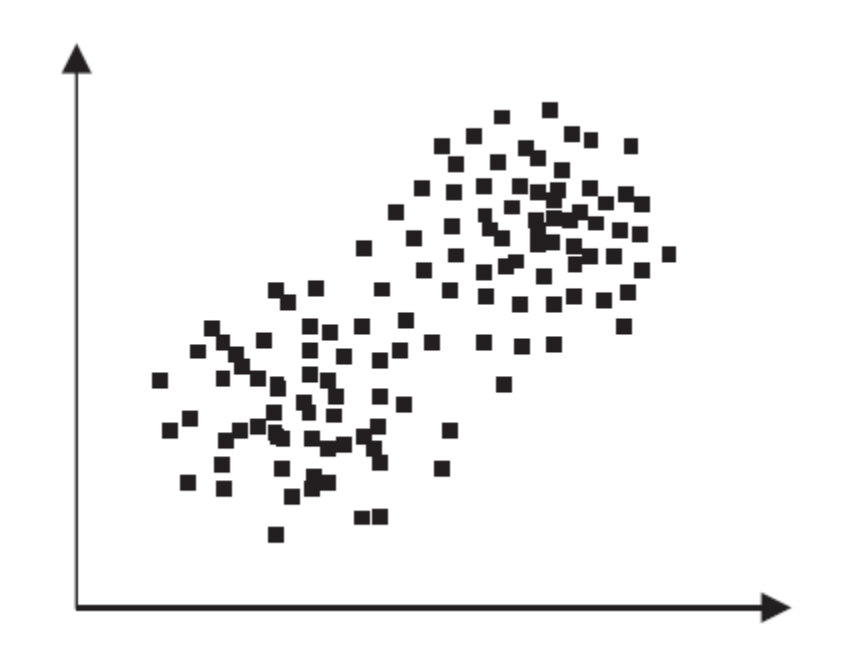
- Consider the following example – what are the main problems for clustering illustrated here?



Solution

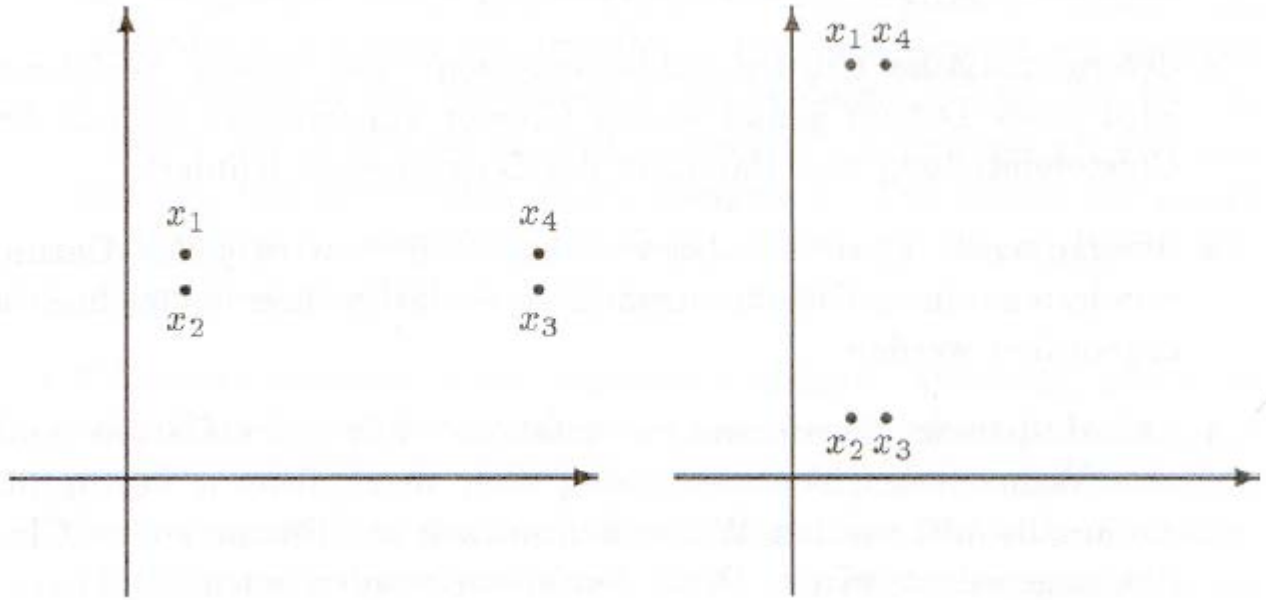
Question:

- How many clusters do you identify in this example?



Problem

- Intuitive cluster assignment depends on the scaling of the individual axes/attributes.
- Example: Same pattern, different scaling



Interesting aspects of clustering:

- Treatment of outliers - Elimination? Formation of own clusters?
- The clustering result depends on the pre-processing of the data (e.g. scaling).
- Interpretation of the clustering results - experts must assign a meaning (label if necessary) to the clusters.
- The number of clusters is sometimes difficult to determine or decide.
- Clustering does not have a "correct" or "optimal" result - the quality of the result usually depends on the task at hand.
- With dynamic data records (time-varying data), the assignment of patterns to clusters can change over time.
- ...?

Terms (according to [Höppner, Klawonn, Kruse 1997]):

- Incomplete cluster analysis methods:
  - Representation or projection techniques
  - Multidimensional data are subject to dimension reduction, e.g. by main component analysis
  - Data is displayed graphically, e.g. 2- or 3-dimensionally.
  - Cluster formation takes place manually by viewing the data
- Deterministic cluster analysis methods:
  - Each pattern is assigned to exactly one cluster.
  - Cluster division defines a partition of data

Terms (continued):

- Overlapping cluster analysis methods:
  - Each sample is assigned to at least one cluster
  - Each sample may be assigned to more than one cluster
- Probabilistic cluster analysis methods:
  - For each sample, a probability distribution over the clusters is determined, indicating the probability with which a sample is assigned to a cluster
  - The sum of the assignment probabilities of a sample must be one
- Possibilistic cluster analysis methods:
  - Pure fuzzy clustering algorithms that work with membership grades instead of probabilities
  - Degree of membership of the sample to the cluster in question
  - Each sample may therefore be assigned to several clusters with certain degrees of membership
  - The secondary condition that the sum of the degrees of membership of a sample to the clusters must be one is often dropped

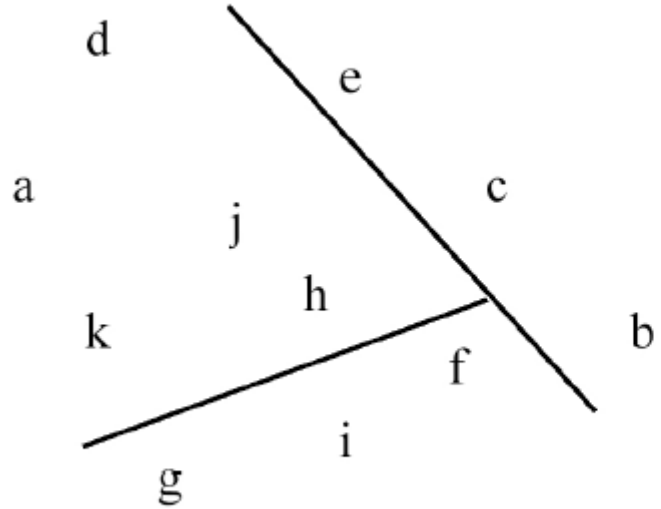
Terms (continued):

