Clustering

Types of Data

Data Matrix

$$\begin{bmatrix} x_{11} & \cdots & x_{1f} & \cdots & x_{1p} \\ \cdot & & \cdot & & \cdot \\ x_{i1} & \cdots & x_{jf} & \cdots & x_{ip} \\ \cdot & & \cdot & & \cdot \\ x_{n1} & \cdots & x_{nf} & \cdots & x_{np} \end{bmatrix}$$

Dissimilarity Matrix

$$\begin{bmatrix} 0 & & & \\ d_{(2,1)} & 0 & & \\ d_{(3,1)} & d_{(3,2)} & 0 & \\ \cdot & \cdot & \cdot & \\ d_{(n,1)} & d_{(n,2)} & \cdots & 0 \end{bmatrix}$$

d(i, j) – difference or dissimilarity between objects x1, x2, …..,xn

Clustering Techniques

- Partitioning methods
- Hierarchical methods
- Density-based

Interval Scaled Variables

- Standardization
  - mean absolute deviation s_f

    $s_f = 1/n (|x_{1f} - m_f| + |x_{2f} - m_f| + ... + |x_{nf} - m_f|)$
  - standardized (normalized) measurement

    $z_{if} = [x_{1f} - m_f] / s_f$

- Compute dissimilarity between objects
  - Euclidean distance

    $$d(i, j) = \sqrt{|x_{i1} - x_{j1}|^2 + |x_{i2} - x_{j1}|^2 + ... + |x_{ip} - x_{jp}|^2}$$
  - Manhattan (city-block) distance

    $$d(i, j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j1}| + ... + |x_{ip} - x_{jp}|$$
  - Minkowski distance

    $$d(i, j) = (|x_{i1} - x_{j1}|^p + |x_{i2} - x_{j1}|^p + ... + |x_{ip} - x_{jp}|^p)^{1/p}$$

Binary Variables (attributes)

There are only two states: 0 (absent) or 1 (present). Ex. *smoker*: yes or no.
Computing dissimilarity between binary variables (attributes):

Dissimilarity matrix (contingency table) if all attributes have the same weight

Object j \ Object i	1	0	Sum
1	q	r	q+r
0	s	t	s+t
Sum	q+s	r+t	p

d(i, j) =[r+s]/[q+r+s]
asymmetric attributes

d(i, j) = [r+s] /[q+r+s+t]
symmetric attributes

i	1	0	0	1	0
j	1	1	1	1	0

q=2, t=1, r=2, s=0

d(I,j)=2+2+0+1
d(I,j)=2+2+0+0 (if last asymmetric)

Example asymmetric:
(1) positive medical test result (2) 1 – smoking, 0 – non-smoking

Nominal Variables

Generalization of binary variable where it can take more than two states. Ex. *color*: red, green, blue.

d(i, j) = [p – m] / p

m – number of matches

p – total number of attributes

Ordinal Variables

Resemble nominal variables except the states are ordered in meaningful sequence. Ex. *medal*: gold, silver, bronze.

