University of Colorado
Boulder

CSCI 4502/5502
Data Mining

Fall 2020
Lecture 15 (Oct 20)

Reminder/Announcement

- Homework 5: due at 9:30am, Thursday, Oct 22
- Midterm exam: Thursday, Oct 29
  - availability survey: Tuesday, Oct 20
  - midterm review: Thursday, Oct 22
  - practice exam: Tuesday, Oct 27

Review: Chap 9: Advanced Classification

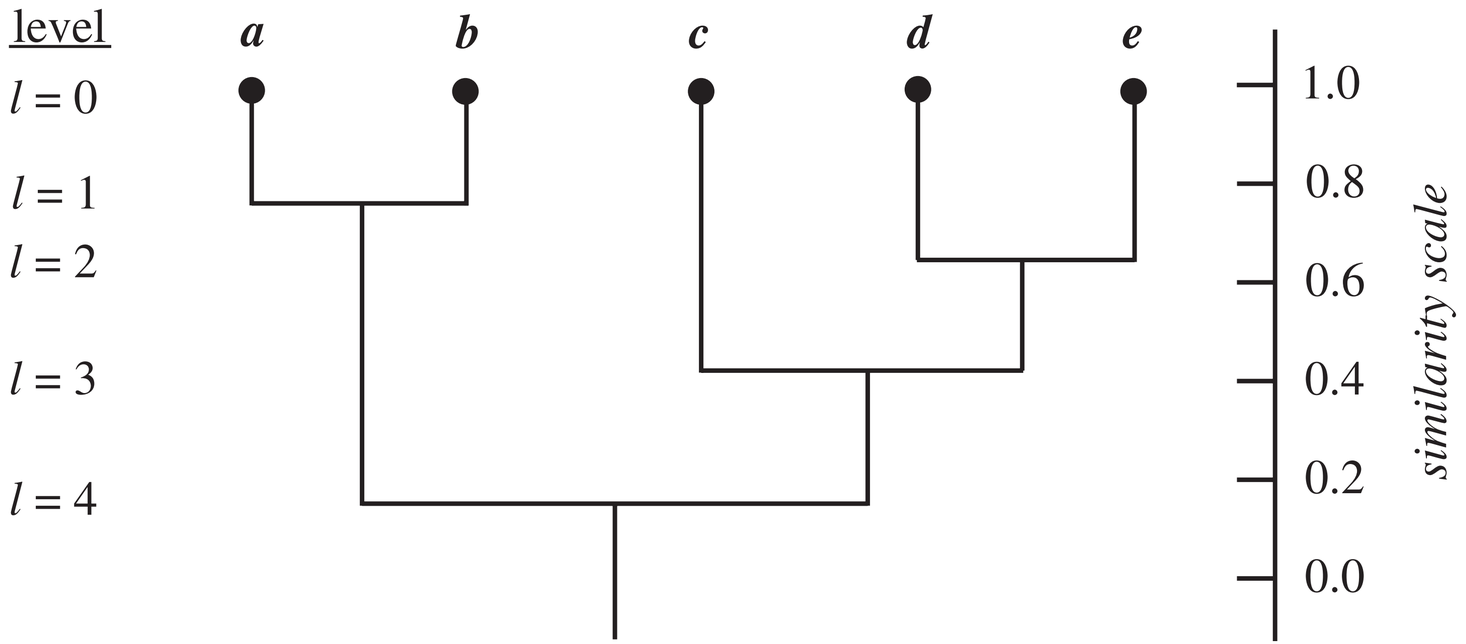
- Bayesian belief networks
- Backpropagation
- Support vector machines
- Lazy learning (or learning from your neighbors)
- Additional topics regarding classification
- Summary

Chapter 10: Cluster Analysis

- Basic concepts
- Partitioning methods
- Hierarchical methods
- Density-based methods
- Grid-based methods
- Evaluation of clustering
- Summary

Hierarchical Clustering (1)

- Groups data objects into a tree of clusters
- Agglomerative
  - bottom-up merging
- Decisive
  - top-down splitting
- Dendrogram



Hierarchical Clustering (2)

- Uses distance matrix as clustering criteria
- Cluster distance
  - centroid distance
  - point distance: min, max, average
- No need to specify k (#clusters)
- Termination condition
  - e.g., cluster distance exceeds a threshold
- Cannot undo previous merge/split decisions

