Hierarchical Clustering

Vijayasaradhi

Winter School on Deep Learning for Vision and Language Modelling
Department of CSE
IIT Guwahati

Introduction

1

Group by department

roll_number	department	pincode	CPI
CS123	CSE	781039	8.7
CS124	CSE	781040	8.2
CS125	CSE	781039	8.8
CS126	CSE	781040	9.1
EE123	EEE	781039	8.1
EE124	EEE	781040	7.2
EE125	EEE	781039	8.8
EE126	EEE	781040	9.2

1

Group by pincode

roll_number	department	pincode	CPI
CS123	CSE	781039	8.7
CS125	CSE	781039	8.8
EE123	EEE	781039	8.1
EE125	EEE	781039	8.8
CS124	CSE	781040	8.2
CS126	CSE	781040	9.1
EE124	EEE	781040	7.2
EE126	EEE	781040	9.2

Introduction 03

1. Data is structured
2. Grouping criteria is unambiguous
3. Obtained result segregate data into groups
4. Number of groups is equal to number of distinct values of the grouped attributed

Introduction 03

1. Many applications, grouping criteria is not fixed
2. One has to experiment with to obtain the criteria
3. Number of groups is not known in advance given criteria is not fixed
4. How to obtain groups given the criteria (algorithm) is also not unique

Clustering Gene Expression Data

Clustering Gene Expression Data 01

1. Microarray data is the numerical output obtained from microarray experiments.
2. This data typically includes Gene expression levels: How actively genes are being transcribed in a given sample.
3. Samples: Biological sources (e.g., tissues, cells) under different conditions.

Clustering Gene Expression Data 02

1. The data is presented as a matrix
2. Rows represent genes.
3. Columns represent samples or experimental conditions.
4. Each cell contains the expression value of a gene in a specific sample.

Clustering Gene Expression Data 03

gene ID	Sample 1	Sample 2	Sample 3
Gene A	10	50	40
Gene B	20	25	30
Gene C	5	100	95
Gene D	15	10	12

1

A gene expression dataset is represented by a real-valued expression matrix

$$E = \begin{pmatrix} e_{11} & e_{12} & \cdots & e_{1d} \\ e_{21} & e_{22} & \cdots & e_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ e_{n1} & e_{n2} & \cdots & e_{nd} \end{pmatrix}$$

; where n is the number of genes and d is the number of samples

Clustering Gene Expression Data 05

1. The rows $\mathbf{g}_i = (e_{i1}, e_{i2}, \cdots, e_{id})$ form the expression patterns of genes
2. The columns $\mathbf{s}_j = (e_{1s}, e_{2s}, \cdots, e_{ds})^T$ form the expression profiles of samples
3. Component e_{ij} represents the measured expression level of i^{th} gene in the j^{th} sample

Clustering Gene Expression Data 06

1. In gene expression data clustering, it is meaningful to cluster genes (rows) as well as samples (columns)
2. In gene-based clustering, genes are treated as objects; samples are treated as features
3. Genes are partitioned into homogeneous groups

Clustering Gene Expression Data 07

1. In sample-based clustering, sample are treated as objects
2. Genes are treated as features.
3. Samples are partitioned into homogeneous groups

Clustering Gene Expression Data 08

1. In gene-based clustering, the goal is to identify coexpressed genes that indicate cofunction and coregulation
2. Challenges in gene-based clustering are
3. Determination of the true number of clusters in the dataset
4. Capability of handling a high level of noise arising from the complex microarray experiments
5. Representation of cluster structures

Clustering Gene Expression Data 09

1. In sample-based clustering, the goal is to identify the phenotype structures or substructures of the sample
2. Genes whose expression levels strongly correlate with the cluster distinction are referred as informative genes
3. That is: Informative genes are those whose expression levels are significantly different across clusters
4. They contribute the most to defining or separating clusters and are essential for the task of clustering

1

Suppose we have two clusters of biological samples (e.g., healthy and diseased) and their gene expression data for three genes: Gene A, Gene B, and Gene C

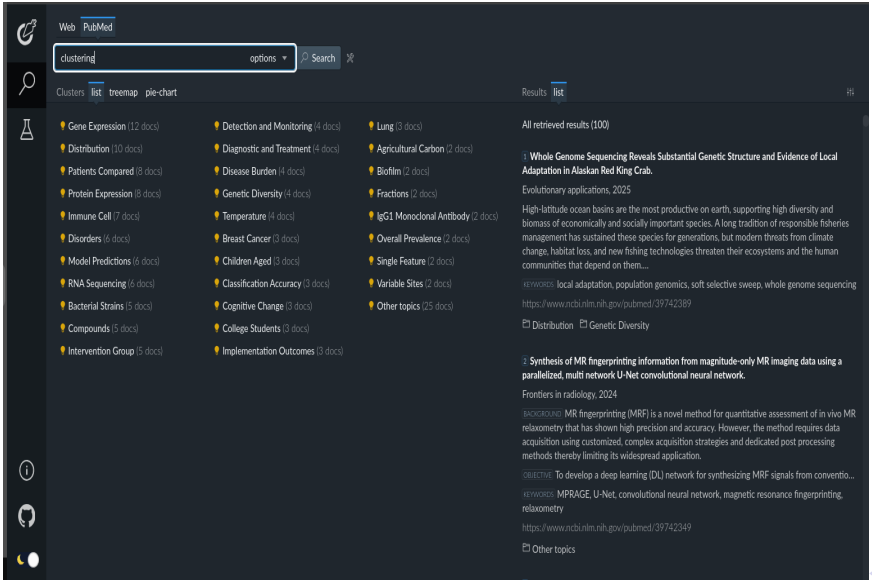
Sample ID	Cluster	Gene A	Gene B	Gene C
1	Healthy	10	5	15
2	Healthy	12	6	14
3	Healthy	11	4	15
4	Diseased	50	5	14
5	Diseased	48	6	13
6	Diseased	49	5	14

Webpages clustering

Introduction

1. Organizing search results into clusters aids quick browsing.
2. Traditional clustering methods struggle with cluster readability.
3. Reformulated problem as a salient phrase ranking task.
4. Supervised learning improves performance using labeled data.

Example


Web PubMed
clustering
options
Search
Clusters list treemap pie-chart
Gene Expression (12 docs)
Distribution (10 docs)
Patients Compared (8 docs)
Protein Expression (6 docs)
Immune Cell (7 docs)
Disorders (6 docs)
Model Predictions (6 docs)
RNA Sequencing (6 docs)
Bacterial Strains (5 docs)
Compounds (5 docs)
Intervention Group (5 docs)
Detection and Monitoring (4 docs)
Diagnostic and Treatment (4 docs)
Disease Burden (4 docs)
Genetic Diversity (4 docs)
Temperature (4 docs)
Breast Cancer (3 docs)
Children Aged (3 docs)
Classification Accuracy (3 docs)
Cognitive Change (3 docs)
College Students (3 docs)
Implementation Outcomes (3 docs)
Lung (3 docs)
Agricultural Carbon (2 docs)
Biofilm (2 docs)
Fractions (2 docs)
IgG1 Monoclonal Antibody (2 docs)
Overall Prevalence (2 docs)
Single Feature (2 docs)
Variable Sites (2 docs)
Other topics (25 docs)
Results list
All retrieved results (100)
1 Whole Genome Sequencing Reveals Substantial Genetic Structure and Evidence of Local Adaptation in Alaskan Red King Crab.
Evolutionary applications, 2025
High-latitude ocean basins are the most productive on earth, supporting high diversity and biomass of economically and socially important species. A long tradition of responsible fisheries management has sustained these species for generations, but modern threats from climate change, habitat loss, and new fishing technologies threaten their ecosystems and the human communities that depend on them....
KEYWORDS: local adaptation, population genomics, soft selective sweep, whole genome sequencing
https://www.ncbi.nlm.nih.gov/pubmed/39742389
Distribution Genetic Diversity
2 Synthesis of MR fingerprinting information from magnitude-only MR imaging data using a parallelized, multi network U-Net convolutional neural network.
Frontiers in radiology, 2024
BACKGROUND: MR fingerprinting (MRF) is a novel method for quantitative assessment of in vivo MR relaxometry that has shown high precision and accuracy. However, the method requires data acquisition using customized, complex acquisition strategies and dedicated post processing methods thereby limiting its widespread application.
OBJECTIVE: To develop a deep learning (DL) network for synthesizing MRF signals from conventio...
KEYWORDS: MPRAGE, U-Net, convolutional neural network, magnetic resonance fingerprinting, relaxometry
https://www.ncbi.nlm.nih.gov/pubmed/39742349
Other topics

1

Similarity

$$s(\mathbf{x}, \mathbf{y}) = s\left(\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{pmatrix}, \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_d \end{pmatrix}\right)$$

1. Symmetric: $s(\mathbf{x}, \mathbf{y}) = s(\mathbf{y}, \mathbf{x})$

2. A metric is a distance function f defined in a set E that satisfies four properties

1. Non-negativity: $f(\mathbf{x}, \mathbf{y}) \geq 0$
2. Reflexivity: $f(\mathbf{x}, \mathbf{y}) = 0 \iff \mathbf{x} = \mathbf{y}$
3. Commutativity: $f(\mathbf{x}, \mathbf{y}) = f(\mathbf{y}, \mathbf{x})$
4. Triangle inequality: $f(\mathbf{x}, \mathbf{y}) \leq f(\mathbf{x}, \mathbf{z}) + f(\mathbf{y}, \mathbf{z})$

1. A dissimilarity function is a metric defined in a set
2. It should satisfy the following properties
3. $0 \leq s(\mathbf{x}, \mathbf{y}) \leq 1$
4. $s(\mathbf{x}, \mathbf{x}) = 1$

1. $s(\mathbf{x}, \mathbf{y}) = s(\mathbf{y}, \mathbf{x})$
2. Where $\mathbf{x}$, $\mathbf{y}$ are two arbitrary data points in the set
3. There are many other similarity and dissimilarity structures
4. Let D be a data set.

