CS724: Topics in Algorithms
Clustering Axiomatization

Prof. Dan A. Simovici

UMASS BOSTON

This is a recapitulation of the relationship between single-link clustering and minimal spanning trees (MSTs) which we covered in our discussion of hierarchical clustering.

Kruskal's Algorithm:

Data: A weighted graph $G = (V, E, c)$;

Result: A minimum spanning tree $T = (V, E', c')$ of G;

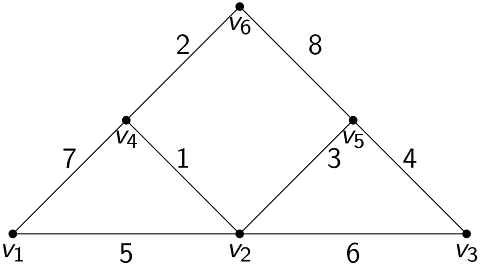
initialize the set of edges U as $U \leftarrow \emptyset$;

insert in U successive edges in the order of increasing weight *provided that the insertion does not create a cycle*; if it does, skip the edge;

stop when all nodes are connected

return: $T = (V, U, c \upharpoonright_U)$

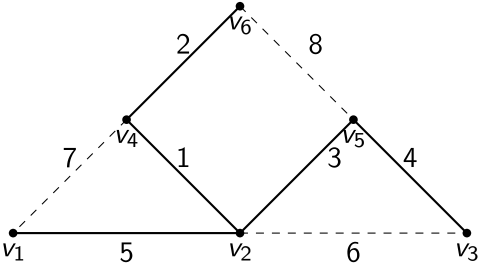
Let $G = (\{v_i \mid 1 \leqslant i \leqslant 6\}, E, c)$ a the weighted graph shown below.



The successive values of the set U are:

$$\begin{array}{l}\emptyset\\ \{\{v_2, v_4\}\}\\ \{\{v_2, v_4\}, \{v_4, v_6\}\}\\ \{\{v_2, v_4\}, \{v_4, v_6\}, \{v_2, v_5\}\}\\ \{\{v_2, v_4\}, \{v_4, v_6\}, \{v_2, v_5\}, \{v_5, v_3\}\}\\ \{\{v_2, v_4\}, \{v_4, v_6\}, \{v_2, v_5\}, \{v_5, v_3\}, \{v_2, v_1\}\}\end{array}$$

The weight of the minimum spanning tree shown is 15.



Notations:

- $\mathcal{DD}_S$ the set of definite dissimilarities on S;
- (S, d) is a dissimilarity space, where d is a dissimilarity on S; in general we assume that $d \in \mathcal{DD}_S$.

Single-link Clustering Algorithm

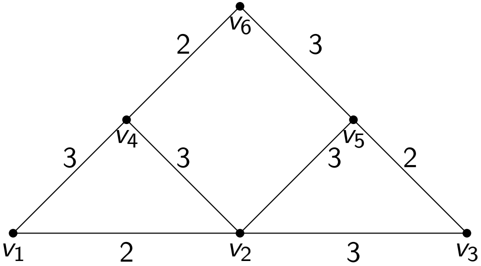
```
Data: A dissimilarity space $(S, d)$;
Result: A single-link clustering;
initialize $\pi \leftarrow \{\{x\} \mid x \in S\}$;
while {stopping condition is not met}{
  seek a pair of clusters $C, C' \in \pi$ such that
  $d(C, C') = \min\{d(x, y) \mid x \in C, y \in C'\}$ is minimal;
  fuse the clusters $C$ and $C'$ into the cluster $C \cup C'$, that is,
    $\pi \leftarrow \pi - \{C, C'\} \cup \{C \cup C'\}$;
}
return $\pi$
```

The most common stopping condition, which we adopt unless specified otherwise is that $\pi = \omega_S$, that is, only one cluster exists.

The single-link algorithm can be presented from the perspective of a minimum spanning tree of the weighted complete graph $\mathcal{G}_d$ whose vertex set is S and for which the weight of edge $\{i, j\}$ is $d(i, j)$.

- List edges in increasing order of their weight.
- Start with the partition of S that consists of singletons and from an MST T of the graph $\mathcal{G}_{S,d}$ labeled by these singletons.
- At each step the algorithm replaces edges in the tree by blocks obtained by fusing the extremities of the edges that have the lowest weight, until a single block partition is obtained.
- As before, the most common stopping condition, which we adopt unless specified otherwise is that $\pi = \omega_S$, that is, only one cluster exists.

