Clustering

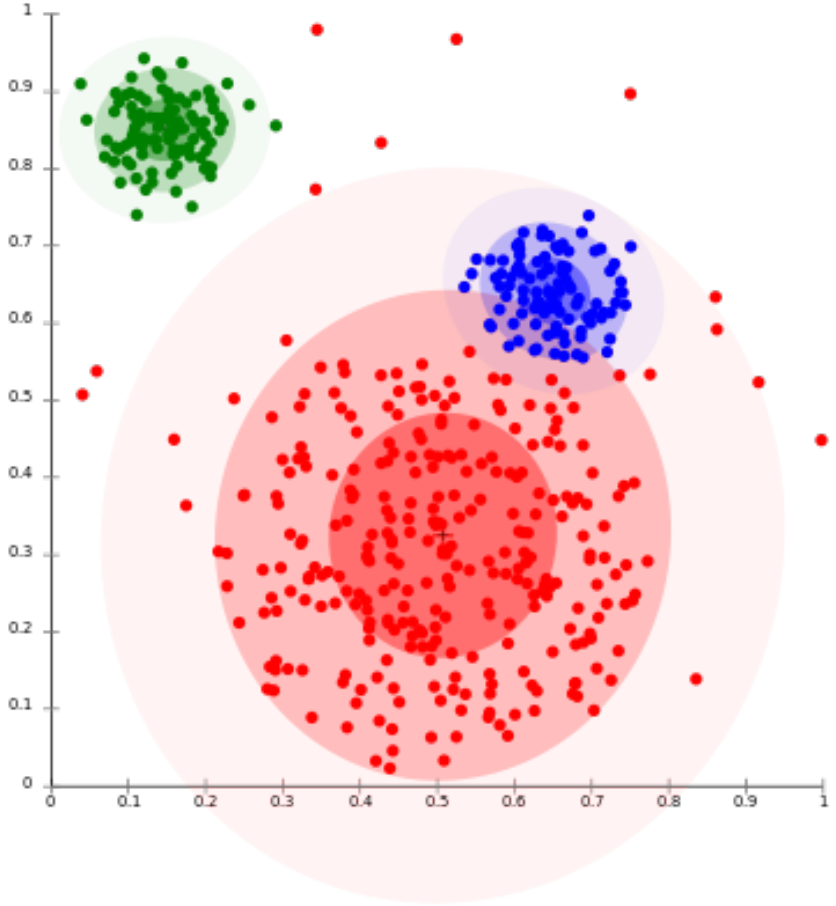
Data Mining:
Data Mining Methods
with **Dr. Qin Lv**

Master of Science in Data Science
UNIVERSITY OF COLORADO **BOULDER**

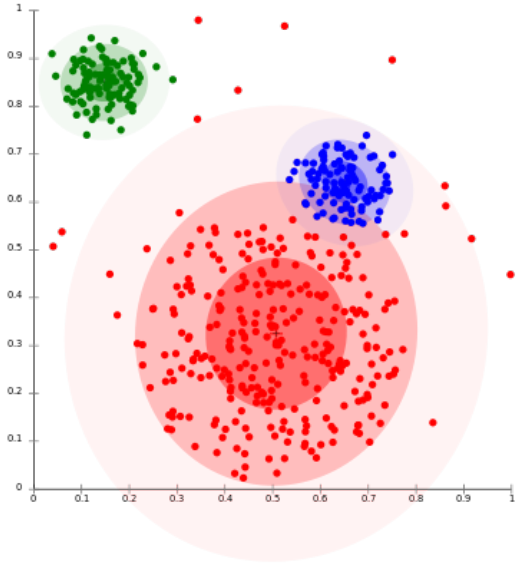
Learning objective: Apply techniques for clustering and explain how they work. Evaluate and compare methods.

Clustering

- No predefined classes
- Intra-cluster similarity
- Inter-cluster dissimilarity

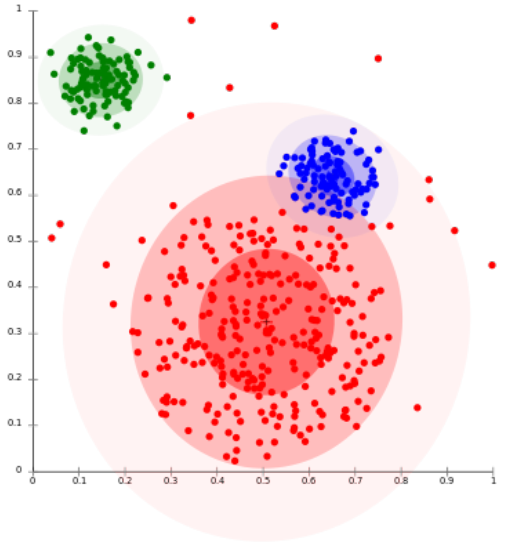


Cluster Analysis



- ➢ Unsupervised learning
  - Group similar objects into clusters
- ➢ Similarity measure
  - Types of objects, similarity/dissimilarity
- ➢ Clustering method
  - Quality, efficiency, incremental