Hartigan (1967) Similarity Structures

1. S defined on $D \times D$ is a Euclidean distance
2. S defined on $D \times D$ is a metric
3. S defined on $D \times D$ is symmetric and real valued
4. S defined on $D \times D$ is real valued
5. S is complete "similarity" order $\leq_S$ on $D \times D$

1. S is partial "similarity" order $\leq_S$ on $D \times D$
2. S is tree on D
3. S is complete 'relative similarity' order $\leq_i$ on D for each $i \in D$
4. S is partial 'relative similarity' order $\leq_i$ on D
5. S is similarity dichotomy on $D \times D$

1. S is similarity trichotomy on $D \times D$

2. S is a partition of D into sets of similar objects

S **defined on** $D \times D$ **is a Euclidean distance**

1

S is a function that computes the straight-line distance between two points in D in a Euclidean space.

2

S measures how far apart the two data points are. This is For two points

3

$$S\left(\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{pmatrix}, \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_d \end{pmatrix}\right) = \sqrt{\sum_{i=1}^{d} (x_i - y_i)^2}$$

Similarity structures - 01 - 02

1. Properties of Euclidean distance are
2. Non-negativity $S(\mathbf{x}, \mathbf{y}) \geq 0$ and $S(\mathbf{x}, \mathbf{y}) = 0$ if $\mathbf{x} = \mathbf{y}$
3. Symmetry $S(\mathbf{x}, \mathbf{y}) = S(\mathbf{y}, \mathbf{x})$
4. Triangle inequality: $S(\mathbf{x}, \mathbf{z}) \leq S(\mathbf{x}, \mathbf{y}) + S(\mathbf{y}, \mathbf{z})$
5. Identity of indiscernible $S(\mathbf{x}, \mathbf{y}) = 0$ if and only if $\mathbf{x} = \mathbf{y}$

1. Let $D = \left(\begin{pmatrix} 1 \\ 2 \end{pmatrix} \quad \begin{pmatrix} 4 \\ 6 \end{pmatrix} \quad \begin{pmatrix} 7 \\ 2 \end{pmatrix} \right)$

2. $S\left(\begin{pmatrix} 1 \\ 2 \end{pmatrix}, \begin{pmatrix} 4 \\ 6 \end{pmatrix} \right) = \sqrt{(1-4)^2 + (2-6)^2} = 5$

3. $S\left(\begin{pmatrix} 1 \\ 2 \end{pmatrix}, \begin{pmatrix} 7 \\ 2 \end{pmatrix} \right) = \sqrt{(1-7)^2 + (2-2)^2} = 6$

S **is partial "similarity" order** $\leq_S$ **on** $D \times D$

1. S is a partial similarity order $\leq_S$ on $D \times D$
2. A partial order differs from complete order
3. Not all pairs of elements comparable
4. For some pairs (**x**, **y**), (**a**, **b**), it may not be possible to determine $S(\mathbf{x}, \mathbf{y}) \leq S(\mathbf{a}, \mathbf{b})$

1. A relation $\leq_S$ derived from S on $D \times D$ is a partial order if it satisfies following properties

2. Reflexivity: $S(\mathbf{x}, \mathbf{y}) \leq_S S(\mathbf{x}, \mathbf{y})$

3. Antisymmetric: if $S(\mathbf{x}, \mathbf{y}) \leq_S S(\mathbf{a}, \mathbf{b})$ and $S(\mathbf{a}, \mathbf{b}) \leq_S S(\mathbf{x}, \mathbf{y})$ then $S(\mathbf{x}, \mathbf{y}) = S(\mathbf{a}, \mathbf{b})$

4. Transitivity: if $S(\mathbf{x}, \mathbf{y}) \leq_S S(\mathbf{a}, \mathbf{b})$, $S(\mathbf{a}, \mathbf{b}) \leq_S S(\mathbf{c}, \mathbf{d})$, then $S(\mathbf{x}, \mathbf{y}) \leq_S S(\mathbf{c}, \mathbf{d})$

1

We will construct a definition for partial similarity $\leq_S$ where some pairs $(\mathbf{x}, \mathbf{y})$ and $(\mathbf{a}, \mathbf{b})$ in the data set are not comparable

Let:

$$D = \left\{ \begin{pmatrix} 1 \\ 2 \end{pmatrix}, \begin{pmatrix} 4 \\ 6 \end{pmatrix}, \begin{pmatrix} 7 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ 6 \end{pmatrix} \right\}$$

1

Similarity function S is Manhattan distance $S(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{d} |x_i - y_i|$

1. Partial similarity order $\leq_S$ definition
2. Define $(\mathbf{x}, \mathbf{y}) \leq_S (\mathbf{a}, \mathbf{b})$ if $(\mathbf{x}, \mathbf{y}) \leq (\mathbf{a}, \mathbf{b})$
3. and if (1) $\mathbf{x}$ and $\mathbf{a}$ share at least one identical coordinate
   (2) $\mathbf{y}$ and $\mathbf{b}$ share at least one identical coordinate
4. Otherwise $(\mathbf{x}, \mathbf{y})$ and $(\mathbf{a}, \mathbf{b})$ are incomparable

1. Apply the condition $s(\mathbf{x}, \mathbf{y}) \leq_S (\mathbf{a}, \mathbf{b})$

2. Compare $\left(\begin{pmatrix} 1 \\ 2 \end{pmatrix}, \begin{pmatrix} 4 \\ 6 \end{pmatrix}\right) \leq_S \left(\begin{pmatrix} 1 \\ 2 \end{pmatrix}, \begin{pmatrix} 1 \\ 6 \end{pmatrix}\right)$;

3. $S\left(\begin{pmatrix} 1 \\ 2 \end{pmatrix}, \begin{pmatrix} 4 \\ 6 \end{pmatrix}\right) = |1-4| + |2-6| = 7$

4. $S\left(\begin{pmatrix} 1 \\ 2 \end{pmatrix}, \begin{pmatrix} 1 \\ 6 \end{pmatrix}\right) = |1-1| + |2-6| = 4$

5. As $S\left(\begin{pmatrix} 1 \\ 2 \end{pmatrix}, \begin{pmatrix} 4 \\ 6 \end{pmatrix}\right) \leq_S S\left(\begin{pmatrix} 1 \\ 2 \end{pmatrix}, \begin{pmatrix} 1 \\ 6 \end{pmatrix}\right)$, this pair is not comparable

1

Compare

$$\left(\begin{pmatrix}1\\2\end{pmatrix},\begin{pmatrix}1\\6\end{pmatrix}\right) \leq_S \left(\begin{pmatrix}1\\2\end{pmatrix},\begin{pmatrix}4\\6\end{pmatrix}\right)$$

$$S\left(\begin{pmatrix}1\\2\end{pmatrix},\begin{pmatrix}4\\6\end{pmatrix}\right) = |1-4| + |2-6| = 7$$

$$S\left(\begin{pmatrix}1\\2\end{pmatrix},\begin{pmatrix}1\\6\end{pmatrix}\right) = |1-1| + |2-6| = 4$$

(1)

$$S\left(\begin{pmatrix}1\\2\end{pmatrix},\begin{pmatrix}1\\6\end{pmatrix}\right) \leq S\left(\begin{pmatrix}1\\2\end{pmatrix},\begin{pmatrix}4\\6\end{pmatrix}\right)$$

(2) First pair has one dimension in common
(3) Second pair has one dimension in common. Therefore this pair holds $\leq_S$

Similarity and Dissimilarity Measures Between Clusters

Similarity measures between clusters 01

1. Many clustering algorithms are hierarchical *i.e.,* is a sequence of nested partitions
2. In an agglomerative hierarchical algorithm, two most similar groups are merged to form a large cluster at each step
3. This process is continued until the desired number of clusters is obtained
4. To merge an object and a cluster, we need to compute the distance between an object and a cluster
5. To merge two clusters, we need to compute the distance between clusters

Similarity measures between clusters 02

1. Let $C_1 = \{\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_r\}$ and $C_2 = \{\mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_s\}$
2. $|C_1| = r$ and $|C_2| = s$

The Mean-Based Distance

Similarity measures between clusters 03

1. To measure the dissimilarity between two clusters:

2. Measure the distance between means of the two clusters.

3. Let $\mu(C_1)$ be the mean of C_1 and $\mu(C_2)$ be the mean of C_2

4. Then
$$D_{mean}(C_1, C_2) = d(\mu(C_1), \mu(C_2))$$

5. Where
$$\mu(C_j) = \frac{1}{|C_j|} \sum_{\mathbf{x} \in C_j} \mathbf{x}, \quad \forall\ j = 1, 2$$

The Nearest Neighbor Distance

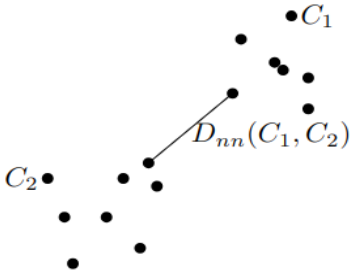
The Nearest Neighbor Distance 01

1

Given a distance function $d(.,.)$, the nearest neighbor distance between C_1 and C_2 with respect to $d(.,.)$ is defined as

$$D_{nn}(C_1, C_2) = \min_{1 \leq i \leq r} \min_{1 \leq j \leq s} d(\mathbf{y}_i, \mathbf{z}_j)$$

The Nearest Neighbor Distance 02



The Farthest Neighbor Distance

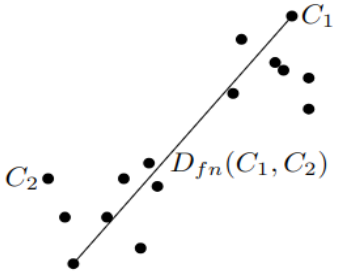
The Farthest Neighbor Distance 01

1

Given a distance function $d(.,.)$, the nearest neighbor distance between C_1 and C_2 with respect to $d(.,.)$ is defined as

$$D_{fn}(C_1, C_2) = \max_{1 \leq i \leq r \;\; 1 \leq j \leq s} ; d(\mathbf{y}_i, \mathbf{z}_j)$$

The Farthest Neighbor Distance 02



The Average Neighbor Distance

The Average Neighbor Distance 01

1

This is defined as

$$D_{ave}(C_1, C_2) = \frac{1}{r \times s} \sum_{i=1}^{r} \sum_{j=1}^{s} d(\mathbf{y}_i, \mathbf{z}_j)$$

