Multivalued and Binary Ultrametrics, and Clusterings
Iasi, Romania

Prof. Dan A. Simovici

University of Massachusetts Boston

* * * * *

dedicated to the memory of
Prof. Grigore C. Moisil

UMASS BOSTON

What is a metric?

- A metric is an abstractization of the notion of distance.
- Metrics were introduced by Maurice Fréchet in 1906 in the paper *Sur quelques points du calcul fonctionnel* in *Rendiconti del Circolo Matematico di Palermo* 22, 1-74.

UMASS BOSTON

The simplest example

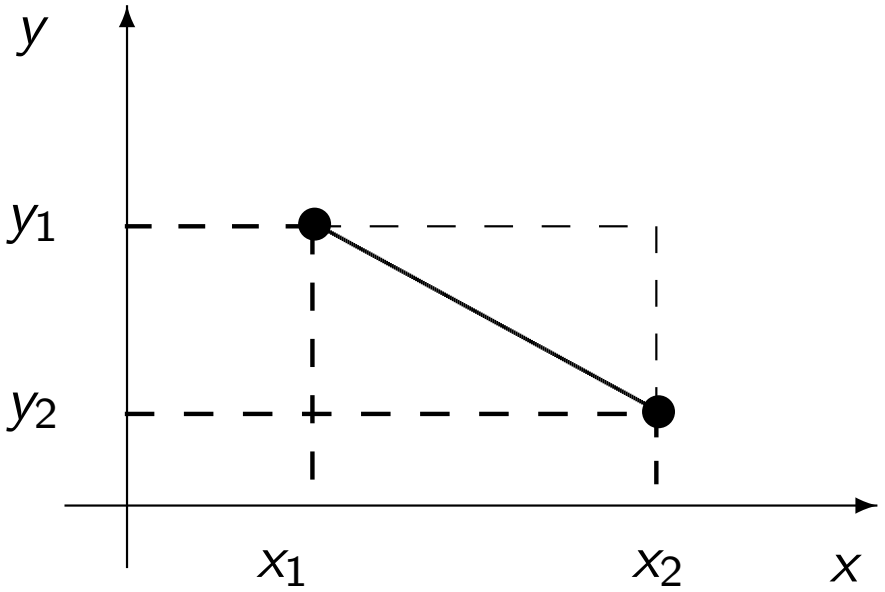
If $\boldsymbol{x}, \boldsymbol{y}$ are two points in $\mathbb{R}^2$,

$$\boldsymbol{x} = \begin{pmatrix} x_1 \\ y_1 \end{pmatrix} \text{ and } \boldsymbol{x} = \begin{pmatrix} x_2 \\ y_2 \end{pmatrix},$$

the distance (metric) between $\boldsymbol{x}$ and $\boldsymbol{y}$ is

$$d(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2},$$

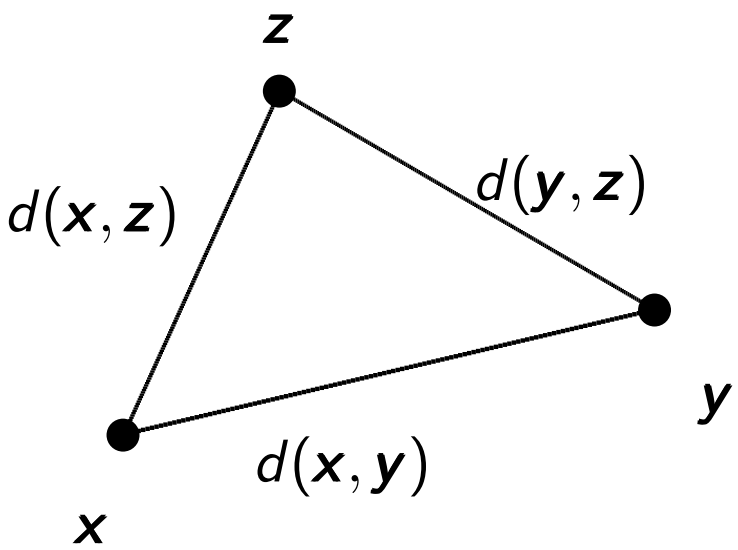
that is, the distance that corresponds to our intuition.



Intuitive Properties of Distances

The pair (S, d) is a *metric space* if

- $d(x, y) \geqslant 0$ (non-negativity);
- $d(x, y) = 0$ if and only if $x = y$;
- $d(x, y) = d(y, x)$ (symmetry);
- $d(x, y) \leqslant d(x, z) + d(z, y)$ (triangular inequality)



Examples - I

- on $\mathbb{R}$, the set of reals: $d(x, y) = |x - y|$;
- on the set of positive numbers: $d(x, y) = |\log \frac{x}{y}|$.

Examples - II

The Minkowski family of metrics on $\mathbb{R}^n$:

$$d_p(\boldsymbol{x}, \boldsymbol{y}) = \left((x_1 - y_1)^p + \cdots + (x_n - y_n)^p\right)^{\frac{1}{p}}$$

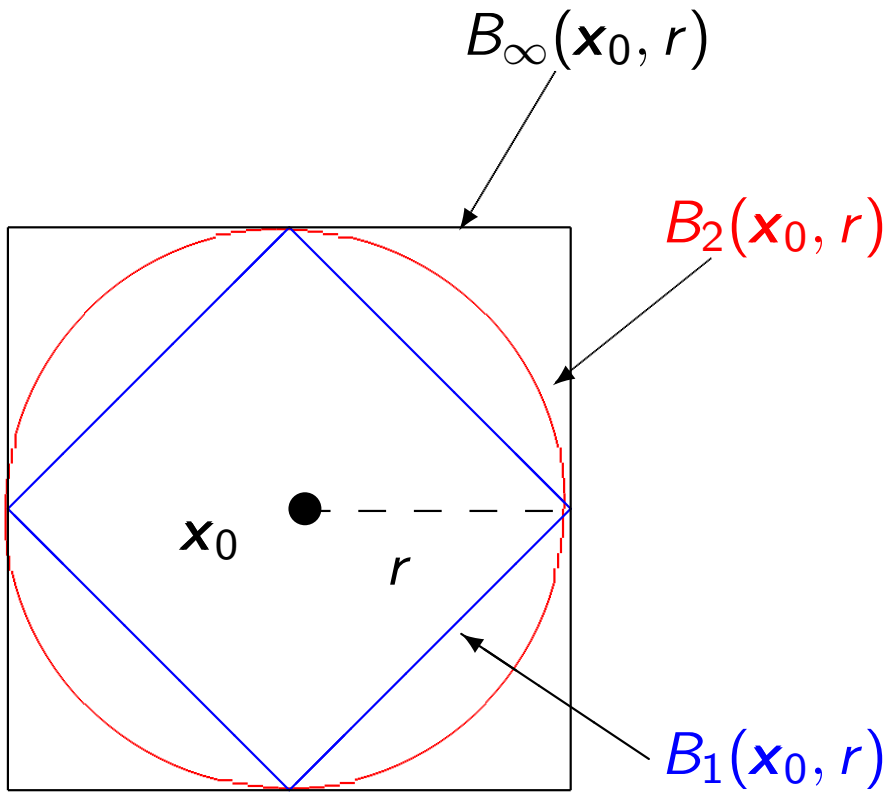
for $p \geqslant 1$.
Special cases:

- $p = 1$: $d_1(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^n |x_i - y_i|$ (the Manhattan metric);
- $p = 2$: $d_2(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{\sum_{i=1}^n |x_i - y_i|^2}$ (the Euclidean metric);
- $p \to \infty$: $d_\infty(\boldsymbol{x}, \boldsymbol{y}) = \max_{1 \leqslant i \leqslant n} |x_i - y_i|$.

How do spheres look in Minkowski spaces?

What is a closed sphere centered in x_0 of radius r in a metric space (S, d)?

$$B(x_0, r) = \{x \in S \mid d(x_0, x) \leqslant r\}$$



Why distances matter for Computer Scientists?

- One of the main problems in machine learning and data mining is clustering. This is grouping objects such that
  - similar objects belong to the same group (cluster);
  - objects that belong to distinct groups (clusters) are different.

  Thus, we have to measure how dissimilar objects are in order to group them properly.
- outlier detection (identifying objects that are "unusual" in a set of objects).

UMASS BOSTON

What is an ultrametric?

Definition

An ultrametric on a set S is a mapping $d : S^2 \longrightarrow \mathbb{R}_{\geqslant 0}$ that has the following properties:

- $d(x, y) = 0$ if and only if $x = y$ for $x, y \in S$;
- $d(x, y) = d(y, x)$ for $x, y \in S$;
- $d(x, y) \leqslant \max\{d(x, z), d(z, y)\}$ for $x, y, z \in S$.

If property (i) is replaced by the weaker requirement that $d(x, x) = 0$ for $x \in S$, then d is a *quasi-ultrametric* on S.

UMASS BOSTON

The Triangular vs. the Ultrametric Inequality

The triangular inequality

$$d(x, y) \leqslant d(x, z) + d(z, y)$$

is replaced by the stronger ultrametric inequality

$$d(x, y) \leqslant \max\{d(x, z), d(z, y)\}.$$

Example

Let $\pi = \{B_1, \ldots, B_m\}$ be an m-block partition of a nonempty set S. Define the mapping $d_\pi : S^2 \longrightarrow \mathbb{R}_{\geqslant 0}$ by

$$d_\pi(x, y) = \begin{cases} 0 & \text{if } \{x, y\} \subseteq B_i \text{ for some } B \in \pi \\ 1 & \text{otherwise,} \end{cases}$$

for $x, y \in S$.

d_π is a quasi-ultrametric. Indeed, $d_\pi(x, x) = 0$ and $d_\pi(x, y) = d_\pi(y, x)$ for $x, y \in S$.

Let x, y, z be three arbitrary elements in S. If $d_\pi(x, y) = 1$, then x and y belong to two distinct blocks of the partition π, say to B_i and B_j, respectively. If $z \in B_i$, then $d_\pi(x, z) = 0$ and $d_\pi(z, y) = 1$; similarly, if $z \in B_j$, then $d_\pi(x, z) = 1$ and $d_\pi(z, y) = 0$. Finally, if $z \in B_k$, where $k \notin \{i, j\}$, then $d(x, z) = 1$ and $d(z, y) = 1$. In all cases, the ultrametric inequality is satisfied.

The Same Example in Terms of Equivalences

Example

If ρ is an equivalence on a set S and $c : S \times S \longrightarrow \{0, 1\}$, is the characteristic function of ρ, then $1 - c$ is a quasi-ultrametric on S.

Example

Let A be a finite set and let $A^{\mathbb{N}}$ be the set of functions of the form $f : \mathbb{N} \longrightarrow A$. For $f, g : \mathbb{N} \longrightarrow A$ define $\delta(f,g) = \min\{n \in \mathbb{N} \mid f(n) \neq g(n)\}$ and

$$d(f,g) = \begin{cases} 0 & \text{if } f = g \\ 2^{-\delta(f,g)} & \text{otherwise.} \end{cases}$$

d is an ultrametric. Let $f, g, h : \mathbb{N} \longrightarrow A$ and let $\delta(f,g) = p$ and $\delta(g,h) = q$. Suppose that $p \leqslant q$. In this case, it follows that $\delta(f,h) = p$. Thus, $d(f,g) = 2^{-p} = d(f,h) \geqslant d(g,h)$ and the ultrametric inequality $d(f,g) \leqslant \max\{d(f,h), d(g,h)\}$ is satisfied. The remaining cases are left to the reader.

