The University of Reading

CSMDM16

Clustering

Dr. Giuseppe Di Fatta
Associate Professor
Department of Computer Science
G.DiFatta@reading.ac.uk

Clustering

- Cluster Analysis
- Similarity
- Types of Clusters
- Clustering Approaches
  - Partitioning
  - Hierarchical
  - Density-based
  - Grid-based
  - Model-based
- Cluster validity

Definition of “Cluster”

Definition of “Cluster” in online dictionaries:

- a number of similar things growing together or of things or persons collected or grouped closely together: BUNCH.
- two or more consecutive consonants or vowels in a segment of speech.
- a group of buildings and esp. houses built close together on a sizable tract in order to preserve open spaces larger than the individual yard for common recreation.
- an aggregation of stars, galaxies, or super galaxies that appear close together in the sky and seem to have common properties (as distance).
- **A cluster is a closely-packed group (of people or things).**

Clustering in Data Mining

- **Cluster Analysis** is the process of partitioning a set of data (or objects) in a set of meaningful sub-classes, called **clusters**.
  - It helps users to understand the natural grouping or structure in a data set.
- Cluster analysis = 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 is unsupervised classification:
  - no predefined classes

General Applications of Clustering

- Pattern Recognition
- Spatial Data Analysis
- Image Processing
- Economic Science (especially market research)
- WWW
  - Document classification
  - Cluster Weblog data to discover groups of similar access patterns

Examples:

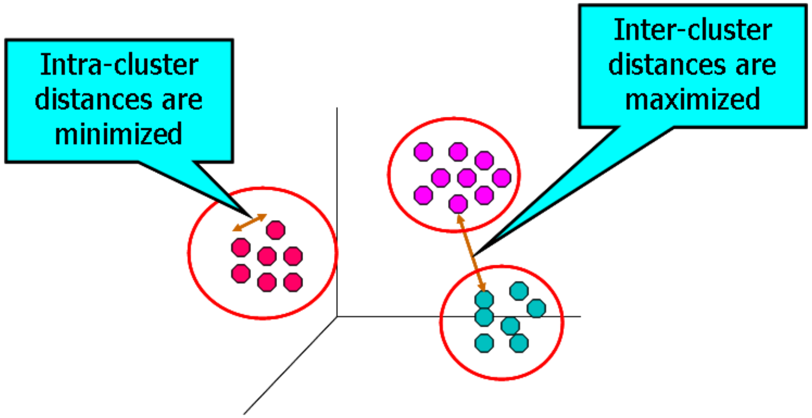
- Marketing: Help marketers discover distinct groups in their customer bases, and then use this knowledge to develop targeted marketing programs
- Land use: Identification of areas of similar land use in an earth observation database
- Insurance: Identifying groups of motor insurance policy holders with a high average claim cost
- City-planning: Identifying groups of houses according to their house type, value, and geographical location
- Earth-quake studies: Observed earth quake epicenters should be clustered along continent faults

What Is Good Clustering?

- A good clustering method will produce high quality clusters with
  - high intra-class similarity
  - low inter-class similarity

- The quality of a clustering result depends on both the similarity measure used by the method and its implementation.

- The quality of a clustering method is also measured by its ability to discover some or all of the hidden patterns.



Measuring Similarity

- Dissimilarity/Similarity metric: similarity is expressed in terms of a “**distance**” function, which is typically a Metric $d(x,y)$.
- There is a separate “**quality**” function that measures the “goodness” of a cluster.
- The definitions of distance functions are usually very different for boolean, nominal, ordinal, interval and ratio variables.
- Weights should be associated with different features based on applications and data semantics.
- It is hard to define “similar enough” or “good enough”
  - the answer is typically highly subjective.

The Notion of a Cluster can be Ambiguous

How many clusters?

Six Clusters

Two Clusters

Four Clusters

Other Distinctions Between Sets of Clusters

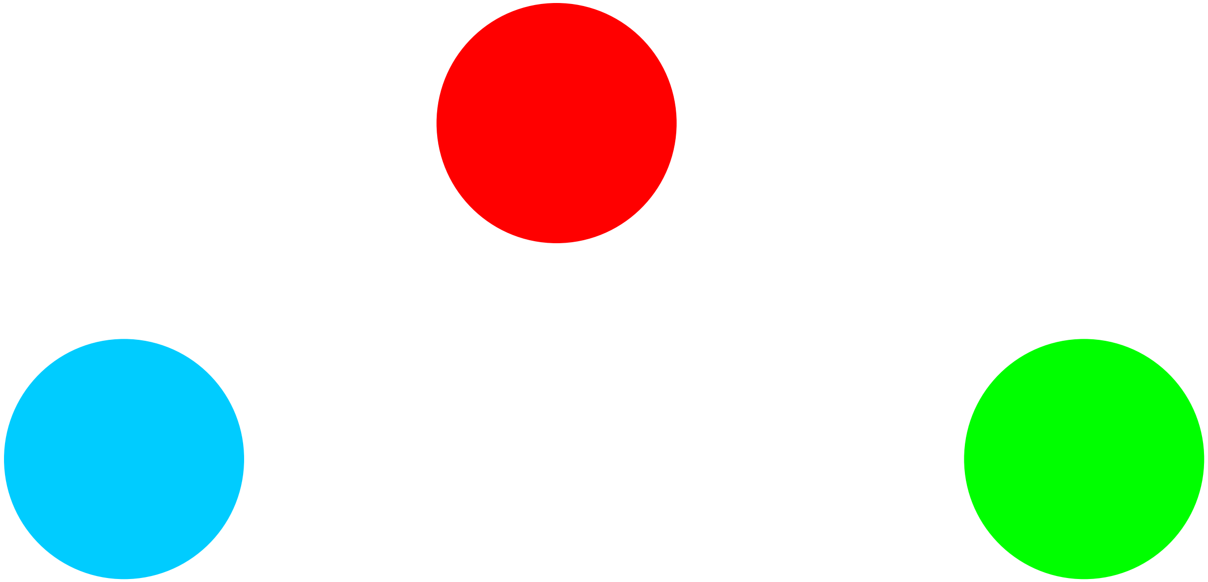
- Exclusive versus non-exclusive
  - In non-exclusive clusterings, points may belong to multiple clusters.
  - Can represent multiple classes or ‘border’ points
- Fuzzy versus non-fuzzy
  - In fuzzy clustering, a point belongs to every cluster with some weight between 0 and 1
  - Weights must sum to 1
  - Probabilistic clustering has similar characteristics
- Partial versus complete
  - In some cases, we only want to cluster some of the data
- Heterogeneous versus homogeneous
  - Cluster of widely different sizes, shapes, and densities