The statistical distance between C_1 and C_2 is defined as

$$D_{stat}(C_1, C_2) = \frac{r \times s}{r + s} (\bar{\mathbf{y}} - \bar{\mathbf{z}})(\bar{\mathbf{y}} - \bar{\mathbf{z}})^T$$

The Average Neighbor Distance 02

1

Where

$$\bar{\mathbf{y}} = \frac{1}{r}\sum_{i=1}^{r}\mathbf{y}_i$$

$$\bar{\mathbf{z}} = \frac{1}{s}\sum_{j=1}^{s}\mathbf{z}_j$$

The Average Neighbor Distance 03

1. Let $C = C_1 \cup C_2$ be the cluster formed by merging C_1 and C_2

2. Let $M_{sca}(C), M_{sca}(C_1)$ and $M_{sca}(C_2)$ be the within-scatter matrices of C, C_1, and C_2

3. Then

$$M_{sca}(C) = M_{sca}(C_1) + M_{sca}(C_2) + \frac{r \times s}{r + s}(\bar{\mathbf{y}} - \bar{\mathbf{z}})^T(\bar{\mathbf{y}} - \bar{\mathbf{z}})$$

Lance-Williams Formula

Lance-Williams Formula 01

1

Lance & Williams propose a recurrence formula as given

$$D(C_k, C_i \cup C_j) = \alpha_i \, D(C_k, C_i) + \alpha_j \, D(C_k, C_j) + \beta \, D(C_i, C_j) + \gamma \, |D(C_k, C_i) - D(C_k, C_j)|$$

Lance-Williams Formula 02

1

Single Linkage

$$D(C_k, C_i \cup C_j) = \frac{1}{2}\, D(C_k, C_i) + \frac{1}{2}\, D(C_k, C_j) + 0 \times D(C_i, C_j) + \frac{-1}{2}\, |D(C_k, C_i) - D(C_k, C_j)|$$

Lance-Williams Formula 03

1

Complete Linkage

$$D(C_k, C_i \cup C_j) = \frac{1}{2}\, D(C_k, C_i) + \frac{1}{2}\, D(C_k, C_j) + 0 \times D(C_i, C_j) + \frac{1}{2}\, |D(C_k, C_i) - D(C_k, C_j)|$$

Agglomerative Hierarchical Clustering

Agglomerative Hierarchical Clustering 01

1. Meaning of Agglomerative: gathered together into a cluster or mass.
2. Meaning of Agglomerative: crowded into a dense cluster
3. Synonyms: accumulate, amass, assemble
4. As the name suggests, hierarchical clustering is built by accumulating smaller clusters.
5. Start with each data point as a cluster and assemble smaller clusters to build larger clusters

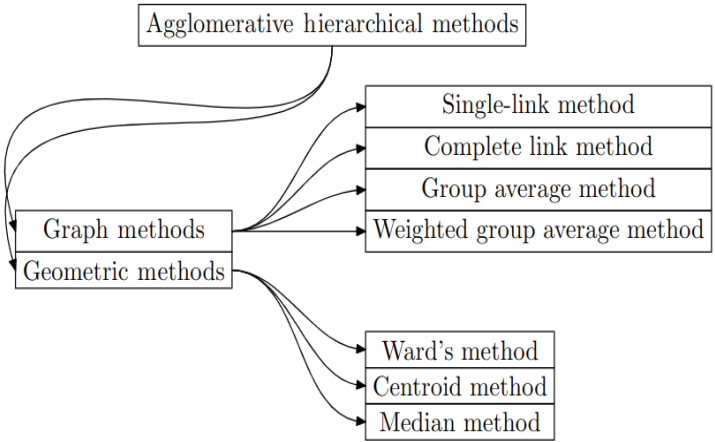
Agglomerative Hierarchical Clustering 02

1. Agglomerative hierarchical methods are subdivided according to different distance measures as given
2. Graph methods - Single-link
3. Graph methods - Complete link
4. Graph methods - Group average method
5. Graph methods - Weighted group average method

Agglomerative Hierarchical Clustering 03

1. Geometric methods - Ward's method
2. Geometric methods - Centroid method
3. Geometric methods - Median method

Agglomerative Hierarchical Clustering 04



The Single Link Method

The Single Link Method 01

1. The single-link method is one of the simplest hierarchical clustering methods.
2. First introduced in 1951 by Florek and later in 1957 by Sneath
3. It employs the nearest neighbor distance to measure dissimilarity between two groups

1

Let C_i, C_j and C_k be three groups of data points. The distance between C_k and $C_i \cup C_j$ is given by Lance-Williams formula as

$$\begin{aligned} D(C_k, C_i \cup C_j) &= \tfrac{1}{2} D(C_k, C_i) \\ &+ \tfrac{1}{2} D(C_k, C_j) \\ &- \tfrac{1}{2} |D(C_k, C_i) - D(C_k, C_j)| \\ &= \min(D(C_k, C_i), D(C_k, D_j)) \end{aligned}$$

The Single Link Method 03

1. Where $D(.,.)$ is a distance between two clusters

2. And is given as

$$D(C_k, C_i) = \min_{\mathbf{x} \in C_k,\ \mathbf{y} \in C_i} d(\mathbf{x}, \mathbf{y})$$

The Single Link Method 04

1

Proof: We need to show

$$D(C_k, C_i \cup C_j) = min\{D(C_k, C_i), D(C_k, C_j)\}$$

The Single Link Method 04

1

$$D(C_k, C_i \cup C_j) = \min_{\mathbf{x} \in C_k,\ \mathbf{y} \in C_i \cup C_j;} d(\mathbf{x}, \mathbf{y})$$

1

Since $\mathbf{y} \in C_i \cup C_j$, $\mathbf{y} \in C_i$ or $\mathbf{y} \in C_j$. We can write

$$D(C_k, C_i \cup C_j) = min \left\{ \min_{\mathbf{a} \in C_k,\, \mathbf{b} \in C_i} dist(\mathbf{a}, \mathbf{b}), \min_{\mathbf{a} \in C_k,\, \mathbf{b} \in C_j} dist(\mathbf{a}, \mathbf{b}) \right\}$$

1

From the definition of distance between two clusters:

$$D(C_k, C_i) = \min_{\mathbf{a} \in C_k,\, \mathbf{b} \in C_i} dist(\mathbf{a}, \mathbf{b})$$

and

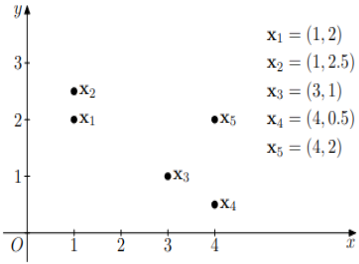
$$D(C_k, C_j) = \min_{\mathbf{a} \in C_k,\, \mathbf{b} \in C_j} dist(\mathbf{a}, \mathbf{b})$$

1

Therefore,

$$D(C_k, C_i \cup C_j) = min\{D(C_k, C_i), D(C_k, C_j)\}$$

Single Link Example 01



Single Link Example 02

	x_1	x_2	x_3	x_4	x_5
x_1	0	0.5	2.24	3.35	3
x_2	0.5	0	2.5	3.61	3.04
x_3	2.24	2.5	0	1.12	1.41
x_4	3.35	3.61	1.12	0	1.5
x_5	3	3.04	1.41	1.5	0

Single Link Example 03

	$\{x_1, x_2\}$	x_3	x_4	x_5
$\{x_1, x_2\}$	0	2.24	3.35	3
x_3	2.24	0	1.12	1.41
x_4	3.35	1.12	0	1.5
x_5	3	1.41	1.5	0

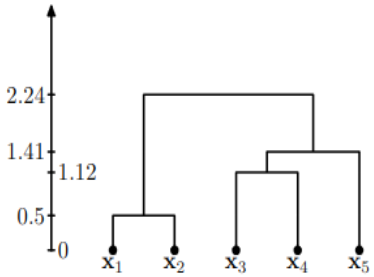
Single Link Example 04

	$\{x_1, x_2\}$	$\{x_3, x_4\}$	x_5
$\{x_1, x_2\}$	0	2.24	3
$\{x_3, x_4\}$	2.24	0	1.41
x_5	3	1.41	0

Single Link Example 05

	$\{x_1, x_2\}$	$\{x_3, x_4, x_5\}$
$\{x_1, x_2\}$	0	2.24
$\{x_3, x_4, x_5\}$	2.24	0

Single Link Example 06



The Complete Link Method

The Complete Link Method 01

1. Unlike the single-link method, the complete link method uses the farthest neighbor distance to measure dissimilarity between two clusters
2. The complete link method is invariant under monotone transformations

The Complete Link Method 02

1

Let C_i, C_j and C_k be three groups of data points. The distance between C_k and $C_i \cup C_j$ is given by Lance-Williams formula as

$$\begin{aligned} D(C_k, C_i \cup C_j) &= \tfrac{1}{2} D(C_k, C_i) \\ &+ \tfrac{1}{2} D(C_k, C_j) \\ &+ \tfrac{1}{2} |D(C_k, C_i) - D(C_k, C_j)| \\ &= \max(D(C_k, C_i), D(C_k, D_j)) \end{aligned}$$

The Complete Link Method 03

1. Where $D(.,.)$ is a distance between two clusters

2. And is given as

$$D(C_k, C_i) = \max_{\mathbf{x} \in C_k,\ \mathbf{y} \in C_i} d(\mathbf{x}, \mathbf{y})$$

The Complete Link Method 04

1

Proof: We need to show

$$D(C_k, C_i \cup C_j) = max\{D(C_k, C_i), D(C_k, C_j)\}$$

1

Since $\mathbf{y} \in C_i \cup C_j$, $\mathbf{y} \in C_i$ or $\mathbf{y} \in C_j$. We can write

$$D(C_k, C_i \cup C_j) = max \left\{ \max_{\mathbf{a} \in C_k,\, \mathbf{b} \in C_i} dist(\mathbf{a}, \mathbf{b}), \max_{\mathbf{a} \in C_k,\, \mathbf{b} \in C_j} dist(\mathbf{a}, \mathbf{b}) \right\}$$

1

From the definition of distance between two clusters:

$$D(C_k, C_i) = \max_{\mathbf{a} \in C_k,\, \mathbf{b} \in C_i} dist(\mathbf{a}, \mathbf{b})$$

and

$$D(C_k, C_j) = \max_{\mathbf{a} \in C_k,\, \mathbf{b} \in C_j} dist(\mathbf{a}, \mathbf{b})$$

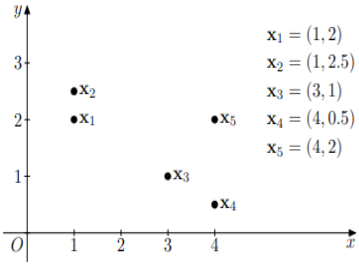
The Single Link Method 08

1

Therefore,

$$D(C_k, C_i \cup C_j) = max\{D(C_k, C_i), D(C_k, C_j)\}$$

Complete Link Example 01



Complete Link Example 02

1

In the first step, we are dealing with individual points, so the clustering algorithm simply merges the pair of closest points based on their Euclidean distance.

