



Big Data Analytics – Week 2: Clustering –

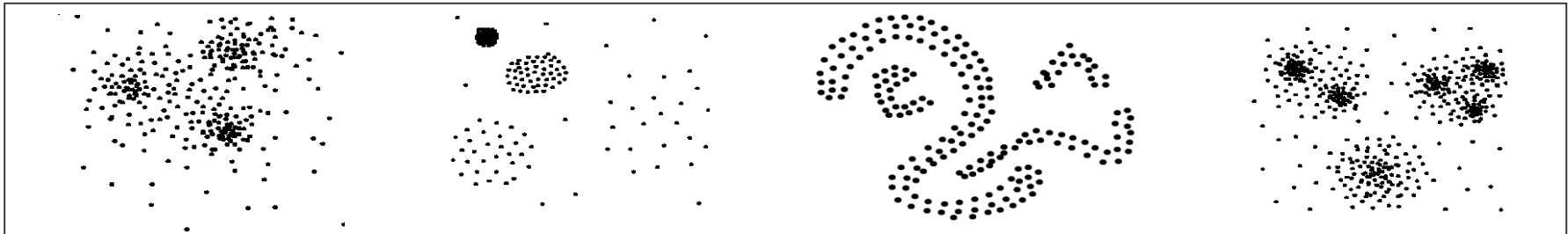
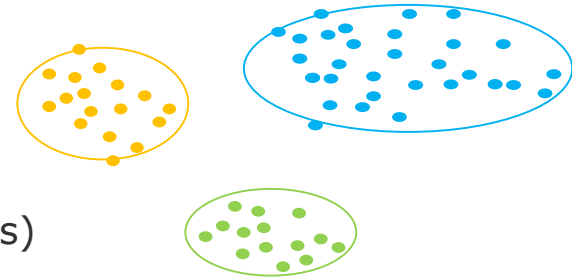
Knowledge Discovery and Data Mining
Prof. Dr. Emmanuel Müller

Content Overview

- Introduction to Clustering
- Partitioning Methods
 - K-Means
 - Variants: K-Medoid, K-Mode, K-Median
 - Choice of parameters: Initialization, Silhouette coefficient
- Probabilistic Model-Based Clusters: Expectation Maximization
- Density-based Methods: DBSCAN
- Hierarchical Methods
 - Agglomerative and Divisive Hierarchical Clustering

What is Clustering?

- Grouping a set of data objects into clusters
 - Cluster: a collection of data objects
 - Similar to one another within the same cluster
 - Dissimilar to the objects in other clusters
- Clustering = unsupervised “classification” (no predefined classes)
- Typical usage
 - As a stand-alone tool to get insight into data distribution
 - As a preprocessing step for other algorithms



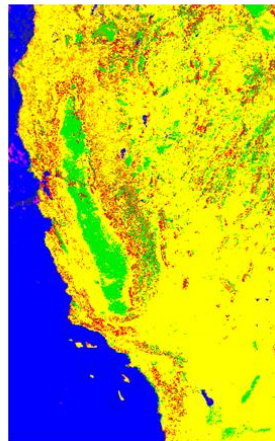
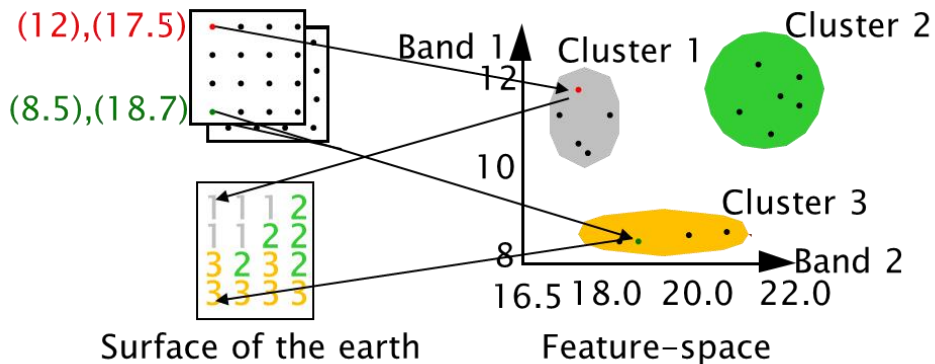
General Applications of Clustering

- **Preprocessing – as a data reduction (instead of sampling)**
 - Image data bases (color histograms for filter distances)
 - Stream clustering (handle endless data sets for offline clustering)
- **Pattern Recognition and Image Processing**
- **Spatial Data Analysis**
 - create thematic maps in GIS (Geographic Information Systems) by clustering feature spaces
 - detect spatial clusters and explain them in spatial data mining
- **Business Intelligence (especially market research)**
- **WWW**
 - Documents (Web Content Mining)
 - Web-logs (Web Usage Mining)
- **Biology**
 - Clustering of gene expression data

Example: Thematic Maps

Satellite images of a region in different wavelengths

- Each point on the surface maps to a high-dimensional feature vector $p = (x_1, \dots, x_d)$ where x_i is the recorded intensity at the surface point in band i .
- Assumption: each different land-use reflects and emits light of different wavelengths in a characteristic way.



Major Clustering Approaches

- **Partitioning algorithms**
 - Find k partitions, minimizing some objective function
- **Probabilistic Model-Based Clustering (EM)**
- **Density-based**
 - Find clusters based on connectivity and density functions
- **Hierarchical algorithms**
 - Create a hierarchical decomposition of the set of objects
- **Other methods**
 - Grid-based
 - Neural networks (SOM's)
 - Graph-theoretical methods
 - Subspace Clustering
 - . . .

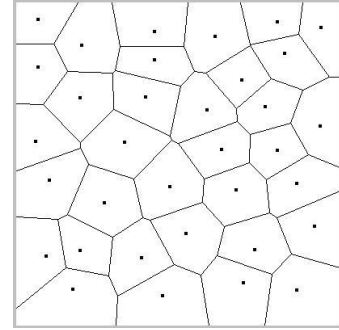
Partitioning: Basic Concept

- Goal: Construct a partitioning of a database D of n objects into a set of k ($k < n$) clusters $C_1, \dots, C_k$ ($C_i \subset D, C_i \cap C_j = \emptyset \Leftrightarrow C_i \neq C_j, \cup C_i = D$) minimizing an objective function.
 - Exhaustively enumerating all possible partitions into k sets in order to find the global minimum is too expensive.
- Popular heuristic methods:
 - Choose k representatives for clusters, e.g., randomly
 - Improve these initial representatives iteratively:
 - Assign each object to the cluster it “fits best” in the current clustering
 - Compute new cluster representatives based on these assignments
 - Repeat until change of objective function is smaller than a threshold
- Examples of representatives for clusters
 - k -means: Each cluster is represented by the center of the cluster
 - k -medoid: Each cluster is represented by one of its objects

Voronoi Model for convex cluster regions

■ Definition: Voronoi diagram

- For a given set of points $P = \{p_i \mid i = 1..k\}$ (here: cluster representatives), a Voronoi diagram partitions the data space in Voronoi cells, one cell per point.
- The cell of a point $p \in P$ covers all points in the data space for which p is the nearest neighbors among the points from P .



■ Observations

- The Voronoi cells of two neighboring points $p_i, p_j \in P$ are separated by the perpendicular hyperplane („Mittelsenkrechte“) between p_i and p_j .
- As Voronoi cells are intersections of half spaces, they are convex regions.