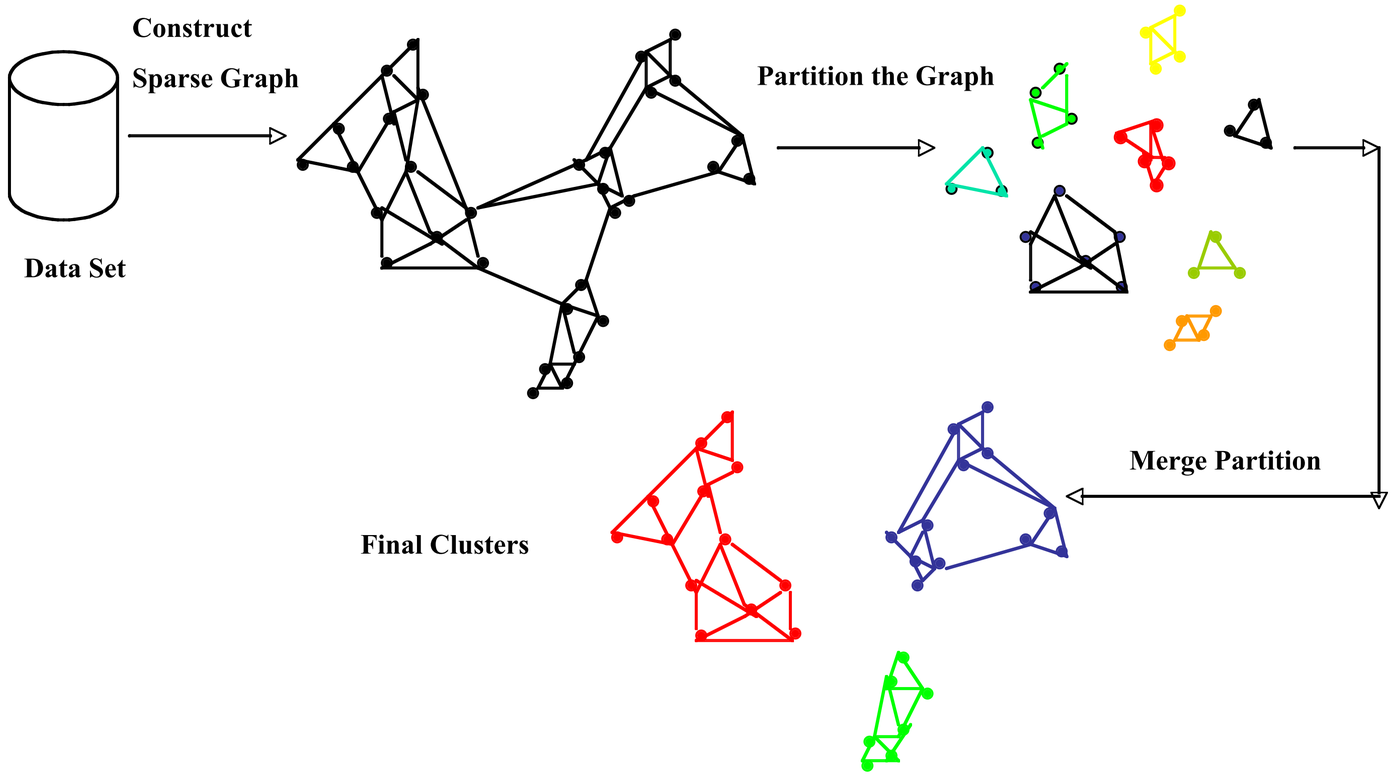
BIRCH

- Clustering large amount of numerical data
- Multi-phase clustering
  - Phase 1: microclustering
    - hierarchical clustering
  - Phase 2: macroclustering
    - e.g., iterative partitioning
- Clustering feature
  - CF = <n, LS, SS>

$$\sum_{i=1}^{N} X_i \qquad \sum_{i=1}^{N} X_i^2$$

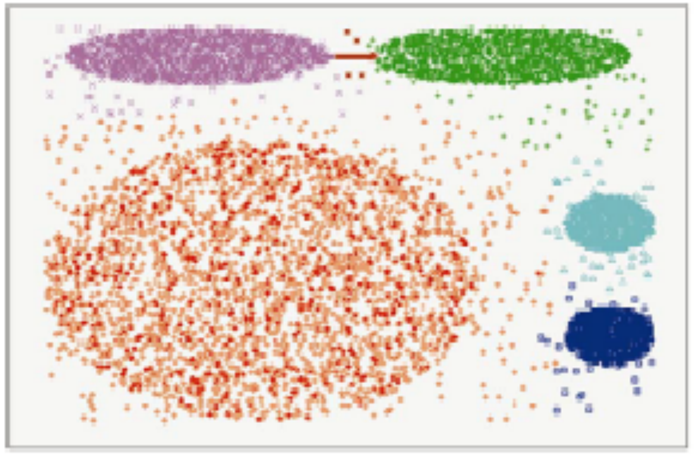
CHAMELEON (1)