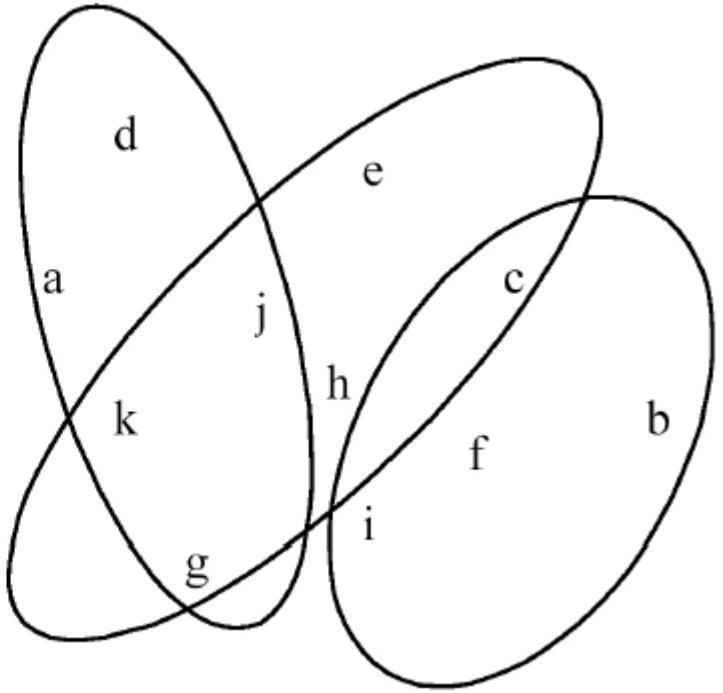
- Hierarchical cluster analysis methods:
  - Procedure bottom-up: Samples are initially assigned to a cluster, clusters are summarised step-by-step.
  - Top-down approach: Initially all samples are combined into one cluster, clusters are broken down step by step.
  - Hierarchical clustering procedures are generally defined in procedural terms, i.e. by rules on when clusters are to be split up or grouped together

Terms (continued):

- Cluster analysis method with objective function:
  - Use an objective function to be optimised that assigns a quality or error value to each possible clustering
  - A set of clusters and the assignments of samples to these is sought that receives the best rating.
  - An optimisation problem must be solved (at least approximately)
- Partitioning cluster analysis techniques:
  - A desired number of clusters is predefined
  - Procedures assign all samples to a cluster in each step and try to gradually improve the quality of clustering by rearranging them.

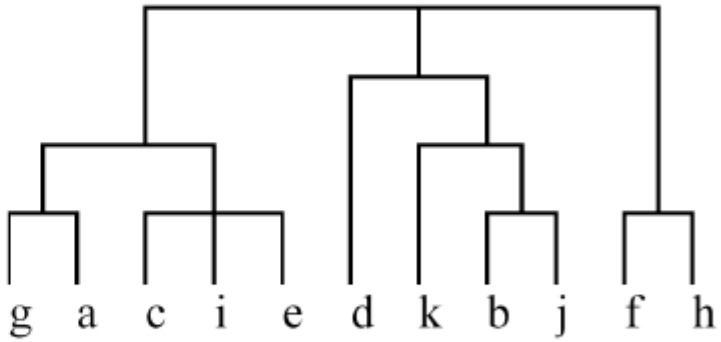
Representation of clusters





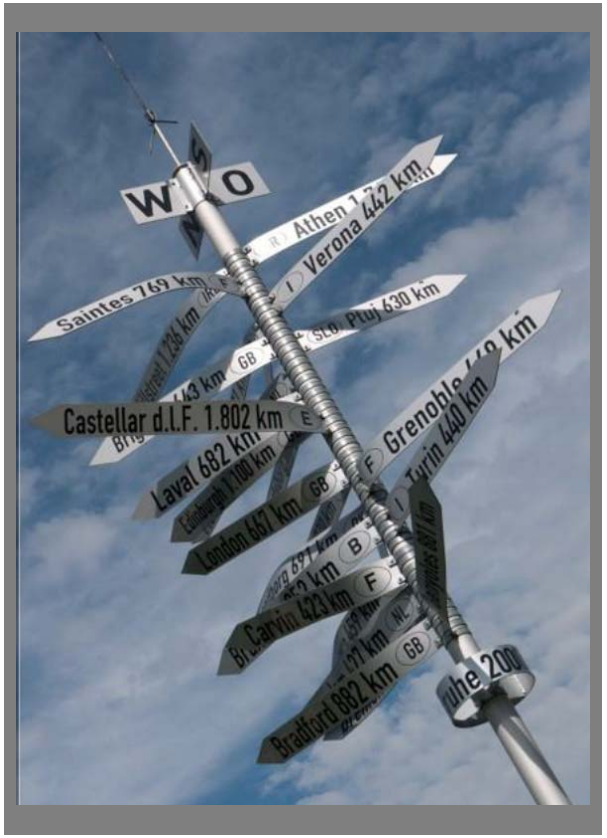
Representation of clusters

	1	2	3
a	0.4	0.1	0.5
b	0.1	0.8	0.1
c	0.3	0.3	0.4
d	0.1	0.1	0.8
e	0.4	0.2	0.4
f	0.1	0.4	0.5
g	0.7	0.2	0.1
h	0.5	0.4	0.1
…			



Agenda

- Basics
- Hierarchical Clustering
- c-Means
- Further techniques
- Cluster evaluation
- Conclusion and further readings



Single-Linkage, Complete-Linkage and Average-Linkage are agglomerative, hierarchical cluster analysis methods.

Hierarchical cluster analysis methods:

- Bottom-up approach: agglomerative
- Top-down approach: dividing, divisive

Agglomerative hierarchical methods:

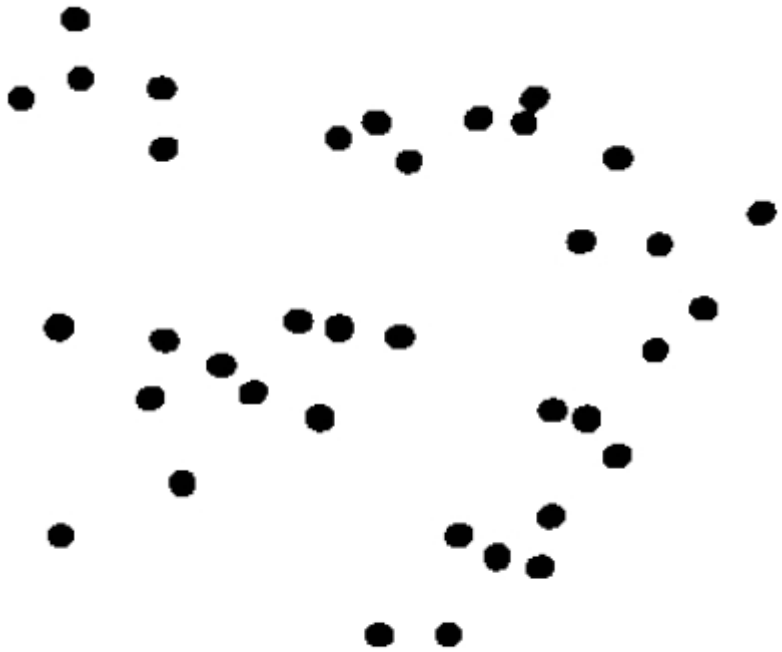
- These methods assign each sample to its own cluster at the beginning.
- Then, in each step of the process, two clusters are searched for a specific optimality criterion, which is then merged to form a cluster.
- This procedure can be continued until the desired number of clusters is reached.

Agglomerative hierarchical methods:

- A major advantage and at the same time a disadvantage of these hierarchical cluster methods is that the large solution space is very quickly narrowed down, but the merger decisions made at an early stage can no longer be revised.
- These algorithms therefore sometimes tend to find poor local optima.

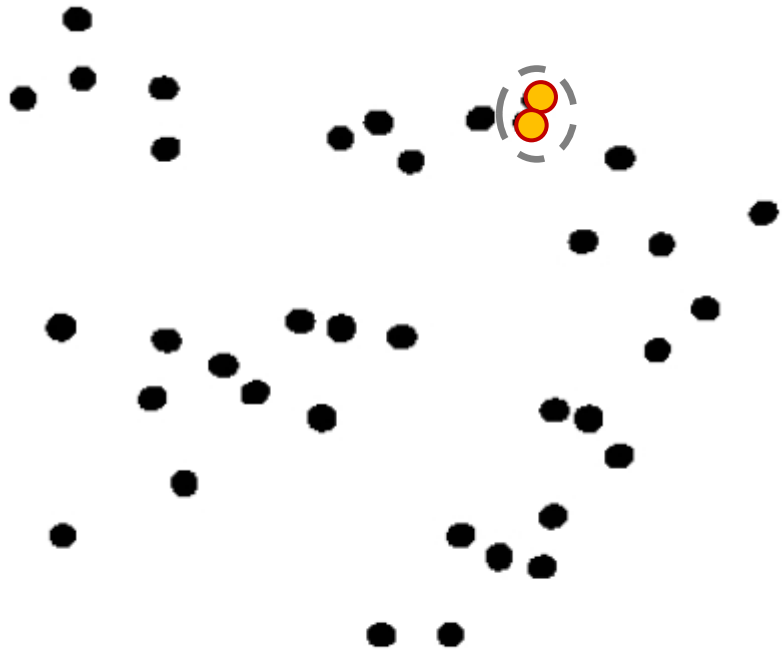
Approach:

1. Assign each point to its own cluster



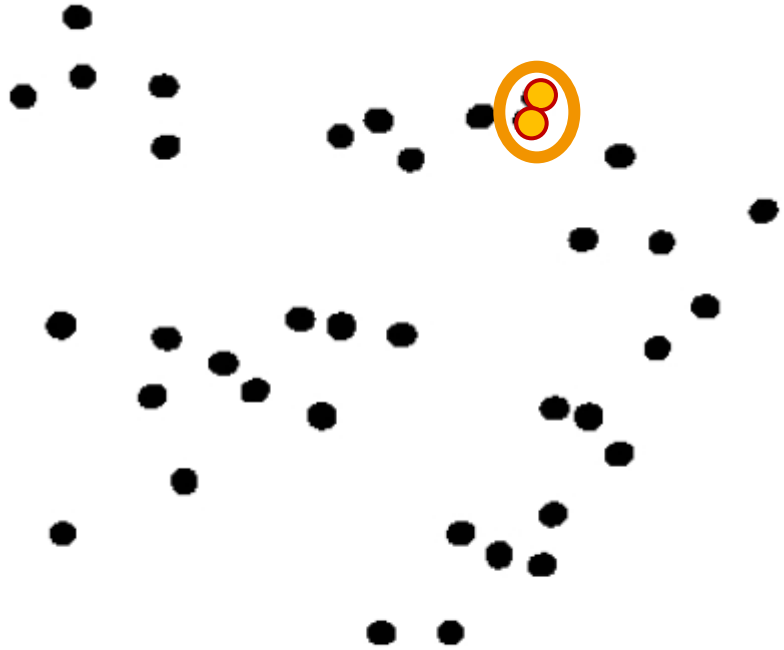
Approach:

1. Assign each point to its own cluster
2. Identify the most similar pair of clusters



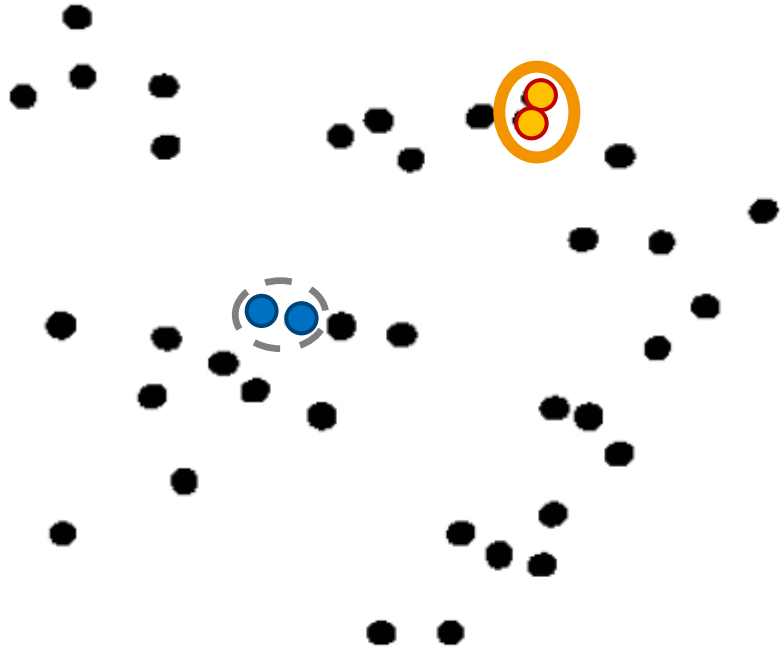
Approach:

1. Assign each point to its own cluster
2. Identify the most similar pair of clusters
3. Create new cluster by merging the two identified ones



Approach:

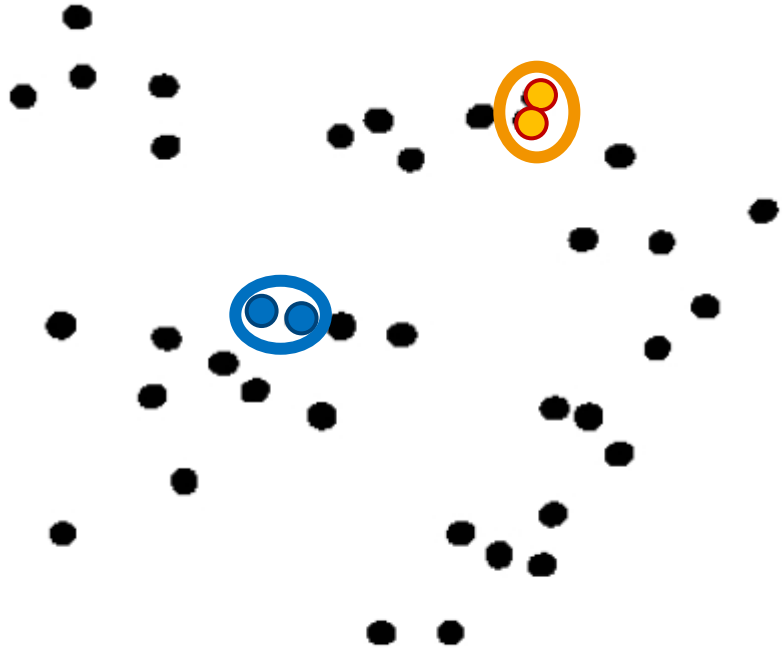
1. Assign each point to its own cluster
2. Identify the most similar pair of clusters
3. Create new cluster by merging the two identified ones
4. Repeat this step



Approach:

1. Assign each point to its own cluster
2. Identify the most similar pair of clusters
3. Create new cluster by merging the two identified ones
4. Repeat this step

… until all samples are part of one large cluster …



What does “similarity of clusters“ mean?

- Let C_i and C_j with $i \neq j$ and $C_i \cap C_j = \emptyset$ be two clusters of samples (data points) $x_n \in \mathbb{R}^D$ with $n = 1, \dots, N$ and $D \in \mathbb{N}$.
- The following criteria are defined to assess the similarity of the two clusters:
  - A) $$\delta_{\min}(\mathcal{C}_i, \mathcal{C}_j) := \min_{\mathbf{x}_n \in \mathcal{C}_i, \mathbf{x}_l \in \mathcal{C}_j} \{\|\mathbf{x}_k - \mathbf{x}_l\|\}$$
  - B) $$\delta_{\max}(\mathcal{C}_i, \mathcal{C}_j) := \max_{\mathbf{x}_k \in \mathcal{C}_i, \mathbf{x}_l \in \mathcal{C}_j} \{\|\mathbf{x}_k - \mathbf{x}_l\|\}$$

– C) $$\delta_{\text{avg}}\left(\mathcal{C}_i, \mathcal{C}_j\right) := \frac{1}{|\mathcal{C}_i| \cdot |\mathcal{C}_j|} \cdot \sum_{\mathbf{x}_k \in \mathcal{C}_i,\, \mathbf{x}_l \in \mathcal{C}_j} \|\mathbf{x}_k - \mathbf{x}_l\|$$

- The criteria δ_{min}, δ_{max}, and δ_{avg} are called ‘nearest neighbour‘, ‘farthest neighbour‘, and ‘average neighbour‘.

Single Linkage

- Given is a set of samples $x_n \in \mathbb{R}^D$ with $n = 1, \ldots, N$ and $D \in \mathbb{N}$ as well as a desired number c of clusters to be detected, where $N \geq c$ applies.

Process:

1. In a first step, a partition of the data consisting of N clusters is formed by assigning a sample to each cluster:

$$C_n := \{x_n\}$$

Single Linkage – Process:

2. These clusters are now gradually merged:
   - Determine the two clusters that are most similar according to the closest neighbour criterion:
     (C_k, C_l) with the property

$$\left(\forall_{(C_{k'}, C_{l'})} : \delta_{\min}(C_k, C_l) \leq \delta_{\min}(C_{k'}, C_{l'})\right)$$

   - Combine the two clusters C_k and C_l to one cluster, so that a new partition of the data is established.
   - If the number of all clusters of the new partition is greater than the desired number c go back to Step 1.

Single Linkage: Evaluation

- Advantage of the process:
  - Outliers are recognised, because they are inserted only very late or not at all into a cluster.
- Disadvantage of the method:
  - Only two close samples at a time deciding on the fusion of two clusters.
  - This can result in concatenations or chain-like structures (chaining effect) in which some samples are less similar to each other than to samples of other clusters.