Ultrametrics on Sets of Words

Example

Let A be a non-empty finite set, whose elements are referred as *symbols*. A word of length n is a sequence $w = (a_0, \ldots, a_{n-1})$ of symbols. The number n is the *length* of the word w, denoted by $|w|$. The set of words over A is denoted by A^*.
The null word λ is the unique sequence of length 0.
If $u, v \in A^*$ let

$$d(u,v) = \begin{cases} 2^{-(p+1)} & \text{if } u \neq v, u = zu', v = zv', z \text{ is the longest} \\ & \text{common prefix of } u, v \text{ and } |z| = p; \\ 0 & \text{otherwise.} \end{cases}$$

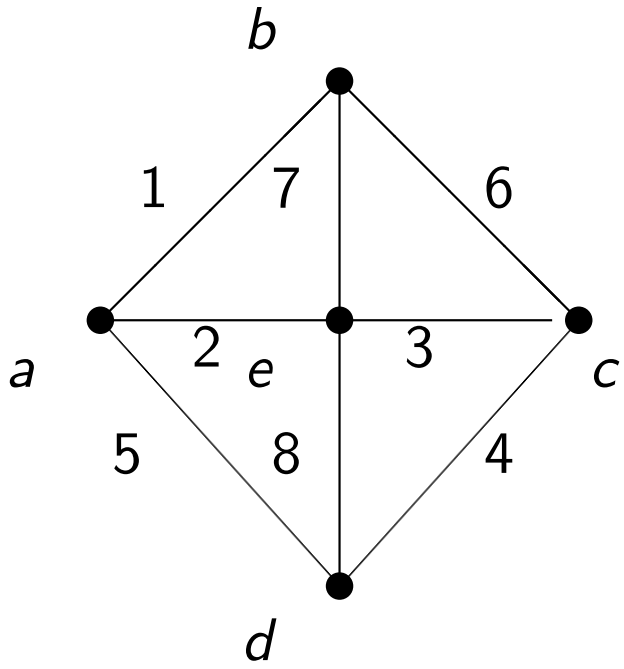
An Example from Weighted Graph

Example

Let $\mathcal{G} = (V, E, w)$ be a weighted undirected finite graph, that is a graph (V, E) equipped with a function $w : E \longrightarrow \mathbb{R}_{\geqslant 0}$. If $\boldsymbol{p}$ is a path in the graph let $\mu(\boldsymbol{p})$ be the largest weight of an edge that occurs in the path $\boldsymbol{p}$. Define $P(x, y)$ as the set of paths in $\mathcal{G}$ and $d(x, y) = \min\{w(\boldsymbol{p}) \mid \boldsymbol{p} \in P(x, y)\}$. We claim that d is an ultrametric on V.
If $\boldsymbol{r} \in P(x, z)$ and $\boldsymbol{s} \in P(z, y)$, then $\boldsymbol{rs} \in P(\boldsymbol{x}, \boldsymbol{z})$. Suppose that $\boldsymbol{r}$ is chosen such that it contains an edge (u, v) with $w(u, v) = \mu(\boldsymbol{r})$ and $\boldsymbol{s}$ is chosen such it contains an edge (u', v') with $w(u', v') = \mu(q)$. Then, the edge with largest weight on $\boldsymbol{rs}$ will have the weight

$$\max\{w(u, v), w(u'v')\} = \max\{d(x, z), d(z, y)\}.$$

Since $d(x, y)$ is the minimal largest value that occurs on a path that joins x and y and $\boldsymbol{rs}$ is one of these paths, it follows that
$d(x, y) \leqslant \max\{d(x, z), d(z, y)\}$.



	a	b	c	d	e
a	0	1	3	4	2
b	1	0	3	4	2
c	3	3	0	4	3
d	4	4	4	0	4
e	2	2	3	4	0

Theorem

Let $a_0, a_1, a_2 \in \mathbb{R}$ be three numbers. If $a_i \leqslant \max\{a_j, a_k\}$ for every permutation (i, j, k) of the set $\{0, 1, 2\}$, then two of the numbers are equal and the third is not larger than the two others.

Proof.

Suppose that a_i is the least of the numbers a_0, a_1, a_2 and a_j, a_k are the remaining numbers. Since $a_j \leqslant \max\{a_i, a_k\} = a_k$ and $a_k \leqslant \max\{a_i, a_j\} = a_j$, it follows that $a_j = a_k \geqslant a_i$. □

Any triangle of an ultrametric space is isosceles

Corollary

Let (S, d) be an ultrametric space. For every $x, y, z \in S$, two of the numbers $d(x, y), d(x, z), d(y, z)$ are equal and the third is not larger than the two other equal numbers.

Any triangle in an ultrametric spaces is isosceles: two of the largest sides are equal.

Theorem

Let $B(x,r)$ be a closed sphere in the ultrametric space (S,d). If $z \in B(x,d)$, then $B(x,r) = B(z,r)$. In other words, in an ultrametric space, a closed sphere has all its points as centers.

Proof.

Suppose that $z \in B(x,r)$, so $d(x,z) \leqslant r$. Let $y \in B(z,r)$. Since $d(y,x) \leqslant \max\{d(y,z), d(z,x)\} \leqslant r$, we have $y \in B(x,r)$. Conversely, if $y \in B(x,r)$, we have $d(y,z) \leqslant \max\{d(y,x), d(x,z)\} \leqslant r$, hence $y \in B(z,r)$. □

Both closed and open spheres in ultrametric spaces are clopen sets as we show next.

Theorem

If d is an ultrametric on S, then any closed sphere $B(t,r)$ and any open sphere $C(t,r)$ are clopen sets in the topological ultrametric space.

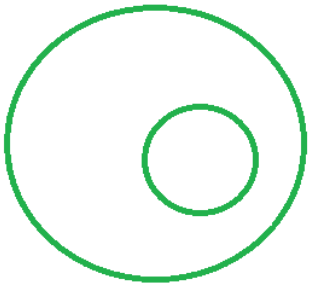
Proof.

We already know that $B(t,r)$ is closed. To prove that this set is also open if d is an ultrametric, let $s \in B(t,r)$. By a previous result, s is a center of the sphere. Therefore, $C\left(s, \frac{r}{2}\right) \subseteq B(t,r)$, so $B(t,r)$ is open. $\square$

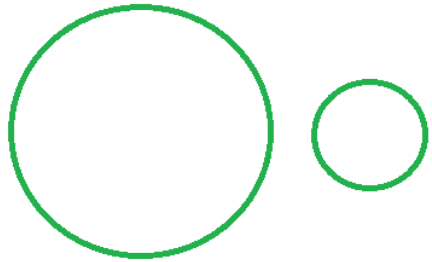
Theorem

If two closed spheres $B(x, r)$ and $B(y, r')$ of an ultrametric space have a point in common, then one of the closed spheres is included in the other. If two spheres of equal radius have a point in common they are equal.

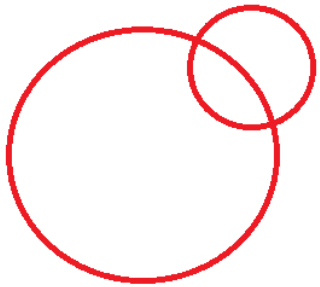
Possible positions of spheres in an ultrametric space



Legitimate position of two spheres



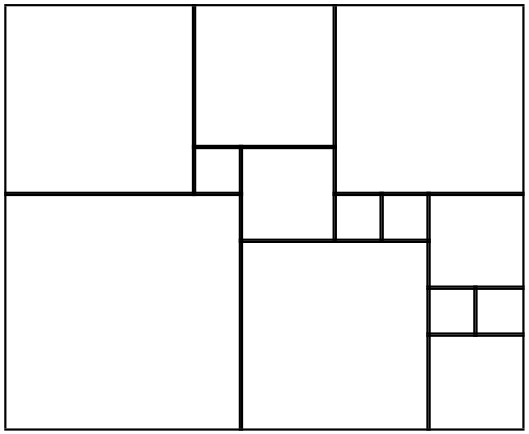
Legitimate position of two spheres



Impossible position in an ultrametric space

Can a metric space be a union of disjoint "spheres"?

YES!



The rectangle is the union of the spheres in the sense of d_∞ (which look like squares)!

In an ultrametric space (S, d), non-empty spheres of radius a constitute a partition of the space! In other words, S is "tiled" by spheres that have the same radii a.

Amplitude of a Sequence

Definition

Let (S, d) be a dissimilarity space and let $S(x, y)$ be the set of all nonnull sequences $\boldsymbol{s} = (s_1, \ldots, s_n) \in \mathbf{Seq}(S)$ such that $s_1 = x$ and $s_n = y$.
The *d-amplitude of* $\boldsymbol{s}$ is the number $amp_d(\boldsymbol{s})$ given by:

$$amp_d(\boldsymbol{s}) = \max\{d(s_i, s_{i+1}) \mid 1 \leqslant i \leqslant n-1\}.$$

UMASS BOSTON

Theorem

Let (S, d) be an ultrametric space, $x, y \in S$, and let $S(x, y) \subseteq \mathbf{Seq}(S)$ be the set of sequences that start with x and end with y. We have $d(x, y) = \min\{amp_d(\boldsymbol{s}) \mid \boldsymbol{s} \in S(x, y)\}$.

Proof: Since d is an ultrametric, we have $d(x, y) \leqslant amp_d(\boldsymbol{s})$ for any nonnull sequence $\boldsymbol{s} = (s_1, \ldots, s_n)$ such that $s_1 = x$ and $s_n = y$. Therefore,

$$d(x, y) \leqslant \min\{amp_d(\boldsymbol{s}) \mid \boldsymbol{s} \in S(x, y)\}.$$

The equality of the theorem follows from the fact that $(x, y) \in S(x, y)$.

Definition

Let S be a set. A *hierarchy* on S is a collection of sets $\mathcal{H} \subseteq \mathcal{P}(S)$ that satisfies the following conditions:

- the members of $\mathcal{H}$ are nonempty sets;
- $S \in \mathcal{H}$;
- for every $x \in S$, we have $\{x\} \in \mathcal{H}$;
- if $H, H' \in \mathcal{H}$ and $H \cap H' \neq \emptyset$, then we have either $H \subseteq H'$ or $H' \subseteq H$.