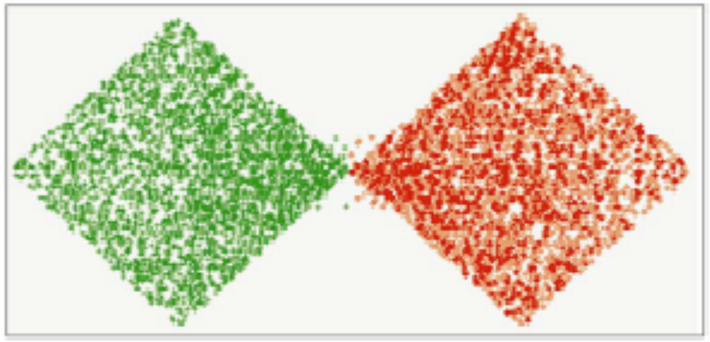
- Overall framework
- k-nearest-neighbor graph
- Graph partition: edge cut $EC(C_i, C_j)$
- Relative interconnectivity (#edge cuts)
- Relative closeness (avg weight of edge cuts)

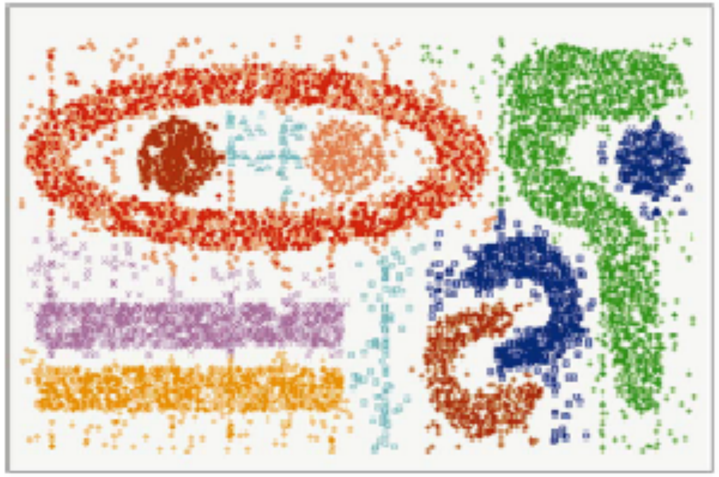


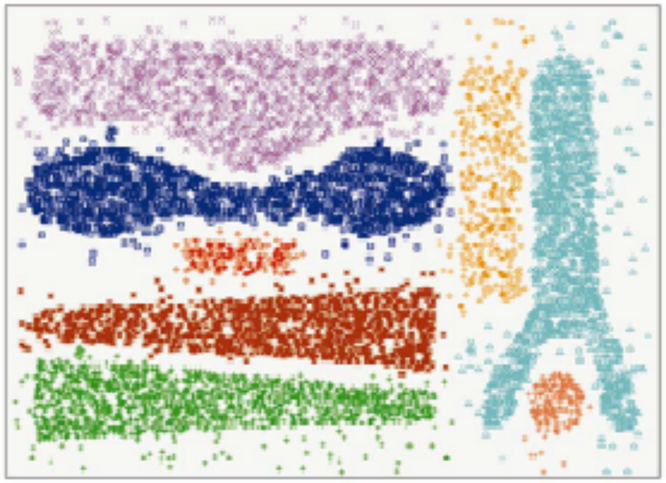
CHAMELEON (2)

- Clustering complex objects









Chapter 10: Cluster Analysis

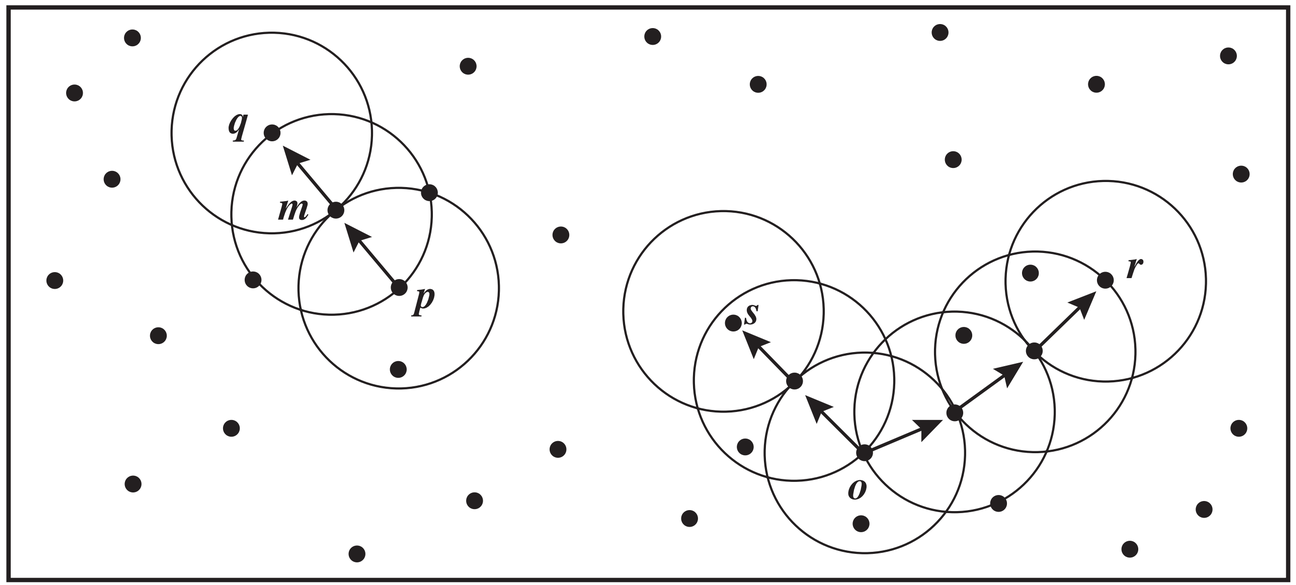
- Basic concepts
- Partitioning methods
- Hierarchical methods
- Density-based methods
- Grid-based methods
- Evaluation of clustering
- Summary