Complete Linkage:

- Process is the same as for single-linkage
- Only difference: Step 2 a) is replaced by:
  - Determine the two clusters that are most similar according to the closest neighbour criterion:
    (C_k, C_l) with the property

$$\left(\forall_{(\mathcal{C}_{k'},\mathcal{C}_{l'})} : \delta_{\max}\left(\mathcal{C}_k, \mathcal{C}_l\right) \leq \delta_{\max}\left(\mathcal{C}_{k'}, \mathcal{C}_{l'}\right)\right)$$

Complete Linkage: Evaluation

- Advantage:
  - Unlike single linkage, chaining cannot occur.
  - This ensures more homogeneous clusters.
- Disadvantage:
  - Outliers are integrated earlier and are no longer so easily detected.

Average Linkage:

- Process is the same as for single/complete linkage
- Only difference: Step 2 a) is replaced by:
  - Determine the two clusters that are most similar according to the average neighbour criterion:
    (C_k, C_l) with the property

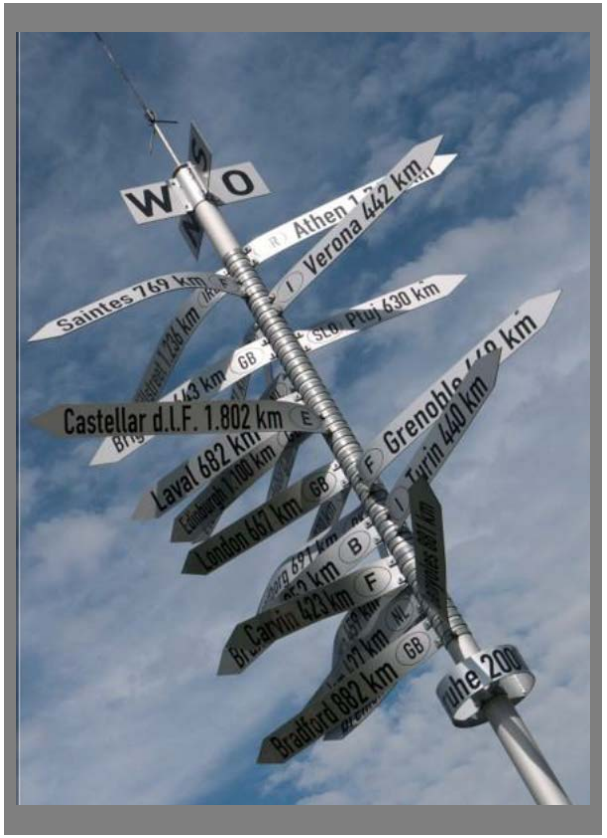
$$\left(\forall_{(\mathcal{C}_{k'},\mathcal{C}_{l'})} : \delta_{\text{avg}}\left(\mathcal{C}_k, \mathcal{C}_l\right) \leq \delta_{\text{avg}}\left(\mathcal{C}_{k'}, \mathcal{C}_{l'}\right)\right)$$

Average Linkage: Evaluation

- In many cases this procedure leads to good results.
- Average linkage is therefore a good compromise.

Agenda

- Basics
- Hierarchical Clustering
- c-Means
- Further techniques
- Cluster evaluation
- Conclusion and further readings



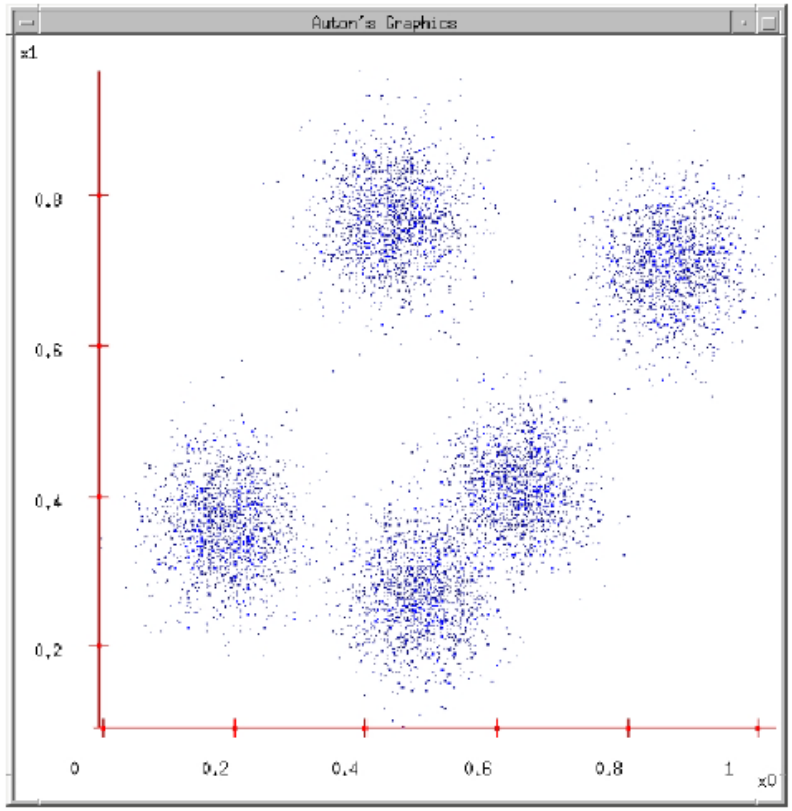
c-Means algorithm

- Also: k-Means or simply clustering
- Was introduced by J. B. MacQueen in 1967.
- Is one of the most common data mining algorithms

Which methods/algorithms did you use for data analysis in 2011? [311 voters]	
Decision Trees/Rules (186)	59.8 %
Regression (180)	57.9 %
Clustering (163)	52.4 %
Statistics (descriptive) (149)	47.9 %
Visualization (119)	38.3 %
Time series/Sequence analysis (92)	29.6 %
Support Vector (SVM) (89)	28.6 %
Association rules (89)	28.6 %
Ensemble methods (88)	28.3 %
Text Mining (86)	27.7 %
Neural Nets (84)	27.0 %
Boosting (73)	23.5 %
Bayesian (68)	21.9 %
Bagging (63)	20.3 %
Factor Analysis (58)	18.7 %
Anomaly/Deviation detection (51)	16.4 %
Social Network Analysis (44)	14.2 %
Survival Analysis (29)	9.32 %
Genetic algorithms (29)	9.32 %
Uplift modeling (15)	4.82 %

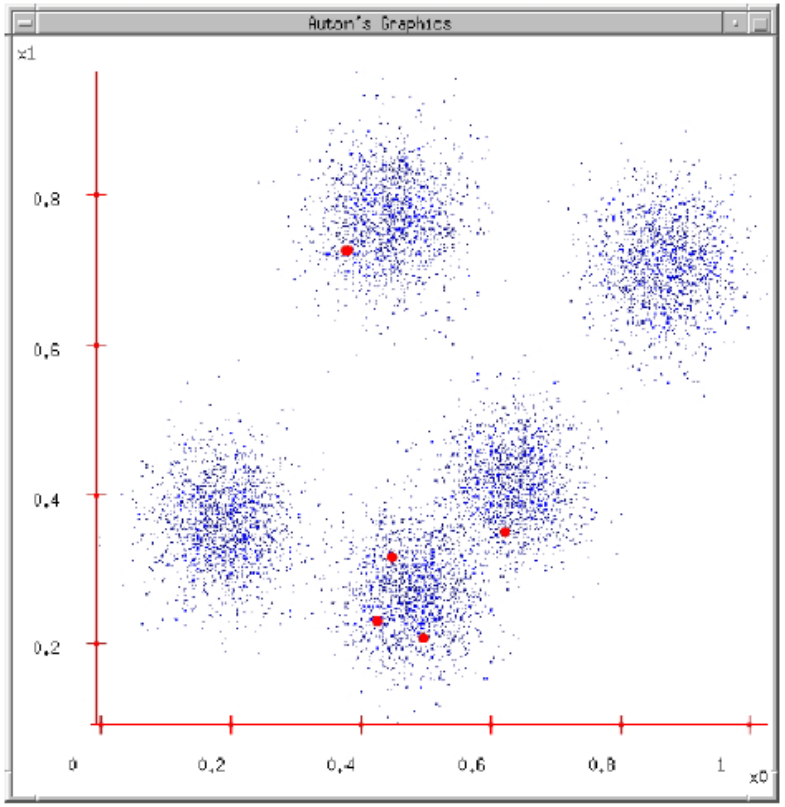
c-Means

- Given is data, ask the user how many clusters to form (here: c = 5).



