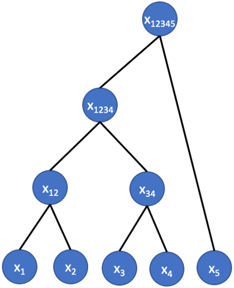
Machine Learning

Deniz Gündüz and Krystian Mikolajczyk

Department of Electrical and Electronic Engineering
Imperial College London

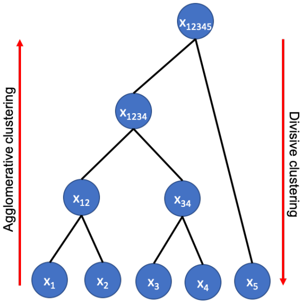
Hierarchical Clustering

- Goal is to create a sequence of nested partitions, which form a tree structure
- Lowest level of the tree (*leaves*) correspond to individual data points as distinct clusters, highest level to all points in a single cluster
- Meaningful clusters obtained at intermediate levels
- Level chosen based on number of clusters



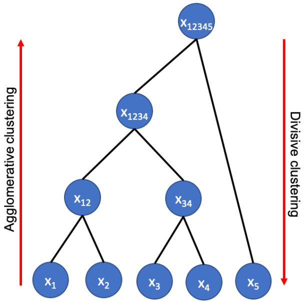
Hierarchical Clustering

- Agglomerative Clustering: Start from bottom; merge most similar pair of clusters until all points form a single cluster
- Divisive Clustering: Start from top; split clusters until each point forms a separate cluster



Hierarchical Clustering

- Agglomerative Clustering: Start from bottom; merge most similar pair of clusters until all points form a single cluster
- Divisive Clustering: Start from top; split clusters until each point forms a separate cluster

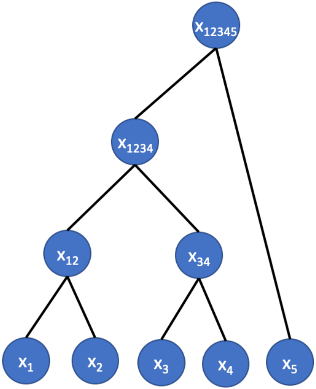


Hierarchical Clustering

Definition

Clustering $\mathcal{A} = \{\mathcal{A}_1, \ldots, \mathcal{A}_r\}$ is said to be nested in another clustering $\mathcal{B} = \{\mathcal{B}_1, \ldots, \mathcal{B}_s\}$ iff $r > s$, and for each cluster $\mathcal{A}_i \in \mathcal{A}$, there exits a cluster $\mathcal{B}_j \in \mathcal{B}$ such that $\mathcal{A}_i \subset \mathcal{B}_j$.

- We obtain a sequence of nested clusterings: $\mathcal{C}^1, \ldots, \mathcal{C}^n$
- Cluster dendrogram captures the nesting structure : we have an edge from cluster $\mathcal{C}_i \in \mathcal{C}^{t-1}$ to cluster $\mathcal{C}_j \in \mathcal{C}^t$ if $\mathcal{C}_i \subset \mathcal{C}_j$



How Many Hierarchical Clusterings Are There?

- Number of hierarchical clusterings = number of dendrograms (binary rooted trees) with n leaves
- Any rooted tree with k vertices has $k-1$ edges
- Any rooted binary tree with n leaves has $n-1$ internal vertices, hence $2n-1$ nodes in total, and $2n-2$ edges
- Consider a dendrogram with n leaves. If we add an extra leave, we can create $2n-1$ new dendrograms: connects either to one of the vertices, or the as the child of a new root.
- Total number of dendrograms with n leaves:

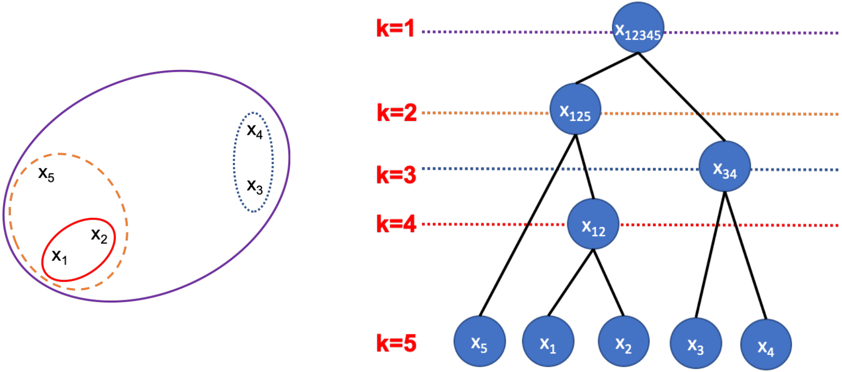
$$\prod_{i=1}^{n-1}(2i-1) = 1 \times 3 \times \cdots \times (2n-3)$$

- Too many clusterings for exhaustive search!

