Clustering

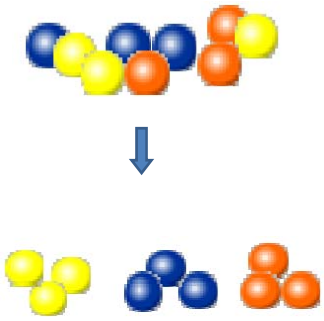
Clustering

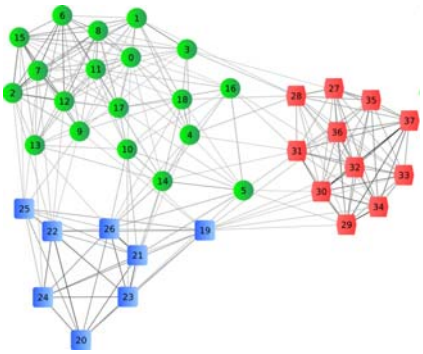
Grouping similar objects

- People with common interests
- Proteins with equal/similar functions
- Web pages in the same topic
- …

Clustering

- Data clustering vs. Graph clustering





3

Data Clustering – Review

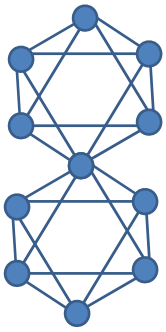
- Find relationships and patterns in the data
- Get insights in underlying characteristics
- Find groups of “similar” genes/proteins/people

- Deal with numerical values
- They have many features (not just color)

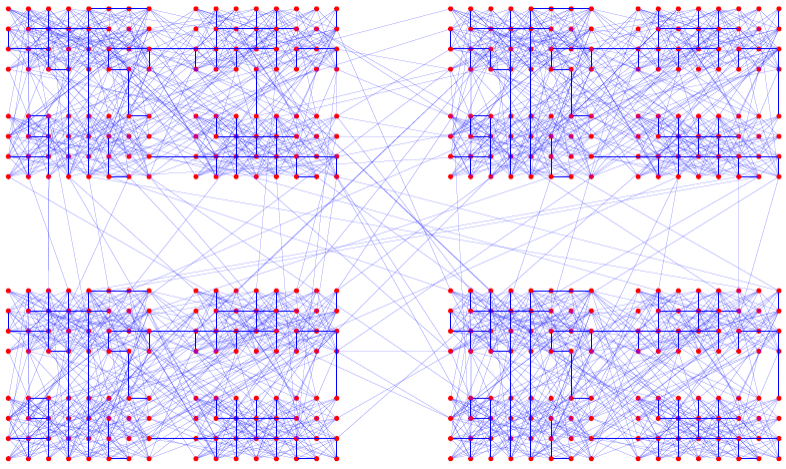
4

Clustering Issues

- Soft vs. hard



- Known vs. unknown number of clusters
- Hierarchy



Sixteen modules with 32 vertices each clearly form four larger clusters.
All vertices have degree 64.

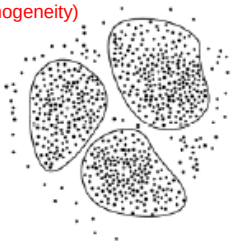
[Lancichinetti, Fortunato, and Kertesz, New J. of Phys., 2009]

Data Clustering – Review

Clustering is a process of partitioning a set of data (or objects) in a set of meaningful sub-classes, called **clusters**.

- Helps users understand the natural grouping or structure in a data set.

- Cluster: a collection of data objects that are "similar" to one another and thus can be treated collectively as one group.
- A good clustering method produces high quality clusters in which:
  - The **intra-class** (that is, **intra-cluster**) similarity is high. (homogeneity)
  - The **inter-class** similarity is low. (separation)
- The **quality** of a clustering result depends on both the similarity measure used and its implementation.
- Clustering = function that maximizes similarity between objects within a cluster and minimizes similarity between objects in different clusters.