Density-Based Clustering

- Clustering based on density (local cluster criterion), e.g., density-connected points
- Typical methods
  - DBSCAN, DENCLUE
- Major features
  - clusters of arbitrary shape, handles noise, single scan, density parameter for termination

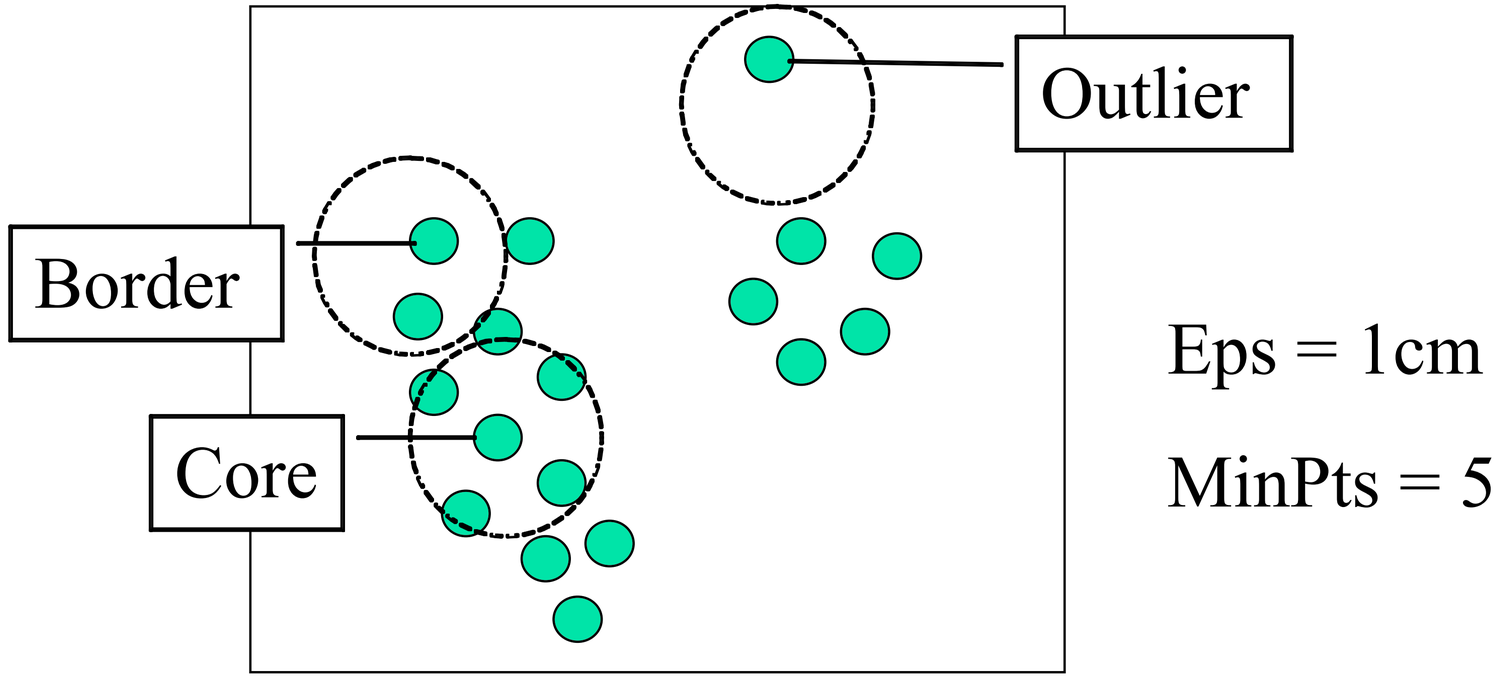
DBSCAN (1)

- ε-neighborhood of p: within radius ε of p
- Core object p: at least MinPts points in the ε-neighborhood of p
- Directly density reachable
- Density reachable
- Density connected



DBSCAN (2)

- Cluster: a maximal set of density-connected points
- Check ε-neighborhood of each point p
- Core object? Border object?



DBSCAN (3)

- Sensitive to parameters: ε and MinPts

Figure 8. DBScan results for DS1 with MinPts at 4 and Eps at (a) 0.5 and (b) 0.4.