Consider the graph

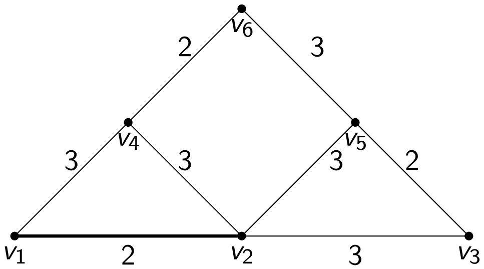


The list of edges in increasing order of the weight:

$$\underset{2}{\{v_1, v_2\}} \quad \underset{2}{\{v_3, v_5\}} \quad \underset{2}{\{v_4, v_6\}} \quad \underset{3}{\{v_1, v_4\}} \quad \underset{3}{\{v_2, v_3\}} \quad \underset{3}{\{v_2, v_4\}} \quad \underset{3}{\{v_2, v_5\}} \quad \underset{3}{\{v_5, v_6\}}$$

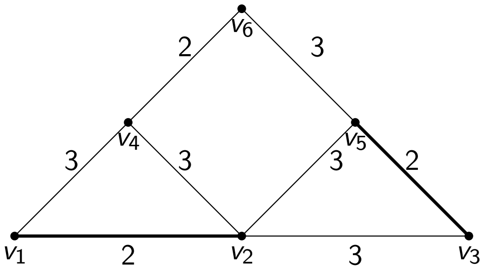
The construction of the single-link clustering proceeds along the by adding the edges whose endpoints are fused in the same cluster (indicated by bold lines).

Single-link Clustering - Recall



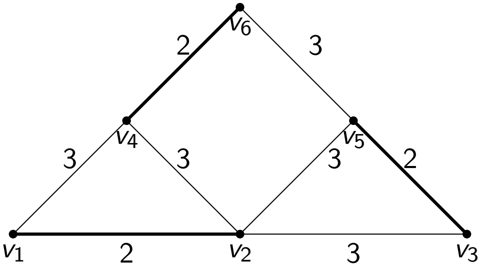
The list of edges in increasing order of the weight:

$\{v_1, v_2\}$	$\{v_3, v_5\}$	$\{v_4, v_6\}$	$\{v_1, v_4\}$	$\{v_2, v_3\}$	$\{v_2, v_4\}$	$\{v_2, v_5\}$	$\{v_5, v_6\}$
2	2	2	3	3	3	3	3



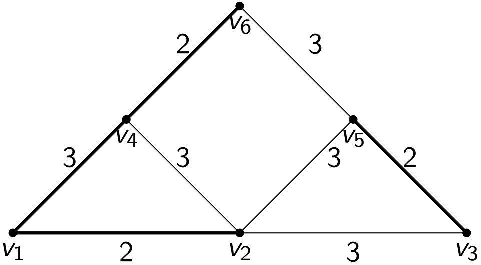
The list of edges in increasing order of the weight:

$\{v_1, v_2\}$	$\{v_3, v_5\}$	$\{v_4, v_6\}$	$\{v_1, v_4\}$	$\{v_2, v_3\}$	$\{v_2, v_4\}$	$\{v_2, v_5\}$	$\{v_5, v_6\}$
2	2	2	3	3	3	3	3



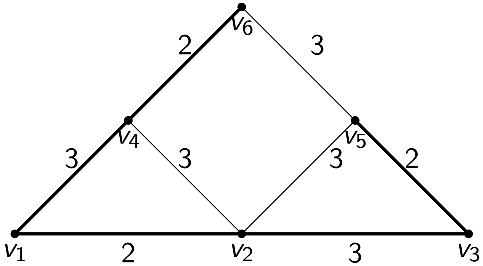
The list of edges in increasing order of the weight:

$$\underset{2}{\{v_1, v_2\}} \quad \underset{2}{\{v_3, v_5\}} \quad \underset{2}{\{v_4, v_6\}} \quad \underset{3}{\{v_1, v_4\}} \quad \underset{3}{\{v_2, v_3\}} \quad \underset{3}{\{v_2, v_4\}} \quad \underset{3}{\{v_2, v_5\}} \quad \underset{3}{\{v_5, v_6\}}$$



The list of edges in increasing order of the weight:

$$\underset{2}{\{v_1, v_2\}} \quad \underset{2}{\{v_3, v_5\}} \quad \underset{2}{\{v_4, v_6\}} \quad \underset{3}{\{v_1, v_4\}} \quad \underset{3}{\{v_2, v_3\}} \quad \underset{3}{\{v_2, v_4\}} \quad \underset{3}{\{v_2, v_5\}} \quad \underset{3}{\{v_5, v_6\}}$$



The list of edges in increasing order of the weight:

$$\underset{2}{\{v_1, v_2\}} \quad \underset{2}{\{v_3, v_5\}} \quad \underset{2}{\{v_4, v_6\}} \quad \underset{3}{\{v_1, v_4\}} \quad \underset{3}{\{v_2, v_3\}} \quad \underset{3}{\{v_2, v_4\}} \quad \underset{3}{\{v_2, v_5\}} \quad \underset{3}{\{v_5, v_6\}}$$

Definition

A *clustering function* on a set S is a mapping $f : \mathcal{DD}_S \longrightarrow PART(S)$, that is, as a function that maps a definite dissimilarity on S to a partition of S.

Definition

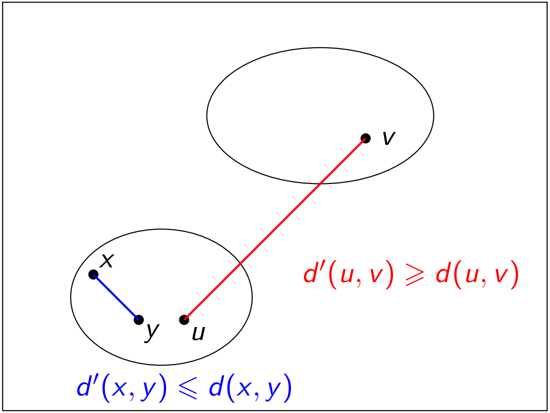
Let $\pi \in PART(S)$ be a partition of the set S. Define the relation $\leqslant_\pi$ on $\mathcal{DD}_S$ as $d \leqslant_\pi d'$ if

- $x \equiv y(\pi)$ implies $d'(x,y) \leqslant d(x,y)$, and
- $x \not\equiv y(\pi)$ implies $d'(x,y) \geqslant d(x,y)$

for $x, y \in S$.