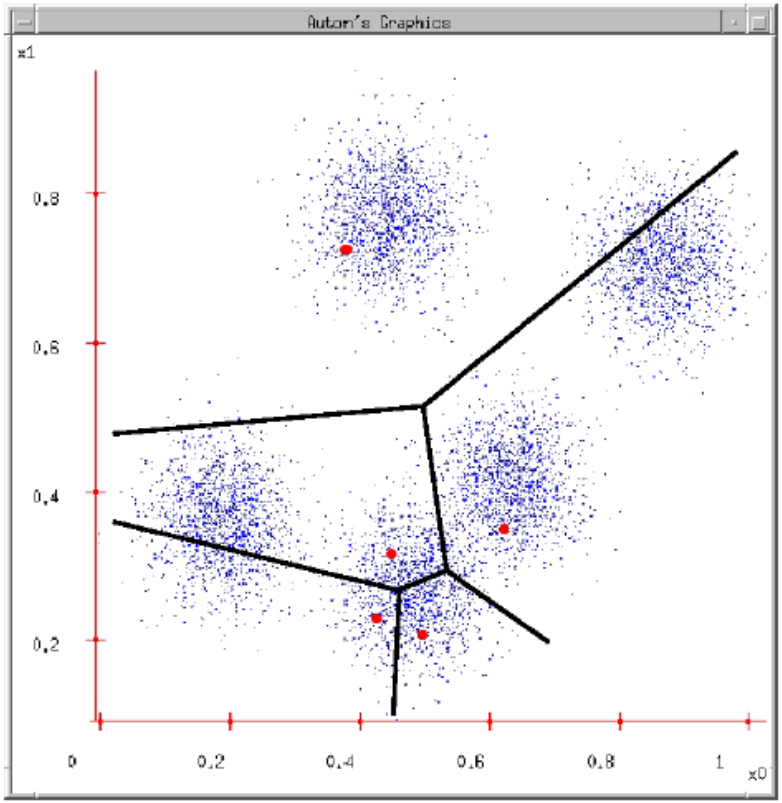
c-Means

- Given is data, ask the user how many clusters to form (here: $c = 5$).
- Choose c initial cluster centres.



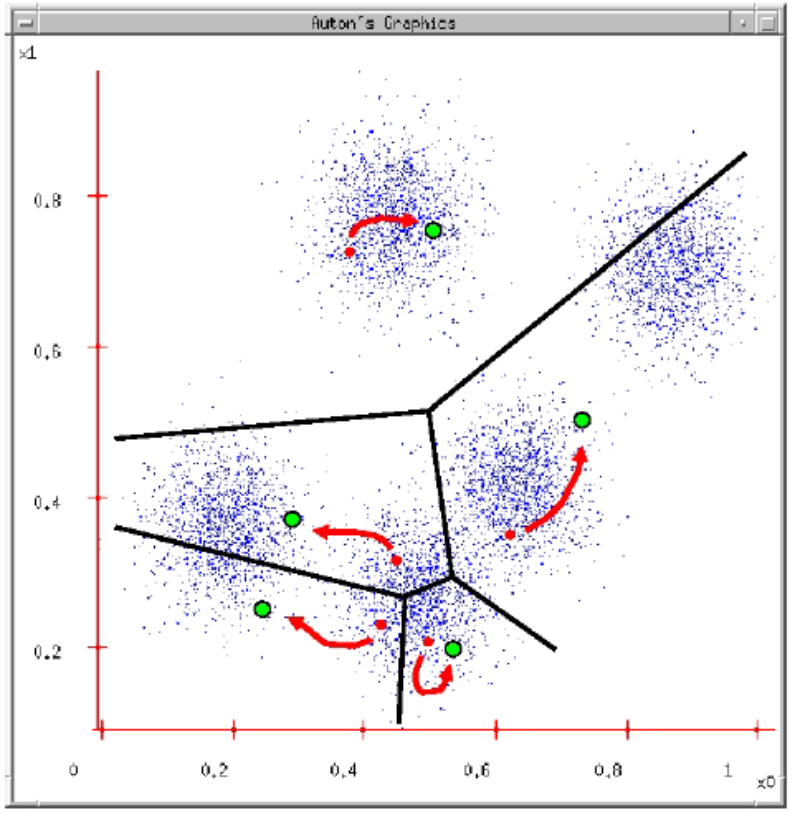
c-Means

- Given is data, ask the user how many clusters to form (here: $c = 5$).
- Choose c initial cluster centres.
- Form clusters by determining the nearest cluster centre for each sample.



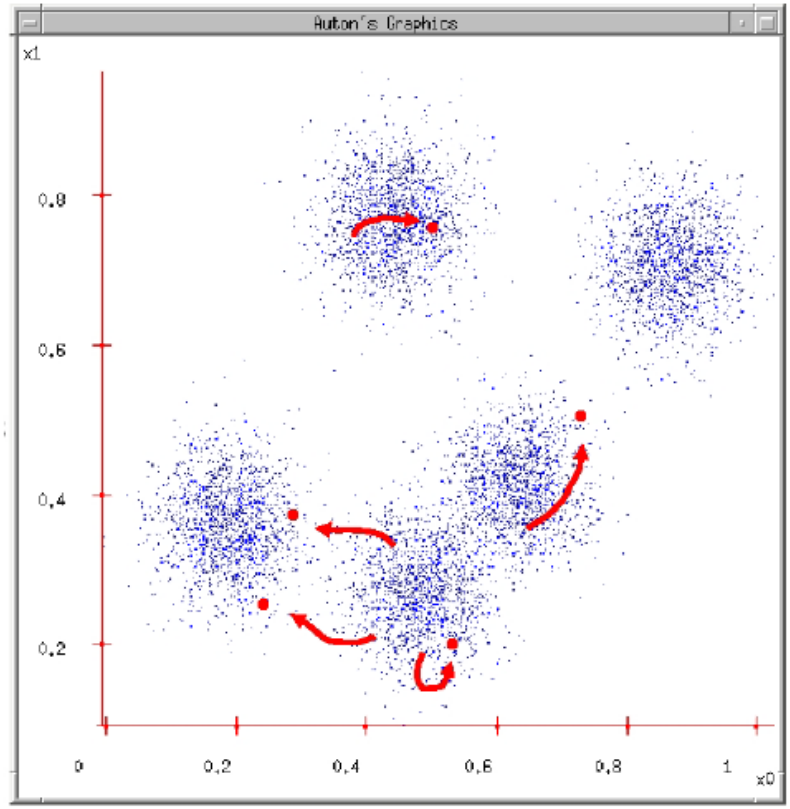
c-Means

- Given is data, ask the user how many clusters to form (here: $c = 5$).
- Choose c initial cluster centres.
- Form clusters by determining the nearest cluster centre for each sample.
- For each cluster, calculate the mean value of the assigned samples.



c-Means

- Given is data, ask the user how many clusters to form (here: $c = 5$).
- Choose c initial cluster centres.
- Form clusters by determining the nearest cluster centre for each sample.
- For each cluster, calculate the mean value of the assigned samples.
- Repeat assignment and centre determination until a stopping criterion is reached.



c-Means

- Advantages:
  - Simple, easy to understand
  - All samples are automatically assigned to clusters
  - Process terminates
- Disadvantages:
  - Convergence in local minimum possible
  - Number of clusters to be determined manually
  - All samples must be assigned to clusters.
  - Sensitive to outliers
  - Sensitive to the choice of initial cluster centres

Questions about c-Means:

- Which objective function is optimised here?
- Does the process terminate?
- Is there an optimal clustering?
- How should the process start?
- How can the number of centres be determined automatically?

Let‘s see how to answer these questions ...!

c-Means

- Given: A set of samples $x_n \in \mathbb{R}^D$ with $n = 1, \ldots, N$ and $D \in \mathbb{N}$ as well as a desired number c of clusters to be detected, where $N \geq c$ applies.
- The clusters are designated C_i, their centres c_i ($c_i \in \mathbb{R}^D$ with $i = 1, \ldots, c$).
- The clusters are initialised with $c_i = \emptyset$.
- A target function value ε is initialised as $\varepsilon := \infty$.