7

Distance Metrics

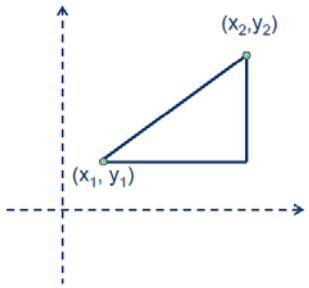
- There are many possible distance metrics between objects
- Theoretical properties of **distance metrics**, d:
  - d(a,b) >= 0
  - d(a,a) = 0
  - d(a,b) = 0 ⇒ a=b
  - d(a,b) = d(b,a) – symmetry
  - d(a,c) <= d(a,b) + d(b,c) – triangle inequality

8

Distance Metrics

Example distances:

- Euclidean (L_2) distance
- Manhattan (L_1) distance
- L_m: $(|x_1-x_2|^m+|y_1-y_2|^m)^{1/m}$
- L_∞: $\max(|x_1-x_2|,|y_1-y_2|)$
- Inner product: $x_1x_2+y_1y_2$
- Correlation coefficient
- For simplicity, we will concentrate on Euclidean distance



9

Similarity Matrix

Distance/Similarity matrices:

- Clustering is based on distances – distance/similarity matrix:
  - Represents the distance between objects
  - Only need half the matrix, since it is symmetric

$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \vdots & \vdots & \vdots & & \\ d(n,1) & d(n,2) & \ldots & \ldots & 0 \end{bmatrix}$$

Hierarchical Clustering

Hierarchical Clustering:

1. Scan the distance matrix for the minimum
2. Join items into one node
3. Update the matrix and repeat from step 1

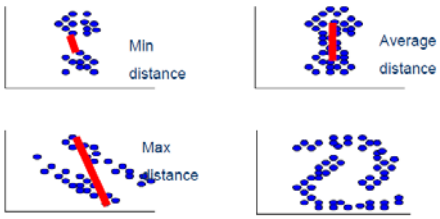
11

Hierarchical Clustering

Hierarchical Clustering:

Distance between two points – easy to compute

Distance between two clusters – harder to compute:
1. Single-Link Method / Nearest Neighbor
2. Complete-Link / Furthest Neighbor
3. Average of all cross-cluster pairs

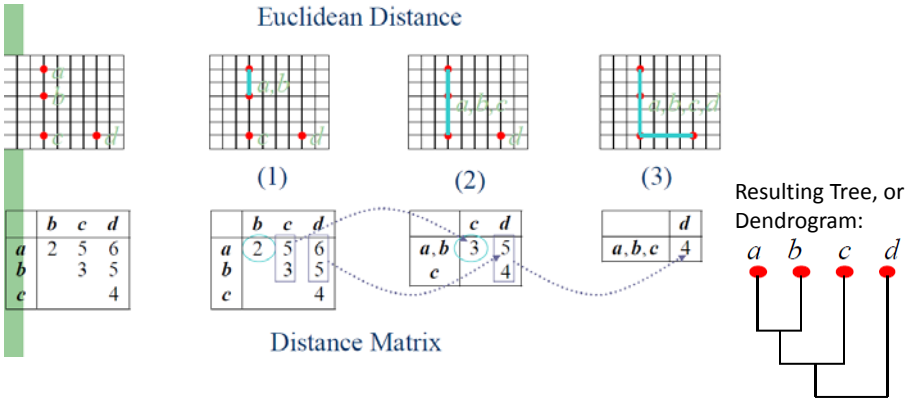


12

Hierarchical Clustering

Hierarchical Clustering:

1. **Example: Single-Link (Minimum) Method**:

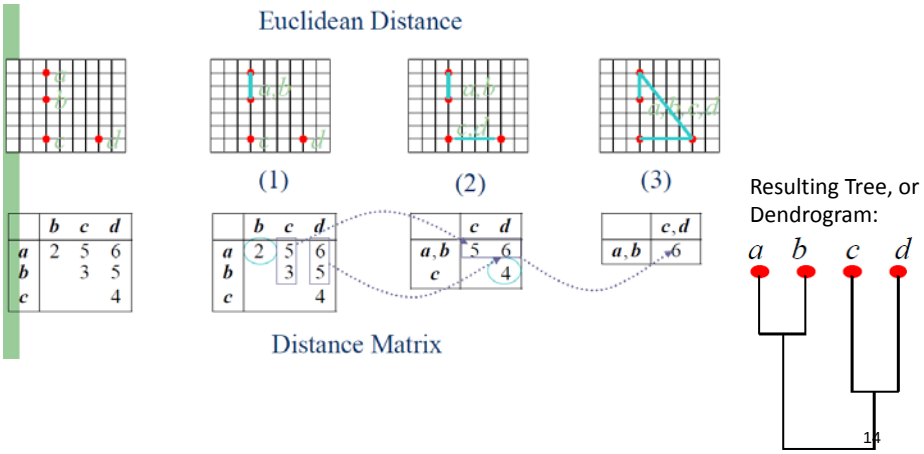


13

Hierarchical Clustering

Hierarchical Clustering:

2. **Example: Complete-Link (Maximum) Method**:



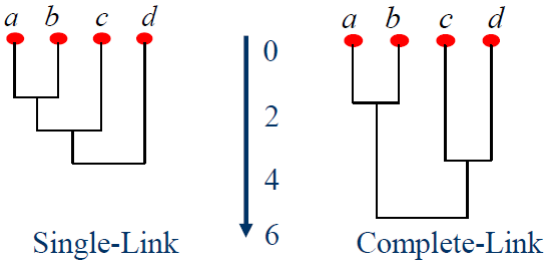
14

Hierarchical Clustering

Hierarchical Clustering:

In a dendrogram, the *length of each tree branch* represents the distance between clusters it joins

Different dendrograms may arise when different linkage methods are used



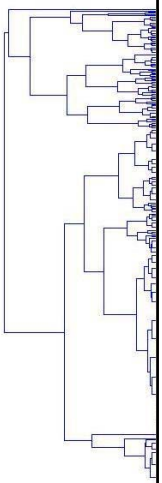
15

Hierarchical Clustering

Hierarchical Clustering:

How do you get clusters from the tree?

Where to cut the tree?



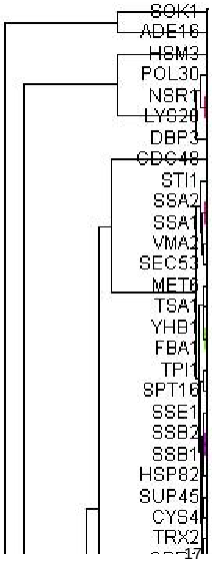
16

Hierarchical Clustering

Hierarchical Clustering:

How do you get clusters from the tree?

Where to cut the tree?



K-Means

K-Means Clustering:

- Basic idea: use cluster centroids (means) to represent cluster
- Assigning data elements to the closest cluster (centroid)
- Goal: minimize intra-cluster dissimilarity

18

K-Means

K-Means Clustering:

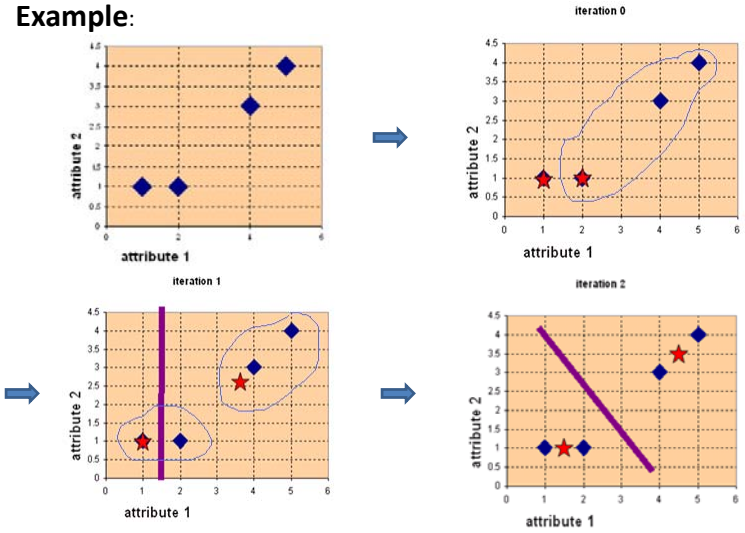
1) Pick K objects as centers of K clusters and assign all the remaining objects to these centers
   - Each object will be assigned to the center that has minimal distance to it
   - Solve any ties randomly
2) In each cluster C, find a new center X so as to minimize the total sum of distances between X and all other objects in C
3) Reassign all objects to new centers as explained in step (1)
4) Repeat the previous two steps until the algorithm converges