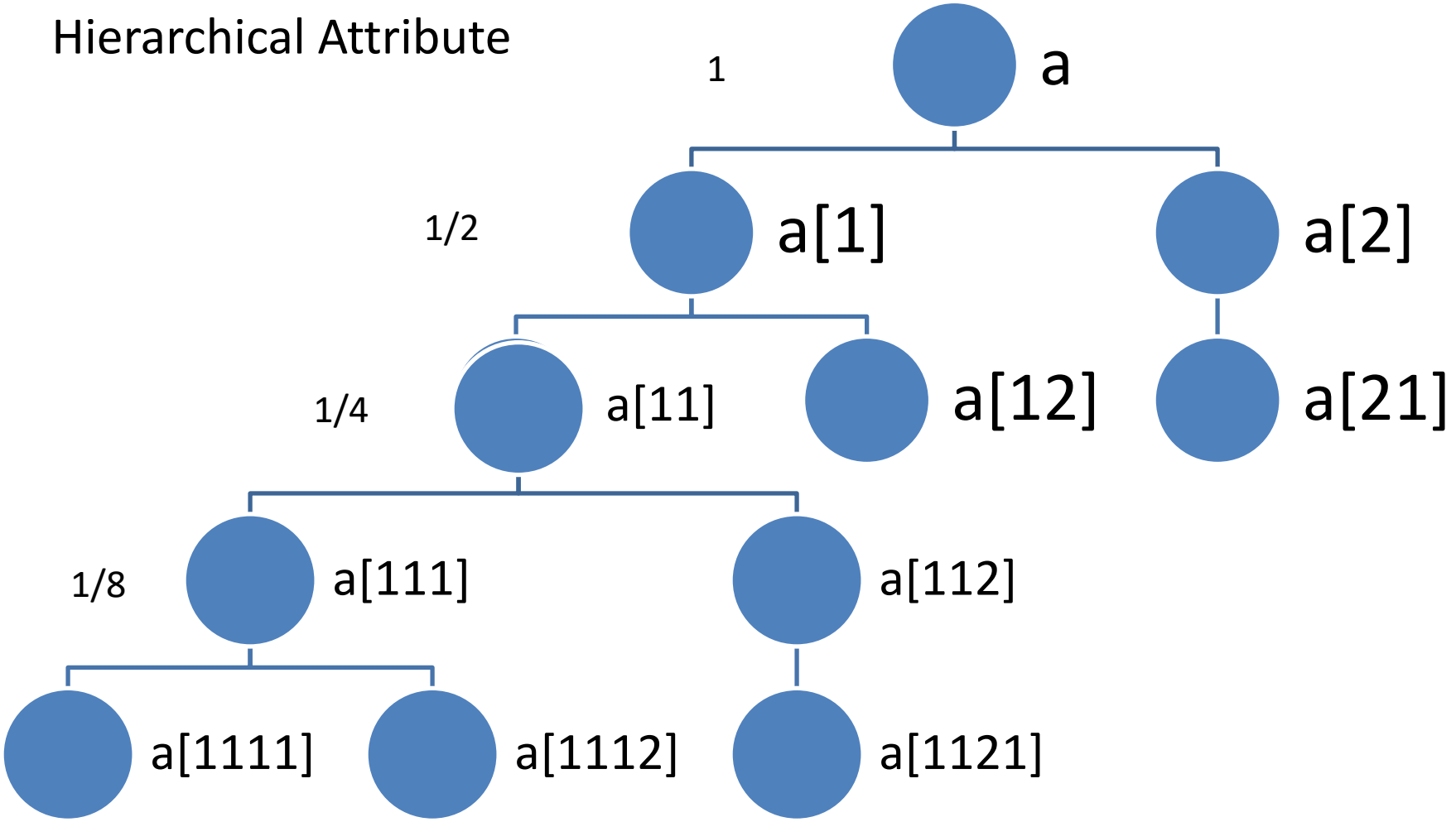
Replace x_{if} by

$$r_{if} \in \{1, \ldots, M_f\}$$

The value of *f* for the *i*th object is x_{if}, and *f* has M_f ordered states, representing the ranking 1, …, M_f. Replace each x_{if} by its corresponding rank.

Example: d(gold, bronze)=1, d(gold,silver)=d(silver,bronze)= 1/2

Hierarchical Attribute



a[1] can be either a[11] or a[12]

d(a[111],a[1121])=1/4, d(a[1],a[11])= [d(a[11],a[11])+d(a[11],a[12])]/2= [½ + ¼]/2=3/8

Clustering Methods

1. Partitioning (k-number of clusters)
2. Hierarchical (hierarchical decomposition of objects)

TV – trees of order k (k=2)

Given: set of N – vectors (based on a concept of active dimension)

Goal: divide these points into maximum I disjoint clusters so that points in each cluster are similar with respect to maximal number of coordinates (called active dimensions).

TV-tree of order 2: (two clusters per node)

Procedure:

Divide set of N points into 2 clusters maximizing the total number of active dimensions. For each cluster repeat the same procedure.

Density-based methods

Can find clusters of arbitrary shape. Can grow (given cluster) as long as density in the neighborhood exceeds some threshold (for each point, neighborhood of given radius contains minimum some number of points).

Partitioning methods

1. K-means method (n objects to k clusters)
Cluster similarity measured in regard to mean value of objects in a cluster (cluster's center of gravity)

- Select randomly k-points (call them means)
- Assign each object to nearest mean
- Compute new mean for each cluster
- Repeat until criterion function converges

$$E = \sum_{i=1}^{K} \sum_{p \in Ci} | p - m_i |^2$$

Squared error criterion
We try to minimize

This method is sensitive to outliers.