Hierarchy Construction

A standard technique for constructing a hierarchy on a set S starts with a rooted tree $(\mathcal{T}, v_0)$ whose nodes are labeled by subsets of the set S. Let V be the set of vertices of the tree $\mathcal{T}$.
The function $\mu : V \longrightarrow \mathcal{P}(S)$, which gives the label $\mu(v)$ of each node $v \in V$, is defined as follows:

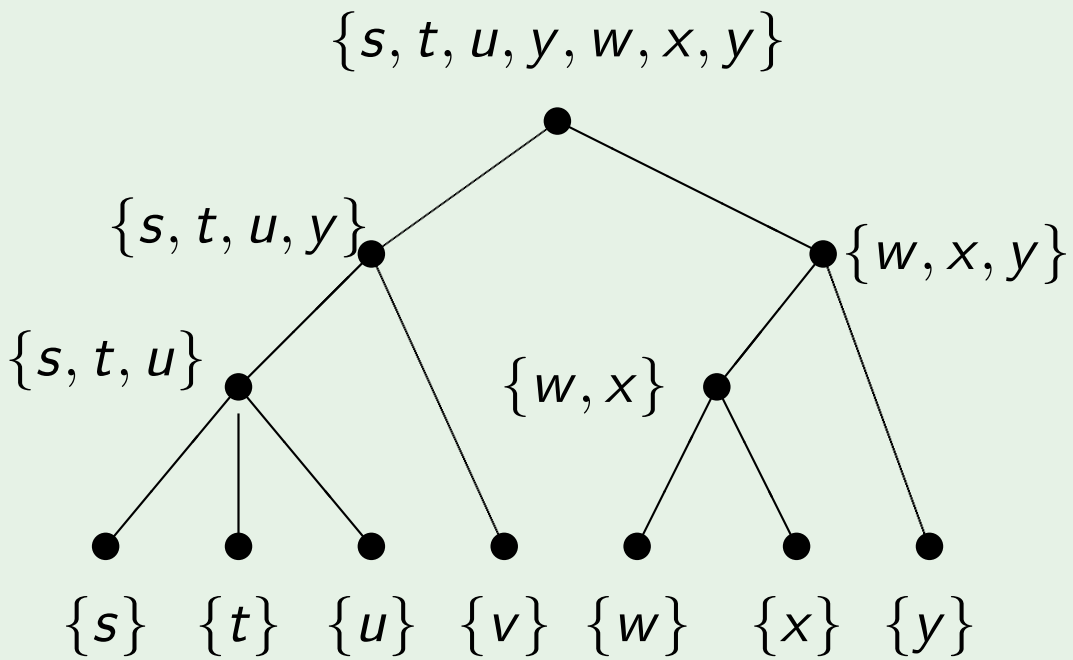
- the tree $\mathcal{T}$ has $|S|$ leaves, and each leaf v is labeled by a distinct singleton $\mu(v) = \{x\}$ for $x \in S$;
- if an interior vertex v of the tree has the descendants $v_1, v_2, \ldots, v_n$, then $\mu(v) = \bigcup_{i=1}^{n} \mu(v_i)$.

Example

Let $S = \{s, t, u, v, w, x, y\}$ and let

$$\begin{aligned}\mathcal{H} &= \{\{s\}, \{t\}, \{u\}, \{v\}, \{w\}, \{x\}, \{y\}, \\ &\quad \{s, t, u\}, \{w, x\}, \{s, t, u, v\}, \{w, x, y\}, \{s, t, u, v, w, x, y\}\}\end{aligned}$$

the hierarchy on the set S that labels the nodes of $\mathcal{T}$.



Definition

Let $\mathcal{H}$ be a hierarchy on a set S. A *grading function* for $\mathcal{H}$ is a function $h : \mathcal{H} \longrightarrow \mathbb{R}$ that satisfies the following conditions:

- $h(\{x\}) = 0$ for every $x \in S$, and
- if $H, K \in \mathcal{H}$ and $H \subset K$, then $h(H) < h(K)$.

If h is a grading function for a hierarchy $\mathcal{H}$, the pair $(\mathcal{H}, h)$ is a *graded hierarchy*.

Example

For the hierarchy $\mathcal{H}$ defined above on the set $S = \{s, t, u, v, w, x, y\}$, the function $h : \mathcal{H} \longrightarrow \mathbb{R}$ given by

$$h(\{s\}) = h(\{t\}) = h(\{u\}) = h(\{v\}) = h(\{w\}) = h(\{x\}) = h(\{y\}) = 0,$$
$$h(\{s, t, u\}) = 1, h(\{w, x\}) = 1, h(\{s, t, u, v\}) = 2, h(\{w, x, y\}) = 2,$$
$$h(\{s, t, u, v, w, x, y\}) = 3,$$

is a grading function and the pair $(\mathcal{H}, h)$ is a graded hierarchy on S.

Theorem

Let $(\mathcal{H}, h)$ be a graded hierarchy on a finite set S. Define the function $d : S^2 \longrightarrow \mathbb{R}$ as $d(x, y) = \min\{h(U) \mid U \in \mathcal{H} \text{ and } \{x, y\} \subseteq U\}$ for $x, y \in S$. The mapping d is an ultrametric on S.

Proof

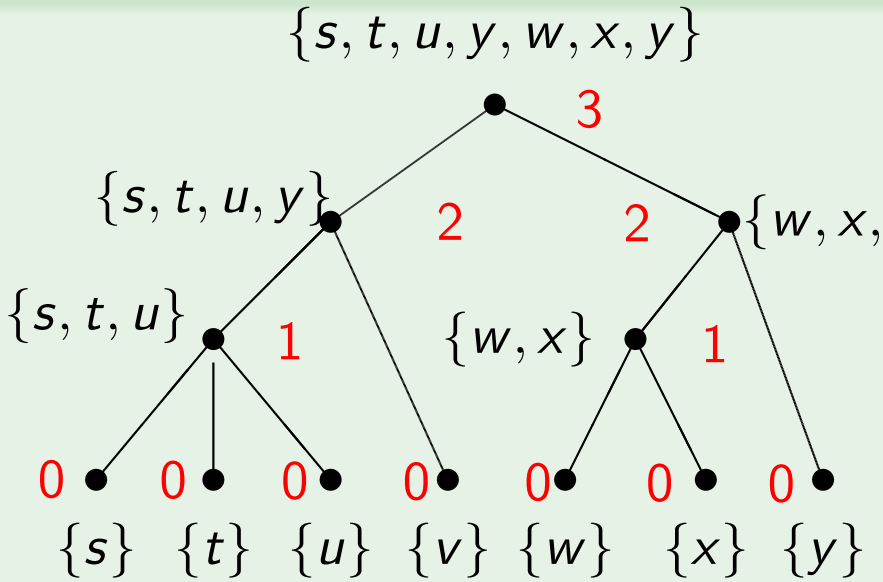
Observe that for every $x, y \in S$ there exists a set $H \in \mathcal{H}$ such that $\{x, y\} \subseteq H$ because $S \in \mathcal{H}$.
Clearly, $d(x, x) = 0$. Conversely, suppose that $d(x, y) = 0$. Then, there exists $H \in \mathcal{H}$ such that $\{x, y\} \subseteq H$ and $h(H) = 0$. If $x \neq y$, then $\{x\} \subset H$, hence $0 = h(\{x\}) < h(H)$, which contradicts the fact that $h(H) = 0$. Thus, $x = y$.
The symmetry of d is immediate.
Let $x, y, z \in S$, and suppose that $d(x, y) = p$, $d(x, z) = q$, and $d(z, y) = r$. There exist $H, K, L \in \mathcal{H}$ such that $\{x, y\} \subseteq H$, $h(H) = p$, $\{x, z\} \subseteq K$, $h(K) = q$, and $\{z, y\} \subseteq L$, $h(L) = r$. Since $K \cap L \neq \emptyset$ (because both sets contain z), we have either $K \subseteq L$ or $L \subseteq K$, so $K \cup L$ equals either K or L and, in either case, $K \cup L \in \mathcal{H}$. Since $\{x, y\} \subseteq K \cup L$, it follows that

$$d(x, y) \leqslant h(K \cup L) = \max\{h(K), H(L)\} = \max\{d(x, z), d(z, y)\},$$

which is the ultrametric inequality.

The ultrametric that corresponds to the grading function

Example



	s	t	u	v	w	x	y
s	0	1	1	2	3	3	3
t	1	0	1	2	3	3	3
u	1	1	0	2	3	3	3
v	2	2	2	0	3	3	3
w	3	3	3	3	0	1	2
x	3	3	3	3	1	0	2
y	3	3	3	3	2	2	0

UMASS BOSTON

From Ultrametrics to Equivalence Relations

Let $d : S \times S \longrightarrow \mathbb{R}_{\geq 0}$ be an ultrametric on the set S.
If $r > 0$, the relation

$$\rho_r = \{(x, y) \in S \times S \mid d(x, y) < r\}$$

is an equivalence on S. Similarly, for every $r \geq 0$, the relation

$$\eta_r = \{(x, y) \in S \times S \mid d(x, y) \leqslant r\}$$

is an equivalence on S.
We verify only the transitivity for ρ_r. Suppose that $(x, z), (z, y) \in \rho_r$.
Then,

$$d(x, y) \leqslant \max\{d(x, z), d(z, y)\} < r,$$

so $(x, y) \in \rho_r$.

Ultrametrics and Equivalences

Theorem

Let S be a finite set and let $d : S \times S \longrightarrow \mathbb{R}_{\geqslant 0}$ be a function whose range is range$(d) = \{r_1, \ldots, r_m\}$, where $r_1 = 0$ such that $d(x, y) = 0$ if and only if $x = y$. Define

$$\eta_r = \{(x, y) \in S \times S \mid d(x, y) \leqslant r\}.$$

The function d is an ultrametric on S if and only if the sequence of relations $\eta_{r_1}, \ldots, \eta_{r_m}$ is an increasing chain of equivalences on S such that $\eta_{r_1} = \iota_S$ and $\eta_{r_m} = \theta_S$.

Proof

Suppose that d is an ultrametric on S. We have $(x, x) \in \eta_{r_i}$ because $d(x, x) = 0$, so all relations η_{r_i} are reflexive. Also, it is clear that the symmetry of d implies $(x, y) \in \eta_{r_i}$ if and only if $(y, x) \in \eta_{r_i}$, so these relations are symmetric.

The ultrametric inequality is essential for proving the transitivity of the relations η_{r_i}. If $(x, y), (y, z) \in \eta_{r_i}$, then $d(x, y) \leqslant r_i$ and $d(y, z) \leqslant r_i$, which implies $d(x, z) \leqslant \max\{d(x, y), d(y, z)\} \leqslant r_i$. Thus, $(x, z) \in \eta_{r_i}$, which shows that every relation η_{r_i} is transitive and therefore an equivalence.

The sequence of relations $\eta_{r_1} \leqslant \eta_{r_2} \leqslant \cdots \leqslant \eta_{r_m}$ is clearly a chain of equivalences.

Conversely, suppose that $\eta_{r_1}, \ldots, \eta_{r_m}$ is an increasing sequence of equivalences on S such that $\eta_{r_1} = \iota_S$ and $\eta_{r_m} = \theta_S$, where $\eta_{r_i} = \{(x, y) \in S \times S \mid d(x, y) \leqslant r_i\}$ for $1 \leqslant i \leqslant m$ and $r_1 = 0$.

Note that $d(x, y) = 0$ is equivalent to $(x, y) \in \eta_{r_1} = \iota_S$, that is, to $x = y$.