Agglomerative Clustering

- Start each data point as a separate cluster: $\{\mathcal{C}_1, \ldots, \mathcal{C}_n\}$
- Find the closest pair of clusters $\mathcal{C}_i$ and $\mathcal{C}_j$
- Replace $\mathcal{C}_i$ and $\mathcal{C}_j$ with $\mathcal{C}_{ij}$
- Repeat until there is a single cluster

Agglomerative Clustering



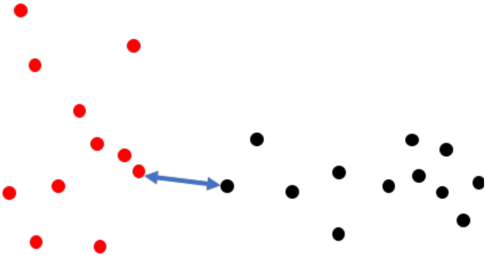
Linkage

- Let $d(\boldsymbol{x}_i, \boldsymbol{x}_j)$ denote the dissimilarity (distance) between any two points
- At first step (each cluster is a single data point) we can identify closest pair using dissimilarity between points
- For following iterations, we need a distance measure between clusters, called linkage

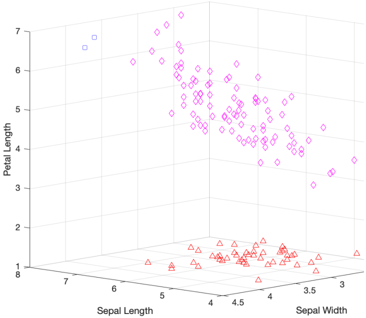
Single Linkage

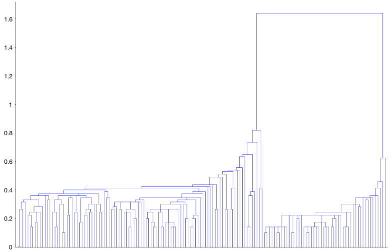
- Dissimilarity between two clusters $\mathcal{C}_i$ and $\mathcal{C}_j$ defined as:

$$\delta(\mathcal{C}_i, \mathcal{C}_j) = \min\{d(\boldsymbol{x}_i, \boldsymbol{x}_j) : \boldsymbol{x}_i \in \mathcal{C}_i, \boldsymbol{x}_j \in \mathcal{C}_j\}$$



Single Linkage Example: Iris Dataset



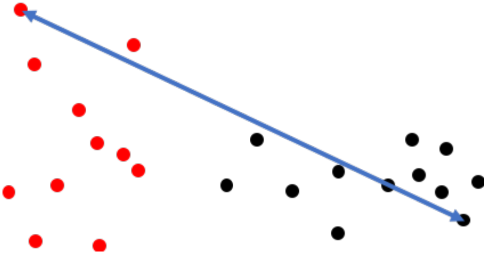


- Cutting the tree at $h = 0.75$
- Interpretation: For each point $\boldsymbol{x}_i$, there is another point in its cluster with dissimilarity ≤ 0.75.

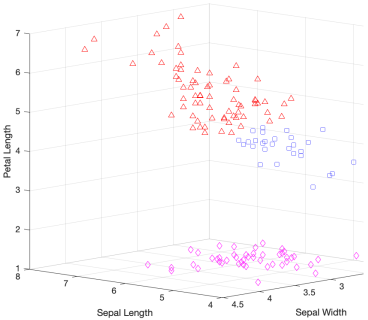
Complete Linkage

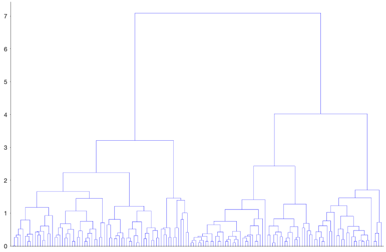
- Dissimilarity between two clusters $\mathcal{C}_i$ and $\mathcal{C}_j$ defined as:

$$\delta(\mathcal{C}_i, \mathcal{C}_j) = \max\{d(\boldsymbol{x}_i, \boldsymbol{x}_j) : \boldsymbol{x}_i \in \mathcal{C}_i, \boldsymbol{x}_j \in \mathcal{C}_j\}$$