19

K-Means

K-Means Clustering

Example:



20

Clustering Algorithms – Review

- **Differences between the two clustering algorithms:**
  - Hierarchical Clustering:
    - Need to select Linkage Method
    - To perform any analysis, it is necessary to partition the dendrogram into *k* disjoint clusters, cutting the dendrogram at some point. A limitation is that it is not clear how to choose this *k*
  - *K*-means
    - Need to select *K*
  - In both cases: Need to select distance/similarity measure
- **K-medoids**
  - Centers are data points
- Hierarchical and k-means clust. implemented in Matlab

21

Nearest Neighbours Clustering

Nearest neighbours "clustering:"

Input:
D={t_1, t_2, …, t_n} // Set of elements
A // matrix showing distance between elements
θ // threshold
Output:
K //Set of k clusters
Nearest-Neighbor algorithm
K_1 ={t_1}; add K_1 to K; // t1 initialized the first cluster
k = 1;
for i =2 to n **do** // for t_2 to t_n add to existing cluster or place in new one
find the t_m in some cluster K_m in K such that d(t_m,t_i) is the smallest;
if d(t_m,t_i) < θ **then**
$K_m = K_m \cup \{t_i\}$ // existing cluster
else
k = k + 1; K_k = {t_i}; add K_k to K // new cluster

22

Nearest Neighbours Clustering

Nearest neighbours "clustering:"

```
Input:
        D={t1, t2, ..., tn}    // Set of elements
        A                      // matrix showing distance between elements
        θ                      // threshold
Output:
        K                      //Set of k clusters
Nearest-Neighbor algorithm
        K1 ={t1}; add K1 to K;         // t1 initialized the first cluster
        k = 1;
        for i =2 to n do       // for t2 to tn add to existing cluster or place in new one
                find the tm in some cluster Km in K such that d(tm,ti) is the smallest;
                if d(tm,ti) < θ then
                        Km = Km U {ti}                    // existing cluster
                else
                        k = k + 1; Kk = {ti}; add Kk to K         // new cluster
```

Example:

- Given: 5 items with the distance between them
- Task: Cluster them using the Nearest Neighbor algorithm with a threshold θ =2

Item	A	B	C	D	E
A	0	1	2	2	3
B		0	2	4	3
C			0	1	5
D				0	3
E					0

-A: K1={A}
-B: d(B,A)=1< θ => K1={A,B}
-C: d(C,A}=d(C,B)=2≤ θ =>K1={A,B,C}
-D: d(D,A)=2, d(D,B)=4, d(D,C)=1 =dmin ≤ θ => K1={A,B,C,D}
-E: d(E,A)=3, d(E,B)=3, d(E, C)=5, d(E, D)=3=dmin> θ =>K2={E}

Pros and cons:

1. No need to know the number of clusters to discover beforehand (different than in k-means and hierarchical).
2. We need to define the threshold θ.