Proof (cont'd)

We claim that $d(x,y) = \min\{r \mid (x,y) \in \eta_r\}$.
Indeed, since $\eta_{r_m} = \theta_S$, it is clear that there is an equivalence η_{r_i} such that $(x,y) \in \eta_{r_i}$. If $(x,y) \in \eta_{r_i}$, the definition of η_{r_i} implies $d(x,y) \leqslant r_i$, so $d(x,y) \leqslant \min\{r \mid (x,y) \in \eta_r\}$. This inequality can be easily seen to become an equality since $(x,y) \in \eta_{d(x,y)}$. This implies immediately that d is symmetric.
To prove the ultrametric inequality, let $p = \max\{d(x,z), d(z,y)\}$. Since $(x,z) \in \eta_{d(x,z)} \subseteq \eta_p$ and $(z,y) \in \eta_{d(z,y)} \subseteq \eta_p$, it follows that $(x,y) \in \eta_p$, due to the transitivity of the equivalence η_p. Thus,
$d(x,y) \leqslant p = \max\{d(x,z), d(z,y)\}$, which proves the triangular inequality for d.

The Size of the Range of an Ultrametric

Theorem

Let d be an ultrametric on a finite set S. If $|S| = n$, then d takes at most $n - 1$ positive values.

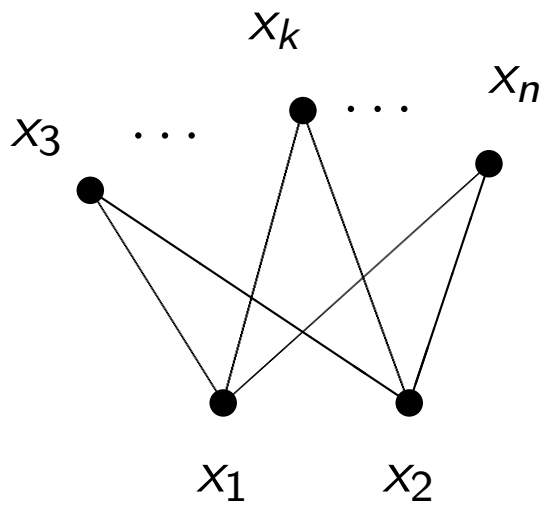
Proof

The proof is by induction on $n \geqslant 2$.

The base case: The base case, $n = 2$ is immediate.

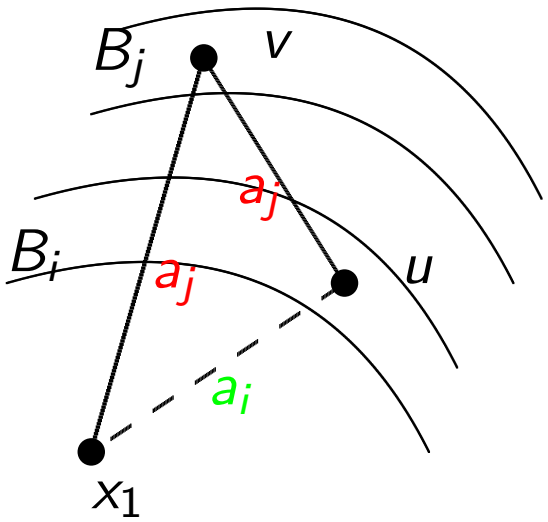
Proof (cont'd)

Let $n \geqslant 3$ and suppose that the statement holds for $m \leqslant n$.
Let $S = \{x_1, x_2, \ldots, x_n\}$. Without loss of generality we may assume that $d(x_1, x_2) \leqslant d(x_i, x_j)$ for all i, j such that $1 \leqslant i \neq j \leqslant n$. Then, $d(x_k, x_1) = d(x_k, x_2)$ for all k such that $3 \leqslant k \leqslant n$.



Proof (cont'd)

Let $\{a_1, \ldots, a_r\} = \{d(x_k, x_1) \mid 3 \leqslant k \leqslant n\}$, where $a_1 < \cdots < a_r$. Define $B_j = \{x_k \mid k \geqslant 3 \text{ and } d(x_k, x_1) = a_j\}$ for $1 \leqslant j \leqslant r$.
The collection $B_1, \ldots, B_r$ is a partition of the set $\{x_3, x_4, \ldots, x_n\}$. Let $m_j = |B_j|$; then $m_1 + \cdots + m_r = n - 2$.
If $u \in B_i$ and $v \in B_j$ with $i < j$, then $d(u, x_1) = a_i < a_j = d(v, x_1)$, hence $d(u, v) = a_j$.



Therefore, the values of $d(u, v)$ for u, v in distinct blocks of the partition belong to the set $\{a_1, \ldots, a_r\}$. By the induction hypothesis, for each k, $1 \leqslant k \leqslant r$, the restriction of d to B_k can take at most $m_k - 1$ distinct positive values; therefore, d can take at most $1 + r + (m_1 - 1) + \cdots + (m_r - 1) = n - 1$ distinct positive values on S.

Multivalued vs. Binary Ultrametrics

Theorem

Let S be a finite set and let $d : S \times S \longrightarrow \mathbb{R}_{\geqslant 0}$ be a an ultrametric whose range is range$(d) = \{r_1, \ldots, r_m\}$, where $r_1 = 0$. Then d is a linear combination with positive coefficients of $m - 2$ binary quasi-ultrametrics,

Proof

Note that we have $m-1$ positive values of d. Assume that $r_1 < r_2 < \cdots < r_m$.

Consider the equivalences ϵ_{r_i} for $2 \leqslant i \leqslant m$ and the corresponding binary ultrametrics d_i given by

$$d_i(x,y) = \begin{cases} 0 & \text{if } d(x,y) < r_i, \\ 1 & \text{otherwise} \end{cases}$$

for $(x,y) \in S \times S$. We claim that

$$d(x,y) = r_2 d_2(x,y) + (r_3 - r_2) d_3(x,y) + \cdots + (r_m - r_{m-1}) d_m(x,y)$$

for $x, y \in S$.

Indeed, note that $d(x,y) \in \{r_1, \ldots, r_m\}$ by the definition of the range of d. Suppose that we have

$$r_1 < r_2 < \cdots < r_k = d(x,y) < r_{k+1} < \cdots < r_m.$$

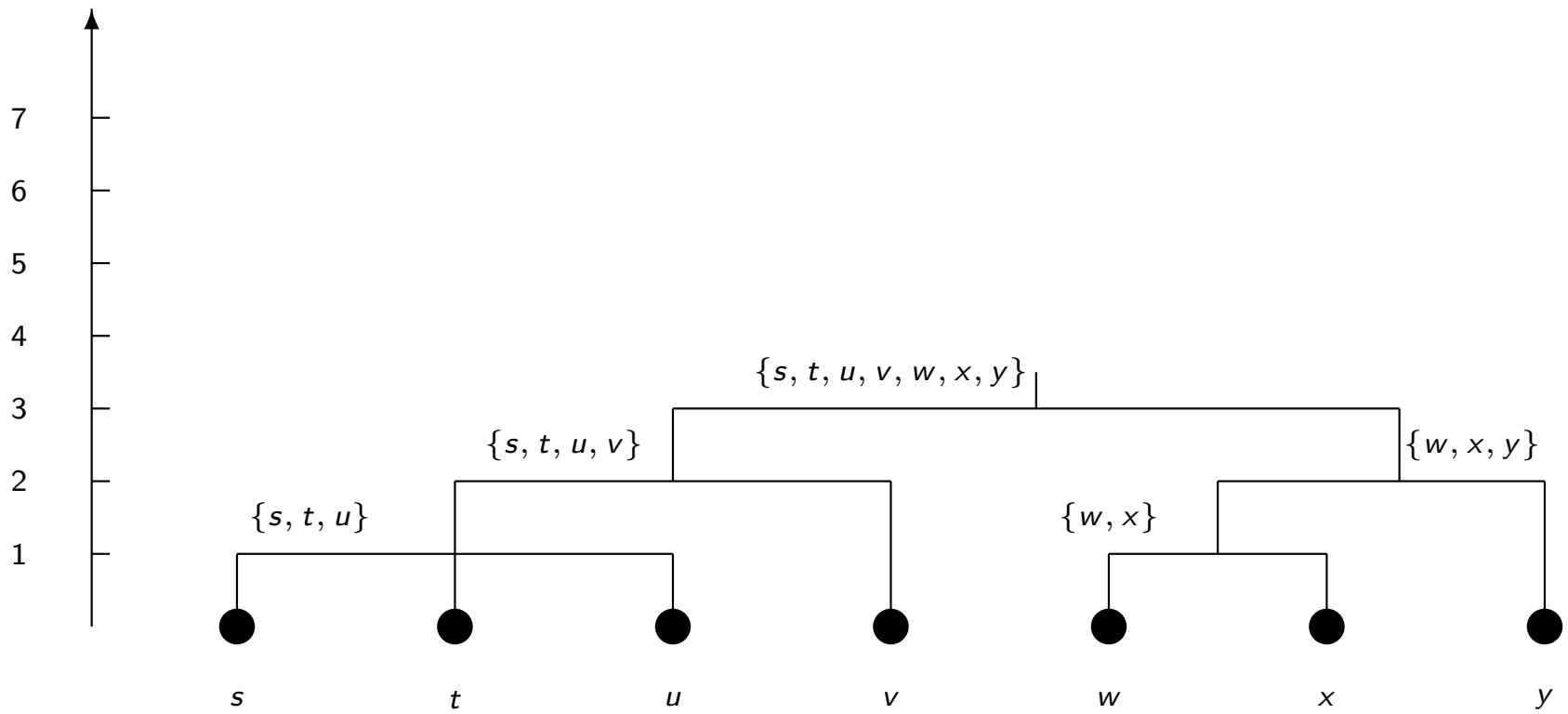
By the definition of the relations ϵ_i we have

$$d_1(x,y) = \cdots = d_k(x,y) = 1,$$

Dendrograms

We shall draw the tree of a graded hierarchy $(\mathcal{H}, h)$ using a special representation known as a *dendrogram*.
In a dendrogram, an interior vertex K of the tree is represented by a horizontal line drawn at the height $h(K)$.
The value $d(x, y)$ of the ultrametric d generated by a hierarchy $\mathcal{H}$ is the smallest height of a member of a hierarchy that contains both x and y.
This allows us to "read" the value of the ultrametric generated by $\mathcal{H}$ directly from the dendrogram of the hierarchy.

A Dendrogram



Hierarchical clustering is a recursive process that

- begins with a metric space of objects (S, d), and
- produces a chain of partitions on the set of objects such that in each partition
  - close objects in the sense of d (that is, similar objects) belong to the same block, and
  - objects that belong to distinct blocks tend to be dissimilar.