Cluster Evaluation



- Clustering tendency
- Cluster cohesion & separation
- #clusters (e.g., silhouette coefficient)
- Comparison with external knowledge
- Comparison of two sets of clusters

Types of Clustering Methods

- Partitioning methods
- Hierarchical methods
- Grid-based methods
- Density-based methods
- Probabilistic methods

Partitioning Methods

- Given n objects and #clusters k
- Partition the n objects into k clusters
- Brute force approach
- Enumerate all possible partitions
- Heuristic methods
- k-means: cluster centroid (mean of objects)
- k-medoids: cluster medoid (“central” object)

k-means Clustering: Method

- 1. Pick k initial centroids (e.g., randomly)
- 2. Assign each object to nearest centroid
- 3. Update each centroid based on objects assigned to its cluster
- Repeat 2. & 3. until centroids are stable
- O(nkt): n objects, k clusters, t iterations

k-means Clustering: Example

- 10 objects: {35, 69, 9, 78, 9, 23, 81, 57, 15, 48}.
- 2 initial centroids: 30, 60
- Sort: {9, 9, 15, 23, 35, 48, 57, 69, 78, 81}
- R1: 30 {9, 9, 15, 23, 35}, 60 {48, 57, 69, 78, 81}
- C_1 = (9 + 9 + 15 + 23 + 35) / 5 = 18.2
- C_2 = (48 + 57 + 69 + 78 + 81) / 5 = 66.6
- R2: 18.2 {9, 9, 15, 23, 35}, 66.6 {48, 57, 69, 78, 81}

k-means Clustering: Features

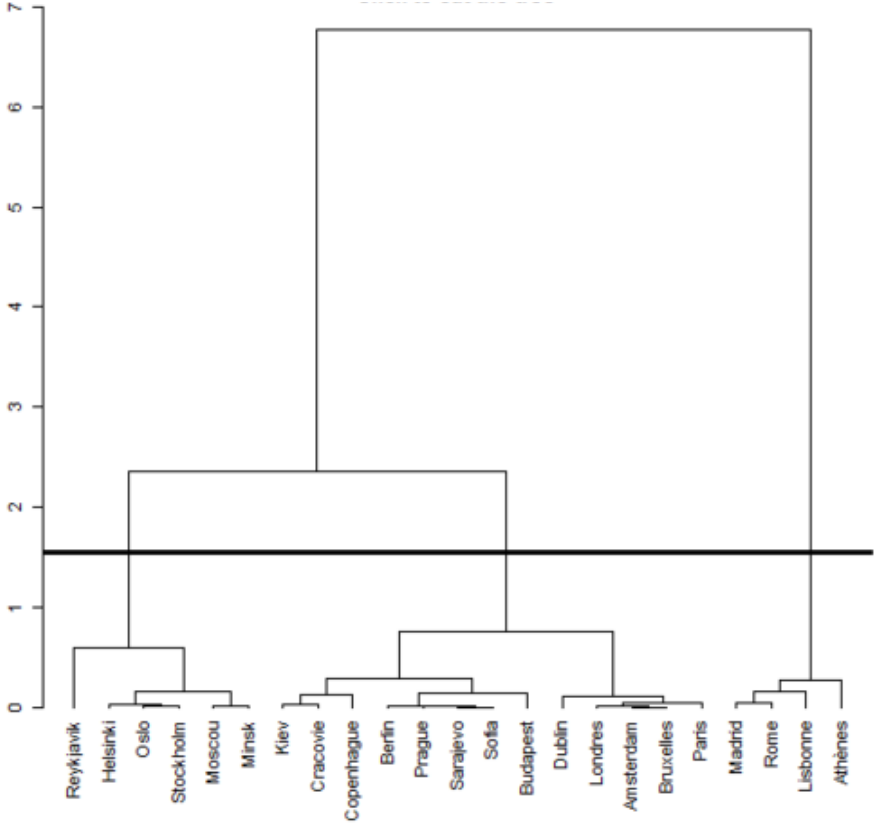
- Widely-used, efficient and good results
- Need to specify k & define centroid
- Choice of initial centroids
- Not suitable for non-convex shapes
- Sensitive to noise & outliers

k-medoids Clustering

- Similar process as k-means
- Cluster medoid: “central” object
- Less sensitive to noise & outliers
- Medoid update: computation expensive
- Speedup using randomized samples