EXAMPLE

2. K-medoids method
Instead of mean, take a medoid (most centrally located object in a cluster)

Hierarchical Methods

Agglomerative hierarchical clustering (bottom-up strategy)

Each object placed in a separate cluster, and then we merge these clusters until certain termination conditions are satisfied.

Divisive hierarchical clustering (top-down strategy)

Distance between clusters (next slide):

- Minimum distance: $d_{min}(C_i, C_j) = min_{p \in Ci,\, p' \in Cj} \mid p - p' \mid$
- Maximum distance: $d_{max}(C_i, C_j) = max_{p \in Ci,\, p' \in Cj} \mid p - p' \mid$
- Mean distance: $d_{mean}(C_i, C_j) = \mid m_i - m_j \mid$
- Average distance: $d_{avg}(C_i, C_j) = 1/n_i n_j \sum_{p \in Ci} \sum_{p' \in Cj} \mid p - p' \mid$

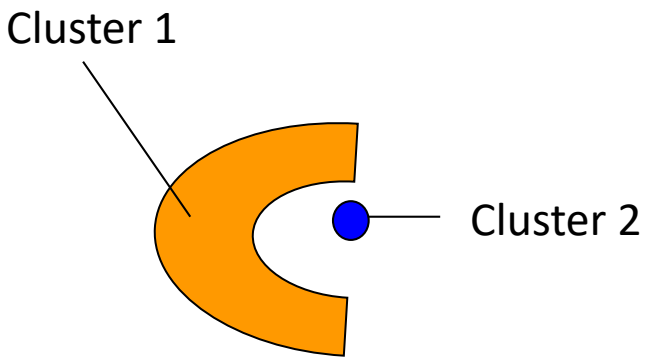
Cluster 1 Cluster 2

Cluster: $K_m = \{t_{m1}, \ldots, t_{mn}\}$

Centroid: $C_m = \sqrt{\sum_{i=1}^{N} t_{mi}} / N$

Radius: $R_m = \sqrt{\sum_{i=1}^{N} (t_{mi} - C_m)^2} / N$

Diameter: $D_m = \sum_{i=1}^{N} \sum_{j=1}^{N} (t_{mi} - t_{mj})^2 / N (N-1)$

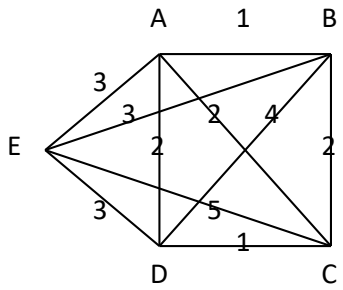


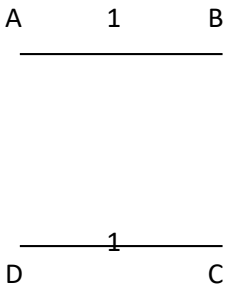
Hierarchical Algorithms

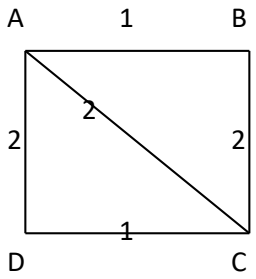
Single Link Technique

(find maximal connected components in a graph)

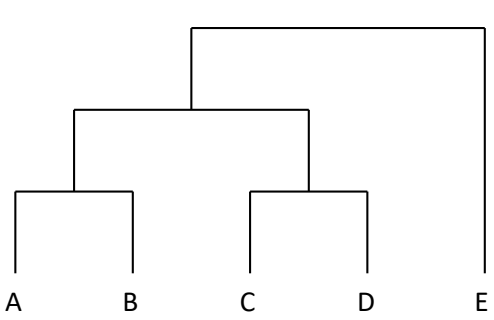
Distances (threshold)

