C | A | U Christian-Albrechts-Universität zu Kiel Technische Fakultät

Inf-KDDM: Knowledge Discovery and Data Mining

Winter Term 2020/21

Lecture 6: Clustering

Lectures: Prof. Dr. Matthias Renz

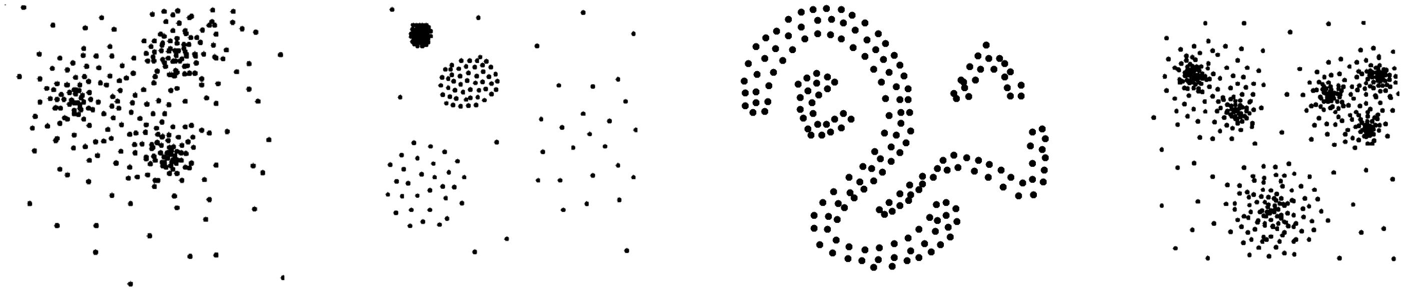
Exercises: Steffen Strohm

Outline

- Unsupervised learning vs supervised learning
- A categorization of major clustering methods
- Partitioning-based clustering
- Hierarchical-based clustering

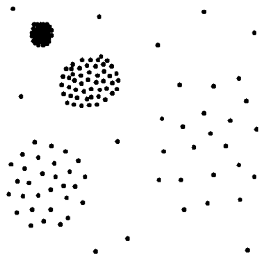
What is cluster analysis?

- Cluster: a collection of data objects
  - Similar to one another within the same cluster
  - Dissimilar to the objects in other clusters
- Cluster analysis
  - Finding similarities between data according to the characteristics found in the data and grouping similar data objects into clusters



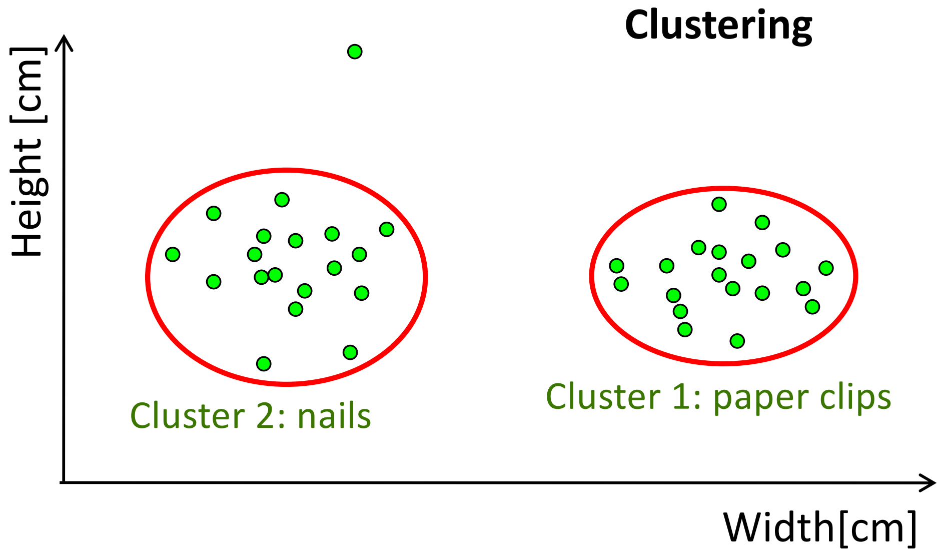
An unsupervised learning task

- Clustering is an unsupervised learning task
  - Given a set of measurements, observations, etc., the goal is to group the data into groups of similar data (clusters)
  - We are given a dataset as input which we want to cluster but there are no class labels
  - We don't know how many clusters exist in the data
  - We don't know the characteristics of the individual clusters