Types of Clusters

- Well-separated clusters
- Center-based clusters
- Contiguous clusters
- Density-based clusters
- Property or Conceptual
- Described by an Objective Function

Types of Clusters: Well-Separated

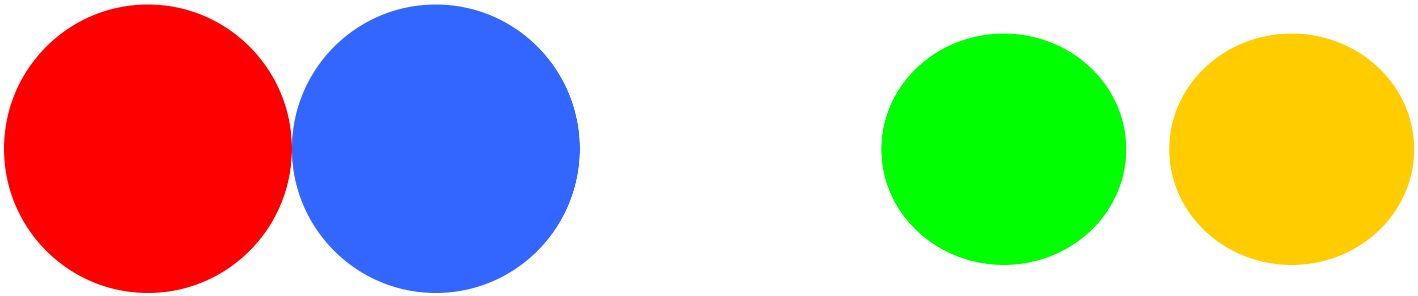
- Well-Separated Clusters:
  - A cluster is a set of points such that any point in a cluster is closer (or more similar) to every other point in the cluster than to any point not in the cluster.



3 well-separated clusters

Types of Clusters: Center-Based

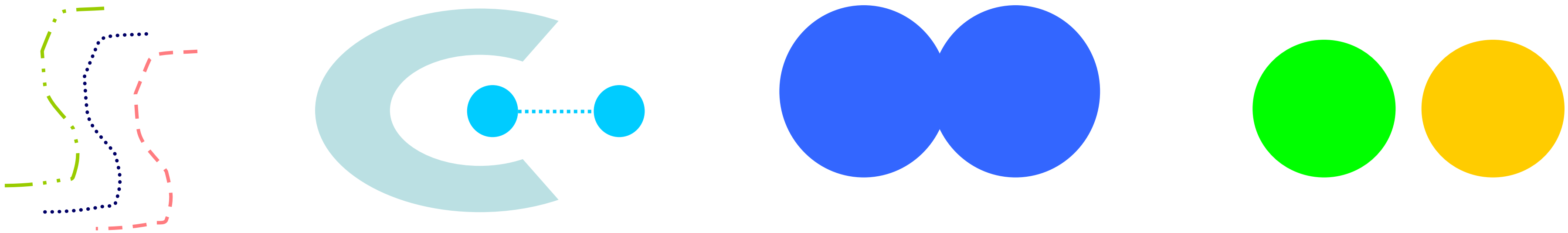
- Center-based
  - A cluster is a set of objects such that an object in a cluster is closer (more similar) to the "center" of a cluster, than to the center of any other cluster
  - The center of a cluster is often a centroid, the average of all the points in the cluster, or a medoid, the most "representative" point of a cluster