If $d \leqslant_\pi d'$ we say that d' is a *π-transformation of d* and we write $d \leqslant_\pi d'$.

Dissimilarity d' is a π-tranformation of dissimilarity d.



Theorem

The relation $\leqslant_\pi$ is a partial order on $\mathcal{DD}_S$.

Proof: It is immediate that $\leqslant_\pi$ is reflexive.
If we have both $d \leqslant_\pi d'$ and $d' \leqslant_\pi d$, then

$$\begin{array}{lll} x \equiv y(\pi) & \text{implies} & d'(x,y) \leqslant d(x,y), \\ x \not\equiv y(\pi) & \text{implies} & d'(x,y) \geqslant d(x,y), \\ x \equiv y(\pi) & \text{implies} & d(x,y) \leqslant d'(x,y), \\ x \not\equiv y(\pi) & \text{implies} & d(x,y) \geqslant d'(x,y), \end{array}$$

hence $d(x,y) = d'(x,y)$ in all cases. This shows that $\leqslant_\pi$ is antisymmetric.

Proof cont'd

Finally, if $d \leqslant_\pi d'$ and $d' \leqslant_\pi d''$, then

$$\begin{array}{lcl} x \equiv y(\pi) & \text{implies} & d'(x,y) \leqslant d(x,y), \\ & \text{and} & d''(x,y) \leqslant d'(x,y), \\ x \not\equiv y(\pi) & \text{implies} & d'(x,y) \geqslant d(x,y), \\ & \text{and} & d''(x,y) \geqslant d'(x,y). \end{array}$$

Thus, $x \equiv y(\pi)$ implies $d''(x,y) \leqslant d(x,y)$ and $x \not\equiv y(\pi)$ implies $d''(x,y) \geqslant d(x,y)$, hence $\leqslant_\pi$ is transitive.

Theorem

The partial ordered set $(\mathcal{DD}_S, \leqslant_\pi)$ *is a lattice.*

Proof

Let $d_1, d_2 \in \mathcal{DD}_S$ such that $d_1 \leqslant_\pi d'$ and $d_2 \leqslant_\pi d'$. We have:

$$\begin{array}{ccc} x \equiv y(\pi) & \text{implies} & d'(x,y) \leqslant d_1(x,y), \\ & \text{and} & d'(x,y) \leqslant d_2(x,y), \\ x \not\equiv y(\pi) & \text{implies} & d'(x,y) \geqslant d_1(x,y), \\ & \text{and} & d'(x,y) \geqslant d_2(x,y). \end{array}$$

This means that $x \equiv y(\pi)$ implies $d'(x,y) \leqslant \min\{d_1(x,y), d_2(x,y)\}$ and $x \not\equiv t(\pi)$ implies $d'(x,y) \geqslant \max\{d_1(x,y), d_2(x,y)\}$. Thus, by defining $d \in \mathcal{DD}_S$ as

$$d(x,y) = \begin{cases} \min\{d_1(x,y), d_2(x,y)\} & \text{if } x \equiv y(\pi), \\ \max\{d_1(x,y), d_2(x,y)\} & \text{if } x \not\equiv y(\pi), \end{cases}$$

we have $d \leqslant_\pi d'$, which shows that d is the infimum of d_1 and d_2 in the partial ordered set $(\mathcal{DD}_S, \leqslant_\pi)$.

Proof cont'd

Similarly, $\tilde{d} \in \mathcal{DD}_S$ defined as

$$\tilde{d}(x,y) = \begin{cases} \max\{d_1(x,y), d_2(x,y)\} & \text{if } x \equiv y(\pi), \\ \min\{d_1(x,y), d_2(x,y)\} & \text{if } x \not\equiv y(\pi), \end{cases}$$

for $x, y \in S$ is the supremum of $\{d_1, d_2\}$

Let $\pi \in PART(S)$ and let a, b be two non-negative numbers such that $a \leqslant b$. Define the mapping $\delta^{\pi}_{a,b} : S \times S \longrightarrow \mathbb{R}_{\geqslant 0}$ as

$$\delta^{\pi}_{a,b}(x,y) = \begin{cases} 0 & \text{if } x = y, \\ a & \text{if } x \equiv y(\pi) \text{ and } x \neq y, \\ b & \text{if } x \not\equiv y(\pi). \end{cases}$$

It is easy to verify that $\delta^{\pi}_{a,b}$ is an ultrametric on S.

Definition

Let a, b be two non-negative numbers and let $\pi \in PART(S)$. A dissimilarity $d \in \mathcal{DD}_S$ is said to *(a, b)-conform* to π if $d \leqslant_\pi \delta^\pi_{a,b}$.