Complete Link Example 03

	$\{x_1, x_2\}$	x_3	x_4	x_5
$\{x_1, x_2\}$	0	2.5	3.61	3.04
x_3	2.5	0	1.12	1.41
x_4	3.61	1.12	0	1.5
x_5	3.04	1.41	1.5	0

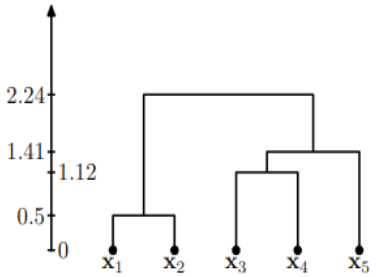
Complete Link Example 04

	$\{x_1, x_2\}$	$\{x_3, x_4\}$	x_5
$\{x_1, x_2\}$	0	3.61	3.04
$\{x_3, x_4\}$	3.61	0	1.5
x_5	3.04	1.5	0

Complete Link Example 05

	$\{x_1, x_2\}$	$\{x_3, x_4, x_5\}$
$\{x_1, x_2\}$	0	3.61
$\{x_3, x_4, x_5\}$	3.61	0

Complete Link Example 06



Hierarchical Clusters representation

Representation of Hierarchical Clusterings 01

1. A hierarchical clustering can be represented by a picture
2. by a list of abstract symbols
3. A picture is easier for humans to interpret
4. A list of abstract symbols is used internally to improve the performance of algorithm

Representation of Hierarchical Clusterings 02

1. n-Tree
2. Dendrogram
3. Banner
4. Pointer representation

Representation of Hierarchical Clusterings 03

1. Loop plot
2. Icicle plot

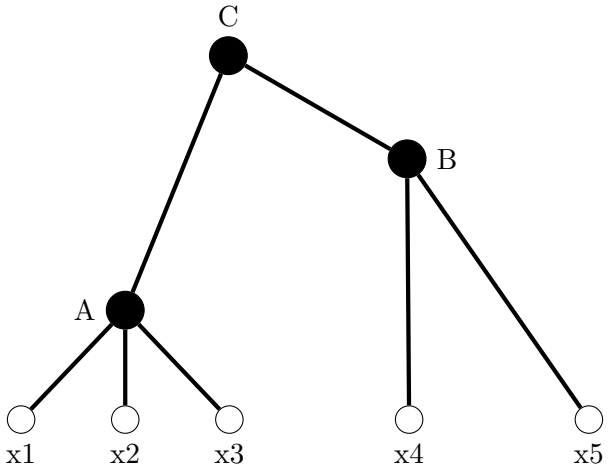
n-Tree

n-Tree 01

1. A hierarchical clustering is generally represented by a tree diagram
2. An n-tree is a simple hierarchically nested tree diagram
3. Let $D = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n\}$
4. n-tree on D is defined to be a set $\mathcal{T} \subseteq D$ satisfying the following conditions

n-Tree 02

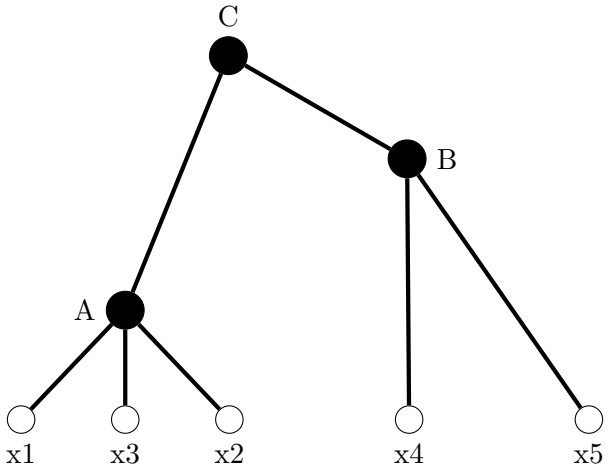
1. $D \in \mathcal{T}$
2. Empty set $\Phi \in \mathcal{T}$
3. $\{\mathbf{x}_i\} \in \mathcal{T}$ for all $i = 1, 2, \cdots, n$
4. If $A, B \in \mathcal{T}$, then $A \cap B \in \{\Phi, A, B\}$

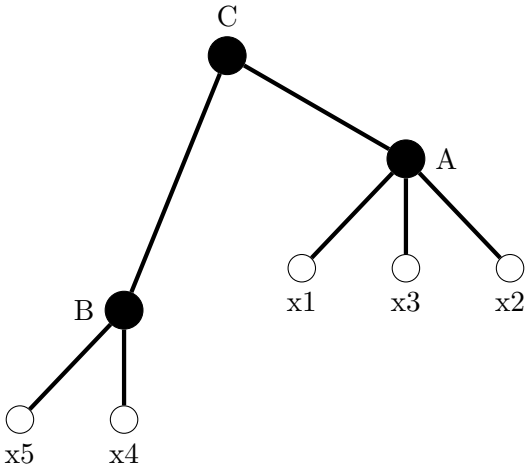

C
B
A
x1
x2
x3
x4
x5

n-Tree 04

1. Leaves are open circles represent a single data point
2. Internal nodes depicted by filled circle represent a group of data points
3. n-trees are also known as non-ranked trees
4. An n-tree with n - 1 internal nodes is a dichotomous tree

1. Tree diagrams contain many indeterminacies
2. Order of leave can be changed
3. Order of internal nodes can be changed


C
B
A
x1
x3
x2
x4
x5

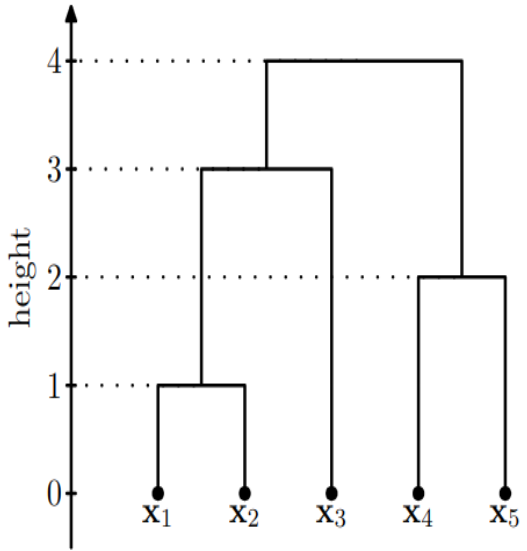

C
A
B
x1
x3
x2
x5
x4

Dendrogram

Dendrogram 01

1. Also called valued tree
2. Is an $n-$tree in which each internal node is associated with a height
3. Height should satisfy the condition
$$h(A) \leq h(B) \Leftrightarrow A \subseteq B$$
4. For all subsets of data points A and B if $A \cap B \neq \Phi$
5. $h(A)$ denote height of A; $h(B)$ denotes height of B.

Dendrogram 02



Dendrogram 03

1. Dendrogram gram with five data points presented
2. Dotted indicates the height of the internal nodes
3. h_{ij} height of internal node specify the smallest cluster to which $\mathbf{x}_i$ and $\mathbf{x}_j$ belong.
4. Small value of h_{ij} indicates a high similarity between $\mathbf{x}_i$ and $\mathbf{x}_j$
5. In the figure, $h_{12} = 1$; $h_{23} = h_{13} = 3$ and $h_{14} = 4$

Dendrogram 04

1. Heights in the dendrogram satisfy the following ultrametric inequality conditions

$$h_{ij} \leq \max\{h_{ik}, h_{jk}\} \ \forall \ i, \ j, \ k \ \in \ \{1, 2, \cdots, n\}$$

2. This condition is required for valid dendrogram construction

3. For h_{23}, assume the following

4. $h_{23} = 3$ (as $\mathbf{x}_2$ and $\mathbf{x}_3$ are joined at height 3)

Dendrogram 05

1. For k = 1: $h_{21} = 1 \;\; h_{31} = 3$

$$h_{23} = 3 \leq \max\{h_{21}, h_{31}\} = \max\{1, 3\} = 3$$

2. For k = 4: $h_{24} = 3 \;\; h_{34} = 3$

$$h_{23} = 3 \leq \max\{h_{24}, h_{34}\} = \max\{3, 3\} = 3$$

3. For k = 5: $h_{25} = 4 \;\; h_{35} = 4$

$$h_{23} = 3 \leq \max\{h_{25}, h_{35}\} = \max\{4, 4\} = 4$$

4. This condition is necessary and sufficient condition for a dendrogram

Dendrogram 06

1. Validate the Ultrametric inequality $h_{ij} \leq \max\{h_{ik}, h_{jk}\}$ for various combinations of i, j, and k.
2. $i = 1, j = 2, k = 3$
3. $h_{12} = 1; h_{13} = 3; h_{23} = 3$
4. $h_{12} = 1 \leq \max\{h_{13}, h_{23}\} = \max\{3,\ 3\} = 3.$

Dendrogram 07

1. Validate the Ultrametric inequality $h_{ij} \leq \max\{h_{ik}, h_{jk}\}$ for various combinations of i, j, and k.
2. $i = 1, j = 3, k = 4$
3. $h_{13} = 3; h_{14} = 4; h_{34} = 4$
4. $h_{13} = 3 \leq \max\{h_{14}, h_{34}\} = \max\{4, 4\} = 4.$