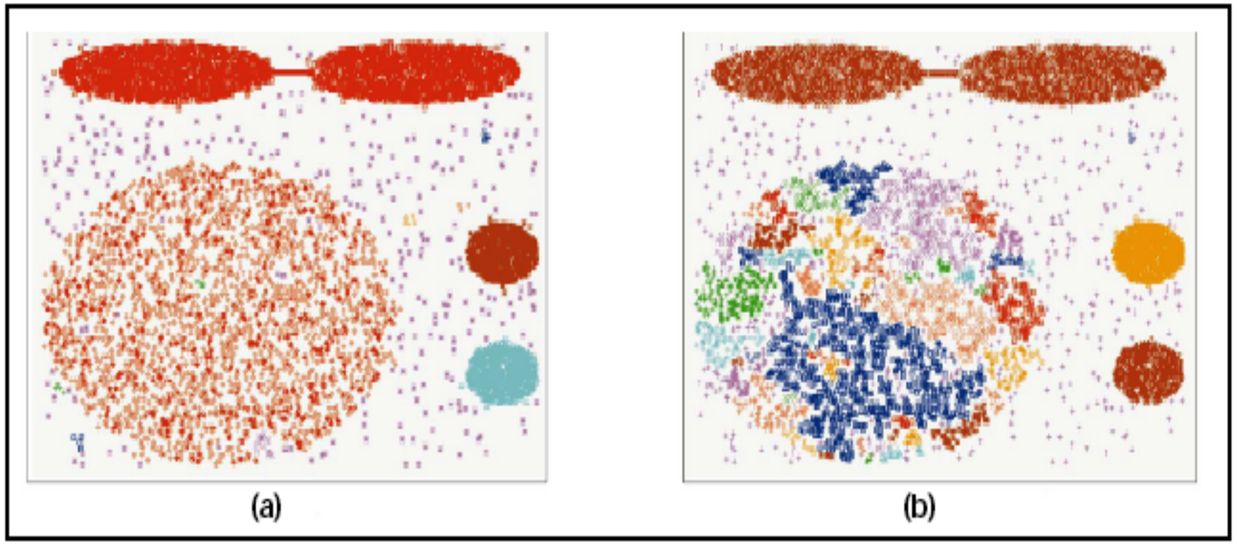
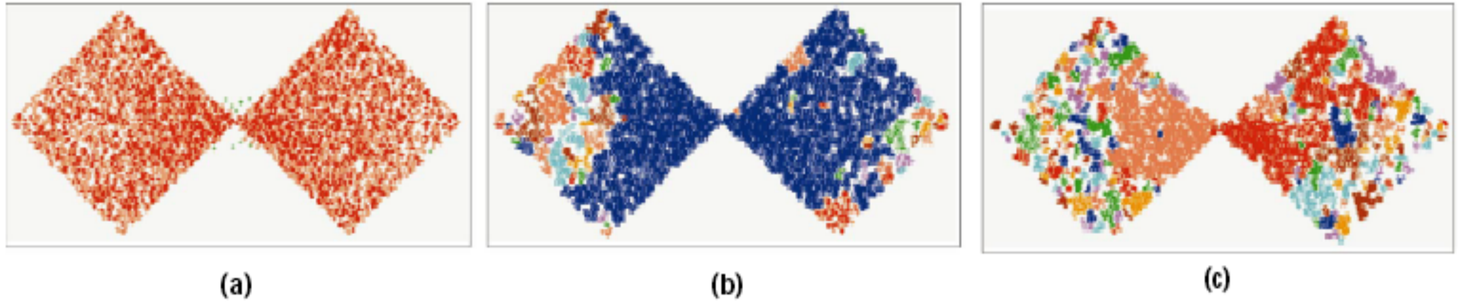


Figure 9. DBScan results for DS2 with MinPts at 4 and Eps at (a) 5.0, (b) 3.5, and (c) 3.0.



DENCLUE (1)

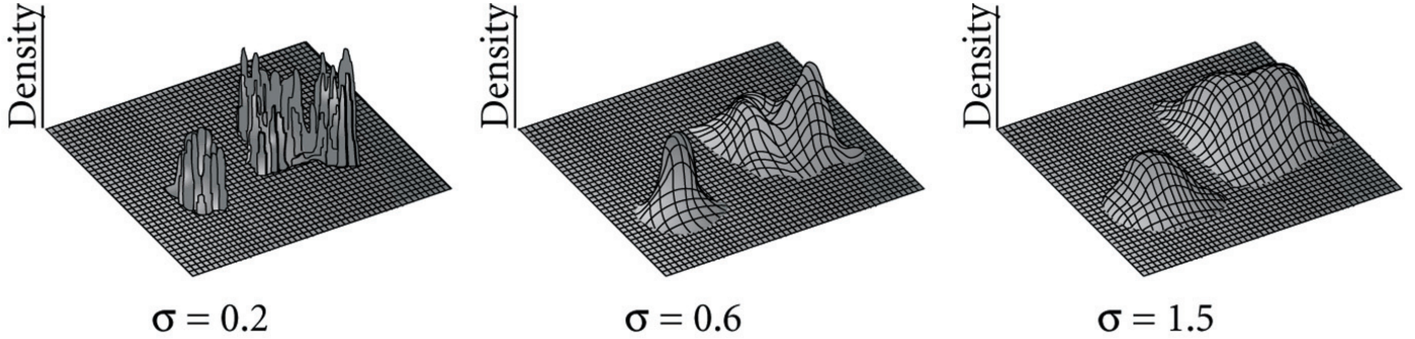
- Uses statistical density functions
- Major features
  - solid mathematical foundation
  - good for data sets with large amounts of noise
  - compact description of arbitrarily-shaped clusters in high-dimensional data sets
  - significantly faster than existing algorithm
  - needs a large number of parameters

