Clustering

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

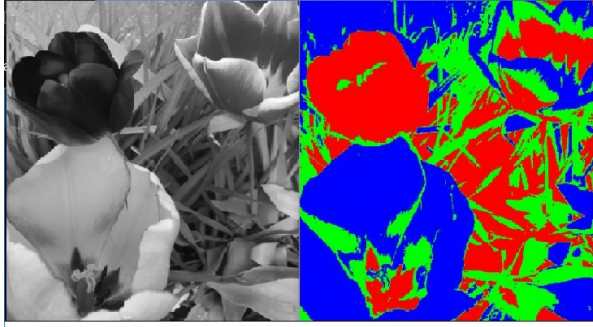
Master of Science in Data Science
UNIVERSITY OF COLORADO **BOULDER**

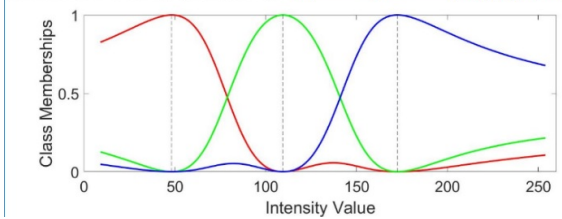
Learning objective: Apply techniques for clustering and explain how they work. Evaluate and compare methods.

Types of Clustering Methods

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

Cluster Membership





- n objects, k clusters
- Single clusters
  - Each object belongs to a single cluster
- Fuzzy clusters
  - Each object has certain probability of belonging to a specific cluster
  - $0 <= w_{ij} <= 1.0$ $(j = 1, \ldots, k)$ & $w_{i1}+\ldots+w_{ik} = 1.0$

Probabilistic Clusters

➢ Hidden categories/cluster models

- E.g., customer groups => purchase behavior
- Each group: probability density function over purchases

➢ Mixture model

- Each object drawn independently from multiple clusters
- E.g., given a customer's purchase behavior, he/she has certain probability of coming from any group

Model-based Clustering

- Assumption: data D generated by a mixture of probabilistic models C
- Goal: optimize the fit between data and models
- find C of k probabilistic clusters s.t. P(D|C) is maximized

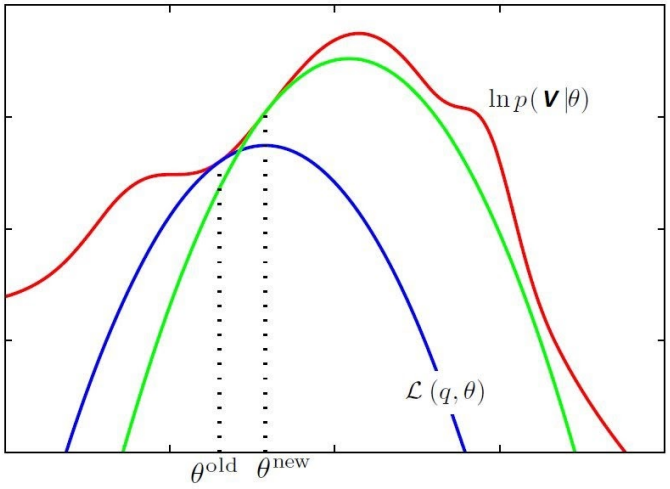
$$P(D|\boldsymbol{C}) = \prod_{i=1}^{n} P(o_i|\boldsymbol{C}) = \prod_{i=1}^{n} \sum_{j=1}^{k} \omega_j f_j(o_i)$$

k-means Clustering: Recap

- 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

Expectation Maximization (EM)

- ➢ Iterative refinement
  - Similar process as k-means
- ➢ Mixture of models
  - E.g., Gaussian models $\theta_j = (\mu_j, \sigma_j)$
- ➢ Probabilistic membership
  - Cluster mean: weighted sum of objects



$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$$

EM Clustering: Method

- E-step: Expectation $P(\Theta_j|o_i,\mathbf{\Theta}) = \frac{P(o_i|\Theta_j)}{\sum_{l=1}^{k} P(o_i|\Theta_j)}$
- M-step: Maximization

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

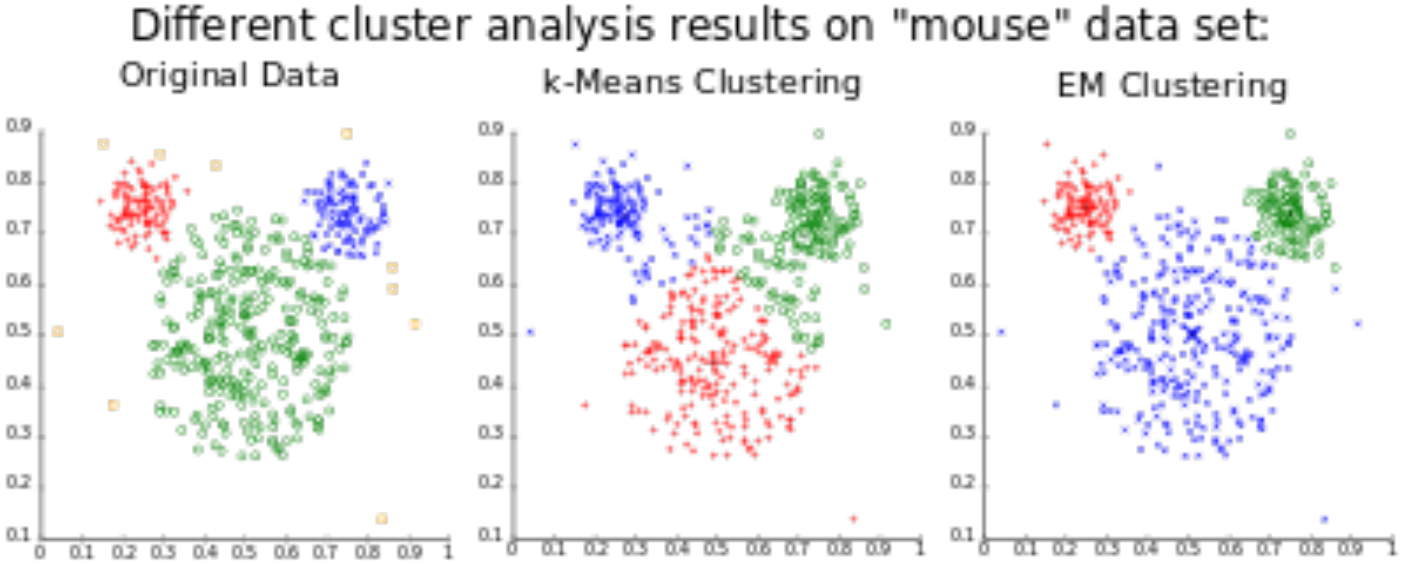
$$\sigma_j = \sqrt{\frac{\sum_{i=1}^{n} P(\Theta_j|o_i,\mathbf{\Theta})(o_i-\mu_j)^2}{\sum_{i=1}^{n} P(\Theta_j|o_i,\mathbf{\Theta})}}$$