Dendrogram 08

1. Validate the Ultrametric inequality $h_{ij} \leq \max\{h_{ik}, h_{jk}\}$ for various combinations of i, j, and k.
2. $i = 4, j = 5, k = 1$
3. $h_{45} = 2; h_{14} = 4; h_{15} = 4$
4. $h_{45} = 2 \leq \max\{h_{14}, h_{15}\} = \max\{4, 4\} = 4.$

Dendrogram 09

1. Validate the Ultrametric inequality $h_{ij} \leq \max\{h_{ik}, h_{jk}\}$ for various combinations of i, j, and k.
2. $i = 1, j = 2, k = 4$
3. $h_{45} = 2; h_{14} = 4; h_{15} = 4$
4. $h_{45} = 2 \leq \max\{h_{14}, h_{15}\} = \max\{4, 4\} = 4.$

Dendrogram 10

1. If the ultrametric inequality holds, a valid dendrogram can be constructed as follows:
2. Arrange the points $\{1, 2, \cdots, n\}$ into a heirarchical structure such that
3. pairs of points (or clusters) merge at heights h_{ij}
4. Use the ultrametric property to ensure the hierarchy satisfies the required constraints:
5. For any three points i, j, k, distance h_{ij} does not violate the maximum distance condition imposed by h_{ik} and h_{jk}

Dendrogram 11

1. The ultrametric inequality ensures that at any level of the hierarchy:
2. The clusters are nested
3. No inconsistencies arise in the merging process
4. Since the ultrametric inequality guarantees consistent merging, the distances can be directly mapped to the heights of a dendrogram
5. Thus, the ultrametric inequality is a sufficient condition for constructing a valid dendrogram.

Dendrogram 12

1

A dendrogram can be represented by a function $c : [0, \infty) \rightarrow E(D)$ that satisfies

$$c(h) \subseteq c(h^{'}) \text{ if } h \leq h^{'}$$

$$c(h) \text{is eventually in} D \times D$$

$$c(h + \delta) = c(h) \text{for some small} \delta > 0$$

1

$$c(h) = \begin{cases} \{(i,i) : i = \{1,2,3,4,5\} & \text{if } 0 \leq h < 1 \\ \{(i,i) : i = \{3,4,5\} \cup & \\ \{(i,j) : i,j = \{1,2\} & \text{if } 1 \leq h < 2 \\ \{(3,3)\} \cup & \\ \{(i,j) : i,j = \{1,2\} & \\ \{(i,j) : i,j = \{4,5\} & \text{if } 2 \leq h < 3 \\ \{(i,j) : i,j = \{4,5\} \cup & \\ \{(i,j) : i,j = \{1,2,3\} & \text{if } 3 \leq h < 4 \\ \{(i,j) : i,j = \{1,2,3,4,5\} & \text{if } 4 \leq h \end{cases}$$

Dendrogram 14

1. A similarity function measures how alike two points are. Higher similarity values mean more similarity.
2. Examples: Cosine similarity, Jaccard similarity, Pearson correlation.
3. Higher similarity values indicate closeness, but dendrograms typically represent distances (dissimilarity).
4. A dissimilarity (or distance) function measures how different two points are. Larger values mean more dissimilarity.
5. Examples: Euclidean distance, Manhattan distance, Hamming distance.

Dendrogram 15

1. To use similarity in dendrograms
2. Convert similarity values to dissimilarity values.
3. A simple approach is to transform similarity into dissimilarity by subtracting from the maximum possible similarity: $d_{ij} = S_{max} - S_{ij}$
4. or $d_{ij} = \frac{1}{S_{ij}}$

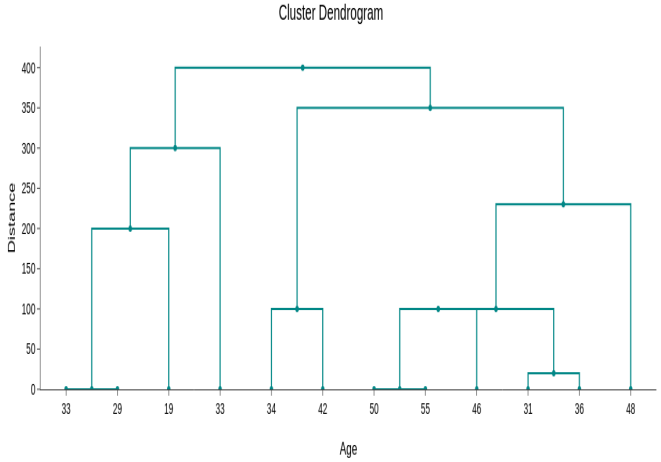
Dendrogram 16

1. Similarity-Based Dendrogram
2. Higher values on the y-axis mean greater similarity.
3. Points merge from higher similarity to lower similarity

Dendrogram 17

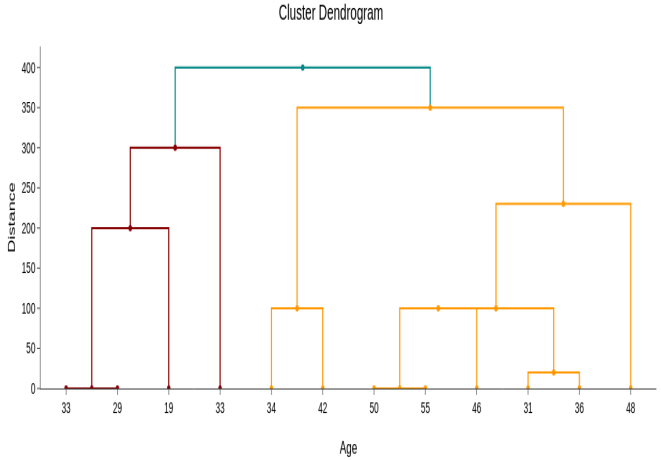
1. Dissimilarity-Based Dendrogram
2. Higher values on the y-axis mean greater dissimilarity
3. Points merge from lower dissimilarity to higher dissimilarity.

Dendrogram with one cluster



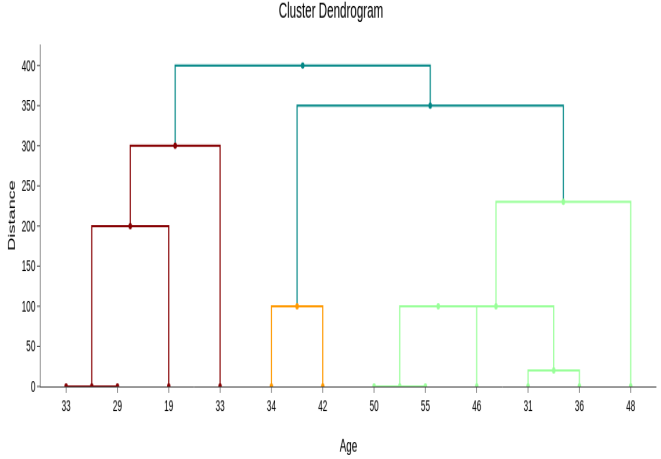
C1: {33, 29,19, 33, 34, 42, 50, 55, 46, 31, 36, 48}

Dendrogram with two clusters



C1: {33, 29,19, 33}
C2: {34, 42, 50, 55, 46, 31, 36, 48}

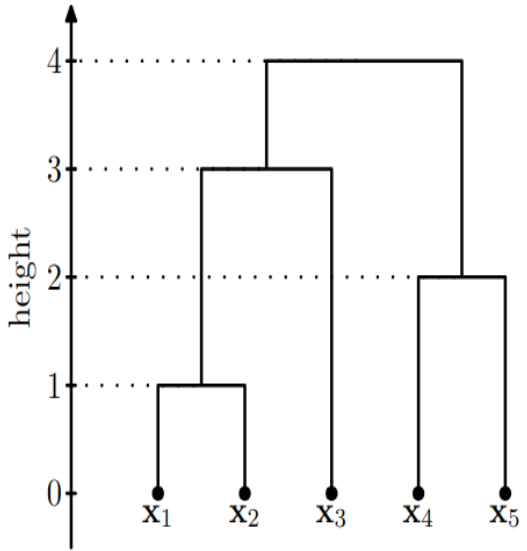
Dendrogram with three clusters

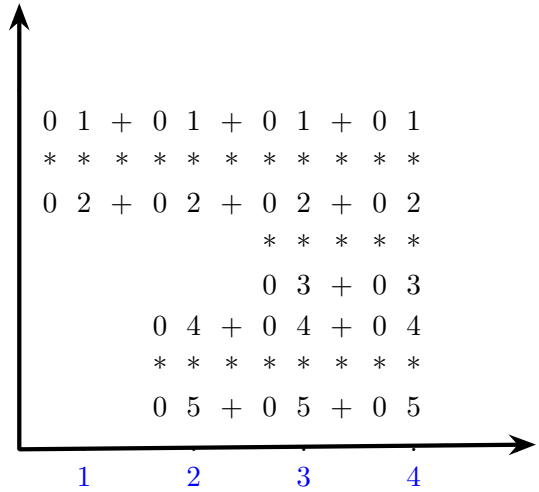


C1: {33, 29,19, 33}
C2: {34, 42}
C3: {50, 55, 46, 31, 36, 48}

Banner

Dendrogram 01




0 1 + 0 1 + 0 1 + 0 1
* * * * * * * * * *
0 2 + 0 2 + 0 2 + 0 2
* * * * *
0 3 + 0 3
0 4 + 0 4 + 0 4
* * * * * * * *
0 5 + 0 5 + 0 5
1
2
3
4

Banner 02

1. A list of symbols and codes representing a hierarchical structure
2. The heights in the dendrogram are represented on horizontal axis
3. Each data point in the banner is assigned a line and a code
4. Separator is denoted with +

Banner 03

1. Presence of “*” between two data points indicate that two points are in the same group

2. Example of banner is as given

Pointer Representation

Pointer Representation 01

1. Pointer representation is a pair of functions containing information on a dendrogram

2. $\pi : \{1, 2, \cdots, n\} \to \{1, 2, \cdots, n\}$

3. $\lambda : \pi(\{1, 2, \cdots, n\}) \to [0, \infty]$

4. These two function satisfy the following properties

Pointer Representation 02

1

$\lambda(i)$: Represents the lowest level (height) at which object i is no longer the last object in its cluster.

$$\lambda(i) = inf\{h : \exists\, j > i \text{ such that } (i, j) \in c(h)\}$$

$\pi(i)$: Represents the last object in the cluster that i joins

$$\pi(i) = \max\{j : (i, j) \in c(\lambda(i))\}$$

$c(h)$: Describes the cluster memberships at height h. As height increases, clusters merge.