4 center-based clusters

Types of Clusters: Contiguity-Based

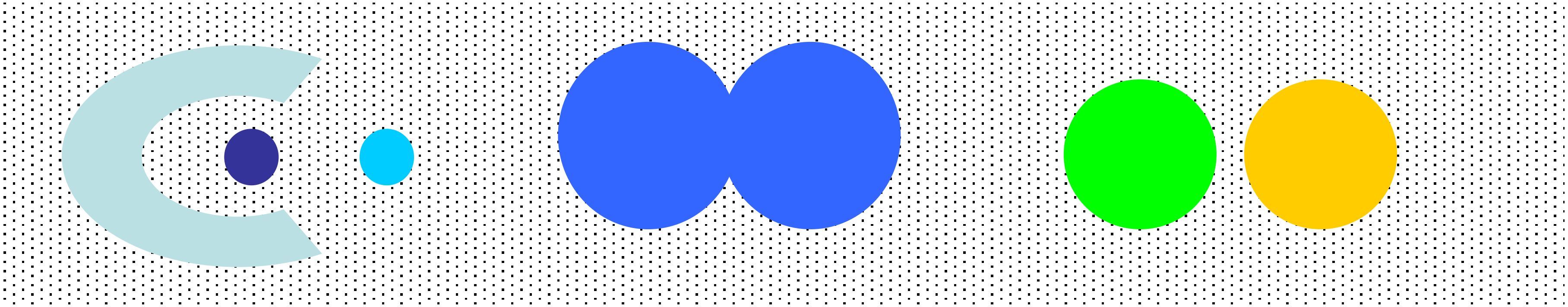
- Contiguous Cluster (Nearest neighbor or Transitive)
  - A cluster is a set of points such that a point in a cluster is closer (or more similar) to one or more other points in the cluster than to any point not in the cluster.



8 contiguous clusters

Types of Clusters: Density-Based

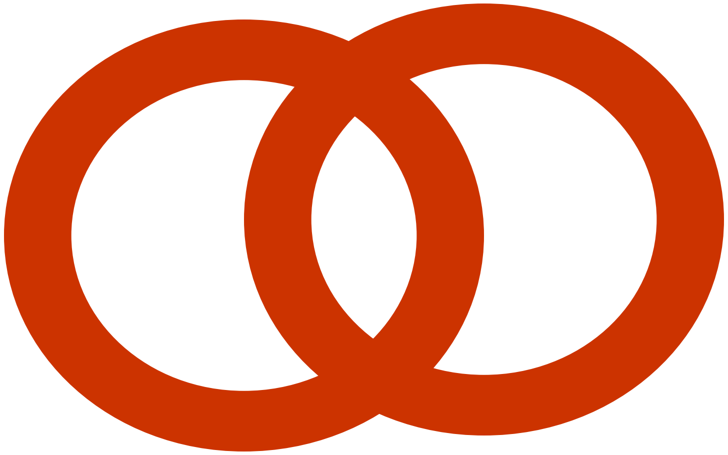
- Density-based
  - A cluster is a dense region of points, which is separated by low-density regions, from other regions of high density.
  - Used when the clusters are irregular or intertwined, and when noise and outliers are present.



6 density-based clusters

Types of Clusters: Conceptual Clusters

- Shared Property or Conceptual Clusters
  - Finds clusters that share some common property or represent a particular concept.



2 Overlapping Circles

Types of Clusters: Objective Clusters

- Clusters Defined by an Objective Function
  - Finds clusters that minimize or maximize an objective function.
  - Enumerate all possible ways of dividing the points into clusters and evaluate the `goodness' of each potential set of clusters.
  - Can have global or local objectives.
    - Hierarchical clustering algorithms typically have local objectives
    - Partitional algorithms typically have global objectives

Clustering Approaches

1. **Partitioning algorithms**: Construct various partitions and then evaluate them by some criterion
2. **Hierarchy algorithms**: Create a hierarchical decomposition of the set of data (or objects) using some criterion
3. **Density-based**: based on connectivity and density functions
4. **Grid-based**: based on a multiple-level granularity structure
5. **Model-based**: A model is hypothesized for each of the clusters and the idea is to find the best fit of that model to each other

Partitioning algorithms

- **Partitioning method:** construct a partition of a database ***D*** of ***n*** objects into a set of ***k*** clusters.

- Given a *k*, find a partition of *k clusters* that optimizes the chosen partitioning criterion.
  – Global optimum: exhaustively enumerate all partitions.
  – Heuristic methods: *k-means* and *k-medoids* algorithms.
    - *k-means* (MacQueen'67): each cluster is represented by the center of the cluster.
    - *k-medoids* or PAM (Partition Around Medoids) (Kaufman & Rousseeuw'87): each cluster is represented by one of the objects in the cluster.

K-means Clustering

- Partitional clustering approach

- Each cluster is associated with a **centroid** (center point)
- Each point is assigned to the cluster with the closest centroid
- Number of clusters, K, must be specified

- The simple basic algorithm:

1: Select K points as the initial centroids.
2: **repeat**
3: Form K clusters by assigning all points to the closest centroid.
4: Recompute the centroid of each cluster.
5: **until** The centroids don't change

K-means Clustering – Details

- Initial centroids are often chosen randomly.
  - Clusters produced vary from one run to another.

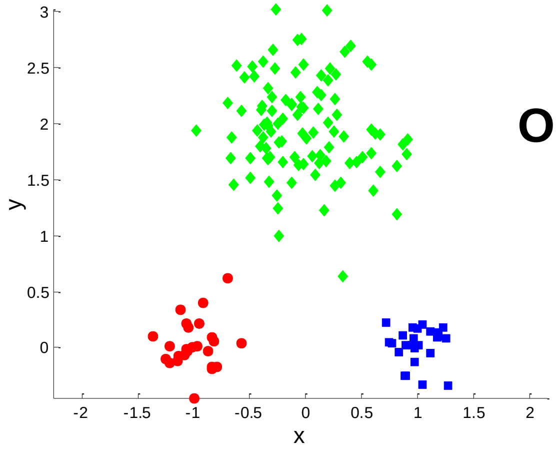
- The centroid is (typically) the mean of the points in the cluster.
- ‘Closeness’ is measured by Euclidean distance, cosine similarity, correlation, etc.