23

k-NN Clustering

k*-nearest neighbors "clustering"** -- ***classification algorithm, but we use the idea here to do **clustering**:

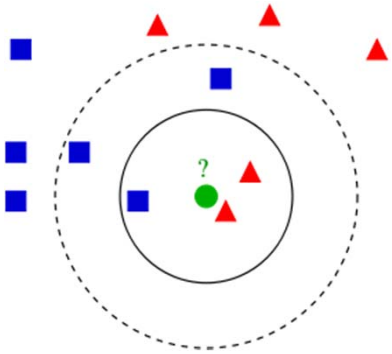
- For point *v*, create the cluster containing *v* and top *k* closest points to *v*,
  e.g., based on Euclidean distance.
- Do this for all points *v*.
- All of the clusters are of size *k*, but they can overlap.

- The challenge: choosing k.

24

k-Nearest Neighbours (k-NN) Classification

- An object is classified by a majority vote of its neighbors
  - It is assigned to the **class** most common amongst its *k* nearest neighbors



Example:
- The test sample (green circle) should be classified either to the first class of blue squares or to the second class of red triangles.
- If *k* = *3* it is classified to the second class (2 triangles vs only 1 square).
- If *k* = *5* it is classified to the first class (3 squares vs. 2 triangles).

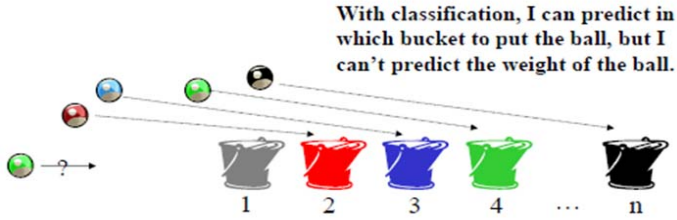
25

What is Classification?

The goal of data classification is to organize and categorize data into distinct classes.

- A model is first created based on the training data (learning).
- The model is then validated on the testing data.
- Finally, the model is used to classify new data.
- Given the model, a class can be predicted for new data.

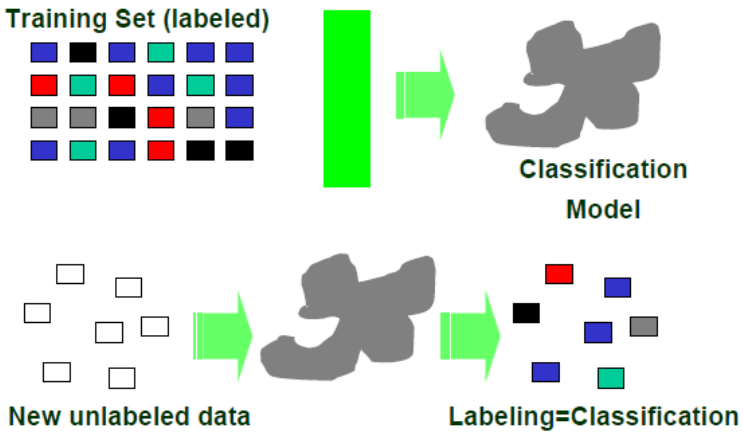
Example:



Application: medical diagnosis, treatment effectiveness analysis, protein function prediction, interaction prediction, etc.

26

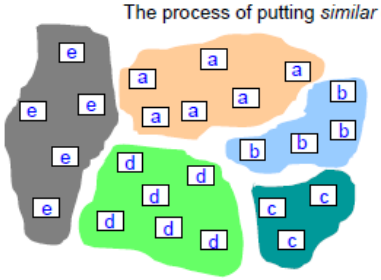
Classification = Learning the Model



27

What is Clustering?

The process of putting *similar* data together.



Grouping
Clustering
Partitioning

- There is no training data (objects are not labeled)
- We need a notion of similarity or distance
- Should we know a priori how many clusters exist?

28

Supervised and Unsupervised

Classification = Supervised approach

- We know the class labels and the number of classes

Clustering = Unsupervised approach

- We do not know the class labels and may not know the number of classes

29

Classification vs. Clustering

Classification	Clustering
• known number of classes	• unknown number of classes
• based on a training set	• no prior knowledge
• used to classify future observations	• used to understand (explore) data
• Classification is a form of supervised learning	• Clustering a form of unsupervised learning

(we can compute it without the need of knowing the correct solution)

30