Pointer Representation 03

1. Cluster representation $c(h)$; at $h \in [0,1) : c(h) = \{(i,i) : i = 1,2,3,4,5\}$
2. At $h \in [2,3) : c(h) = \{(3,3) \cup \{(i,j) : i,j = 1,2\} \cup \{(i,j) : i,j = 4,5\}\}$
3. At $h \in [3,4) : c(h) = \{\{(i,j) : i,j = 4,5\} \cup \{(i,j) : i,j = 1,2,3\}\}$
4. $h \geq 4 : c(h) = \{\{(i,j) : i,j = 1,2,3,4,5\}\}$

Pointer Representation 04

1. Find $\lambda(1)$: i $= 1$ is no longer the last object in its cluster when it joins a cluster with $j > 1$

2. At $h = 1$, $\mathbf{x}_1$ and $\mathbf{x}_2$ merge.

3. Hence, $\lambda(1) = 1$

Pointer Representation 04

1. Find $\pi(1)$:
2. At $\lambda(1) = 1$, the cluster contains $\{\mathbf{x}_1, \mathbf{x}_2\}$
3. The last object in this cluster is $\mathbf{x}_2$
4. Hence $\pi(1) = 2$

Pointer Representation 06

1. Find $\lambda(2)$: i $= 2$ is no longer the last object in its cluster when it joins a cluster with $j > 2$
2. At $h = 3$, cluster $\{\mathbf{x}_1, \mathbf{x}_2\}$ merges with $\mathbf{x}_3$
3. Hence, $\lambda(2) = 3$

Pointer Representation 07

1. Find $\pi(2)$:
2. At $\lambda(2) = 3$, the cluster contains $\{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3\}$
3. The last object in this cluster is $\mathbf{x}_3$
4. Hence $\pi(2) = 3$

Pointer Representation 08

1. Find $\lambda(3)$: i $= 3$ is no longer the last object in its cluster when it joins a cluster with $j > 3$
2. At $h = 4$, cluster $\{\mathbf{x}_3\}$ joins the cluster with $\{\mathbf{x}_4, \mathbf{x}_5\}$
3. Hence, $\lambda(3) = 4$

Pointer Representation 07

1. Find $\pi(3)$:
2. At $\lambda(3) = 4$, the cluster contains $\{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4, \mathbf{x}_5\}$
3. The last object in this cluster is $\mathbf{x}_5$
4. Hence $\pi(3) = 5$

Pointer Representation 10

1. Find $\lambda(4)$: i $= 4$ is no longer the last object in its cluster when it joins a cluster with $j > 4$
2. At $h = 2$, $\{\mathbf{x}_4\}$ and $\{\mathbf{x}_5\}$ merge
3. Hence, $\lambda(4) = 2$

Pointer Representation 11

1. Find $\pi(4)$:
2. At $\lambda(4) = 2$, the cluster contains $\{\mathbf{x}_4, \mathbf{x}_5\}$
3. The last object in this cluster is $\mathbf{x}_5$
4. Hence $\pi(4) = 5$

Pointer Representation 11

1. Find $\lambda(5)$: i $= 5$ is always the last object in its cluster. So $\lambda(5)$ is undefined or ∞

2. Find $\pi(5)$: As $\lambda(5) = \infty$, $\pi(5) = \infty$

Pointer Representation 12

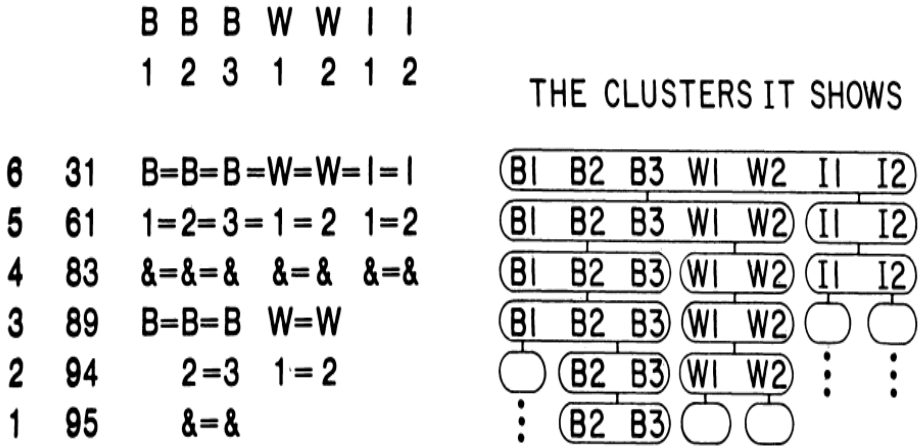
i	$\lambda(i)$	$\pi(i)$
1	1	2
2	3	3
3	4	5
4	2	5
5	∞	∞

Icicle Plot

Icicle plot 01

1. Proposed by Kruskal and Landwehr in 1983
2. Easier to program than tree plots
3. The plot type is line printer
4. Icicle plots can be much enhanced using sophisticated plotter

Icicle plot 02



Icicle plot 03

1. First row: each data point correspond to a label
2. Data point 1 and its label is B
3. Data point 2 and its label is B
4. Data point 3 and its label is B

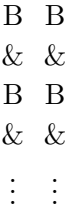
Icicle plot 04

1. First row: each data point correspond to a label
2. Data point 4 and its label is W
3. Data point 5 and its label is W
4. Data point 6 and its label is I
5. Data point 7 and its label is I

Icicle plot 05

1. Second row onwards data point labels are repeated with separators
2. Numbers in this example are given to distinguish between label B or label W or label I
3. Separator is indicated with “&”

Icicle plot 06



Icicle plot 07

1. Objects in the same cluster are joined by the symbol '='
2. Clusters are separated by SPACE in the row

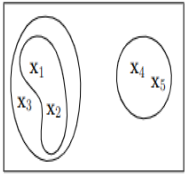
Icicle plot 08

B = B = B W = W

There are two clusters.
One cluster has data points {B, B, B};
Another cluster has data points {W, W}
Remaining objects I and I are placed in two separate clusters

Loop Plot

Loop plot 01



Hierarchical Clusters Schemes

Hierarchical clustering schemes 01

1. Partitioning objects into homogeneous groups based on similarity received a lot of attention
2. A useful correspondence between any hierarchical system and distance measure is explored
3. This correspondence gives rise to 2 methods of clustering

Hierarchical clustering schemes 02

1. In many empirical fields there is an increasing interest in identifying groups or clusters of objects
2. Those groups best represent certain empirically measured relations of similarity
3. Often large arrays of data are collected, but strong theoretical structures are lacking
4. The problem is then one of discovering whether there is any structure inherent in the data themselves

Hierarchical clustering schemes 03

1. The seminal work in biological sciences has gone under the name `numerical taxonomy`
2. Techniques described are general and are applicable to any field such as biology, medicine and psychology.
3. Objects may be human or animal subjects
4. Use measures of similarities among objects to classify objects into homogeneous groups

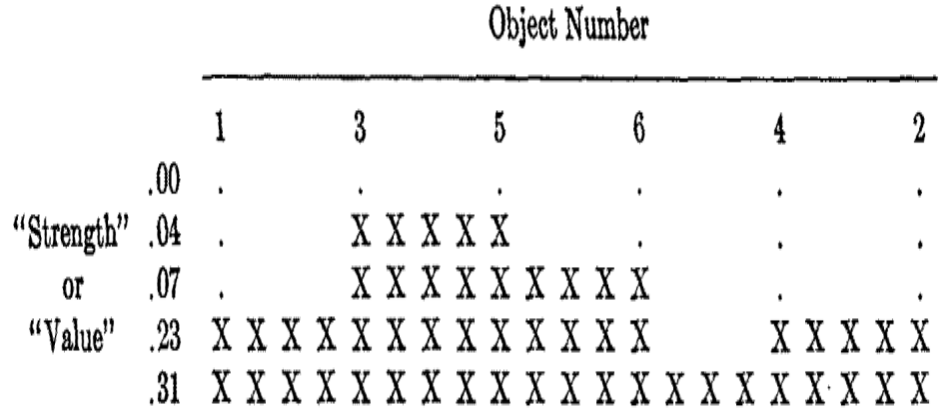
Hierarchical clustering schemes 04

1. Suitable data on similarities among object may be obtained directly or indirectly
2. If the number of objects is large, the underlying structure in the similarities is not evident from inspection alone
3. A procedure when applied to such array of similarity measures constructs a hierarchical system of clustering

Hierarchical clustering schemes 05

1. The input consists of $\binom{n}{2}$ similarities
2. There should be a clear, explicit and intuitive description of clustering
3. The clustering procedure should be invariant under monotone transformations of the similarity data

Clusterings and Metrics


Object Number
1
3
5
6
4
2
.00
.04
.07
.23
.31
"Strength"
or
"Value"

Hierarchical clustering schemes 07

1. Top row is a weak clustering where each object is a cluster.
2. This results in 6 clusters each containing one data point
3. The is given the “value” or “rating” 0.00
4. Second row has 5 clusters; {3, 5} and {1}, {2}, {4}, {6}
5. This is given the value 0.4

Hierarchical clustering schemes 08

1. At level 0.7 we have 4 clusters: {1}, {4}, {2} and {3, 5, 6}
2. At level 0.23 we have two clusters: {1, 3, 5, 6} and {2, 4}
3. At level 0.31 we have “strong” clustering with all objects in the same cluster

Hierarchical clustering schemes 09

1. We examine the following relevant features of this model
2. The values start at 0 and increase strictly as we read down
3. The clusterings “increase” hierarchically
4. Each clustering is obtained by merging of clusters at the previous level
5. If level 0.23 had had clusters {1, 3}, {5, 6, 4} and {2}. The cluster {1, 3} cannot be obtained by merging any of the 0.07 level clusters