Agglomerative Hierarchical Clustering (AHC)

AHC begins with the unit partition $\pi^1 = \alpha_S$ whose blocks are singletons.
If the partition constructed at step k is

$$\pi^k = \{U_1^k, \ldots, U_{m_k}^k\},$$

then two distinct blocks U_p^k and U_q^k of this partition are selected using a *selection criterion*. These blocks are fused and a new partition

$$\pi^{k+1} = \{U_1^k, \ldots, U_{p-1}^k, U_{p+1}^k, \ldots, U_{q-1}^k, U_{q+1}^k, \ldots, U_p^k \cup U_q^k\}$$

is formed. Clearly, we have $\pi^k \prec \pi^{k+1}$.
The process must end because the poset $(PART(S), \leqslant)$ is of finite height.
The algorithm halts when the one-block partition ω_S is reached.

The Cohesiveness of Groups of Objects

For a group of object U let $\boldsymbol{c}_U$ be their centroid given by

$$\boldsymbol{c}_U = \frac{1}{|U|}\sum\{\boldsymbol{o} \in U\}.$$

The sum of square errors of U is the number

$$sse(U) = \sum\{d^2(\boldsymbol{o}, \boldsymbol{c}_U) \mid \boldsymbol{o} \in U\}.$$

The smaller $sse(U)$ the more cohesive U is.

The Quality of Partitions

If two blocks U and V of a partition π are fused into a new block W to yield a new partition π' that covers π, then the variation of the sum of squared errors is given by

$$\begin{aligned} sse(\pi') - sse(\pi) &= \sum\{d^2(\boldsymbol{o}, \boldsymbol{c}_W) | \boldsymbol{o} \in U \cup V\} \\ &\quad - \sum\{d^2(\boldsymbol{o}, \boldsymbol{c}_U) | \boldsymbol{o} \in U\} - \sum\{d^2(\boldsymbol{o}, \boldsymbol{c}_V) | \boldsymbol{o} \in V\}. \end{aligned}$$

The centroid of the new cluster W is given by

$$\boldsymbol{c}_W = \frac{1}{|W|} \sum\{\boldsymbol{o} | \boldsymbol{o} \in W\} = \frac{|U|}{|W|} \boldsymbol{c}_U + \frac{|V|}{|W|} \boldsymbol{c}_V.$$

This allows us to evaluate the increase in the sum of squared errors:

$$\begin{aligned} sse(\pi') - sse(\pi) &= \sum\{d^2(\boldsymbol{o}, \boldsymbol{c}_W) \mid \boldsymbol{o} \in U \cup V\} \\ &\quad - \sum\{d^2(\boldsymbol{o}, \boldsymbol{c}_U) \mid \boldsymbol{o} \in U\} - \sum\{d^2(\boldsymbol{o}, \boldsymbol{c}_V) \mid \boldsymbol{o} \in V\} \\ &= \sum\{d^2(\boldsymbol{o}, \boldsymbol{c}_W) - d^2(\boldsymbol{o}, \boldsymbol{c}_U) \mid \boldsymbol{o} \in U\} \\ &\quad + \sum\{d^2(\boldsymbol{o}, \boldsymbol{c}_W) - d^2(\boldsymbol{o}, \boldsymbol{c}_V) \mid \boldsymbol{o} \in V\}. \end{aligned}$$

Observe that:

$$\begin{aligned}
&\sum\{d^2(\boldsymbol{o},\boldsymbol{c}_W) - d^2(\boldsymbol{o},\boldsymbol{c}_U) \mid \boldsymbol{o} \in U\} \\
&= \sum_{\boldsymbol{o}\in U} \left((\boldsymbol{o}-\boldsymbol{c}_W)'(\boldsymbol{o}-\boldsymbol{c}_W) - (\boldsymbol{o}-\boldsymbol{c}_U)'(\boldsymbol{o}-\boldsymbol{c}_U)\right) \\
&= |U|(\boldsymbol{c}_W^2 - \boldsymbol{c}_U^2) + 2(\boldsymbol{c}_U - \boldsymbol{c}_W)' \sum_{\boldsymbol{o}\in U} \boldsymbol{o} \\
&= |U|(\boldsymbol{c}_W^2 - \boldsymbol{c}_U^2) + 2|U|(\boldsymbol{c}_U - \boldsymbol{c}_W)'\boldsymbol{c}_U \\
&= |U|(\boldsymbol{c}_W - \boldsymbol{c}_U)^2.
\end{aligned}$$

Using the equality $\boldsymbol{c}_W - \boldsymbol{c}_U = \frac{|U|}{|W|}\boldsymbol{c}_U + \frac{|V|}{|W|}\boldsymbol{c}_V - \boldsymbol{c}_U = \frac{|V|}{|W|}(\boldsymbol{c}_V - \boldsymbol{c}_U)$, we obtain $\sum\{d^2(\boldsymbol{o},\boldsymbol{c}_W) - d^2(\boldsymbol{o},\boldsymbol{c}_U) \mid \boldsymbol{o} \in U\} = \frac{|U||V|^2}{|W|^2}(\boldsymbol{c}_V - \boldsymbol{c}_U)^2$.
Similarly, we have

$$\sum\{d^2(\boldsymbol{o},\boldsymbol{c}_W) - d^2(\boldsymbol{o},\boldsymbol{c}_V) \mid \boldsymbol{o} \in V\} = \frac{|U|^2|V|}{|W|^2}(\boldsymbol{c}_V - \boldsymbol{c}_U)^2,$$

so,

$$sse(\pi') - sse(\pi) = \frac{|U||V|}{|W|}(\boldsymbol{c}_V - \boldsymbol{c}_U)^2. \qquad (1)$$

Merging Clusters

- In each phase of hierarchical clustering two of the "closest" clusters are merged.
- The notion of closest clusters is dependent on the specific dissimilarity between clusters considered in each variant of the clustering algorithm.

Variants of the AHC

If U and V are two clusters, the dissimilarity between them is be defined using one of the following real-valued, two-argument functions defined on the set of subsets of S:

$$\begin{aligned}
sl(U,V) &= \min\{d(u,v)|u \in U, v \in V\};\\
cl(U,V) &= \max\{d(u,v)|u \in U, v \in V\};\\
gav(U,V) &= \frac{\sum\{d(u,v)|u \in U, v \in V\}}{|U| \cdot |V|};\\
cen(U,V) &= (\boldsymbol{c}_U - \boldsymbol{c}_V)^2;\\
ward(U,V) &= \frac{|U||V|}{|U|+|V|}(\boldsymbol{c}_V - \boldsymbol{c}_U)^2.
\end{aligned}$$

- The names of the functions *sl*, *cl*, *gav*, and *cen* defined above are acronyms of the terms "single link", "complete link", "group average", and "centroid", respectively.
- In the case of the *ward* function the value equals the increase in the sum of the square errors when the clusters U, V are replaced with their union.
- The specific selection criterion for fusing blocks defines the clustering algorithm.
  At the leaf-level, when clusters are sigletons, all methods produce exactly the same result. At higher levels the results diverge.

- All algorithms store the dissimilarities between the current clusters $\pi^k = \{U_1^k, \ldots, U_{m_k}^k\}$ in an $m_k \times m_k$-matrix $D^k = (d_{ij}^k)$, where d_{ij}^k is the dissimilarity between the clusters U_i^k and U_j^k.
- As new clusters are created by merging two existing clusters, the distance matrix must be adjusted to reflect the dissimilarities between the new cluster and existing clusters.

General Agglometaive Clustering Algorithm

Input: the initial dissimilarity matrix D^1
Output: the cluster hierarchy on the set of objects S, where $|S| = n$
$k = 1$;
initialize clustering: $\pi^1 = \alpha_S$;
while (π^k contains more than one block){
    merge a pair of two of the closest clusters;
    output new cluster;
    $k++$;
    compute the dissimilarity matrix D^k;
}

Space and Time Complexity

- the algorithm must handle the matrix of the dissimilarities between objects, and this is a symmetric $n \times n$-matrix having all elements on its main diagonal equal to 0; in other words, the algorithm needs to store $\frac{n(n-1)}{2}$ numbers;
- to keep track of the clusters, an extra space that does not exceed $n-1$ is required. Thus, the total space required is $O(n^2)$.

UMASS BOSTON

Computation of the dissimilarity between a new cluster and existing clusters

Theorem

Let U and V be two clusters of the clustering π that are joined into a new cluster W. Then, if $Q \in \pi - \{U, V\}$, we have

$$
\begin{aligned}
sl(W,Q) &= \frac{1}{2}sl(U,Q) + \frac{1}{2}sl(V,Q) - \frac{1}{2}\Big|sl(U,Q) - sl(V,Q)\Big|; \\
cl(W,Q) &= \frac{1}{2}cl(U,Q) + \frac{1}{2}cl(V,Q) + \frac{1}{2}\Big|cl(U,Q) - cl(V,Q)\Big|; \\
gav(W,Q) &= \frac{|U|}{|U|+|V|}gav(U,Q) + \frac{|V|}{|U|+|V|}gav(V,Q); \\
cen(W,Q) &= \frac{|U|}{|U|+|V|}cen(U,Q) + \frac{|V|}{|U|+|V|}cen(V,Q) \\
&\quad - \frac{|U||V|}{(|U|+|V|)^2}cen(U,V); \\
ward(W,Q) &= \frac{|U|+|Q|}{\qquad\qquad}ward(U,Q) + \frac{|V|+|Q|}{\qquad\qquad}ward(V,Q)
\end{aligned}
$$

- The variant using *sl* is known as the *single-link* clustering. It tends to favor elongated clusters.
- The variant using *cl* is the *complete link* clustering.
- The *group average method*, which makes use of the *gav* function generates an intermediate approach between the single-link and the complete-link method.