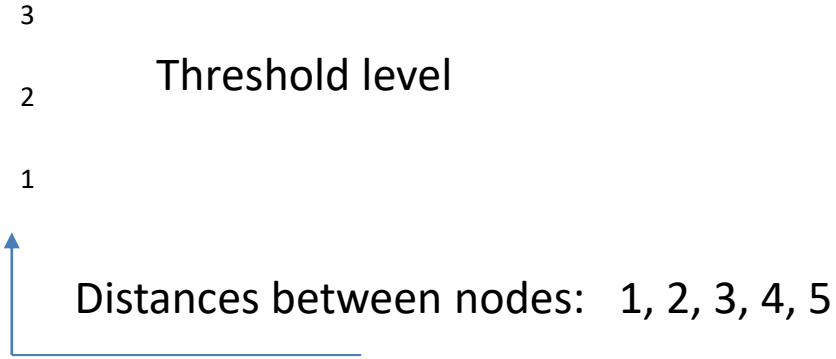






Dendogram



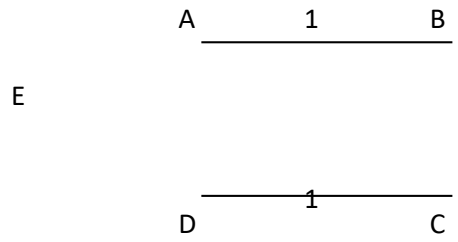


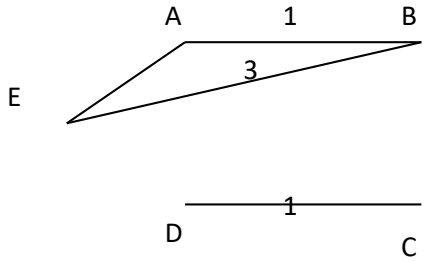
Complete Link Technique

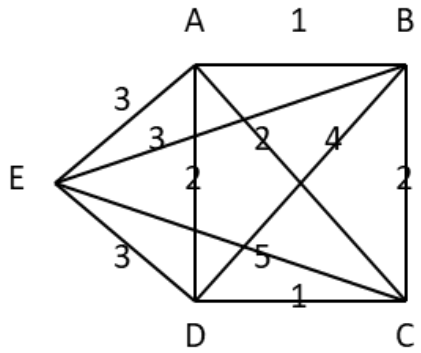
(looks for cliques – maximal graphs in which there is an edge between any two vertices)

Distances (threshold)

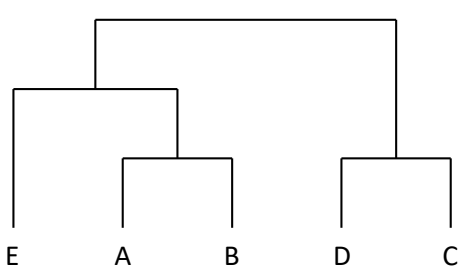
1 **2** **3** **4** ….







Dendogram



5

3

1

(5, {EABCD})

(3, {EAB}, {DC})

(1, {AB}, {DC}, {E})

(0, {E}, {A}, {B}, {C}, {D})

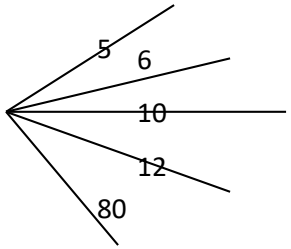
Partitioning Algorithms

Minimum Spanning Tree (MST)

Given: n – points
k – clusters

Algorithm:

- Start with complete graph
- Remove largest inconsistent edge (its weight is much larger than average weight of all adjacent edges)



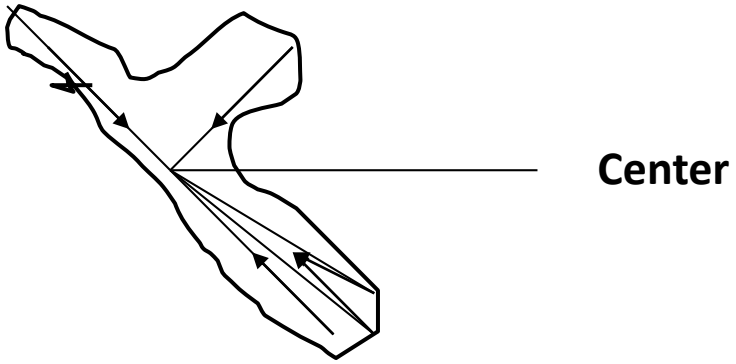
5+6+10+12+80=113

113/5 = 20.26

- Repeat

CURE (Clustering Using Representatives)

Idea: handling clusters of different shapes



Algorithm:

- Constant number of points are chosen from each cluster
- These points are shrunk toward the cluster's centroid
- Clusters with closest pair of representative points are merged