Hierarchical clustering schemes 10

1. First clustering is the weak clustering
2. Last clustering is the strong clustering
3. Assume n data points represented by integers 1 to n
4. We have a sequence of $m+1$ clustering: $C_0, C_1, \cdots, C_m$

Hierarchical clustering schemes 11

1. Each clustering C_i we have a number α_i (height), its value
2. C_0 a weak clustering has a value $\alpha_0 = 0$
3. We require that number increase $\alpha_{j-1} \leq \alpha_j \quad \forall j = 1, 2, \cdots, m$
4. $C_{i-1} < C_i$ means every cluster in C_i is the merging of clusters C_{i-1}
5. This general agreement is referred as hierarchical clustering scheme

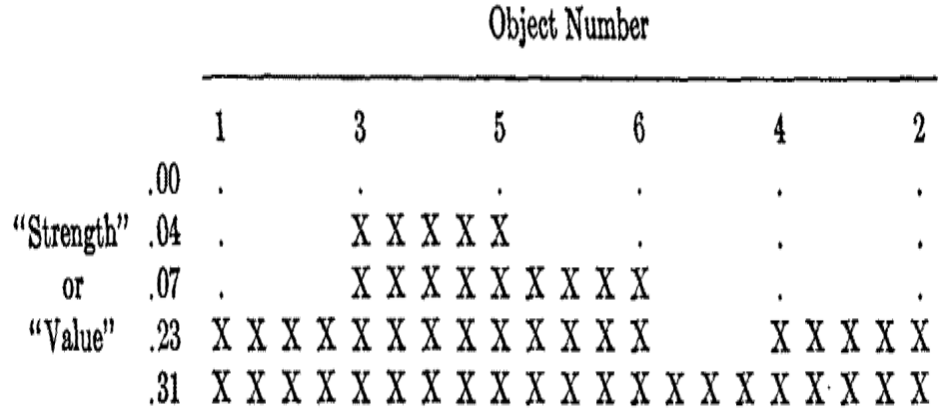
Hierarchical clustering schemes 12

1. Every HCS gives rise to particular kind of distance or metric between objects
2. Conversely, given such a metric, we may recover the HCS from it (distance metric)
3. This reduces the study of HCS's to the study of these metrics
4. First we shall assume that we are given an HCS
5. A sequence of clusterings $C_0, C_1, \cdots, C_m$

Hierarchical clustering schemes 13

1. A sequence of clusterings $C_0, C_1, \cdots, C_m$
2. Also the values $\alpha_0, \alpha_1, \cdots, \alpha_m$
3. For each pair of objects $\mathbf{x}, \mathbf{y}$, define $d(\mathbf{x}, \mathbf{y})$
4. We prove that d(.) is a metric

1. Define d as follows:
2. Given two objects $\mathbf{x}$ and $\mathbf{y}$
3. We notice that in C_m $\mathbf{x}$ and $\mathbf{y}$ are in the same cluster
4. Let j be the least integer in $\{0, 1, \cdots, m\}$ such that $C_j, \mathbf{x}$ and $\mathbf{y}$ are in the same cluster.
5. Define $d(\mathbf{x}, \mathbf{y}) = \alpha_j$ as explained


Object Number
1
3
5
6
4
2
.00
.04
.07
.23
.31
"Strength"
or
"Value"

1. Define $d(\mathbf{x}, \mathbf{y}) = \alpha_j$ as explained
2. For example, in the figure, we have d(o, 5) = 0.04 as objects 3 and 5 are clustered at level 0.04
3. d(1, 4) = 0.31
4. d(1, 6) = 0.23
5. d(4, 2) = 0.23

Distance matrix corresponding to HCS

o	1	2	3	4	5	6
1	0.00	0.31	0.23	0.31	0.23	0.23
2	0.31	0.00	0.31	0.23	0.31	0.31
3	0.23	0.31	0.00	0.31	0.04	0.07
4	0.31	0.23	0.31	0.00	0.31	0.31
5	0.23	0.31	0.04	0.31	0.00	0.07
6	0.23	0.31	0.07	0.31	0.07	0.00

$$d(\mathbf{x}, \mathbf{x}) = 0$$

1. objects $\mathbf{x}$ and $\mathbf{x}$ are in the same cluster for all C_j
2. 0 is the smallest j.
3. So by definition, $d(\mathbf{x}, \mathbf{x}) = \alpha_0 = 0.00$
4. Conversely, if $d(\mathbf{x}, \mathbf{y}) = 0$ for some $\mathbf{x}$, and $\mathbf{y}$, it imply that $\mathbf{x}$ and $\mathbf{y}$ are in the same cluster in C_0

1. but C_0 being the weak clustering, the only element in the same cluster with $\mathbf{x}$ is $\mathbf{x}$

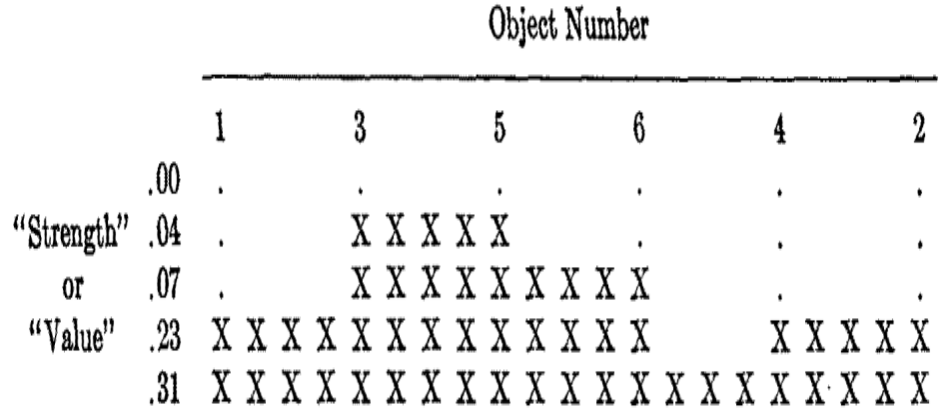
2. That is $d(\mathbf{x}, \mathbf{y}) = 0$ implies $\mathbf{x} = \mathbf{y}$.

3. Thus $d(\mathbf{x}, \mathbf{y}) = 0 \iff \mathbf{x} = \mathbf{y}$

$$d(\mathbf{x}, \mathbf{y}) = d(\mathbf{y}, \mathbf{x})$$

1. If $\mathbf{x}$ and $\mathbf{y}$ belong to a C_i with level α_i
2. Then $\mathbf{y}$ and $\mathbf{x}$ also belong to the same cluster
3. This implies $d(\mathbf{x}, \mathbf{y}) = d(\mathbf{y}, \mathbf{x}) = \alpha_i$

Triangle inequality


Object Number
1
3
5
6
4
2
.00
.04
.07
.23
.31
"Strength"
or
"Value"

Hierarchical clustering schemes 19

1. Let $\mathbf{x}, \mathbf{y}$ and $\mathbf{z}$ be three objects
2. Let $d(\mathbf{x}, \mathbf{y}) = \alpha_j$
3. $\mathbf{x}$ and $\mathbf{y}$ are in the same cluster C_j
4. Let $d(\mathbf{y}, \mathbf{z}) = \alpha_k$
5. $\mathbf{y}$ and $\mathbf{z}$ are in the same cluster C_k

1. As the clusterings are hierarchical, one of these clusters include the other

2. In fact that cluster corresponding to the larger of j and k

3. Let this be $\ell = \max(j, k)$. Then $C_\ell, \mathbf{x}, \mathbf{y}$ and $\mathbf{z}$ are all in the same cluster

4. From the definition of $d(.,.)$ we have

$$d(\mathbf{x}, \mathbf{y}) \leq \alpha_\ell$$

1. From the definition of $d(.,.)$ we have

$$d(\mathbf{x}, \mathbf{y}) \leq \alpha_\ell$$

2. $\alpha_\ell = \max(\alpha_j, \alpha_k)$

3. $d(\mathbf{x}, \mathbf{z}) \leq \max(d(\mathbf{x}, \mathbf{y}), d(\mathbf{y}, \mathbf{z}))$

4. This is classed the ultrametric inequality

1. $d(.,.)$ satisfies ultrametric inequality
2. It is stronger than the triangle inequality
3. Which would merely
$$d(\mathbf{x}, \mathbf{z}) \leq d(\mathbf{x}, \mathbf{y}) + d(\mathbf{y}, \mathbf{z})$$
4. $$\max(d(\mathbf{x}, \mathbf{y}), d(\mathbf{y}, \mathbf{z})) \leq d(\mathbf{x}, \mathbf{y}) + d(\mathbf{y}, \mathbf{z})$$

1. We have proved given a HCS, a metric d is obtained on the objects satisfying ultrametric inequality
2. We now do the converse
3. Given a distance matrix representing *some metric* d which satisfies ultrametric inequality, we will construct a HCS from it

Converse proof

1. At level 0, we have the weak clustering
2. That is six clusters. Each clustering having one object
3. The smallest element of the distance matrix ($\neq 0$) is 0.04 that appears between objects 3 and 5
4. Create a clustering with value $\alpha = 0.04$ with 3 and 5 in the same cluster
5. Objects 1, 2, 4 and 6 continue to be singleton clusters

Distance matrix corresponding to HCS

o	1	2	3	4	5	6
1	0.00	0.31	0.23	0.31	0.23	0.23
2	0.31	0.00	0.31	0.23	0.31	0.31
3	0.23	0.31	0.00	0.31	0.04	0.07
4	0.31	0.23	0.31	0.00	0.31	0.31
5	0.23	0.31	0.04	0.31	0.00	0.07
6	0.23	0.31	0.07	0.31	0.07	0.00

1. An interesting observation from the similarity table
2. $d(3, \mathbf{x}) = d(5, \mathbf{x})$ for all $\mathbf{x} = \{1, 2, 4, 6\}$
3. This gives rise to computing distance from $\mathbf{x}$ to cluster $\{3, 5\}$
4. Or computing distance from cluster $\{\mathbf{x}\}$ to cluster $\{3, 5\}$