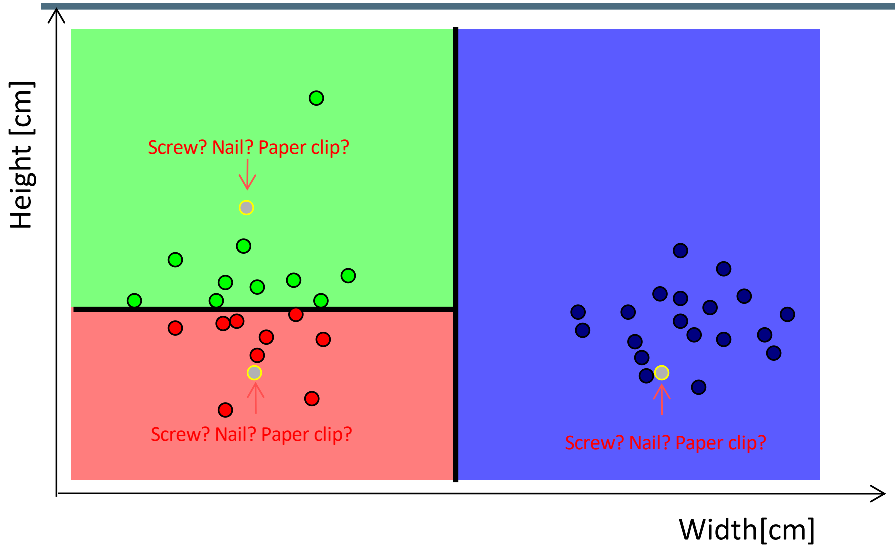
- In contrast to classification, which is a supervised learning task
  - Supervision: The training data (observations, measurements, etc.) are accompanied by *labels* indicating the *class* of the observations
  - New data is classified based on the training set

Unsupervised learning example



Question:
Is there any structure in data (based on their characteristics, i.e., width, height)?

Supervised learning example



Classification model

● Screw

● Nails

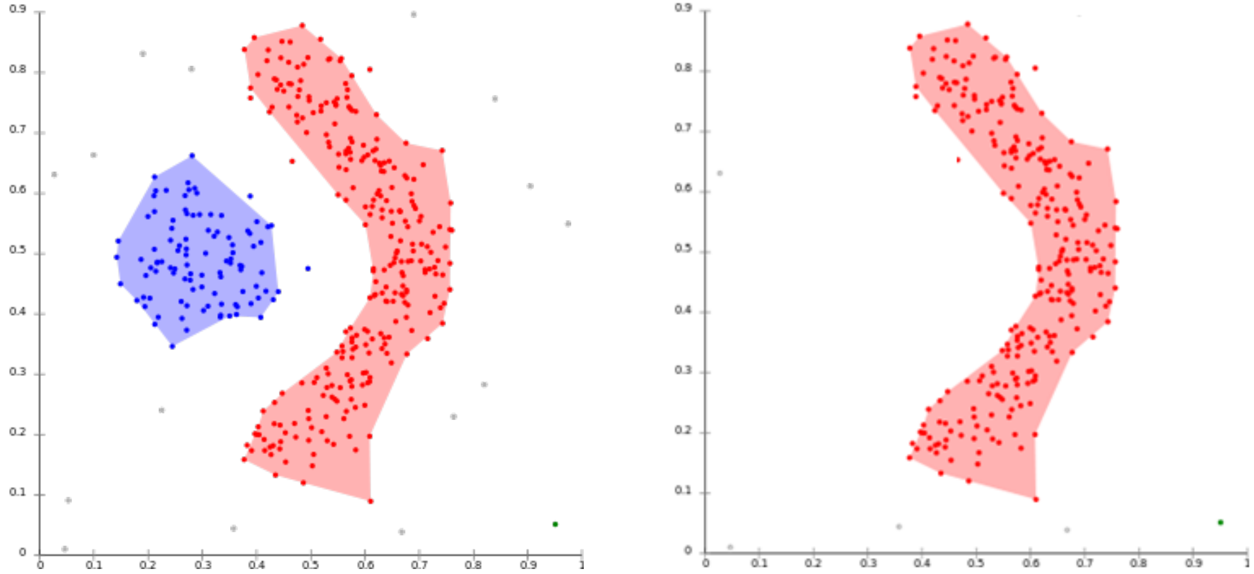
● Paper clips

● New object (unknown class)

Question:
What is the class of a new object???
Screw, nail or paper clip?

Why clustering?