In other words, a dissimilarity $d \in \mathcal{DD}_S$ is said to *(a, b)-conform* to π if

- if $x \equiv_\pi y$ then $d(x, y) \leqslant a$, and
- if $x \not\equiv_\pi y$ then $d(x, y) \geqslant b$.

for all $x, y \in S$.

Observe that d is (a, b)-conform to π if

$$\begin{aligned} M(\pi) = \max\{d(x, y) \mid x \equiv y(\pi)\} &\leqslant a, \text{ and} \\ m(\pi) = \min\{d(x, y) \mid x \not\equiv y(\pi)\} &\geqslant b. \end{aligned}$$

Note that if d is (a, b)-conform to π and $e \leqslant_\pi d$, then e is also (a, b)-conform to π.

Definition

A pair of positive real numbers (a, b) is *π-forcing relative to a clustering function f* if for all $d \in \mathcal{DD}_S$ that are (a, b)-conform to π we have $f(d) = \pi$.

Equivalently, (a, b) is a π-forcing pair relative to f if

$$d \leqslant_\pi \delta^\pi_{a,b} \text{ implies } f(d) = \pi.$$

SYNOPSIS

- d is *(a, b)-conforms* to π if $d \leqslant_\pi \delta^\pi_{a,b}$.
- (a,b) is *π-forcing* relative to f if when $d \leqslant_\pi \delta^\pi_{a,b}$ (that is, d (a,b)-conforms to π) then $f(d) = \pi$.
- f is *consistent* if $d \leqslant_{f(d)} d'$ implies $f(d') = f(d)$.

Kleinberg considers three desirable and natural properties of clustering functions: scale-invariance, richness, and consistency.
Namely, a clustering function f is:

- *scale-invariant*, if for any dissimilarity function d we have $f(ad) = f(d)$ if $a > 0$;
- *rich*, if it is surjective, that is, for any partition $\pi \in PART(S)$ there exists $d \in \mathcal{DD}_S$ such that $f(d) = \pi$;
- *consistent*, if $d \leqslant_{f(d)} d'$ then $f(d) = f(d')$.

Variants of single-link clustering

Besides the common halting condition for the single-link algorithm ($\pi = \omega_S$) there are several alternatives:

- k-`cluster stopping condition:` Stop adding edges when the partition first consists of k blocks. (This condition is well-defined when the number of points is at least k.)
- `dissimilarity-`r `stopping condition:` Fuse clusters C, C' only if $d(C, C') \leqslant r$;
- `scale-`α `stopping condition:` Let $\alpha \in (0, 1)$ and let d^* denote the maximum pairwise dissimilarity; i.e. $d^* = \max\{d(x, y) \mid x, y \in V\}$. Then, fuse clusters C, C' only if $d(C, C') \leqslant \alpha d^*$.

- By choosing a stopping condition for the single-link procedure, one obtains a clustering function, which maps the dissimilarity function to the set of connected components that results at the end of the procedure.
- For any two of the three properties considered above one can choose a single-link stopping condition so that the resulting clustering function satisfies exactly these two properties.

Theorem

- *For any $k \geqslant 1$, and $n \geqslant k$, single-link with the k-cluster stopping condition satisfies scale-invariance and consistency but fails richness.*
- *For any $r > 0$, and any $n \geqslant 2$, single-link with the dissimilarity-r stopping condition satisfies richness and consistency but fails scale-invariance.*
- *For any positive $\alpha < 1$, and any $n \geqslant 3$, single-link with the scale α-stopping condition satisfies scale-invariance and richness but fails consistency.*

Proof

Single-link with the k-cluster stopping condition satisfies scale-invariance and consistency but fails richness.

This function fails the richness condition because not every partition has k-clusters.

It is immediate that f is scale invariant.

To prove that f it is consistent suppose that $f(d) = \pi$ and that $d \leqslant_{\pi} d'$. If x, y belong to the same cluster of π, that is, if $x \equiv y(\pi)$, then $d'(x, y) \leqslant d(x, y)$, which means that $x \equiv y(\pi')$ because the unordered pair $\{x, y\}$ is added to the MST that corresponds to d' before the same edge is added to the MST that corresponds to d.

Proof cont'd

For any $r > 0$, and any $n \geqslant 2$, single-link with the dissimilarity-r stopping condition satisfies richness and consistency but fails scale-invariance.
Scale invariance is not satisfied because by multiplying the dissimilarity by an appropriate constant we obtain the clustering that consists only of singletons. The stopping condition means that $x \equiv y(f(d))$ implies $d(x, y) \leqslant r$.
Richness follows from the fact that the constant r and the dissimilarity d can be chosen such that $f(\pi)$ equals any partition on the set of objects.