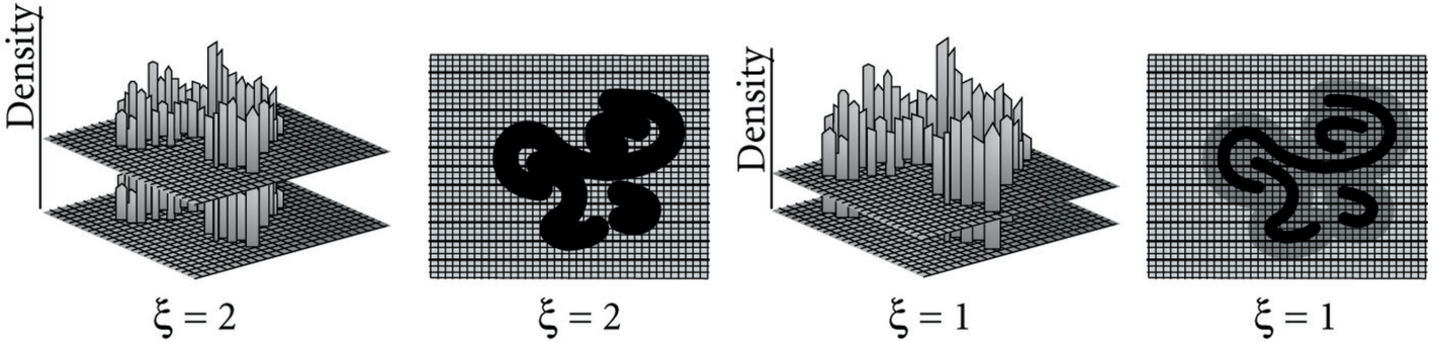
DENCLUE (2)

- Influence function: impact of a data point within its neighborhood
- Overall density: sum of the influence function of all data points
- Density attractors: local maximal of overall density function
- Clusters can be determined mathematically by identifying density attractors

DENCLUE (3)

- Center-defined or arbitrary-shaped clusters





University of Colorado
Boulder

Chapter 11:
Advanced Cluster Analysis

Chapter 11: Advanced Cluster Analysis

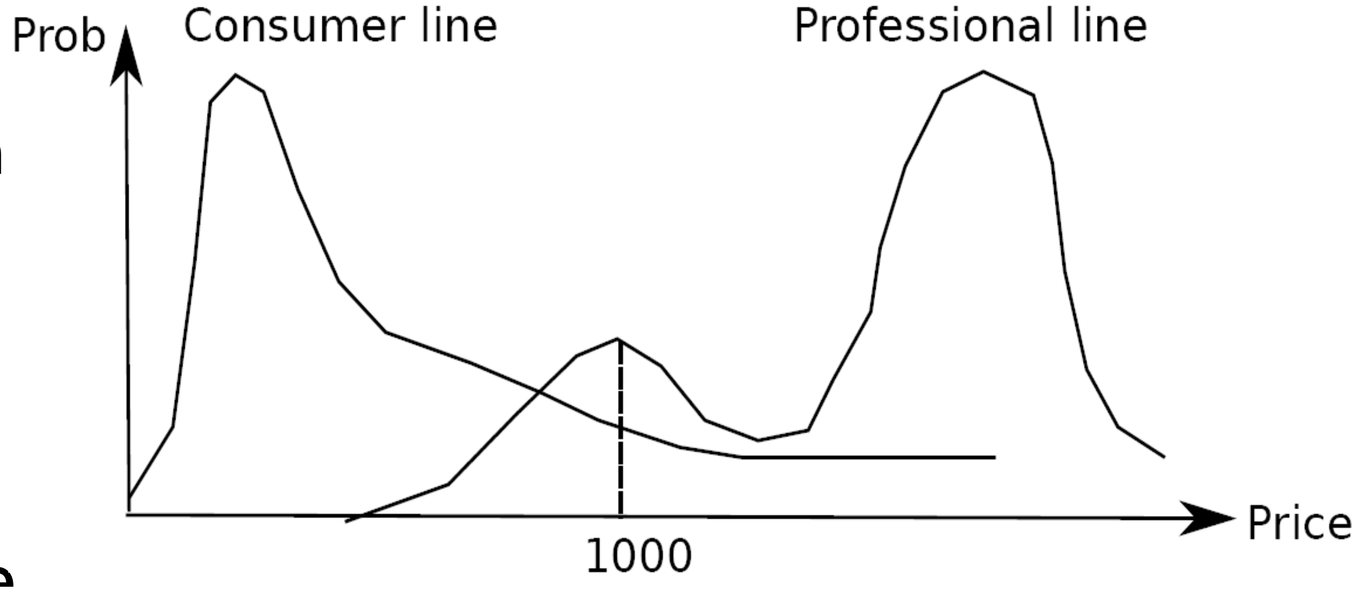
- probabilistic model-based clustering
- clustering high-dimensional data
- clustering graph and network data
- clustering with constraints

