COMP229: Introduction to Data Science
Lecture 17: Introduction to clustering, Hierarchical Agglomerative Clustering

Olga Anosova, O.Anosova@liverpool.ac.uk
Autumn 2023, Computer Science department
University of Liverpool, United Kingdom

Lecture plan

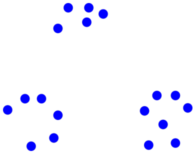
- Clustering: setting the problem
- Clustering types
- Hierarchical Agglomerative clustering
- Distances and similarity measures
- Dynamic Time Warping

Reminder: a metric

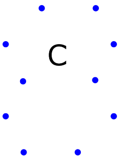
- A metric (distance) $d : C \times C \to \mathbb{R}$ should satisfy the axioms
  identity: $d(p, q) = 0$ if and only if $p = q$,
  symmetry: $d(p, q) = d(q, p)$ for any $p, q \in C$,
  triangle inequality: $d(p, q) + d(q, r) \geqslant d(p, r)$ for any $p, q, r \in C$.
- $L_s(p, q) = \left(\sum_{i=1}^{n} |p_i - q_i|^s \right)^{1/s}$ in $\mathbb{R}^n$, $s \geqslant 1$, where $s \in \mathbb{R}, s \geqslant 1$ and $p, q \in \mathbb{R}^n$.

What is clustering?

Wikipedia: clustering is grouping a set of objects in such a way that objects in the same group (called a *cluster*) are more similar (in some sense) to each other than to those in other groups (clusters).



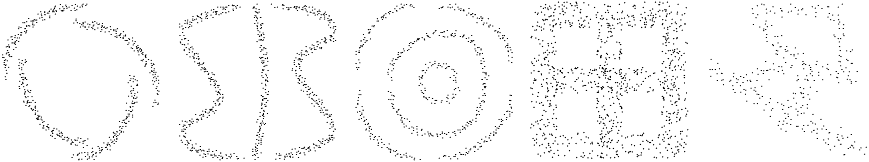
3 clusters on the left can be clear for a human, but not for a robot



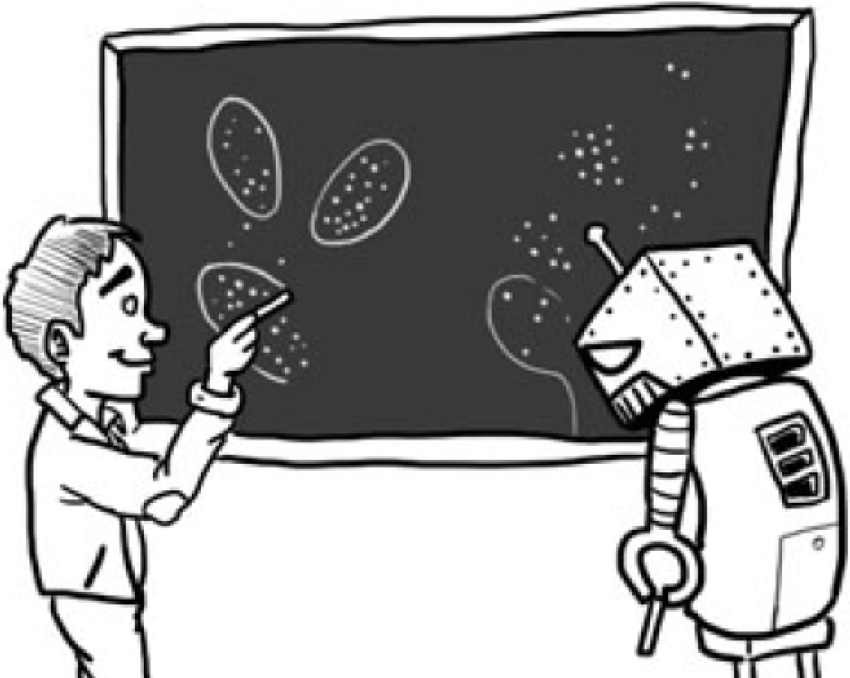
Can you see an interesting cluster in the last image?

Many clustering problems

To make the clustering problem exact, one needs to define in what sense objects are close, and also add conditions, e.g. on a number of clusters or a cost function to minimise. There are 1000s of algorithms that solve specific instances of clustering problems, some can be found in the classic 2015 overview paper.



There is no universal algorithm, hence clustering algorithms are often (re-)invented for new data.



Clustering task and process

General principles:

- In the same cluster instances must be similar.
- Instances in the different clusters must be different.
- Measurement for similarity and dissimilarity must be clear and have the practical meaning.

The standard process:

- *Feature extraction and selection:* choose the most representative ones from the data set;
- *Clustering algorithm design:* to fit the problem;
- *Result evaluation:* evaluate the clustering result and judge the validity of the algorithm;
- Result explanation: give a practical explanation.