- Clustering is widely used as:
  - As a stand-alone tool to get insight into data distribution
  - As a preprocessing step for other algorithms



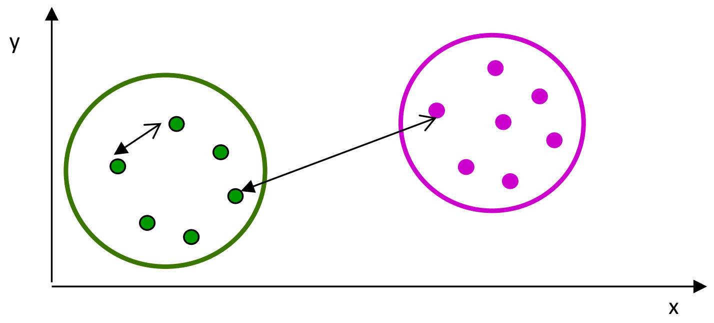
http://en.wikipedia.org/wiki/Cluster_analysis

Example applications

- Marketing:
  - Help marketers discover distinct groups in their customer bases, and then use this knowledge to develop targeted marketing programs
- Telecommunications:
  - Build user profiles based on usage and demographics and define profile specific tariffs and offers
- Land use:
  - Identification of areas of similar land use in an earth observation database
- City-planning:
  - Identifying groups of houses according to their house type, value, and geographical location
- Bioinformatics:
  - Cluster similar proteins together (similarity wrt chemical structure and/or functionality etc)
- Web:
  - Cluster users based on their browsing behavior
  - Cluster pages based on their content (e.g. News aggregators)

The clustering task

- **Goal:** Group objects into groups so that the objects belonging in the same group are similar (high intra-cluster similarity), whereas objects in different groups are different (low inter-cluster similarity)
- A good clustering method will produce high quality clusters with
  - high intra-cluster similarity
  - low inter-cluster similarity



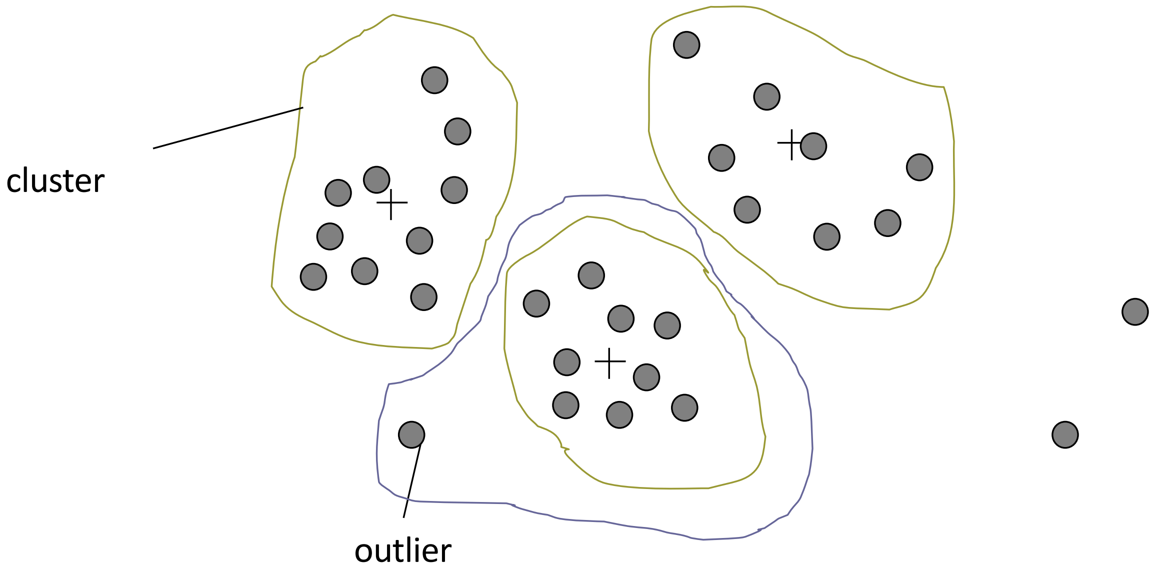
- The quality of a clustering method is also measured by its ability to discover some or all of the hidden patterns
- The quality of a clustering result depends on both the similarity measure used by the method and its implementation

Requirements for clustering

- Discovery of clusters with arbitrary shape
- Minimal requirements for domain knowledge to determine input parameters
- Able to deal with noise and outliers
- Incorporation of user-specified constraints
- Interpretability and usability
- Insensitive to order of input records
- Scalability
- Ability to deal with different types of attributes
- Ability to handle dynamic data
- High dimensionality

Outliers

- There might be objects that do not belong to any cluster



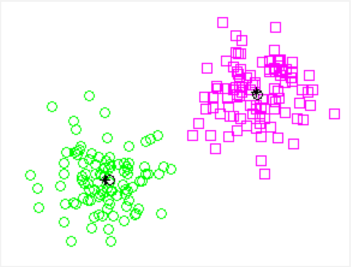
- There are cases where we are interested in detecting outliers not clusters
- Outlier analysis is related to clustering but considered as a different problem!