The Monotonicity Property

Theorem

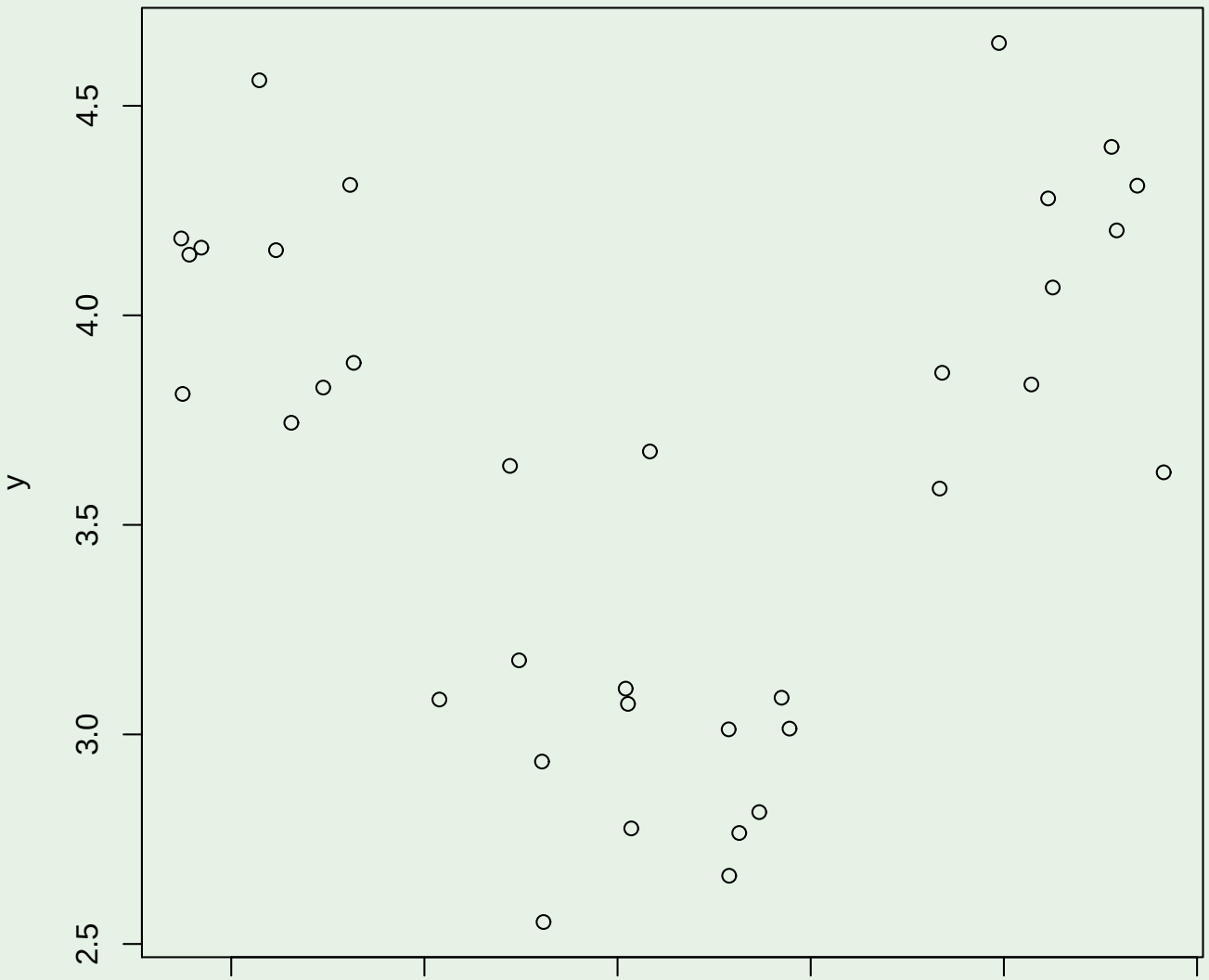
Let (S, d) be a finite metric space and let $D^1, \ldots, D^m$ be the sequence of matrices constructed by any of the first three hierarchical methods (single, complete, or average link), where $m = |S|$. If μ_i is the smallest entry of the matrix D^i for $1 \leqslant i \leqslant m$, then $\mu_1 \leqslant \mu_2 \leqslant \cdots \leqslant \mu_m$. In other words, the dissimilarity between clusters that are merged at each step is nondecreasing.

The centroid method and the Ward method of clustering. lack the monotonicity properties.

UMASS BOSTON

Example

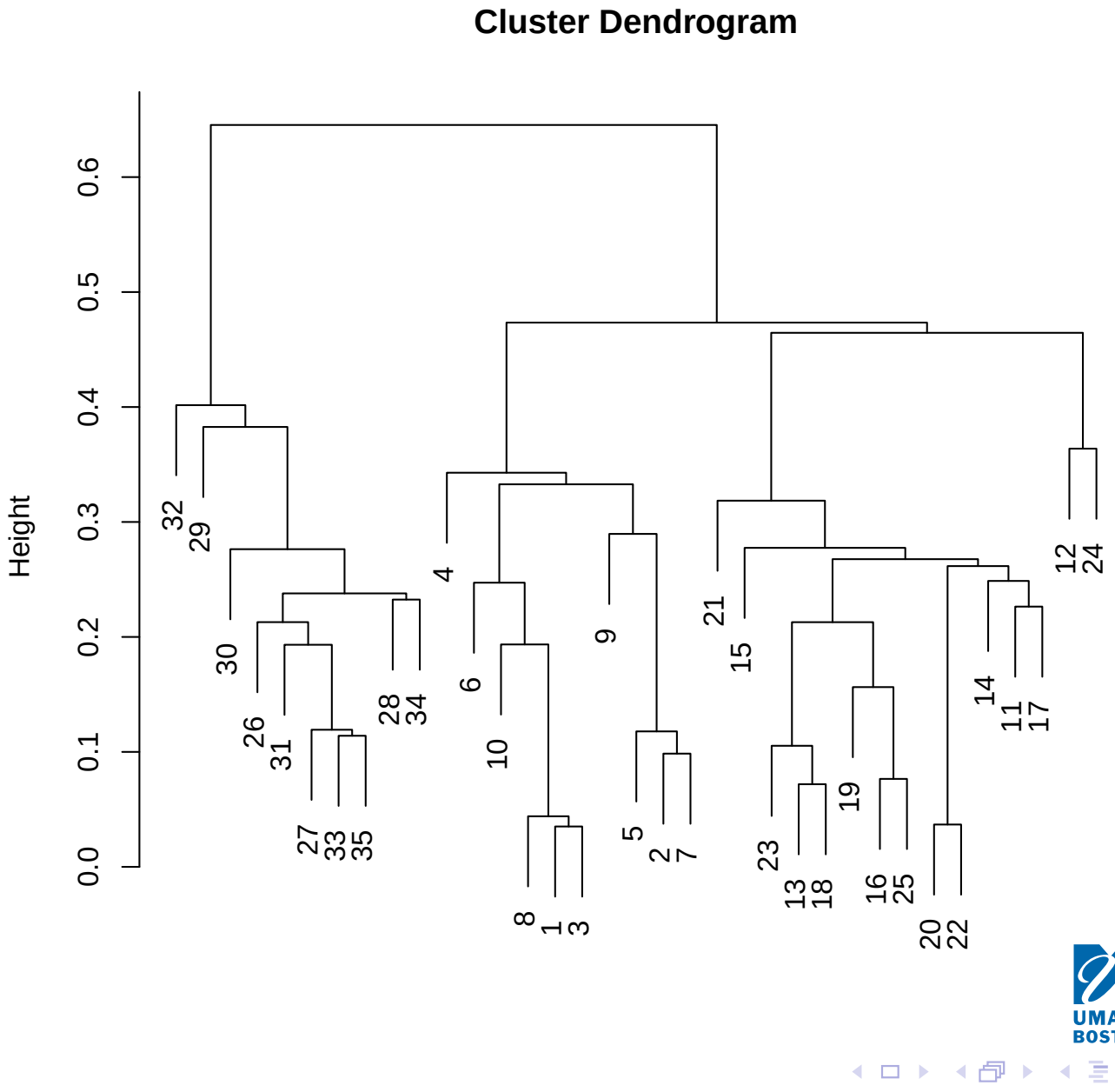
Let S be a synthetic data set that contains 35 points generated in R.



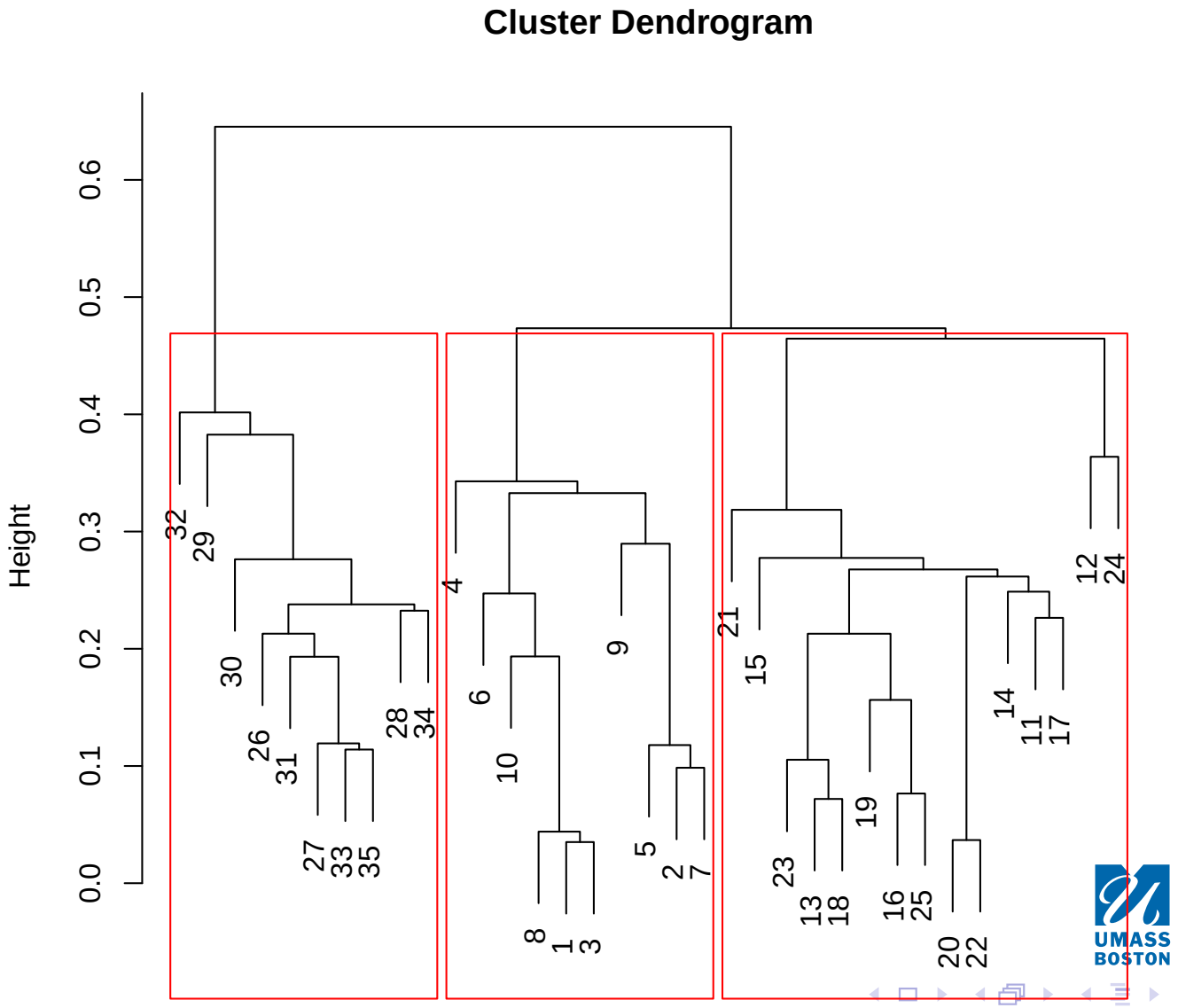
Starting from the matrix of S a `dist` object is produced by `d<-dist(S)`. Next, the function `hclust` is applied in order to produce the single-link hierarchical clustering `sLink`:

$$sLink \leftarrow hclust(d,method="single")$$

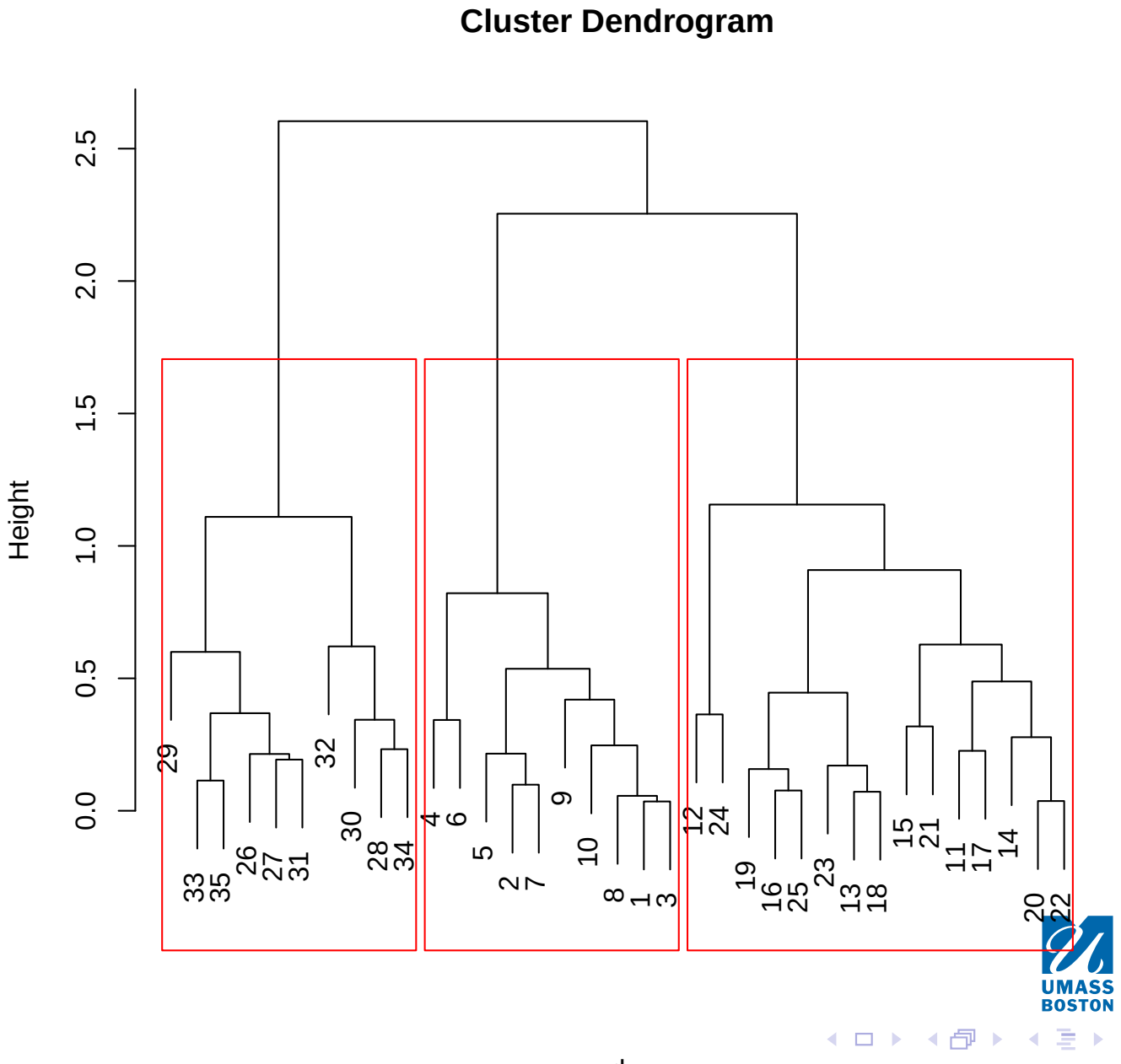
Dendrogram of Single-Link



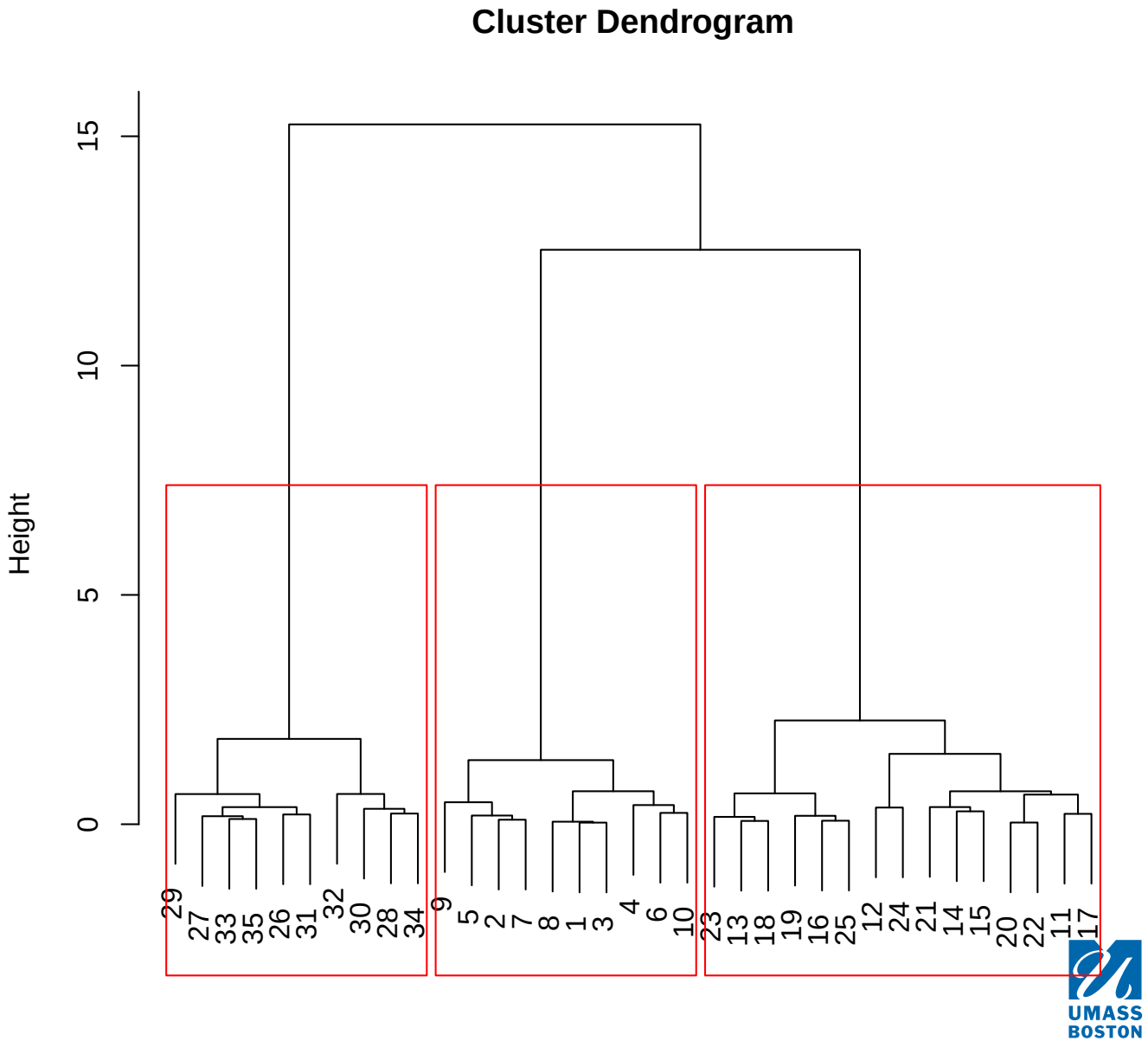
To obtain three clusters, the dendrogram is "cut" at an appropriate level using the function call `rect.hclust(sLink,3)` generating the representation shown below:



Clustering Obtained by Complete-link



Clustering Obtained by Ward Method



The Largest Ultrametric Dominated by a Dissimilarity

Let S be a set and let $\mathcal{D}_S$ the set of dissimilarities defined on S. The set of ultrametrics defined on S is $\mathcal{U}_S$.
A partial order ca be defined on $\mathcal{D}$ by writing $d \leqslant e$ if

$$d(x, y) \leqslant e(x, y) \text{ for every } x, y \in S.$$

Theorem

Let d be a dissimilarity on a set S and let U_d be the set of ultrametrics $U_d = \{e \in \mathcal{U}_S \mid e \leqslant d\}$. The set U_d has a largest element in the poset $(\mathcal{U}_S, \leqslant)$.

Proof

$U_d \neq \emptyset$ because d_0 given by $d_0(x, y) = 0$ for every $x, y \in S$ is an ultrametric and $d_0 \leqslant d$.

Since the set $\{e(x, y) \mid e \in U_d\}$ has $d(x, y)$ as an upper bound, it is possible to define the mapping $e_1 : S^2 \longrightarrow \mathbb{R}_{\geqslant 0}$ as

$$e_1(x, y) = \sup\{e(x, y) \mid e \in U_d\}$$

for $x, y \in S$. It is clear that $e \leqslant e_1$ for every ultrametric e. We claim that e_1 is an ultrametric on S.

We prove only that e_1 satisfies the ultrametric inequality. Suppose that there exist $x, y, z \in S$ such that e_1 violates the ultrametric inequality; that is,

$$\max\{e_1(x, z), e_1(z, y)\} < e_1(x, y).$$

This is equivalent to

$$\sup\{e(x, y) \mid e \in U_d\} > \max\{\sup\{e(x, z) \mid e \in U_d\}, \sup\{e(z, y) \mid e \in U_d\}\}$$

Thus, there exists $\hat{e} \in U_d$ such that

$$\hat{e}(x, y) > \sup\{e(x, z) \mid e \in U_d\}, \text{ and } \hat{e}(x, y) > \sup\{e(z, y) \mid e \in U_d\}.$$

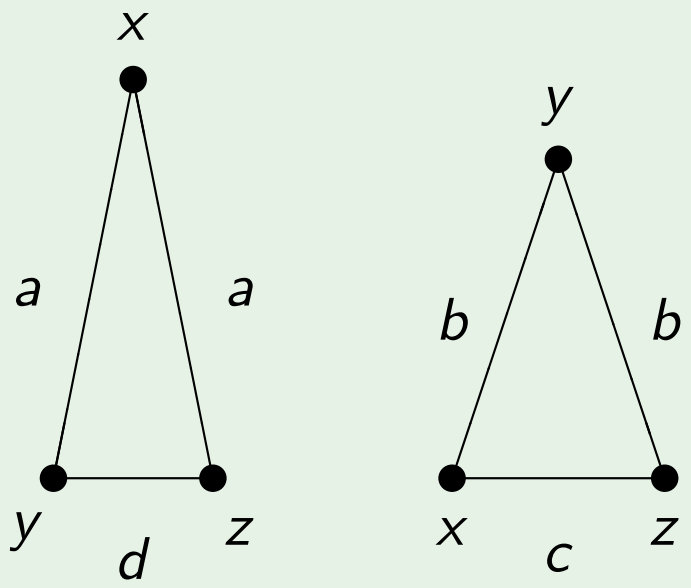
- The ultrametric e_1 defined above is known as the *maximal subdominant ultrametric for the dissimilarity d*.
- The situation is not symmetric with respect to the infimum of a set of ultrametrics because, in general, the infimum of a set of ultrametrics is not necessarily an ultrametric.

UMASS BOSTON

The Infimum of Ultrametrics May Fail to be an Ultrametric

Example

Let $S = \{x, y, z\}$, and let a, b, c, d be such that $a > b > c > d$ and the ultrametrics d and d' defined by the triangles below.