Proof cont'd

Let d, d' be dissimilarities such that $d \leqslant_{f(d)} d'$. We need to prove that $f(d') = f(d)$, or equivalently, that $x \equiv y(f(d))$ if and only if $x \equiv y(f(d'))$. Since both partitions $f(d)$ and $f(d')$ are obtained by the application of the single-link with the dissimilarity-r stopping condition it follows that

$$x \equiv y(f(d)) \text{ implies } d(x,y) \leqslant r \text{ and } x \equiv y(f(d')) \text{ implies } d'(x,y) \leqslant r.$$

If $x \equiv y(f(d))$ we have $d'(x,y) \leqslant d(x,y) \leqslant r$ so $x \equiv y(f(d'))$. Suppose now that $x \equiv y(f(d'))$ but $x \not\equiv y(f(d))$. Since $d \leqslant_{f(d)} d'$, we have $d'(x,y) \geqslant d(x,y)$ and $r \geqslant d'(x,y)$. Thus, $r > d(x,y)$, which contradicts the fact that $x \not\equiv y(f(d))$. Therefore, $x \equiv y(f(d'))$ implies $x \equiv y(f(d))$, hence $f(d) = f(d')$.

Proof cont'd

For any $0 < \alpha < 1$, and any $n \geqslant 3$, single-link with the scale α-stopping condition satisfies scale-invariance and richness but fails consistency.
Recall that clusters are fused when $d(C, C') \leqslant \alpha \max\{d(x,y) \mid x, y \in V\}$.
Scale-invariance is immediate since both the values of the dissimilarities and the values of the threshold are multiplied at the same rate. Richness is also immediate.
However, consistency fails.

Let $V = \{x_1, x_2, x_3\}$ and let d be defined by

$$d(x_1, x_2) = a, d(x_2, x_3) = b, d(x_1, x_3) = c,$$

where $a < b < c$. Thus, the maximum dissimilarity is c.
Choose α such that $a < \alpha c < b$, or $\frac{a}{c} < \alpha < \frac{b}{c}$. The resulting partition is $\pi = f(d) = \{\{x_1, x_2\}, \{x_3\}\}$.

If d' is such that $d \leqslant_{\pi} d'$ then

$$\begin{aligned} d'(x_1, x_2) &\leqslant d(x_1, x_2) = a, \\ d'(x_2, x_3) &\geqslant d(x_2, x_3) = b, \\ d'(x_1, x_3) &\geqslant d(x_1, x_3) = c. \end{aligned}$$

There conditions are satisfied by d' defined as

$$d'(x_1, x_2) = a, d'(x_2, x_3) = b, d'(x_1, x_3) = kc.$$

Choose k such that $b < \alpha k c$. We have $f(d') = \{\{x_1, x_2, x_3\}\}$. Since $d \leqslant_{\pi} d'$ but $f(d') \neq f(d)$, consistency fails.

Lemma

Let f be a consistent clustering function on a dissimilarity space (S, d). For any $\pi \in Ran(f)$ there exist positive numbers a, b such that the pair (a, b) is π-forcing relative to f.

Proof

Since $\pi \in \text{Ran}(f)$ there exists d such that $f(d) = \pi$. Let

$$\begin{aligned} a' &= \min\{d(x,y) \mid x \equiv y(\pi)\}, \\ b' &= \max\{d(x,y) \mid x \not\equiv y(\pi)\}, \end{aligned}$$

and let a, b be two numbers such that $a \leqslant a' \leqslant b' \leqslant b$. Since d' (a,b)-conforms to $\pi = f(d)$, we have $f(d') = \pi$ by the consistency property. It follows that the pair (a,b) is π-forcing relative to f.

Theorem

If a clustering function $f : \mathcal{DD}_S \longrightarrow PART(S)$ *is scale-invariant and consistent, then its range is an antichain in the partially ordered set of partitions of* S*.*

Proof

Suppose that f is scale-invariant and that exist distinct partitions $\pi_0, \pi_1 \in \mathrm{Ran}(f)$ such that π_0 is a refinement of π_1, that is, $\pi_0 < \pi_1$. Let (a_0, b_0) be a π_0 forcing pair and let (a_1, b_1) be a π_1 forcing pair relative to f, where $a_0 < b_0$ and $a_1 < b_1$.
Let a_2 be such that $a_2 \leqslant a_1$, and let ϵ such that $0 < \epsilon < \frac{a_0 a_2}{b_0}$.
Since $\pi_0 < \pi_1$ define a dissimilarity $d \in \mathcal{D}\mathcal{D}_S$ such that:

- if $x \equiv y(\pi_0)$, then $d(x, y) \leqslant \epsilon$;
- if $x \equiv y(\pi_1)$ but $x \not\equiv y(\pi_0)$, then $a_2 \leqslant d(x, y) \leqslant a_1$;
- if $x \not\equiv y(\pi_1)$ then $d(x, y) \geqslant b_1$.

Proof cont'd

The dissimilarity d (a_1, b_1)-conforms to π_1 and so $f(d) = \pi_1$.
Set $\alpha = \frac{b_0}{a_2}$ and define $d' = \alpha d$. By scale invariance we have $f(d') = f(d) = \pi_1$.
For $x \equiv y(\pi_0)$ we have $d'(x, y) \leqslant \frac{\epsilon b_0}{a_2} < a_0$, while for $x \not\equiv y(\pi_0)$ we have

$$d'(x, y) \geqslant a_2 b_0 a_2^{-1} = b_0.$$

Thus, d' (a_0, b_0) conforms to π_0 and so we have $f(d') = \pi_0$. Since $\pi_0 \neq \pi_1$. this is a contradiction.

Theorem

For every antichain of partitions $\mathcal{A}$, there is a clustering function that is scale-invariant and consistent such that $Ran(f) = \mathcal{A}$.