Outline

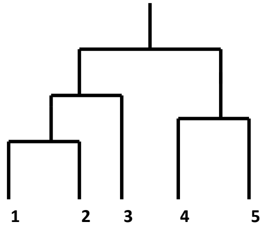
- Unsupervised learning vs supervised learning
- A categorization of major clustering methods
- Partitioning-based clustering
- Hierarchical-based clustering

Major clustering methods 1/2

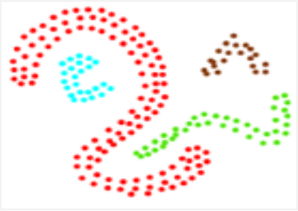
- Partitioning approach:
  - Construct various partitions and then evaluate them by some criterion, e.g., minimizing the sum of square errors
  - Typical methods: k-means, k-medoids, CLARANS



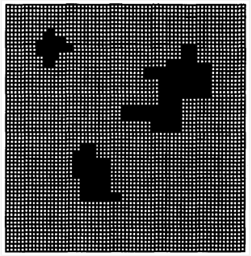
- Hierarchical approach:
  - Create a hierarchical decomposition of the set of data (or objects) using some criterion
  - Typical methods: Diana, Agnes, BIRCH, ROCK, CHAMELEON



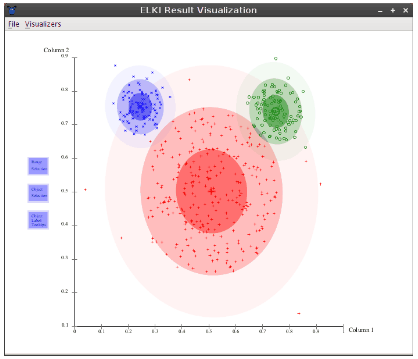
- Density-based approach:
  - Based on connectivity and density functions
  - Typical methods: DBSCAN, OPTICS, DenClue



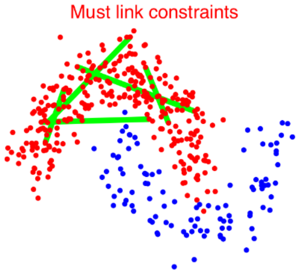
Major clustering methods 2/2



- Grid-based approach:
  - based on a multiple-level granularity structure
  - Typical methods: STING, WaveCluster, CLIQUE
- Model-based:
  - A model is hypothesized for each of the clusters and tries to find the best fit of that model to each other
  - Typical methods: EM, SOM, COBWEB



- Frequent pattern-based:
  - Based on the analysis of frequent patterns
  - Typical methods: pCluster
- User-guided or constraint-based:
  - Clustering by considering user-specified or application-specific constraints
  - Typical methods: COD (obstacles), constrained clustering



Cluster descriptors (numerical data)

- Centroid: the "middle" of a cluster

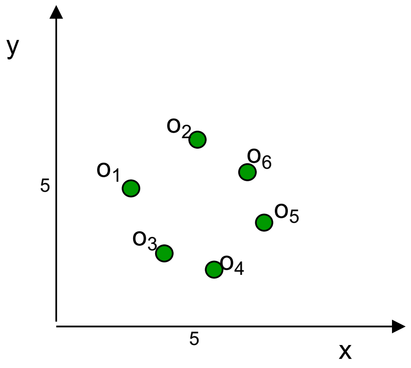
$$c_m = \frac{\sum_{i=1}^{n} p_i}{n}$$

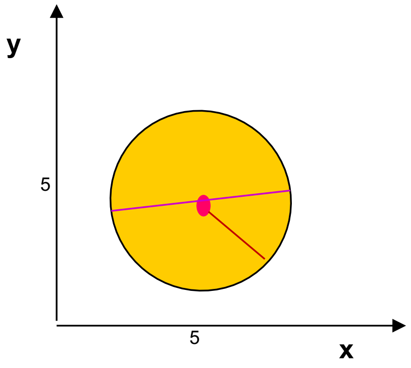
- Radius: square root of average distance from any point of the cluster to its centroid

$$r_m = \sqrt{\frac{\sum_{i=1}^{n} (p_i - c_m)^2}{n}}$$

- Diameter: square root of average mean squared distance between all pairs of points in the cluster

$$d_m = \sqrt{\frac{\sum_{i=1}^{n} \sum_{i=1}^{n} (p_i - p_j)^2}{n(n-1)}}$$





Outline

- Unsupervised learning vs supervised learning
- A categorization of major clustering methods
- Partitioning-based clustering
- Hierarchical-based clustering