Fuzzy Clusters

- Cluster membership of each object
  - belongs to a single cluster
  - weighted distribution in multiple clusters
- Fuzzy clusters (soft clusters)
  - n objects, k clusters
  - w_ij: probability of object i belonging to cluster j
  - 0 <= w_ij <= 1

Probabilistic Clusters

- Hidden categories (probabilistic clusters)
  - each represented by a probability density function over the data space
- Mixture model: observed data instances drawn independently from multiple clusters



Model-Based Clustering

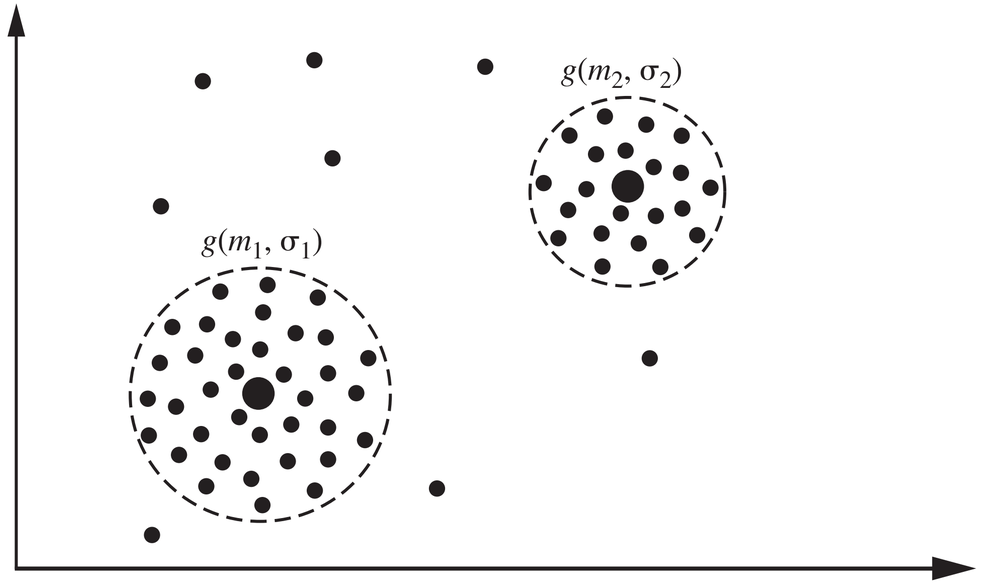
- Assumption: Data are generated by a mixture of underlying probability distributions

$$P(D|\boldsymbol{C}) = \prod_{i=1}^{n} P(o_i|\boldsymbol{C}) = \prod_{i=1}^{n} \sum_{j=1}^{k} \omega_j f_j(o_i)$$

- Attempt to optimize the fit between data and some mathematical model
  - find a set C of k probabilistic clusters s.t. P(D|C) is maximized

EM: Expectation Maximization

- A popular iterative refinement algorithm
- An extension to k-means
  - assign each object to a cluster according to a weight (probability distribution)
  - new means computed on weighted sum
- Mixture of k distributions
  - distribution => cluster
  - e.g., Gaussian distr. $\theta_j = (\mu_j, \sigma_j)$



The EM Algorithm (1)

- Expectation step (E-step)

$$P(\Theta_j|o_i,\mathbf{\Theta}) = \frac{P(o_i|\Theta_j)}{\sum_{l=1}^{k} P(o_i|\Theta_l)}$$

- Maximization step (M-step)

$$\mu_j = \sum_{i=1}^{n} o_i \frac{P(\Theta_j|o_i,\mathbf{\Theta})}{\sum_{l=1}^{n} P(\Theta_j|o_l,\mathbf{\Theta})} = \frac{\sum_{i=1}^{n} o_i P(\Theta_j|o_i,\mathbf{\Theta})}{\sum_{i=1}^{n} P(\Theta_j|o_i,\mathbf{\Theta})}$$

$$\sigma_j = \sqrt{\frac{\sum_{i=1}^{n} P(\Theta_j|o_i,\mathbf{\Theta})(o_i-u_j)^2}{\sum_{i=1}^{n} P(\Theta_j|o_i,\mathbf{\Theta})}}$$