Process:

1. In a first step, the cluster centres are determined by assigning a randomly selected sample to each centre:

   $$c_i := x_r$$

   Each sample is selected at most once.

c-Means process

2. Each of the N samples is assigned to that cluster, where the distance to the centre head is minimal:

   $C_i \coloneqq C_i \cup x_n$ - exactly in case of:

$$\forall_{c_j}: \|x_n - c_i\| \leq \|x_n - c_j\|$$

3. Afterwards, all c cluster centres are updated (re-calculated):

$$c_i \coloneqq \frac{1}{|C_i|} \sum_{x_n \in C_i} x_n$$

c-Means process

4. If the result of the target value function has been decreased:

$$\varepsilon = \sum_{i=1,\ldots,c} \sum_{x_n \in C_i} \|x_n - c_i\|$$

Go to Step 2. Again: $c_i = \emptyset$.
Otherwise: Stop.

Remark:

Other distance measures (e.g. Manhattan, ...) are possible, too!

c-Means properties:

- Cluster centres are only moved if the value of the target function decreases as a result.
  → Process is descending concerning the target function value.
- Number of possibilities to assign patterns to clusters is finite.
  → Process terminates.
- Process is "greedy" in the sense that in each run the currently best assignment of samples to centres is selected.
  → Process can converge in a local minimum.

Number c of clusters:

- Start with a low value of c and repeat the process with increasing c.
- In doing so, study the improvements in the target function value.
- Decision for c, e.g., with knee/elbow approach/rule.

Choice of starting points:

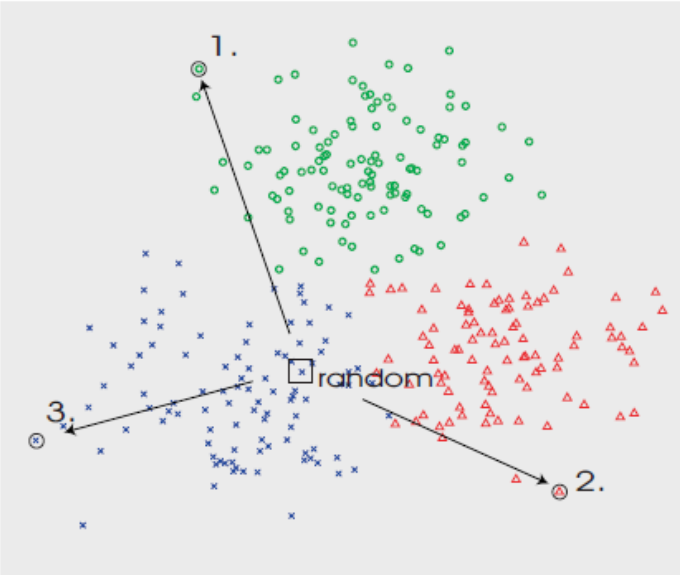
- Repeat the process with different centre initialisations.
- “Skilful“ selection of the initial centres.

The heuristic to select the initial centres:

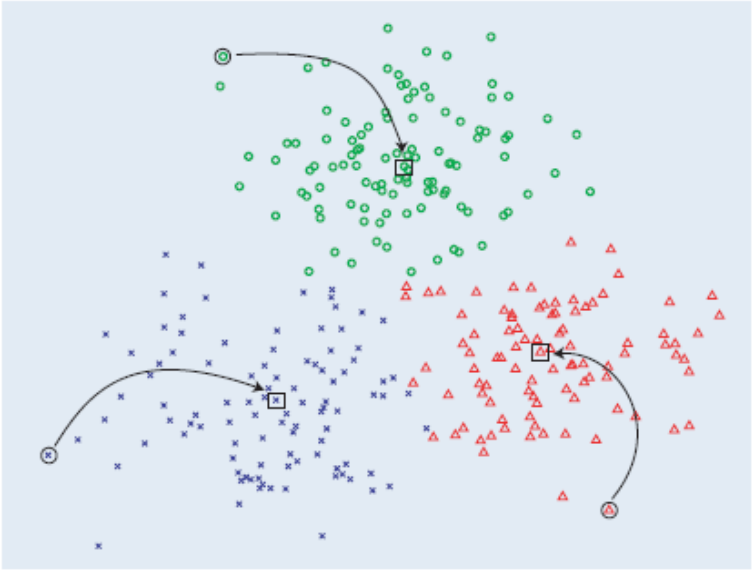
- Select a sample randomly
- As the first initial cluster centre, select a sample that is far away from the first selected sample.
- ...
- Select as i-th initial cluster centre a sample that is far from the previously selected samples
- ...
- ... until all c initial cluster centres are determined.

The heuristic for choosing cluster centres

Choosing initial cluster centres



Clustering result



Remarks:

- Application of c-means to recognise structures in data with one-dimensional data is a promising method for quantisation (buckets)
- Related processes:
  - Vector quantisation
  - Learning Vector Quantisation (LVQ)
  - Competition learning / Online-c-means
  - Self-Organising Maps (SOM)

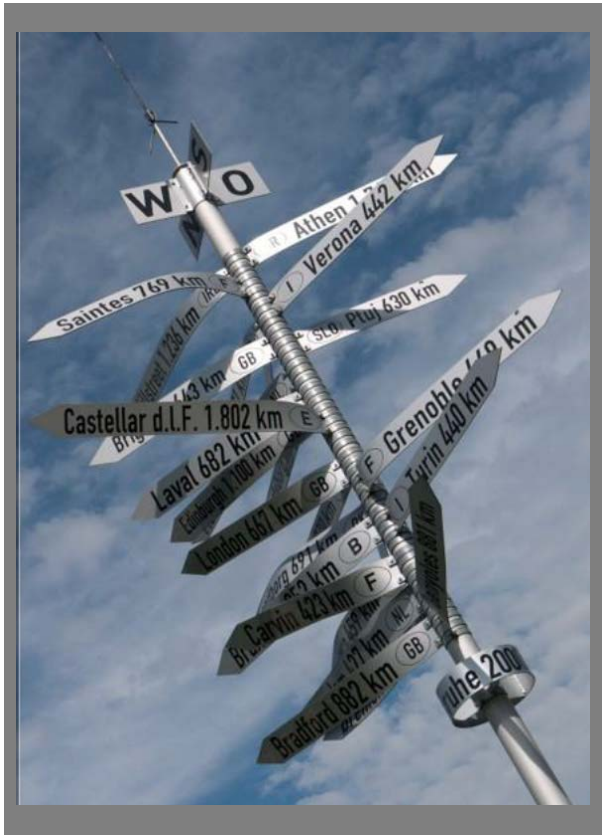
c-Medoids

- Problem: Result of c-means can be significantly affected by outliers.
- Idea: use medians instead of averages
- Example:
  - Average value of 1, 3, 5, 7, 9 is 5
  - Average value of 1, 3, 5, 7, 1009 is 205
  - Median of 1, 3, 5, 7, 1009 is 5

- Other variants (e.g. Fuzzy c-Means) known ...

Agenda

- Basics
- Hierarchical Clustering
- c-Means
- Further techniques
- Cluster evaluation
- Conclusion and further readings



Nearest Neighbour Clustering

- Cf. hierarchical, agglomerative clustering
- The process starts with a randomly selected sample assigned to a newly created cluster.
- For each sample a different sample is searched in an existing cluster so that the distance to this sample is minimal.
- If this distance is smaller than a specified threshold value, the corresponding clusters are combined, otherwise a new cluster is formed.

Clustering with Gaussian Mixture Models:

- It is assumed that clusters are generated by Gaussian random processes, i.e. they can also be modelled by a combination of several Gaussian distributions.
- Instead of using degrees of affiliation (like in fuzzy logic) one works with probabilities
- Parameters of the Gauss functions are determined again in an iterative procedure (EM: Expectation Maximisation) (see c-means).

We will discuss GMM in detail later on in the chapters …

DBSCAN

Everything is related to everything else,

but near things are more related than distant things.

[First geographical law, Waldo Tobler, 1970]

DBSCAN:

- Density-Based Spatial Clustering of Applications with Noise
- Density-based spatial cluster analysis with noise
- Basic idea similar to Nearest Neighbour Clustering, i.e. neighbouring samples are combined to form a cluster.
- Also however, the number of adjacent samples is taken into account and individual samples at the edge are explicitly recognised as noise (or outliers).

We will discuss this technique in detail in the chapter about anomaly / novelty detection…!