Proof

Let $\mathcal{A}$ be an antichain of partitions of the set S. An *$\mathcal{A}$-sum-of-pairs clustering function* f is defined as $f(d) = \pi$, where π is the partition that minimizes the sum

$$\Phi_d(\pi) = \sum \{d(x,y) \mid x \equiv y(\pi)\}$$

over partitions π in $\mathcal{A}$.
Since $\Phi_{\alpha d}(\pi) = \alpha \Phi_d(\pi)$ it is clear that f is scale-invariant.

For $\pi \in \mathcal{A}$ let d be the dissimilarity on the set S with $|S| = n$ having the following properties:

- $d(x, y) < \frac{1}{n^3}$ for $x \equiv y(\pi)$;
- $d(x, y) \geqslant 1$ for $x \not\equiv y(\pi)$.

We have $\Phi_d(\pi) < 1$; moreover, $\Phi_d(\pi') < 1$ only for partitions π' such that $\pi' \leqslant \pi$. Since $\mathcal{A}$ is an antichain, π minimizes Φ_d over all partitions in $\mathcal{A}$, hence $f(d) = \pi$.

To prove consistency suppose that $f(d) = \pi$ and let d' be such that $d \leqslant_\pi d'$. For any partition π' let $\Delta(\pi') = \Phi_d(\pi') - \Phi_{d'}(\pi')$. It suffices to show that for any $\pi' \in \mathcal{A}$ we have $\Delta(\pi) \geqslant \Delta(\pi')$.

Note that

$$\begin{aligned}
\Delta(\pi) &= \sum \{d(x,y) - d'(x,y) \mid x \equiv y(\pi)\}, \\
\Delta(\pi') &= \sum \{d(x,y) - d'(x,y) \mid x \equiv y(\pi')\} \\
&\leqslant \sum \{d(x,y) - d'(x,y) \mid x \equiv y(\pi \wedge \pi')\} \\
&\leqslant \Delta(\pi),
\end{aligned}$$

where both inequalities follow from $d \leqslant_\pi d'$ (for the first, only terms that correspond to pairs in the same cluster of π are non-negative; for the second, every term corresponding to a pair in the same cluster of π is non-negative). This concludes the argument.

Kleinberg's Main Result

Corollary

For each $n \geqslant 2$, there is no clustering function that satisfies scale-invariance, richness and consistency.

Proof: Suppose that $f : \mathcal{DD}_S \longrightarrow PART(S)$ is a clustering function that satisfies scale-invariance and consistency. By a previous theorem, the range of f is an antichain in $(PART(S), \leqslant)$, so f cannot be a surjective. Therefore, f fails the richness property, which contradicts the initial assumption.

In centroid-based clustering k input points are selected as tentative centroids followed by the definition of clusters by assigning each point in S to its nearest centroid.
The aim is to choose centroids such that each point in S is close to at least one of them.

Example

A choice is to select centroids such that the sum of dissimilaritiess to its assigned points is minimal (Fermat points or k-median).
An alternative, used in the case of k-means is to seek centroids such that the sum of the squares of dissimilarities to its assigned points is minimal.

Kleinberg proved that for a general class of centroid-based clustering functions, including k-means and k-median, none of the functions in the class satisfies the consistency property. This contrasts with with the results for single-link and sum-of-pairs.

For $k \in \mathbb{N}$, $k \geqslant 2$ and any continuous, non-decreasing, and unbounded function $g : \mathbb{R}_{\geqslant 0} \longrightarrow \mathbb{R}_{\geqslant 0}$, define the (k, g)-centroid clustering function as follows.

Choose the subset T of S consisting of k centroid for which the objective function $\lambda_d^g(T) = \sum_{x \in S} g(d(x, T))$ is minimized. (Here $d(x, T) = \min_{c \in T} d(x, c)$). Then, define a partition of S into k clusters by assigning each point to the element of T closest to it.

- the k-median function is obtained by setting g to be the identity function;
- the objective function underlying k-means clustering is obtained by setting $g(d) = d^2$.

Theorem

For every $k \geqslant 2$ and every function g chosen as above, and for n sufficiently large relative to k, the (k, g)-centroid clustering function fails the consistency property.

Proof

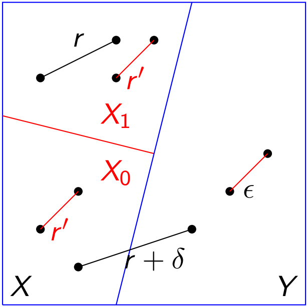
Suppose that $k = 2$; the argument for $k \geqslant 2$ is similar. Let $\pi_\gamma = \{X, Y\} \in PART(S)$, where $|X| = m$ and $|Y| = \gamma m$ for $\gamma > 0$. Assume that the dissimilarity between points in X is r, the dissimilarities between points in Y are equal to ϵ, where $\epsilon < r$, and the dissimilarity between x in X and y in Y is $r + \delta$, for some $\delta > 0$.

Proof cont'd

By choosing γ, r, ϵ and δ appropriately, the optimal choice of 2 centroids will consist of one point from X and one from Y, and the resulting partition π will have clusters X and Y.