EM Clustering: Features

- Good performance in many applications
- Easy to implement: (μ_j, σ_j) parameters
- Converges quickly, may not be optimal
- Not good if #objects is small
- Computation-intensive for large #clusters

k-means vs. EM: Example

- Choose the right method and model

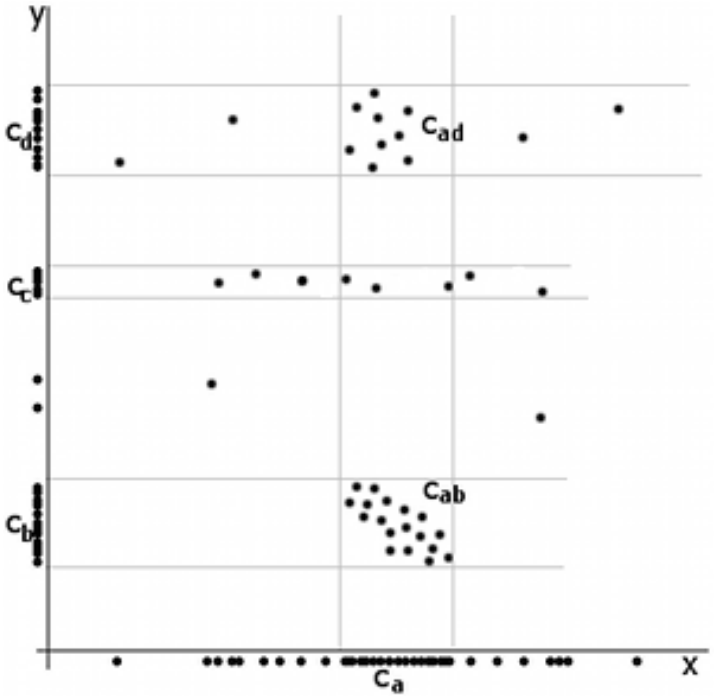


Clustering High-dimensional Data

- ➢ High dimensionality in many applications
  - E.g., text, audio, video, scientific data
- ➢ Curse of dimensionality
  - Sparse, noisy, equi-distance, irrelevant dimensions
- ➢ Methods
  - Dimensionality reduction, subspace clustering

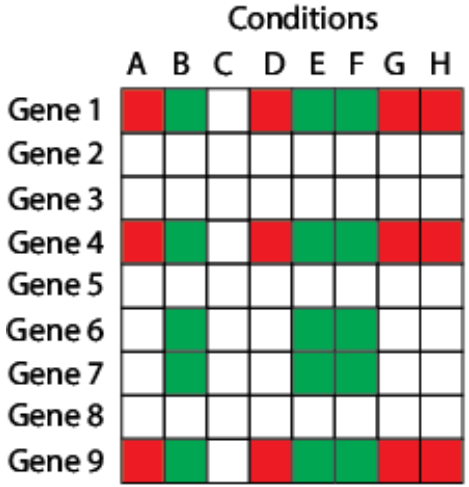
Subspace Clustering

- Clusters in subspaces
- Dimension-growth
- Dimension-reduction
- Frequent pattern based



Bi-clustering

- Cluster both objects & attributes



Graph Clustering

- ➢ Wide use of graph data in the real world
  - E.g., social networks, co-authorship, protein interaction
  - Graph clusters: well-connected substructures
- ➢ Clustering graphs
  - Generic clustering methods
  - Graph-specific clustering: e.g., community detection

Constraint-based Clustering (1)

➢ Benefits

- Focused mining, domain knowledge, efficiency

➢ Objects

- E.g., sales in specific location/time/category

➢ Distance functions

- E.g., weighted attributes, obstacles

Constraint-based Clustering (2)

- ➢ Clustering parameters
  - E.g., #clusters, ε-neighborhood, MinPts
- ➢ Domain knowledge
  - E.g., initial clusters, bird sightings, scientific topics
- ➢ Semi-supervised
  - E.g., certain objects in the same/different clusters

Summary: Clustering

➢ Unsupervised learning

- No predefined classes, exploratory analysis

➢ Methods

- Partitioning, hierarchical, grid, density, probabilistic

➢ Model evaluation

- Clustering tendency, cohesion, separation, #clusters