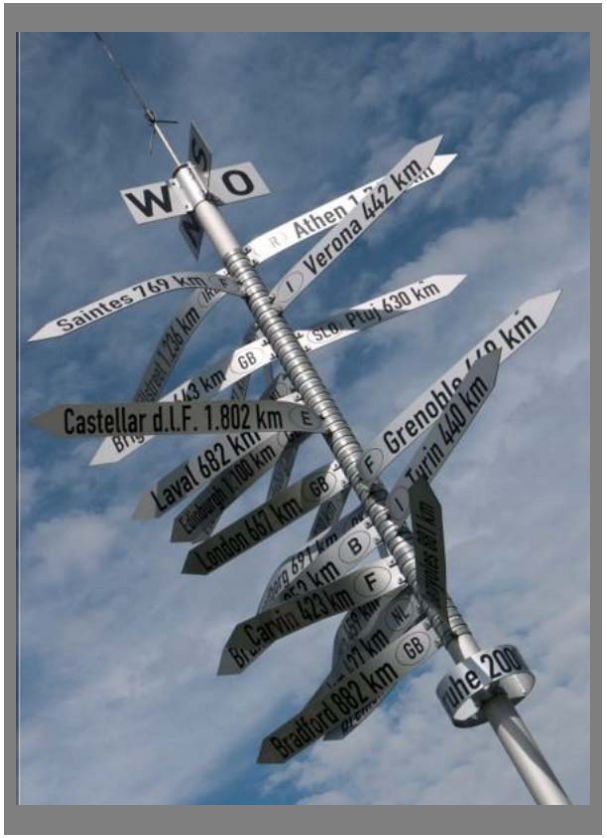
OPTICS

- Ordering Points To Identify the Clustering Structure
- Based on DBSCAN
- Also density-based algorithm for cluster analysis, further development of DBSCAN
- Fixes weakness of DBSCAN: can detect clusters of different density
- Eliminates to a large extent the ε parameter (neighbourhood length) of DBSCAN
- Idea: Arranges the points of the data set linearly, so that spatially adjacent points in this order follow each other closely, and determines the so-called "reachability distance".

Agenda

Christian-Albrechts-Universität zu Kiel

- Basics
- Hierarchical Clustering
- c-Means
- Further techniques
- Cluster evaluation
- Conclusion and further readings

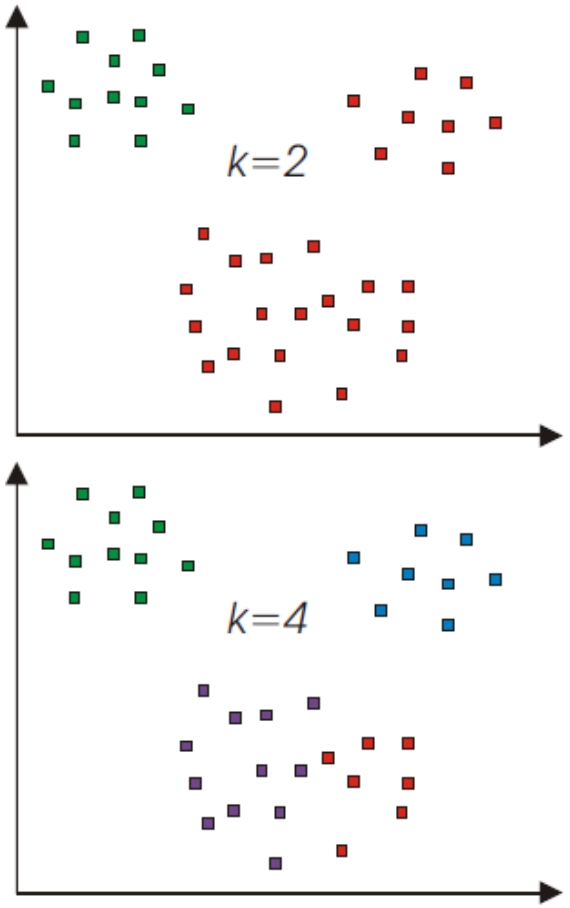


Evaluation of clustering methods:

- Usually: Evaluation by experts or with regard to a target function with test data
- Sensible evaluation is often only possible in the application
- Some algorithms already work with target functions (e.g. c-means), others do not (e.g. single linkage).
- Examples of criteria:
  - Dunn's Index,
  - Davies Bouldin Index,
  - SD Index,
  - Minimal Description Length,
  - Category utility,
  - ...

Evaluation criteria for clusters

- Evaluation of the clustering results: Which partitioning "fits" best with the data?
- Quantitative evaluation of the clustering results
- Search for the optimal set of parameters (depending on clustering method):
  - number of clusters (c-means, c-medoids, fuzzy-c-means, ...)
  - Radii (BIRCH, DBSCAN, DENCLUE,...)
  - Maximum number of neighbours (CLARANS, ...)



Cluster evaluation can be subdivision into:

- External criteria
  - Comparison of the cluster structure found with a predefined structure
  - Assumptions about underlying data are necessary
- Internal criteria
  - Only the clustering itself and the data are taken into account.
- Relative criteria (commonly used)
  - Evaluation of several (successive) cluster structures
  - Use of the same algorithm, but with different parameter sets

Relative criteria

- Basic idea
  - For a clustering algorithm, P is the set of parameters.
  - Among the generated cluster structures, search for the best configuration of P about the selected evaluation criterion f.

- Distinguish between two cases:
  1. P does not contain the number of clusters
     - Start the clustering algorithm for a wide range of parameters.
     - Select the largest range for which the number of k clusters remains constant.
     - Select the mean value from this range.

Relative criteria

- Distinguish between two cases:
  2. P does contain the number of clusters
     - Start the clustering algorithm for all c at a user-specified interval $[c_{min}, c_{max}]$.
     - For each c, vary the remaining parameters.
     - Define the function $f(c)$ that contains the optimal value for each c.
     - Determine the optimum based on the function $f(c)$, e.g., by visualising the function $f(c)$.
- If the evaluation criterion has no rising or falling trend
  → Search for the maximum or the minimum!
- In the presence of a trend:
  - Search for a significant change in the curve of $f(c)$: strong local increase or decrease, so-called 'knee'.
  - No knee can be an indicator that there is no cluster structure!

Examples:

- Dunn Index
- Davies Bouldin Index

Dunn Index

- Presented by Dunn, 1974
- Search for compact and well-separated clusters
- Very popular, widely used
- Basic idea: For compact and well-separated clusters, the distance between the clusters should be large (high inter-cluster distance) and the diameter of the respective cluster should be small (small intra-cluster distance).

Dunn Index

- Dunn Index D_c is defined for c cluster:

$$D_c := \min_{i=1,\ldots,c} \left\{ \min_{j=i+1,\ldots,c} \left\{ \frac{\delta(\mathcal{C}_i, \mathcal{C}_j)}{\max_{m=1,\ldots,c} \operatorname{diam}(\mathcal{C}_m)} \right\} \right\}$$

- It makes use of the dissimilarity measure:

$$\delta(\mathcal{C}_i, \mathcal{C}_j) := \min_{\mathbf{x} \in \mathcal{C}_i, \mathbf{y} \in \mathcal{C}_j} d(\mathbf{x}, \mathbf{y})$$

- $d(x, y)$ is a distance measure which determines the distance (i.e. the dissimilarity= of two samples x and y.
- $diam(C_m)$ is a function that determines the diameter of the cluster C_m
- It can be defined as:

$$\operatorname{diam}(\mathcal{C}_m) := \max_{\mathbf{x}, \mathbf{y} \in \mathcal{C}_m} d(\mathbf{x}, \mathbf{y})$$

Dunn Index

- Analysis and interpretation:
  - Large values for Dunn Index indicate compact and well-separated clusters.
  - The c, for which D_c is the largest, is assumed to be the best cluster structure

- Disadvantages
  - High computational effort (quadratic in the number of clusters)
  - High susceptibility to outliers, as outliers, can easily influence $\max_{m=1,\ldots,c} diam(C_m)$.
  - Improvements are possible, e.g., the reduction of runtime and lower sensitivity to outliers.

DB Index

- Davies & Bouldin, 1979
- Basic idea as with Dunn Index: Individual clusters should be as compact as possible with the greatest possible distance between the clusters at the same time.
- Very popular and widely used criterion.
- Approach: Consider scattering within clusters with the distance between clusters.