Suppose we partition X into sets X_0 and X_1 of equal size, and reduce the dissimilarities between points in the same X_i to be $r' < r$ (keeping all other dissimilarities the same). This yields the dissimilarity d'.

Proof cont'd



This can be done, for r' small enough, so that the optimal choice of two centroids will now consist of one point from each X_i, yielding a different partition of S.

As our second dissimilarity is a π-transformation of the first, this violates consistency.

The notion of *partitioning function*, a modification of the notion of clustering function is considered.

Definition

A *partitioning function* on a definite dissimilarity space is a function $f : \mathcal{DD}_S \times \{1, \ldots, |S|\} \longrightarrow PART(S)$ such that $f(d, k)$ is a partition of S having k blocks.

One could consider properties of partitioning functions similar to the ones previously introduced by Kleinberg for clustering functions.
Namely, a partitioning function f is:

- *scale-invariant*, if for any dissimilarity $d \in \mathcal{DD}_S$ and number of clusters k (such that $1 \leqslant k \leqslant |S|$) we have $f(ad, k) = f(d, k)$ if $a > 0$;
- *rich*, if for any number of clusters k such that $1 \leqslant k \leqslant |S|$, $\text{Ran}(f(\cdot, k))$ equals the set of all partitions that have k blocks;
- *order-consistent*, if for any d, d' and k the order of edges of G is identical for d and d', then $f(d, k) = f(d', k)$;

Order-consistency means that the only way that the partition function uses edge weights is by comparing them against each other. Note that order-consistency implies scale invariance.

Definition

A partitioning function $f : \mathcal{DD}_S \times \{1, \ldots, |S|\} \longrightarrow PART(S)$ is *consistent* if $f(d,k) = \pi$ and $d \leqslant_\pi d'$ implies $f(d',k) = \pi$.

The main result discussed here is that the four properties enumerated above: scale invariance, k-richness, order-consistency, and consistency are satisfiable. To present this result we shall revisit the single-link clustering. The single-link algorithm on a dissimilarity space (S,d) can be discussed in the context of a complete weighted graph $G = (S, E, d)$, where the weight of an edge $\{x,y\}$ is $d(x,y)$. If $S = \{x_1, \ldots, x_n\}$, the dissimilarity d is specified by a list L_d of numbers in non-decreasing order

$$L_d = (d_1, d_2, \ldots, d_{\binom{n}{2}}),$$

of the weights of the edges of G.

An edge $\{x, y\}$ is *redundant* if x and y are connected via a path whose edges have smaller weight than $d(\{x, y\})$. The following algorithm constructs the single-link clustering κ, where C_x is the cluster that contain x.

Data: A dissimilarity space (S, d), given by the list L_d and a number k, where $1 \leqslant k \leqslant |S|$.
Result: A single-link clustering that consists of no more than k clusters.

```
$\pi \leftarrow \{\{x_i\} \mid 1 \leqslant i \leqslant |S|\}$;
$i \leftarrow 1$;
while $\{|\pi| > k\}\{$
    let $e_i = \{x, y\}$;
    let $C_x \in \pi$ such that $x \in C_x$;
    let $C_y \in \pi$ such that $y \in C_x$;
    if$\{C_x \neq C_y\}\{$
        merge $C_x$ and $C_y$;
        $\pi \leftarrow \pi - \{C_x, C_y\} \cup \{C_x \cup C_y\}$;
    $\}$
    $i \leftarrow i + 1$;
$\}$ return $\pi$
```

Theorem

The partitioning function computed by the previous single-link algorithm is scale invariant, k-rich, order-consistent, and consistent.

Proof

Single-link is order-consistent because if its decisions are based on comparing two edges to determine which dissimiarities are smaller or larger. Scale-invariance follows from order-consistency.
To obtain a k-partition π it suffices to set intra-block dissimilarities to 1 and the inter-block dissimilarities to 2 to have the algorithm return π.

To show the consistency of the algorithm, let $f(d,k) = \pi$. An edge $e = \{x, y\}$ is an *inner* edge if $x \equiv y(\pi)$ and an *outer* edge if $x \not\equiv y(\pi)$. To construct π the algorithm sorts all edges of the graph and then examines every edge. While there are more than k clusters, the algorithm transforms the smallest outer edge into an inner edge (thereby reducing the number of clusters by 1). An inner edge that is larger than any outer edge is refered to as a *redundant* inner edge. Such an edge is not considered for merging; however, it becomes an inner edge by transitivity.

If the edges of the graph are listed as $\mathbf{e} = (e_1, e_2, \ldots, e_{\binom{n}{2}})$ in ascending order of the corresponding dissimilarities, each of these edges may be an outer edge, a non-redundant inner edge, or a redundant inner edge. By the definition of the algorithm there is a prefix $\mathbf{p}$ of $\mathbf{e}$ which consists of inner edges and suffices to define π. If $k = n$, $\mathbf{p}$ will be empty as there are no inner edges.

Consider now the π-transformations of d. If we shrink a non-redundant inner edge of d, then $\mathbf{p}$ will not change and the algorithm will still produce π. If we shrink a redundant inner edge, $\mathbf{p}$ may change to $\mathbf{p}'$, but the clustering produced will not change as a result of transitivity. Finally, if we exapand an outer edge, again $\mathbf{p}$ will not change leaving π intact. Thus, for all posssible π-transformations d' of d we will obtain the same clustering.