Complete Linkage Example: Iris Dataset





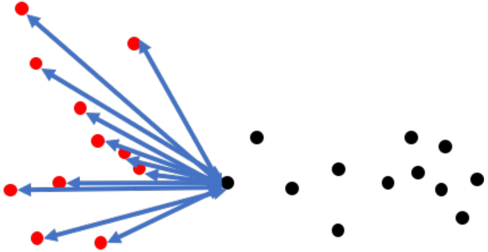
- Cutting the tree at $h = 3.5$
- Interpretation: For each point $\boldsymbol{x}_i$, every other point in its cluster has dissimilarity ≤ 3.5.

Average Linkage

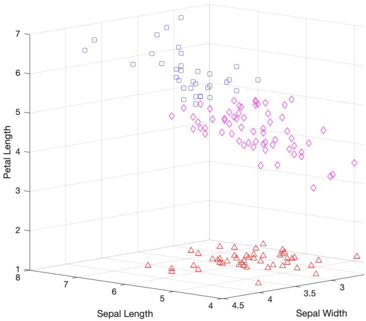
- Dissimilarity between two clusters $\mathcal{C}_i$ and $\mathcal{C}_j$ defined as:

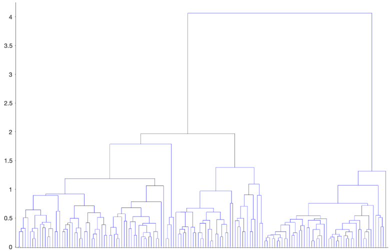
$$\delta(\mathcal{C}_i, \mathcal{C}_j) = \frac{1}{n_i \cdot n_j} \sum_{\boldsymbol{x}_i \in \mathcal{C}_i} \sum_{\boldsymbol{x}_j \in \mathcal{C}_j} d(\boldsymbol{x}_i, \boldsymbol{x}_j)$$

where $n_i = |\mathcal{C}_i|$.



Average Linkage Example

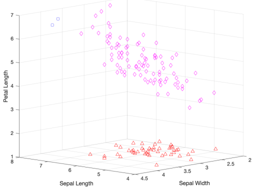


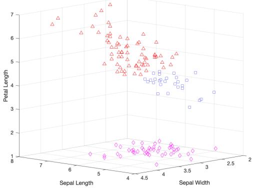


- No intuitive interpretation.

Limitations

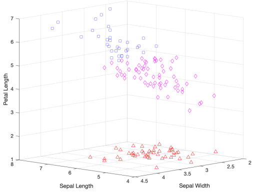
- Single linkage suffers from chaining: Clusters can be too much spread out, and not compact enough
- Complete linkage suffers from crowding: A point can be closer to points in other clusters than those in its own cluster. Clusters are compact, but not far enough apart.



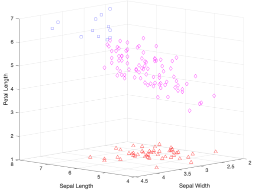


Limitations of Average Linkage

- No intuitive interpretation
- Clustering may change if a monotone increasing transformation is applied to the dissimilarity measure; i.e., $d \rightarrow d^2$ or $d \rightarrow e^d/(1+e^d)$



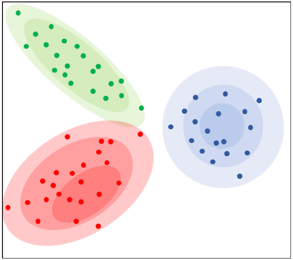
Average linkage with Eucledian.



Average linkage with Eucledian squared.

Expectation Maximization Clustering

- We have seen hard-clustering algorithms
- Instead we can use soft assignment of points to clusters, where each point belongs to each cluster with some probability



$$
\begin{aligned}
f_i(\mathrm{x}) &= f(\mathrm{x}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \\
&= \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}_i|^{1/2}} \exp\left\{ \frac{(\mathrm{x}-\boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1} (\mathrm{x}-\boldsymbol{\mu}_i)}{2} \right\}
\end{aligned}
$$