1. We create a new object $\{3, 5\}$ in distance matrix computation
2. We remove singleton clusters $\{3\}$ and $\{5\}$ and include new object $\{3, 5\}$
3. We get next clustering using the matrix given as presented in the next slide
4. By taking smallest nonzero entry in the new matrix (0.07, between $\{3, 5\}$ and $\mathbf{x}$)
5. As 0.07 is the smallest level in the new matrix

Distance matrix corresponding to HCS

o	1	2	{3, 5}	4	6
1	0.00	0.31	0.23	0.31	0.23
2	0.31	0.00	0.31	0.23	0.31
{3, 5}	0.23	0.31	0.00	0.31	0.07
4	0.31	0.23	0.31	0.00	0.31
6	0.23	0.31	0.07	0.31	0.00

1. By taking smallest nonzero entry $d(\{3,5\},\{\mathbf{x}\});\mathbf{x} = 1,2,4,6$

2. $\min(d(\{3,5\},1),d(\{3,5\},2),d(\{3,5\},4),d(\{3,5\},6)) = \min(0.23,0.31,0.31,0.07) = 0.07$

3. Therefore merge {6} with {3, 5}

1. Repeat this process
2. Merge {3, 5} and {6} by retaining maximum distance.
3. That is max(d({3, 5}, {1}), d({6}, {1}))
4. That is max(d({3, 5}, {2}), d({6}, {2}))
5. That is max(d({3, 5}, {3, 5}), d({6}, {3, 5}))

Distance matrix corresponding to HCS

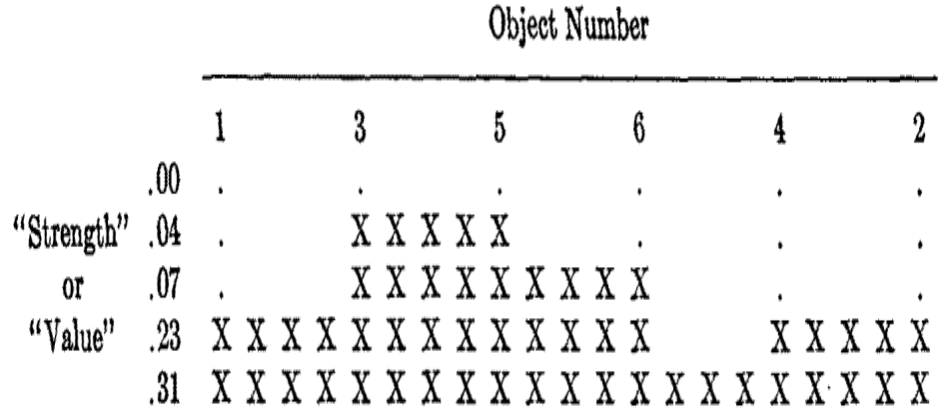
o	1	2	{3, 5, 6}	4
1	0.00	0.31	0.23	0.31
2	0.31	0.00	0.31	0.23
{3, 5, 6}	0.23	0.31	0.00	0.31
4	0.31	0.23	0.31	0.00

1. Take smallest nonzero entry $d(\{3,5,6\},\{\mathbf{x}\}); \mathbf{x} = 1, 2, 4$

2. $\min(d(\{3,5,6\},1), d(\{3,5,6\},2), d(\{3,5,6\},4)) = \min(0.23, 0.31, 0.31) = 0.23$

3. Distance 0.23 is found while merging {1} with {3, 5, 6} and {2} with {4}

4. Therefore the clusters are: {1, 3, 5, 6} and {2, 4}


Object Number
1
3
5
6
4
2
.00
.04
.07
.23
.31
"Strength"
or
"Value"

Distance matrix corresponding to HCS

o	1	2	{3, 5, 6}	4
1	0.00	0.31	0.23	0.31
2	0.31	0.00	0.31	0.23
{3, 5, 6}	0.23	0.31	0.00	0.31
4	0.31	0.23	0.31	0.00

- Remove objects {2} and {4}
- Place a new object {2, 4} whose distance is max(d(2, x), d(4, **x**))where **x** = {1, 2, {3, 5, 6}, 4}
- Remove objects {1} and {3, 5, 6}
- Place a new object {1, 3, 5, 6} whose distance is max(d({1}, **x**), d({3, 5, 6}, **x**) where **x** = 1, 2, {3, 5, 6}, 4

Distance matrix corresponding to HCS

o	{2, 4}	{1, 3, 5, 6}
{2, 4}	0.00	0.31
{3, 5, 6}	0.31	0.00

1. The key to the above process is being able to replace two or more objects by a cluster
2. Still being able to define the distance between such clusters and other objects or clusters
3. This property in turn depends on two essential facts
4. d satisfies the ultrametric inequality
5. At each stage we cluster the minimum distances

From d To HCS - A General Method

Hierarchical clustering schemes 32

1. Step 1: Clustering C_0 with value 0 is the weak clustering
2. Step 2: Assume we are given the clustering C_{i-1} with distance matrix.
3. Step 2: Let α_i be the smallest non-zero entry in the matrix
4. Step 2: Merge the pair of points or clusters with distance α_i. Create C_i with value α_i

1. Step 3: Create new distance matrix
2. Step 3: If $\mathbf{x}$ and $\mathbf{y}$ are two objects or clusters at level C_{i-1} and if $d(\mathbf{x}, \mathbf{y}) = \alpha_i$
3. Step 3: If $\mathbf{z}$ is any other point at level C_{i-1} then $d(\mathbf{x}, \mathbf{z}) = d(\mathbf{y}, \mathbf{z})$

1

2

Assume $d(\mathbf{x}, \mathbf{z}) > d(\mathbf{y}, \mathbf{z})$; Ultrametric inequality demands

$$\begin{aligned} d(\mathbf{x}, \mathbf{z}) &\leq \max(d(\mathbf{x}, \mathbf{y}), d(\mathbf{y}, \mathbf{z})) \\ &\leq \max(\alpha_i, d(\mathbf{y}, \mathbf{z})) \end{aligned}$$

As $d(\mathbf{y}, \mathbf{z}) < d(\mathbf{x}, \mathbf{z})$, it follows that

$$d(\mathbf{x}, \mathbf{z}) \leq \alpha_i$$

1. But α_i chosen to be the least non-zero distance in the matrix

2. Thus,
$$d(\mathbf{x}, \mathbf{z}) = \alpha_i$$

3. This turns out that
$$d(\mathbf{y}, \mathbf{z}) < d(\mathbf{x}, \mathbf{z}) = \alpha_i$$

4. This is contradiction. Therefore
$$d(\mathbf{x}, \mathbf{z}) = d(\mathbf{y}, \mathbf{z})$$

The Two Methods

The two methods 01

1. We have illustrated a way of going from a metric d to an HCS
2. d must satisfy ultrametric inequality
3. In general the similarity matrix does not satisfy the ultrametric inequality
4. Thus we need a method to obtain reasonable clusterings in this case

The two methods 02

1. In the Step 3 of the above procedure, we defined $d(\{\mathbf{x}, \mathbf{y}\}, \mathbf{z})$

2. Through ultrametric inequality we proved that $d(\mathbf{x}, \mathbf{z}) = d(\mathbf{y}, \mathbf{z})$

3. This lead to natural definition

$$d(\{\mathbf{x}, \mathbf{y}\}, \mathbf{z}) = d(\mathbf{x}, \mathbf{z}) = d(\mathbf{y}, \mathbf{z})$$

4. In general we may not expect $d(\mathbf{x}, \mathbf{z}) = d(\mathbf{y}, \mathbf{z})$

The two methods 03

1. Still we define $d(\{\mathbf{x}, \mathbf{y}\}, \mathbf{z})$ some function of $d(\mathbf{x}, \mathbf{z})$ and $d(\mathbf{y}, \mathbf{z})$. That is

$$d(\{\mathbf{x}, \mathbf{y}\}, \mathbf{z}) = f(d(\mathbf{x}, \mathbf{z}), d(\mathbf{y}, \mathbf{z}))$$

2. The functions *max* or *min* are considered

3. These functions are monotone invariant clustering methods

Monotone Transformations

The two methods 04

1

A monotone transformation is any function f that preserves the order of pairwise dissimilarities.
If d_{ij} and $d_{k\ell}$ are dissimilarities between pairs of points, a monotone transformation satisfies

$$d_{ij} \leq d_{k\ell} \implies f(d_{ij}) \leq f(d_{k\ell})$$

. This ensures that the relative ordering of dissimilarities is preserved.
A clustering method is monotone invariant if it produces the same dendrogram for any dissimilarity matrix transformed by a monotone function f

min Method

Minimum method 01

1. Step 1: Clustering C_0 with value 0 is the weak clustering
2. Step 2: Assume we are given the clustering C_{i-1} with distance matrix.
3. Step 2: Let α_i be the smallest non-zero entry in the matrix
4. Step 2: Merge the pair of points or clusters with distance α_i. Create C_i u value α_i

Minimum method 02

1. Step 3: Create new distance matrix for C_i as follows

2. Step 3: If $\mathbf{x}$ and $\mathbf{y}$ are two objects or clusters at level C_i and NOT in C_{i-1} then defined d as:

$$d(\{\mathbf{x}, \mathbf{y}\}, \mathbf{z}) = \min(d(\mathbf{x}, \mathbf{z}), d(\mathbf{y}, \mathbf{z}))$$

max Method

Maximum method 01

1. Step 1: Clustering C_0 with value 0 is the weak clustering
2. Step 2: Assume we are given the clustering C_{i-1} with distance matrix.
3. Step 2: Let α_i be the smallest non-zero entry in the matrix
4. Step 2: Merge the pair of points or clusters with distance α_i. Create C_i u value α_i

Maximum method 02

1. Step 3: Create new distance matrix for C_i as follows

2. Step 3: If $\mathbf{x}$ and $\mathbf{y}$ are two objects or clusters at level C_i and NOT in C_{i-1} then defined d as:

$$d(\{\mathbf{x}, \mathbf{y}\}, \mathbf{z}) = \max(d(\mathbf{x}, \mathbf{z}), d(\mathbf{y}, \mathbf{z}))$$

Hierarchical Clustering Schemes

1. We will discuss the *min* method with examples
2. We will discuss the `max` method with examples
3. Discuss the classification of hierarchical clustering methods