Types of clustering algorithms

- Hierarchical clustering outputs a hierarchy of many outputs, a tree or a dendrogram similar to a classification of biological species.
- Centroid-based clustering optimises centres of clusters, e.g. we'll discuss k-means clustering.
- Density-based clustering defines clusters as areas of higher density, e.g. DBSCAN.
- Distribution-based clustering, e.g. using continuous normal densities around points.

Potential inputs for clustering

Often data points are given by coordinates. An input can be considered as a cloud of points in $\mathbb{R}^m$.

Some algorithms use only distances (similarities between points), not coordinates of data points.

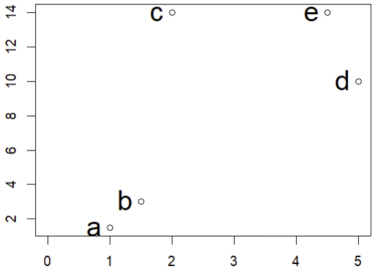
If all points are numbered, say from 1 to n, distances can be given or kept in the *distance matrix*, where each entry d_{ij} is the distance from the point i to the point j.
If the full matrix is too large, one can keep a smaller graph with lengths (weights) of edges.

Hierarchical Agglomerative Clustering

Hierarchical clustering can be *Divisive* (top-down) or *Agglomerative* (bottom-up).

Input: nomalised distance matrix. Normalisation is needed to avoid situations when points $A = (6ft, 75kg)$, $B = (6ft, 77kg)$, $C = (8ft, 75kg)$ produce equal distance of 2 units between A, B and A, C.

Start with the data like this:



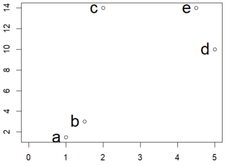
HAC algorithm

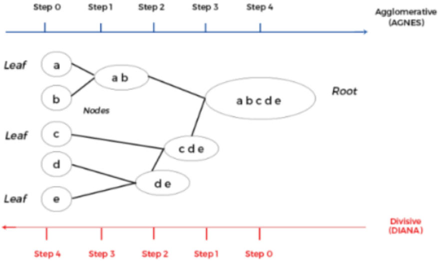
Input: threshold $t \in \mathbb{R}$ and the distance matrix:

	1	2	3	4
2	1.581139			
3	9.394147	7.826238		
4	12.539936	11.011358	5.000000	
5	12.980755	11.401754	4.031129	2.500000

Algorithm:

- Assign each item to its own cluster.
- Merge the closest pair of clusters into a single cluster.
- Repeat until all items are clustered into a single cluster.

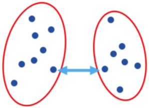




Cluster linkage

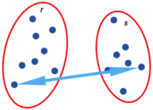
What is the distance between clusters?

Single Linkage:



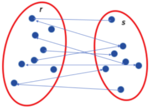
$L(r,s) = \min(D(x_{ri}, x_{sj}))$

Complete Linkage:



$L(r,s) = \max(D(x_{ri}, x_{sj}))$

Average Linkage:

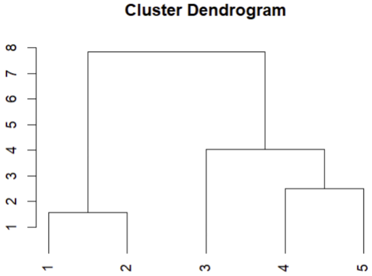


$$L(r,s) = \frac{1}{n_r n_s} \sum_{i=1}^{n_r} \sum_{j=1}^{n_s} D(x_{ri}, x_{sj})$$

Linkage influences the shape of the clusters.

When to stop?

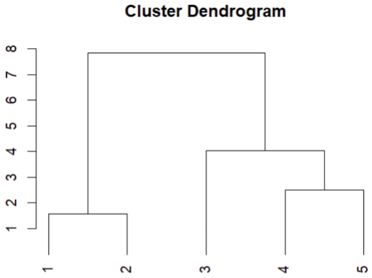
If we know the numbers of cluster that we need, we stop when we reach this number.



Otherwise we cut the dendrogram tree with a horizontal line where the line can move the max up-down without intersecting the node, i.e. when the clusters are most stable.

In this example, between heights 4 and 8, hence we'll have 2 clusters.

Clustering result



Another parameter that defines the results is the choice of a distance metric.

Metrics vs similarity measures

If the triangle inequality fails, Rass et al proved that the popular clustering algorithms, including k-means, can output any predetermined clusters.

Many comparisons output 'similarity' measures, e.g. the *Tanimoto* similarity between molecules is $T(A,B) = \dfrac{|A \cap B|}{|A \cup B|}$, where A, B can be any sets or strings of chemical compositions, $|A|$ denotes a size of A. Then T is symmetric, but fails the identity axiom: $A \neq B$ can be disjoint with $A \cap B = \varnothing$.