Davies Bouldin Index DB_c for c cluster

$$DB_c := \frac{1}{c}\sum_{i=1}^{c} R_i$$

with

$$R_i := \max_{j=1,\ldots,c \text{ and } i \neq j} R_{ij}$$

for

$$i = 1, \ldots, c$$

Remark:

- For R_{ij} must hold: The smaller the value, the better the separation of clusters.

DB Index

- Let d_{ij} be a measure of dissimilarity (distance measure) for two clusters C_i and C_j.
- E.g. measured by the distance of the cluster centres (mean values) c_i and c_j:

$$d_{ij} := \|c_i - c_j\|$$

- Let s_i be the scattering (or spread) of the cluster C_i, e.g., the mean distance of all samples (data points) of a cluster to the corresponding cluster centre.

DB Index

- R_{ij} must be a similarity measure for two clusters that meets the following requirements:
  1. $R_{ij} \geq 0$
  2. $R_{ij} = R_{ji}$
  3. If $s_i = 0$ and $s_j = 0$ then $R_{ij} = 0$
  4. If $s_j > s_k$ and $d_{ij} = d_{ik}$ then $R_{ij} > R_{ik}$
  5. If $s_j = s_k$ and $d_{ij} < d_{ik}$ then $R_{ij} > R_{ik}$
- R_{ij} is therefore especially non-negative and symmetrical.

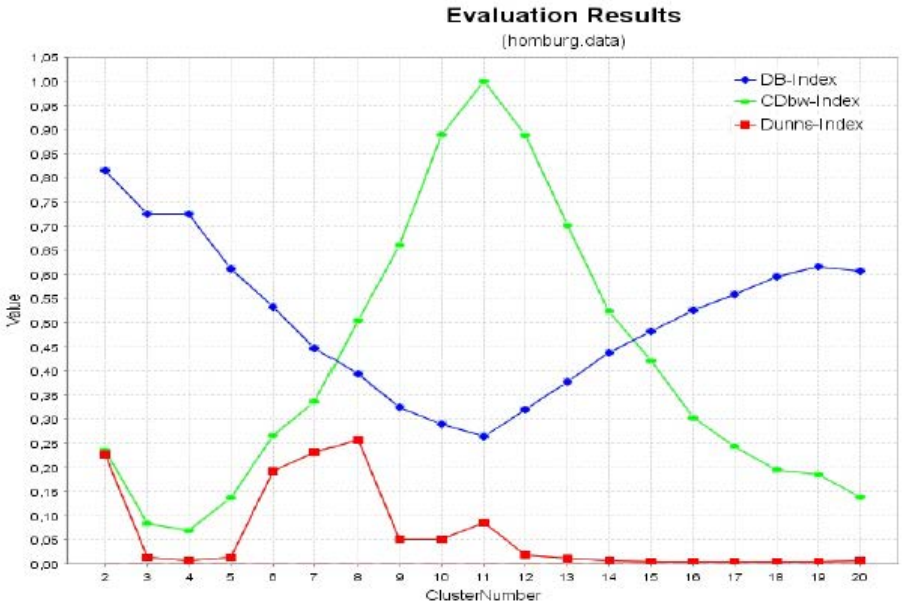
DB Index

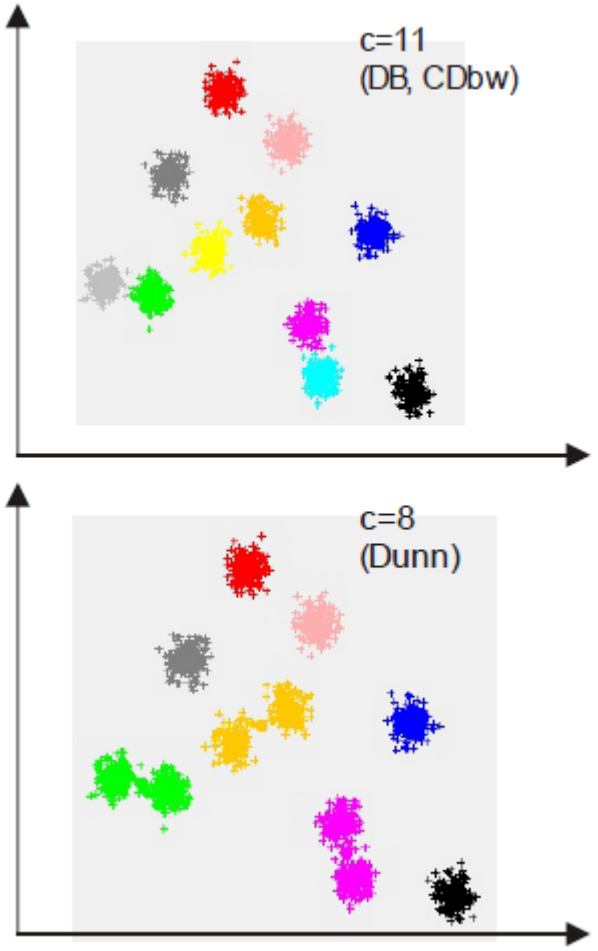
- Example of a simple measure that meets the conditions:

$$R_{ij} := \frac{s_i + s_j}{d_{ij}}$$

- Analysis and interpretation
- DB_c measures the average similarity between a cluster and ist most similar cluster.
- The goal is to find clusters that are as dissimilar as possible, i.e. search for the minimum DB_c.
- This comes with high computational effort.
- For further alternatives considering the scattering within clusters and the spacing of clusters, see [Davies & Bouldin, 1979]

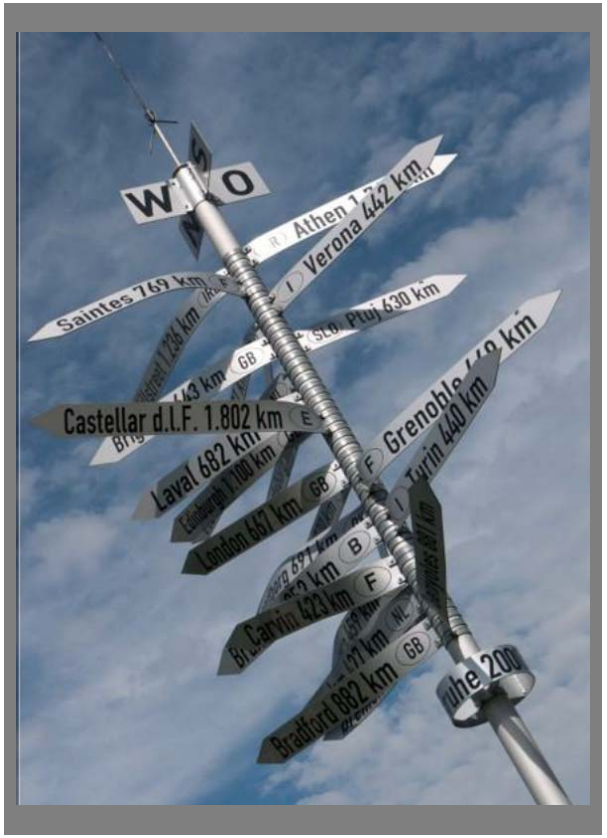
Example: 11 clusters in 2 dimensions





Agenda

- Basics
- Hierarchical Clustering
- c-Means
- Further techniques
- Cluster evaluation
- Conclusion and further readings



Summary of the chapter

- Basic idea of clustering: Finding natural groups in data
- First basic approach: Hierarchical clustering
- Probably the most popular approach: c-Means (and c-Medoids)
- Other different clustering methods
- Evaluation of clustering results using appropriate metrics: In particular Dunn Index and Davies-Bouldin Index

Further readings

- [EKSX96] Ester, Kriegel, Sander, Xu, A density-based algorithm for discovering clusters in large spatial databases with noise, 1996
- [HKK97] Höppner, Frank, Frank Klawonn, and Rudolf Kruse. Fuzzy-Clusteranalyse: Verfahren für die Bilderkennung, Klassifizierung und Datenanalyse. Vieweg+ Teubner Verlag, 1997
- [MQ67] MacQueen, James. "Some methods for classification and analysis of multivariate observations." *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*. Vol. 1. No. 14. 1967
- [Dunn74] Dunn, Joseph C. "Well-separated clusters and optimal fuzzy partitions." *Journal of cybernetics* 4.1 (1974): 95-104.
- [DB79] Davies, David L., and Donald W. Bouldin. "A cluster separation measure." *IEEE transactions on pattern analysis and machine intelligence* 2 (1979): 224-227.

- Any questions…?