(a) (b)

$d_0 = \min\{d(u, v), d'(u, v)\}$ given by

$$d_0(x, y) = b, d_0(y, z) = d, \text{ and } d_0(x, z) = c,$$

is not an ultrametric because the triangle xyz is not isosceles.

An Algorithm

We give an algorithm for computing the maximal subdominant ultrametric for a dissimilarity defined on a finite set S.
We define inductively an increasing sequence of partitions $\pi_1 \prec \pi_2 \prec \cdots$ and a sequence of dissimilarities $d_1, d_2, \ldots$ on the sets of blocks of $\pi_1, \pi_2, \ldots$, respectively.

An Algorithm (cont'd)

- $\pi_1 = \alpha_S$ and $d_1(\{x\}, \{y\}) = d(x, y)$ for $x, y \in S$;
- Suppose that d_i is defined on π_i.
  If $B, C \in \pi_i$ is a pair of blocks such that $d_i(B, C)$ has the smallest value, define the partition π_{i+1} by

$$\pi_{i+1} = (\pi_i - \{B, C\}) \cup \{B \cup C\}.$$

  In other words, to obtain π_{i+1}, we replace two of the closest blocks B and C, of π_i (in terms of d_i) with new block $B \cup C$. Clearly, $\pi_i \prec \pi_{i+1}$ in $PART(S)$ for $i \geqslant 1$.
- The dissimilarity d_{i+1} is given by
  $d_{i+1}(U, V) = \min\{d(x, y) \mid x \in U, y \in V\}$ for $U, V \in \pi_{i+1}$.

The Hierarchy of Partition Blocks

The collection of blocks of the partitions π_i forms a hierarchy $\mathcal{H}_d$ on the set S.
We introduce a grading function h_d on the hierarchy defined by this chain of partitions starting from the dissimilarity d. The definition is done for the blocks of the partitions π_i by induction on i.
For $i = 1$ the blocks of the partition π_1 are singletons; in this case we define $h_d(\{x\}) = 0$ for $x \in S$.
Suppose that h_d is defined on the blocks of π_i, and let D be the block of π_{i+1} that is generated by fusing the blocks B and C of π_i. All other blocks of π_{i+1} coincide with the blocks of π_i. The value of the function h_d for the new block D is given by $h_d(D) = \min\{d(x, y) \mid x \in B, y \in C\}$.

For a set U of $\mathcal{H}_d$, define

$$\begin{aligned} p_U &= \min\{i \mid U \in \pi_i\} \\ q_U &= \max\{i \mid U \in \pi_i\} \end{aligned}$$

p_U is the first index i such that U is a block of π_i, and q_U is the last i such that U is a block of π_i.
If $H, K \in \mathcal{H}_\lceil$ and $H \subseteq K$, this means that both H and K are blocks of some partitions π_h and π_k and

$$p_H \leqslant q_H \leqslant p_K \leqslant q_K,$$

so $q_H \leqslant p_K$.

The construction of the sequence of partitions implies that there are $H_0, H_1 \in \pi_{p_H - 1}$ and $K_0, K_1 \in \pi_{p_K - 1}$ such that $H = H_0 \cup H_1$ and $K = K_0 \cup K_1$. Therefore,

$$\begin{aligned} h_d(H) &= \min\{d(x,y) \mid x \in H_0, y \in H_1\}, \\ h_d(K) &= \min\{d(x,y) \mid x \in K_0, y \in K_1\}. \end{aligned}$$

Since H_0 and H_1 were fused (to produce the block H of the partition π_{p_H}) before K_0 and K_1 were (to produce the block K of the partition π_{p_K}), it follows that $h_d(H) < h_d(K)$.
Let e be the ultrametric defined by the graded hierarchy $(\mathcal{H}_d, h_d)$.
e is the maximal subdominant ultrametric for d.

Since

$$e(x, y) = \min\{h_d(W) \mid \{x, y\} \subseteq W\}$$

and $h_d(W)$ is the least value of $d(u, v)$ such that $u \in U, v \in V$ if $W \in \pi_{p_W}$ is obtained by fusing the blocks U and V of $\pi_{p_W - 1}$, it follows that we have neither $\{x, y\} \subseteq U$ nor $\{x, y\} \subseteq V$. Thus, we have either $x \in U$ and $y \in V$ or $x \in V$ and $y \in U$, so $e(x, y) \leqslant d(x, y)$.

We now prove that:

$$e(x,y) = \min\{amp_d(\boldsymbol{s}) \mid \boldsymbol{s} \in S(x,y)\}$$

for $x, y \in S$.
Let D be the minimal set in $\mathcal{H}_d$ that includes $\{x, y\}$. Then, $D = B \cup C$, where B and C are two disjoint sets of $\mathcal{H}_d$ such that $x \in B$ and $y \in C$.
Case I: If $\boldsymbol{s}$ is a sequence included in D, then there are two consecutive components of $\boldsymbol{s}$, s_k and s_{k+1}, such that $s_k \in B$ and $s_{k+1} \in C$. This implies

$$\begin{aligned} e(x,y) &= \min\{d(u,v) \mid u \in B, v \in C\} \\ &\leqslant d(s_k, s_{k+1}) \\ &\leqslant amp_d(\boldsymbol{s}). \end{aligned}$$

Case II: If $\boldsymbol{s}$ is not included in D, let s_q and s_{q+1} be two consecutive components of $\boldsymbol{s}$ such that $s_q \in D$ and $s_{q+1} \notin D$.
Let E be the smallest set of $\mathcal{H}_d$ that includes $\{s_q, s_{q+1}\}$. We have $D \subseteq E$ (because $s_k \in D \cap E$) and therefore $h_d(D) \leqslant h_d(E)$. If E is obtained as the union of two disjoint sets E' and E'' of $\mathcal{H}_d$ such that $s_k \in E'$ and $s_{k+1} \in E''$, we have $D \subseteq E'$. Consequently,

$$h_d(E) = \min\{d(u, v) \mid u \in E', v \in E''\} \leqslant d(s_k, s_{k+1}),$$

which implies

$$e(x, y) = h_d(D) \leqslant h_d(E) \leqslant d(s_k, s_{k+1}) \leqslant amp_d(\boldsymbol{s}).$$

Therefore, we conclude that $e(x, y) \leqslant amp_d(\boldsymbol{s})$ for every $\boldsymbol{s} \in S(x, y)$.

There is a sequence $\boldsymbol{w} \in S(x, y)$ such that $e(x, y) \geqslant amp_d(\boldsymbol{w})$, which implies the equality $e(x, y) = amp_d(\boldsymbol{w})$. To this end, we prove that for every $D \in \pi_k \subseteq \mathcal{H}_d$ there exists $\boldsymbol{w} \in S(x, y)$ such that $amp_d(\boldsymbol{w}) \leqslant h_d(D)$.
The argument is by induction on k.
For $k = 1$, the statement obviously holds.

The Inductive Step of the Argument

Suppose that it holds for $1, \ldots, k-1$, and let $D \in \pi_k$. The set D belongs to π_{k-1} or D is obtained by fusing the blocks B, C of π_{k-1}. In the first case, the statement holds by inductive hypothesis. The second case has several subcases:

- If $\{x, y\} \subseteq B$, then by the inductive hypothesis, there exists a sequence $\boldsymbol{u} \in S(x, y)$ such that $amp_d(\boldsymbol{u}) \leqslant h_d(B) \leqslant h_d(D) = e(x, y)$.
- The case $\{x, y\} \subseteq C$ is similar to the first case.
- If $x \in B$ and $y \in C$, there exist $u, v \in D$ such that $d(u, v) = h_d(D)$. By the inductive hypothesis, there is a sequence $\boldsymbol{u} \in S(x, u)$ such that $amp_d(\boldsymbol{u}) \leqslant h_d(B)$ and there is a sequence $\boldsymbol{v} \in S(v, y)$ such that $amp_d(\boldsymbol{v}) \leqslant h_d(C)$. This allows us to consider the sequence $\boldsymbol{w}$ obtained by concatenating the sequences $\boldsymbol{u}, (u, v), \boldsymbol{v}$; clearly, $\boldsymbol{w} \in S(x, y)$ and $amp_d(\boldsymbol{w}) = \max\{amp_d(\boldsymbol{u}), d(u, v), amp_d(\boldsymbol{v})\} \leqslant h_d(D)$.

To complete the argument, we need to show that if e' is another ultrametric such that $e(x,y) \leqslant e'(x,y) \leqslant d(x,y)$, then $e(x,y) = e'(x,y)$ for every $x, y \in S$. By the previous argument, there exists a sequence $\boldsymbol{s} = (s_0, \ldots, s_n) \in S(x,y)$ such that $amp_d(\boldsymbol{s}) = e(x,y)$. Since $e'(x,y) \leqslant d(x,y)$ for every $x, y \in S$, it follows that $e'(x,y) \leqslant amp_d(\boldsymbol{s}) = e(x,y)$. Thus, $e(x,y) = e'(x,y)$ for every $x, y \in S$, which means that $e = e'$. This concludes our argument.

Open Questions

- How close is a metric to an ultrametric?
- Transformations that alter ultrametricity: how can we use these to improve data mining algorithms?

UMASS BOSTON

Evaluating Ultrametricity

Let $p \geqslant 1$ and $\mathcal{D}_p(S)$ be the set of dissimilarities defined on S that satisfy the inequality

$$d(x,y)^p \leqslant d(x,z)^p + d(z,y)^p$$

for $x, y, z \in X$.

- every dissimilarity belongs to $\mathcal{D}_0$;
- a dissimilarity in $\mathcal{D}_1$ is a semimetric.

p can be used to evaluate the ultrametricity of d.

A Global Measure of Ultrametricity

Define $\mathcal{D}_\infty = \bigcap_{p \geqslant 0} \mathcal{D}_p$. If $d \in \mathcal{D}_\infty$, then d is an ultrametric. Indeed, let $d \in \mathcal{D}_\infty$ and assume that $d(x,y) \geqslant d(x,z) \geqslant d(z,y)$. Then

$$d(x,y) \leqslant d(x,z)\left(1 + \left(\frac{d(y,z)}{d(x,z)}\right)^p\right)^{\frac{1}{p}}$$

for every $p \geqslant 0$. Since

$$\lim_{p \to \infty} d(x,z)\left(1 + \left(\frac{d(y,z)}{d(x,z)}\right)^p\right)^{\frac{1}{p}} = d(x,z),$$

it follows that $d(x,y) \leqslant d(x,z) = \max\{d(x,z), d(z,y)$ for $x, y, z \in S$, which allows us to conclude that d is an ultrametric.
Since $p \leqslant q$ implies

$$(d(x,z)^p + d(z,y)^p)^{\frac{1}{p}} \geqslant (d(x,z)^q + d(z,y)^q)^{\frac{1}{q}}$$

Thus, if $p \leqslant q$ we have the inequality $\mathcal{D}_q \subset \mathcal{D}_p$.

Thanks for your attention!

Slides can be found at
www.cs.umb.edu/∼dsim