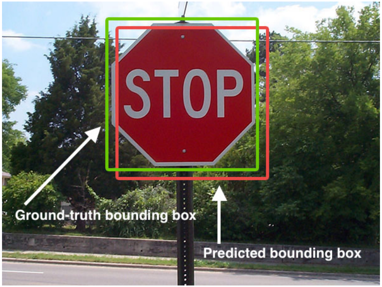
Metrics vs similarity measures

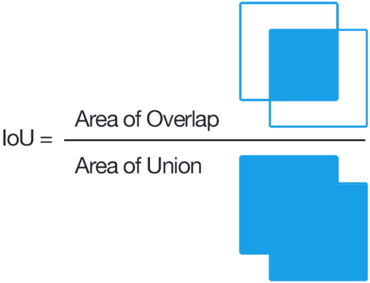
If the triangle inequality fails, Rass et al proved that the popular clustering algorithms, including k-means, can output any predetermined clusters.

Many comparisons output 'similarity' measures, e.g. the *Tanimoto* similarity between molecules is $T(A, B) = \dfrac{|A \cap B|}{|A \cup B|}$, where A, B can be any sets or strings of chemical compositions, $|A|$ denotes a size of A. Then T is symmetric, but fails the identity axiom: $A \neq B$ can be disjoint with $A \cap B = \emptyset$.

Claim 17.1. The Jaccard distance $J(A, B) = 1 - T$ measures *dissimilarity*, and it is a metric. A proof isn't needed for the exam.

Jaccard index in image recognition





Metric axioms are often assumed

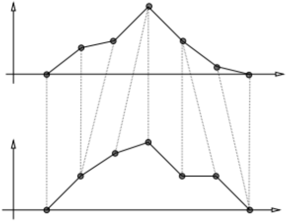
Problem 17.2. Let $|A|$ be the area of a subset $A \subset \mathbb{R}^2$. Find $J(A, B)$ if $A = [-1, 1]^2$, $B = [0, 2]^2$.

Solution 17.2. $A \cap B = [0, 1]^2$ has area 1. Then $|A \cup B| = |A| + |B| - |A \cap B| = 2^2 + 2^2 - 1^2 = 7$, so $J(A, B) = 1 - \dfrac{|A \cap B|}{|A \cup B|} = 1 - \dfrac{1}{7} = \dfrac{6}{7}$.

Many clustering algorithms, e.g. a 'fast k-means' in the next lecture, assume that input distances satisfy the triangle inequality, so a metric should be used.

Dynamic Time Warping (DTW)

DTW finds an optimal match between time series: finite sequences $x = (x_1, \ldots, x_n)$, $y = (y_1, \ldots, y_m)$.



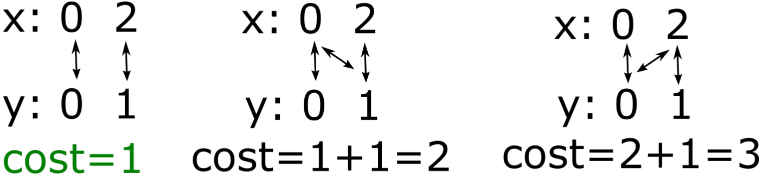
Definition 17.7. A match M is a set of pairs (x_i, y_j) including (x_1, y_1), (x_n, y_m), every x_i is paired with some y_j and every y_j is paired with some x_i

in a monotone way, i.e. there are no 'intersecting' pairs (x_i, y_j), (x_k, y_l) with $i < k$ and $j > l$.

$$DTW(x, y) = \sum_{(x_i, y_j) \in M} |x_i - y_j| \text{ minimised over } M.$$

DTW example computations

Problem 17.3. Find all pairwise DTW between the sequences $x = (0, 2)$, $y = (0, 1)$, $z = (0, 1, 1)$.



Solution 17.3. For x, y, the pairs $(0, 0)$ and $(2, 1)$ should be included into any match. Due to the monotonicity there are only two more possible pairs: $(0, 1)$ or $(2, 0)$. The minimum is $DTW(x, y) = 1$.

DTW pluses and minuses

Here are optimal matches for other sequences:

y: 0 1
z: 0 1 1
min cost=0

x: 0 2
z: 0 1 1
min cost=1+1=2

$$DTW(y,z) = 0,$$
$$DTW(x,z) = 2.$$

Problem 17.4. Is DTW a metric?

Solution 17.4. 1st axiom fails for y, z. 3rd axiom fails: $DTW(x,y) + DTW(y,x) < DTW(x,z)$. One plus: DTW is computed in linear time $O(m+n)$ proportional to the sum of lengths n, m of x, y.

Summary and final question

- Clustering is grouping a set of objects, it requires choice of measurement for similarity and dissimilarity.
- Each clustering is problem-specific.
- The are multiple types of clustering to choose from:
  - connectivity-based (hierarchical)
  - centroid-based (k-means)
  - density-based (DBSCAN)
  - distribution-based.
- Each of clustering methods depends on the cheoice of the distance (or similarity measure).

Problem 17.5. In how many ways can we split n points into k subsets containing $n_1, \ldots, n_k$ points?