- K-means will converge for common similarity measures mentioned above.
- Most of the convergence happens in the first few iterations.
  - Often the stopping condition is changed to ‘Until relatively few points change cluster’

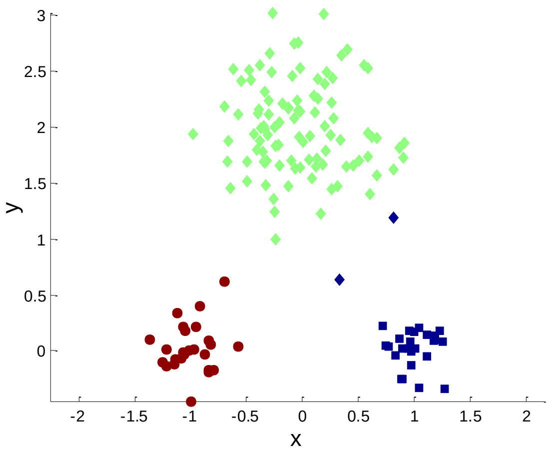
- Complexity is O(n * K * I * d)
  - n = number of points, K = number of clusters,
    I = number of iterations, d = number of attributes

Two different K-means Clusterings

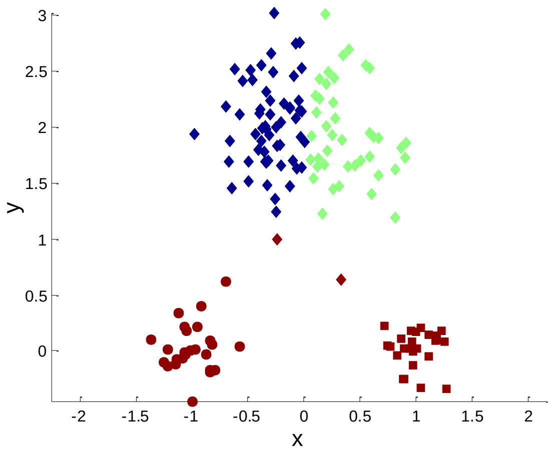
- Importance of choosing initial centroids



Original Points



Optimal Clustering



Sub-optimal Clustering

Evaluating K-means Clusters

- Most common measure is Sum of Squared Error (SSE)
  - For each point, the error is the distance to the nearest cluster
  - The objective function is:

$$SSE = \sum_{i=1}^{K} \sum_{x \in C_i} dist^2(m_i, x)$$

  - x is a data point in cluster C_i and m_i is the representative point for cluster C_i
    - it can be shown that m_i corresponds to the center (mean) of the cluster
  - Given two clusters, we can choose the one with the smallest error
  - One easy way to reduce SSE is to increase K, the number of clusters
    - A good clustering with smaller K can have a greater SSE than a bad clustering with higher K.

Solutions to Initial Centroids Problem

- Multiple runs
  - Helps, but probability is not on your side
- Sample and use hierarchical clustering to determine initial centroids
- Select more than k initial centroids and then select among these initial centroids
  - Select most widely separated
- Post-processing attempts to “fix-up” the clustering
- Bisecting K-means
  - Not as susceptible to initialization issues

Handling Empty Clusters

- Basic K-means algorithm can yield empty clusters

- Several strategies
  - Choose the point that contributes most to SSE
  - Choose a point from the cluster with the highest SSE
- If there are several empty clusters, the above can be repeated several times.