We present now an axiomatization of measures of clustering quality developed by M. Ackerman and S. Ben-David. This is an alternative approach in the attempt to axiomatize clustering and leads to a consistent system of axioms.

Definition

A *clustering quality measure* is a function $m(S, d, \pi)$ ranging over $\mathbb{R}_{\geqslant 0}$, where (S, d) form a dissimilarity spaces and $\pi \in PART(S)$.

The quality measure m is

- *scale invariant* if for every $\lambda > 0$ we have $m(S, \lambda d, \pi) = m(S, d, \pi)$;
- *consistent* if $d \leqslant^{\pi}_{a,b} d'$ implies $m(S, d', \pi) \leqslant m(S, d, \pi)$;
- *rich* if for every $\pi_0 \in PART(S)$ with $\pi_0 \notin \{\alpha_S, \omega_S\}$ there exists a dissimilarity d such that $\pi_0 = \arg\max_{\pi} m(S, d, \pi)$.

For center-based clustering it is possible to formulate a quality measure that satisfies all requirements of the previous definition. We assume that the dissimilarity distance is a metric and thus, it is possible to define cluster centers (either as medians or as means). This makes centers invariant to scaling.

Definition

Let (S, d) be a dissimilarity space and let $\pi = \{C_1, \ldots, C_k\} \in PART(S)$ be a clustering.
A subset K is a *representative set* for π if $K \cap C_i$ contains a unique element c_i for each block C_i of π and K is invariant under scaling. It is clear that $|K| = k$. Denote by $\text{REP}(\pi)$ the set of possible representative sets for π.

Define the *point margin* of $x \in S$ relative to K as

$$\mathrm{pom}_{\pi,d}(x) = \frac{d(x, c_x)}{d(x, e_x)},$$

where $c_x \in K$ is the closest representative to x, and e_x is the second closest representative to x.

The smaller the value of the point margin, the better the clustering is.

The *relative margin of a clustering* π is the number $\mathrm{relm}(\pi)$ defined as

$$\mathrm{relm}(S, d, \pi) = \min_{K \in \mathrm{REP}(\pi)} \mathrm{avg}_{x \in S-K} \mathrm{pom}_{\pi,d}(x).$$

Theorem

The relative margin relm is scale-invariant, consistent and rich.

Proof

The scale-invariance of $\text{relm}(S, d, \pi)$ follows from the fact that K is invariant under scaling.

Let d' be a π-transformation of d, that is, $d \leqslant^{\pi}_{a,b} d'$. Since x and c_x belong to the same cluster of π and x, e_x belong to two distinct clusters, we have $d'(x, c_x) \leqslant d(x, c_x)$ and $d(x, e_x) \leqslant d'(x, e_x)$, which implies

$$\text{pom}_{\pi,d'}(x) = \frac{d'(x, c_x)}{d'(x, e_x)} \leqslant \frac{d(x, c_x)}{d(x, e_x)} = \text{pom}_{\pi,d}(x).$$

This implies $\text{relm}(S, d', \pi) \leqslant \text{relm}(S, d, \pi)$, so relm is consistent.

Starting with a non-trivial clustering π on S consider the ultrametric $\delta^{\pi}_{a,b}$ where $a < b$. Then $\pi = \text{relm}(S, \delta^{\pi}_{a,b}, \pi)$. Thus, relm is rich.

Previous theorem shows that the system of axioms introduced is consistent (which means that the set of objects that satifies this system is non-void).

Definition

Let (S, d) be a dissimilarity space. The clusterings $\pi, \sigma \in PART(S)$ are *isomorphic* if there is a bijection $h : S \longrightarrow S$ such that $x \equiv y(\pi)$ if and only if $h(x) \equiv h(y)(\sigma)$. This is denoted by $\pi \sim_d \sigma$.
A clustering quality measure m is *isomorphic invariant* if if all $\pi, \sigma \in PART(S)$ such that $\pi \sim_d \sigma$ we have $m(S, d, \pi) = m(S, d, \sigma)$.

If we add isomorphic invariance to the system of axioms introduced previously, the system remains consistent because it is easily seen that relm satisfies this extra axiom.

An example of clustering quality measure that satisfies scale invariance, consistency, richness and isomorphic invariance.

Definition

Let (S, d) be a dissimilarity space and let $G = (S, \mathcal{P}_2(S), d)$ be the weighted graph of (S, d). For $\pi \in PART(S)$, a cluster $C \in \pi$ consider the subgraph G_C and the set of paths $paths_C$ in G_C.
Let $x, y \in C$. The *weakest point link* of C is the number $\mathrm{wlp}_\pi(x, y) = d_{G_C}(x, y)$, where d_{G_C} is the ultrametric earlier defined.

In other words, wlp_π is the least maximum value of dissimilarity encountered on a path in C that joins x to y.

Definition

The *weakest link of the clustering* π is the number $\text{wl}(\pi)$ given by

$$\text{wl}(\pi) = \frac{\max\{\text{wlp}_\pi(x,y) \mid x \equiv y(\pi)\}}{\min\{d(x,y) \mid x \not\equiv y(\pi)\}}$$

wl satisfies all axioms.