The EM Algorithm (2)

- Simple, easy to implement
- Can be characterized by a few parameters
- Generally converges quickly but may not reach the global optima
- Computationally expensive if number of distributions is large or data set contains very few observed data points

Clustering High-Dimensional Data

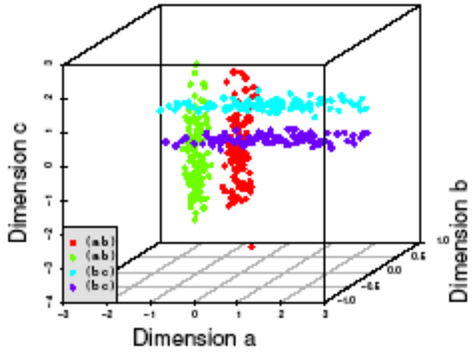
- High-dimensional data
  - e.g., text documents, DNA micro-array data
- Methods
  - subspace clustering, dimensionality reduction
- Challenges
  - many irrelevant dimensions may mask clusters
  - distance measure dominated by noises
  - clusters may exist only in some subspaces

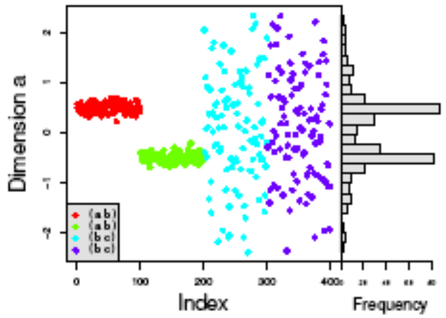
The Curse of Dimensionality

- Data in only one dimension is relatively packed
- Adding a dimension "stretch" the points across that dimension, making them further apart
- Adding more dimensions will make the points further apart
  - high-dimensional data is extremely sparse
- Distance measure becomes meaningless
  - due to equi-distance

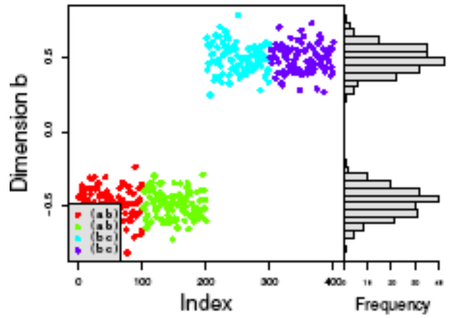
Why SubSpace Clustering?

- Clusters may exist only in some subspaces
- Subspace clustering: find clusters in all the subspaces

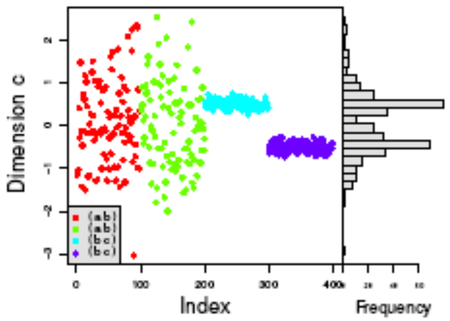




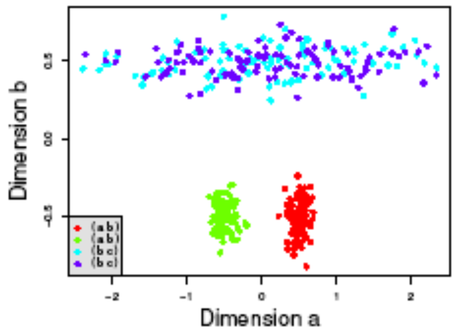
(a) Dimension a



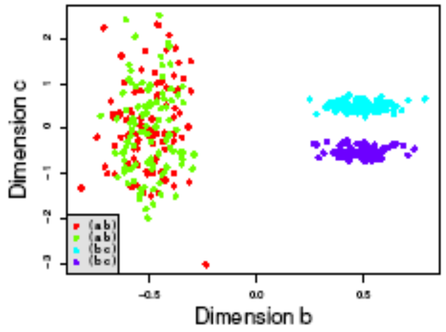
(b) Dimension b



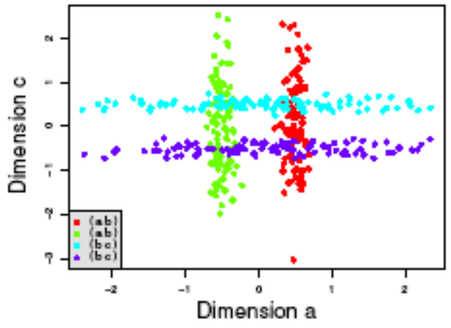
(c) Dimension c



(a) Dims a & b



(b) Dims b & c



(c) Dims a & c