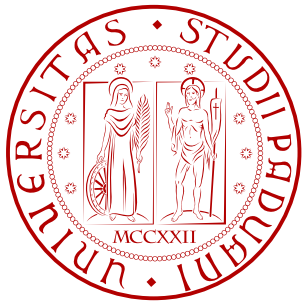
CLUSTERING
“GENERAL CONCEPTS”

Michele Rossi
rossi@dei.unipd.it

Dept. of Information Engineering
University of Padova, IT



Overview

- General concepts
  - Objective
  - Metrics
  - Approaches

- Techniques
  - Flat clustering (k-means, and “soft” k-means)
  - Divisive clustering (“hierarchical”)
  - Online clustering (Self Organizing Maps - SOM)
  - Density-based clustering (DBSCAN)

Informal goal

- Given a set of objects and a measure of similarity
- Group similar objects together

- Questions
  - What do we mean by “similar”?
  - What good grouping looks like?
  - Computation time/quality tradeoff

Applications

- Many, in all fields
  - Biology
  - Astronomy
  - Information organization
  - Pattern recognition and analysis
  - Marketing
  - …

Issues

- What attributes represent items for clustering purpose?
- What is measure of similarity between items?
  - General objects and matrix of pairwise similarities $S(o_i,o_j)$
  - Objectives with specific properties that allow other measures
    - Most common objects are d-dimensional vectors
    - Most common distance is Euclidean distance

Issues continued

- Clustering objectives?
  - Number of clusters?
  - Flat or hierarchical clustering?
  - Cohesiveness of clusters?
- How shall we evaluate cluster results?
  - Measure of closeness within cluster elements
- Efficiency of clustering algorithm
  - Large data sets: online *vs* offline clustering
- Best clustering algorithm?
  - There are many
  - Size of dataset (complexity), online *vs* offline, type of measure

General types of clustering

- "Soft" vs "hard" clustering
  - "hard": partition the objects
    - Each object in exactly one partition
  - "soft": assign degree to which object in each cluster
    - View as a probability or "score"

- "Flat" vs "hierarchical" clustering
  - "Hierarchical": clusters within clusters
  - A cluster "hierarchy" is constructed

Hierarchical clustering

- "agglomerative" vs "divisive" algorithms
  - "agglomerative": bottom-up
    - Build up clusters from single objects
  - "divisive": top-down
    - Break up clusters containing all objects into smaller clusters
- Both approaches lead to hierarchies

How clustering progresses

- "constructive" vs "iterative" improvement
  - "constructive": decide to which cluster each object belongs to and do not change this choice
    - Often faster
  - "iterative" improvement: start with a clustering solution and move objects around to see if improvements are possible
    - Often slower but leads to better results

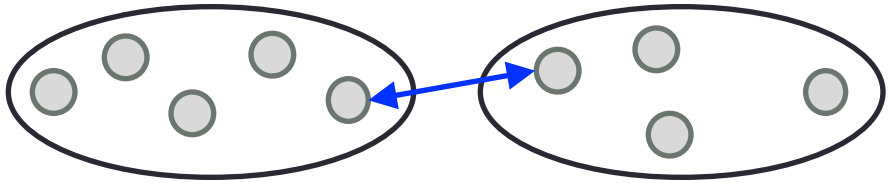
Quality of clustering

- In applications, the quality of clustering depends on "how well it solves the problem at hand"
- Algorithm uses measure of quality that can be optimized, but that may or may not do a good job in capturing application needs
- Underlying graph-theoretic problems usually NP-complete
  - e.g., graph partitioning
  - usually algorithms do not find optimal clustering

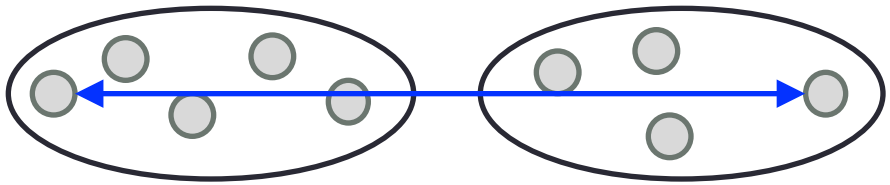
Distance between two clusters (1/2)

- Possible approaches

1) Distance between closest objects in the clusters
   - Called *single link*



2) Distance between the furthest away objects (one per cluster)
   - Called *complete linkage*

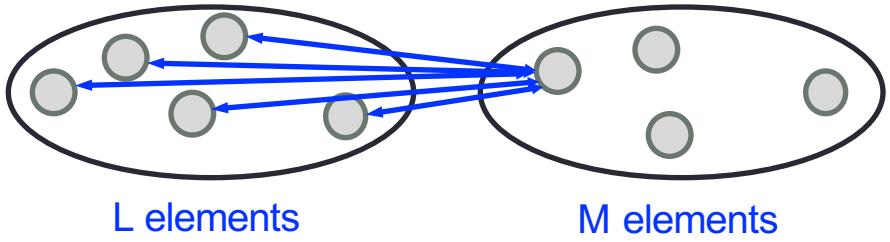


Distance between two clusters (2/2)

- Possible approaches

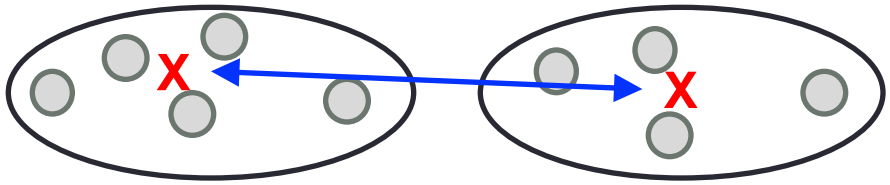
3) Average of pairwise distance between all pairs of objects, one for each cluster

- More computation



4) If there exists a measure, e.g., Euclidean

- Centroids can be computed and used to evaluate distance



CLUSTERING
“GENERAL CONCEPTS”

Michele Rossi
rossi@dei.unipd.it

Dept. of Information Engineering
University of Padova, IT