Hierarchical Clustering: Method

- ➢ Dendrogram
  - Tree of clusters
- ➢ Agglomerative
  - Bottom-up merging
- ➢ Decisive
  - Top-down splitting

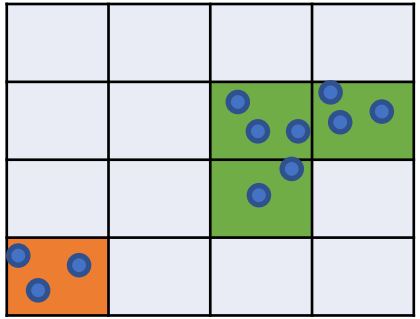


Hierarchical Clustering: Features

- Useful in many real-world applications
- No need to specify #clusters
- Need to define cluster distance
- Multi-level clustering
- Cannot undo cluster merge/split

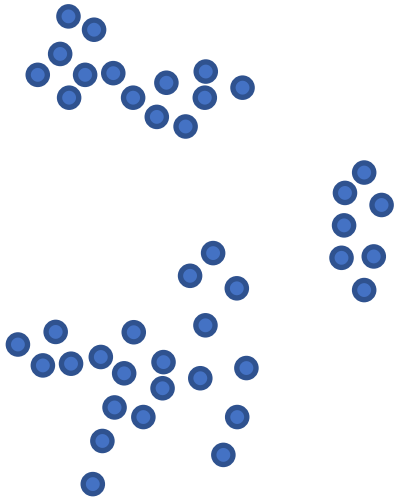
Grid-based Clustering

- ➢ Multi-resolution grid structure
- Clusters of different resolutions
- Horizonal & vertical cluster boundaries
- ➢ Object space => grid cells
- Depends on #cells, easy to parallelize
- ➢ Statistical information of grid cells
- Incremental processing



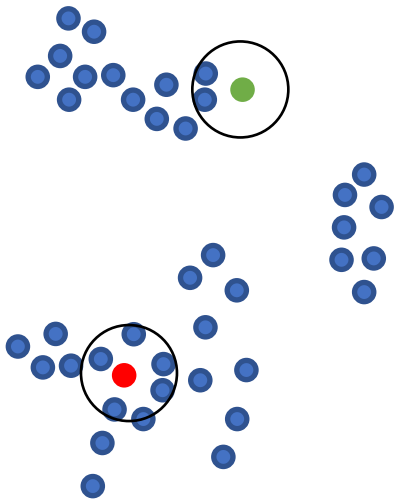
Density-based Clustering

- Local clusters with high density
  - DBSCAN: connected dense neighborhood
  - DENCLUE: sum of local influence functions
- Key features
  - Arbitrary cluster shape, noise tolerant
  - Single scan, adjustable density parameters

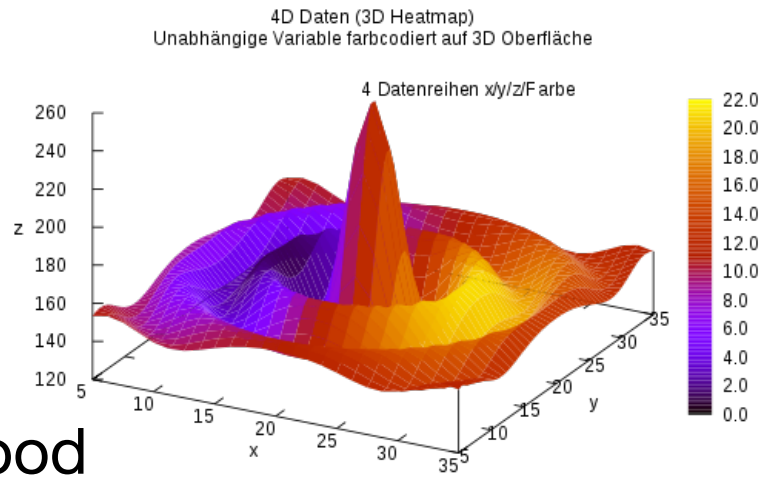


DBSCAN

- Two key parameters
  - ε-neighborhood: within radius ε of p
  - MinPts: min #points in p's ε-neighborhood for p to be considered a core object
- Clustering
  - Core objects, border objects
  - Density-connected, density-reachable



DENCLUE



➢ Influence function

- Object's impact in its neighborhood

➢ Overall density

- Sum of all objects' influence function

➢ Density attractors

- Clusters correspond to local maxima

Types of Clustering Methods

- Partitioning methods
- Hierarchical methods
- Grid-based methods
- Density-based methods
- Probabilistic methods