Pre-processing and Post-processing

- Pre-processing
  - Normalize the data
  - Eliminate outliers

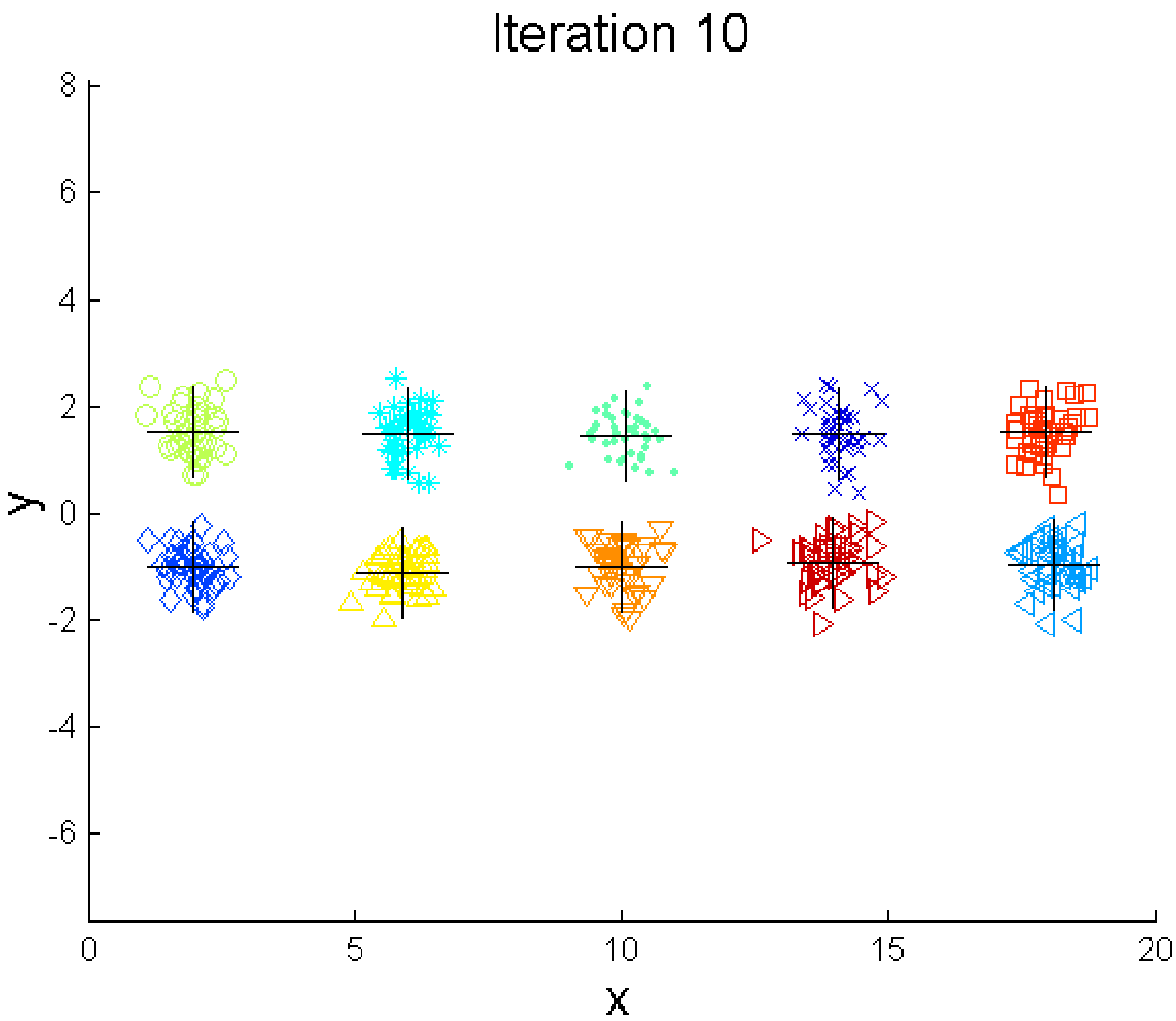
- Post-processing
  - eliminate “small” clusters since they may represent groups of outliers
  - split ‘loose’ clusters, i.e., clusters with relatively high SSE
  - merge clusters that are ‘close’ and that have relatively low SSE
  - can also use these steps <u>during</u> the clustering process
    - e.g., the ISODATA algorithm (Jensen, 1996)

Bisecting K-means

- Bisecting K-means algorithm
  - Variant of K-means that can produce a partitional or a hierarchical clustering

Initialize the list of clusters to contain the cluster containing all points.
repeat
    Select a cluster from the list of clusters
    for $i = 1$ to $number\_of\_iterations$ **do**
        Bisect the selected cluster using basic K-means
    end for
    Add the two clusters from the bisection with the lowest SSE to the list of clusters.
until Until the list of clusters contains K clusters

Bisecting K-means Example



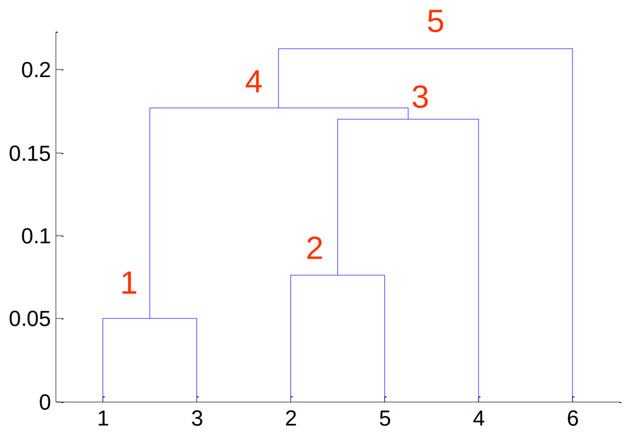
Other Limitations of K-means

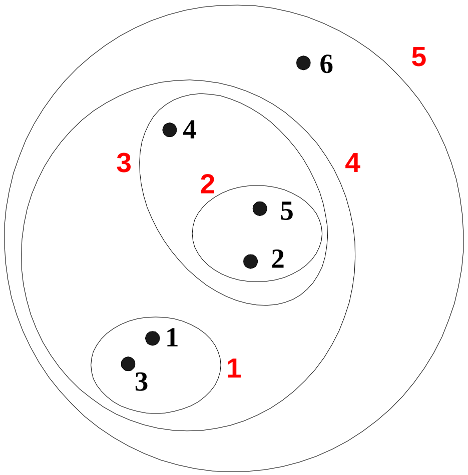
- K-means has problems when clusters are of differing
  - sizes
  - densities
  - non-globular shapes

- K-means has problems when the data contains outliers.

- Hierarchical clustering approach

- Produces a set of nested clusters organized as a hierarchical tree
- Can be visualized as a dendrogram
  - A tree like diagram that records the sequences of merges or splits

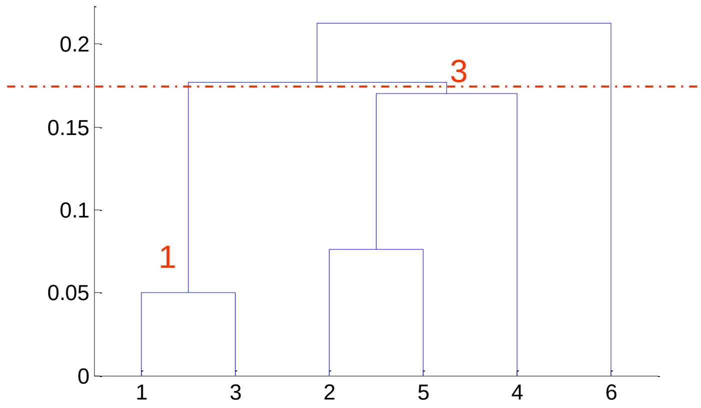


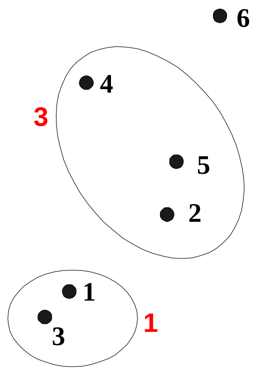


Strengths of Hierarchical Clustering

- Do not have to assume any particular number of clusters
  - Any desired number of clusters can be obtained by 'cutting' the dendogram at the proper level

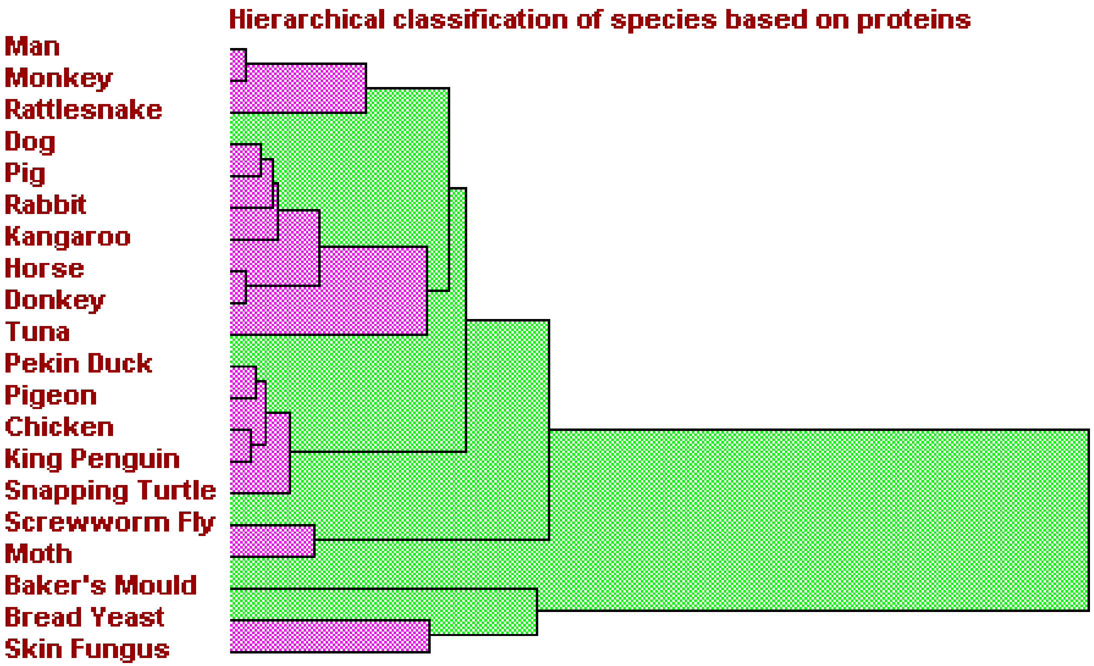
- They may correspond to meaningful taxonomies
  - Example in biological sciences (e.g., animal kingdom, phylogeny reconstruction, …)





Hierarchical Classification of Species

- Hierarchical clusters may correspond to meaningful taxonomies



Hierarchical Clustering

- Two main types of hierarchical clustering:
  - **Agglomerative**:
    - Start with the points as individual clusters
    - At each step, merge the closest pair of clusters until only one cluster (or k clusters) left
  - **Divisive**:
    - Start with one, all-inclusive cluster
    - At each step, split a cluster until each cluster contains a point (or there are k clusters)
- Traditional hierarchical algorithms use a similarity (proximity) or distance matrix
  - Merge or split one cluster at a time

Cluster Validity

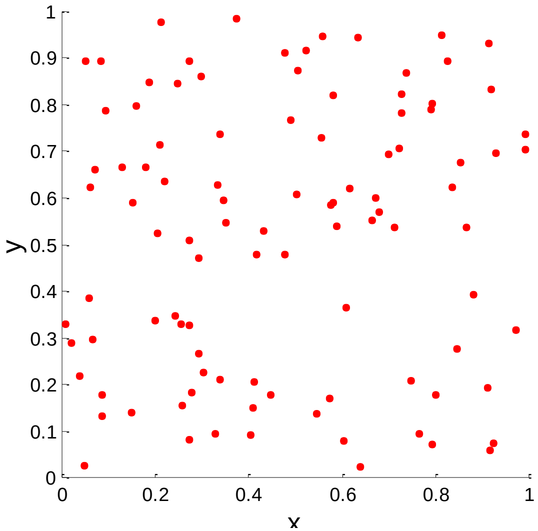
Cluster Validity

Clustering is an unsupervised learning task

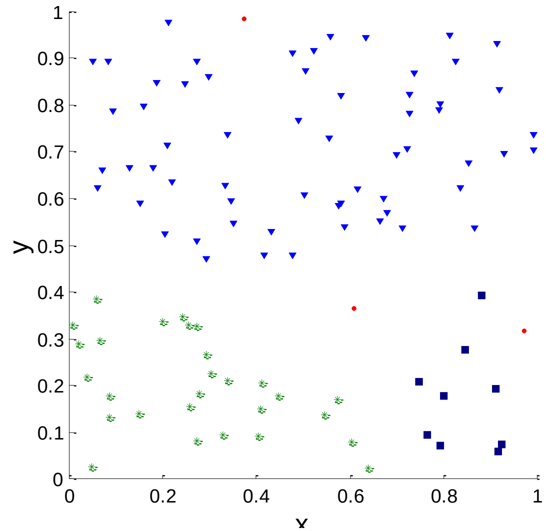
- For cluster analysis, the analogous question is how to evaluate the “goodness” of the resulting clusters?
- But “clusters are in the eye of the beholder”!
- Then why do we want to evaluate them?
  - To avoid finding patterns in noise
  - To compare clustering algorithms
  - To compare two clusters
  - To compare two sets of clusters

Clusters found in Random Data

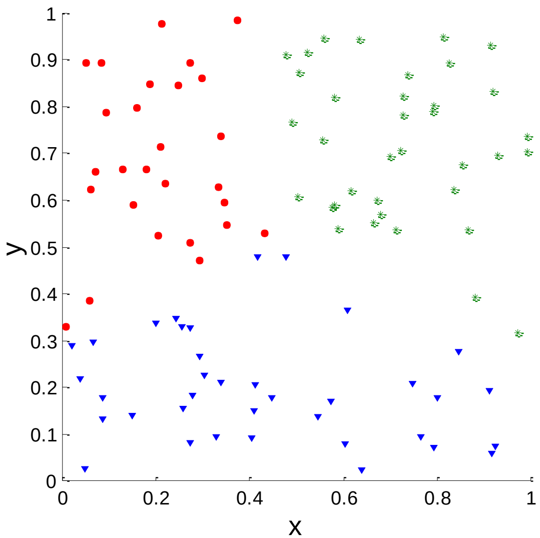
Random Points



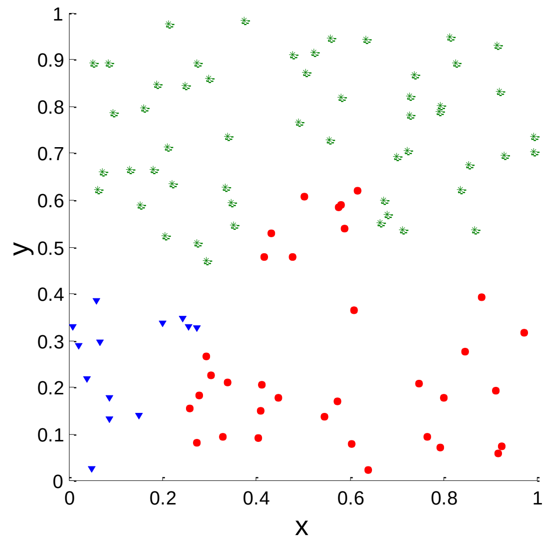
DBSCAN



K-means



Complete Link



Measures of Cluster Validity

- Numerical measures that are applied to judge various aspects of cluster validity, are classified into the following three types.

  – External Index: Used to measure the extent to which cluster labels match externally supplied class labels.
    - Entropy

  – Internal Index: Used to measure the goodness of a clustering structure *without* respect to external information.
    - Sum of Squared Error (SSE)

  – Relative Index: Used to compare two different clusterings or clusters.
    - Often an external or internal index is used for this function, e.g., SSE or entropy

- Sometimes these are referred to as criteria instead of indices
  – However, sometimes criterion is the general strategy and index is the numerical measure that implements the criterion.

Internal Measures: Cohesion and Separation

- Cluster Cohesion: Measures how closely related are objects in a cluster
  - Example: SSE
- Cluster Separation: Measure how distinct or well-separated a cluster is from other clusters
- Example: Squared Error
  - Cohesion is measured by the **W**ithin cluster **S**um of **S**quares (SSE)

$$WSS = \sum_i \sum_{x \in C_i} (x - m_i)^2$$

  - Separation is measured by the **B**etween cluster **S**um of **S**quares

$$BSS = \sum_i |C_i| (m - m_i)^2$$

    - Where $|C_i|$ is the size of cluster i, m_i is the cluster mean, m is the overall mean.

External Measures of Cluster Validity: Entropy and Purity

Entropy: $$H = -\sum_i p_i \log(p_i)$$

Table 5.9. K-means Clustering Results for LA Document Data Set

Cluster	Entertainment	Financial	Foreign	Metro	National	Sports	Entropy	Purity
1	3	5	40	506	96	27	1.2270	0.7474
2	4	7	280	29	39	2	1.1472	0.7756
3	1	1	1	7	4	671	0.1813	0.9796
4	10	162	3	119	73	2	1.7487	0.4390
5	331	22	5	70	13	23	1.3976	0.7134
6	5	358	12	212	48	13	1.5523	0.5525
Total	354	555	341	943	273	738	1.1450	0.7203

entropy For each cluster, the class distribution of the data is calculated first, i.e., for cluster j we compute p_{ij}, the 'probability' that a member of cluster j belongs to class i as follows: $p_{ij} = m_{ij}/m_j$, where m_j is the number of values in cluster j and m_{ij} is the number of values of class i in cluster j. Then using this class distribution, the entropy of each cluster j is calculated using the standard formula $e_j = \sum_{i=1}^{L} p_{ij} \log_2 p_{ij}$, where the L is the number of classes. The total entropy for a set of clusters is calculated as the sum of the entropies of each cluster weighted by the size of each cluster, i.e., $e = \sum_{i=1}^{K} \frac{m_i}{m} e_j$, where m_j is the size of cluster j, K is the number of clusters, and m is the total number of data points.

purity Using the terminology derived for entropy, the purity of cluster j, is given by $purity_j = \max p_{ij}$ and the overall purity of a clustering by $purity = \sum_{i=1}^{K} \frac{m_i}{m} purity_j$.

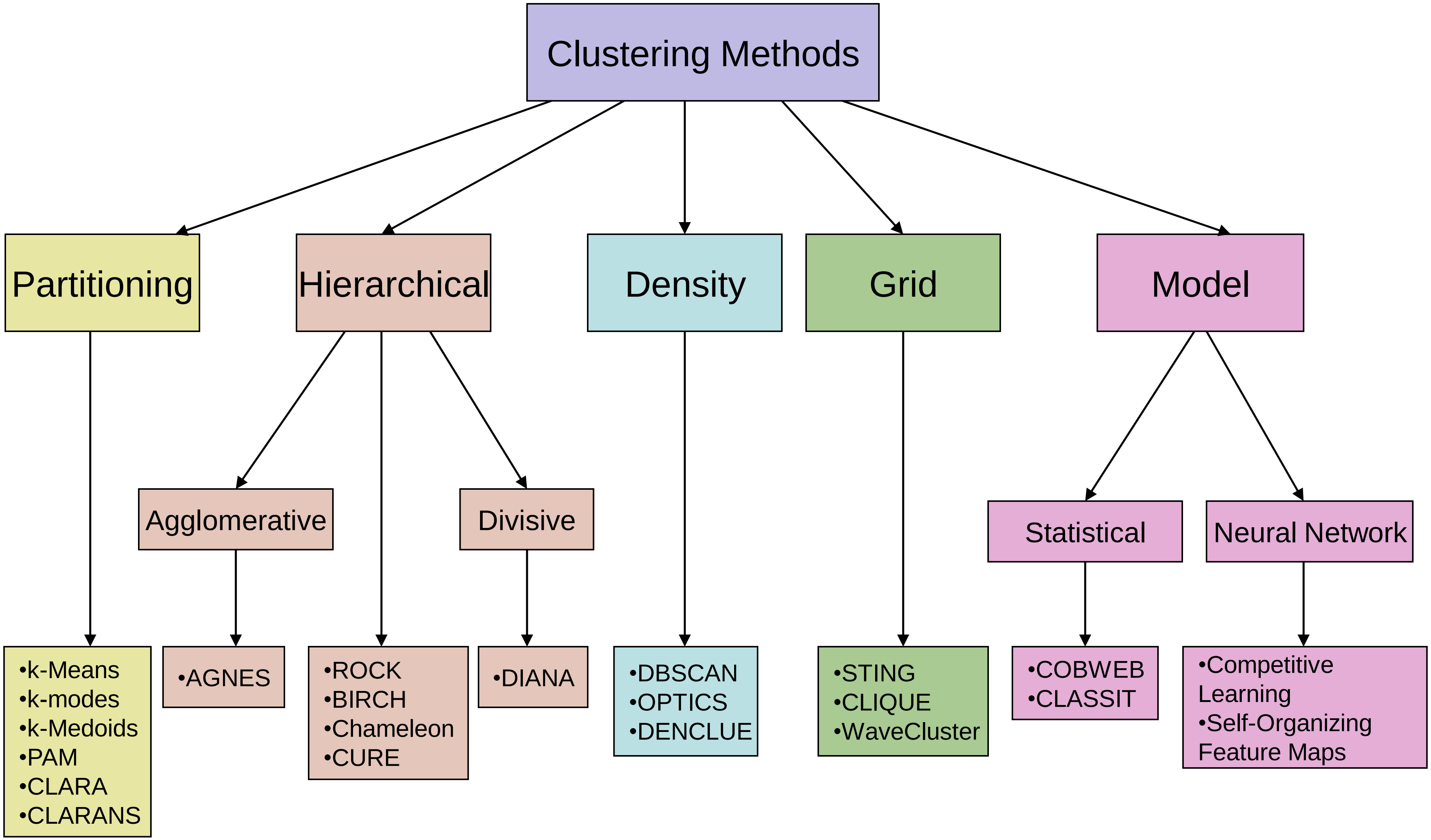
Final Comment on Cluster Validity

"The validation of clustering structures is the most difficult and frustrating part of cluster analysis.

Without a strong effort in this direction, cluster analysis will remain a black art accessible only to those true believers who have experience and great courage."

From "Algorithms for Clustering Data", Jain and Dubes

Taxonomy


Clustering Methods
Partitioning
Hierarchical
Density
Grid
Model
Agglomerative
Divisive
Statistical
Neural Network
•k-Means
•k-modes
•k-Medoids
•PAM
•CLARA
•CLARANS
•AGNES
•ROCK
•BIRCH
•Chameleon
•CURE
•DIANA
•DBSCAN
•OPTICS
•DENCLUE
•STING
•CLIQUE
•WaveCluster
•COBWEB
•CLASSIT
•Competitive Learning
•Self-Organizing Feature Maps

Summary

- Cluster Analysis
- Similarity and Metrics
- 6 Types of Clusters
- 5 Clustering Approaches
- Cluster Validity
  - Cohesion and Separation
  - Entropy and Purity