香港城市大學
City University of Hong Kong
專業 創新 胸懷全球
Professional·Creative
For The World

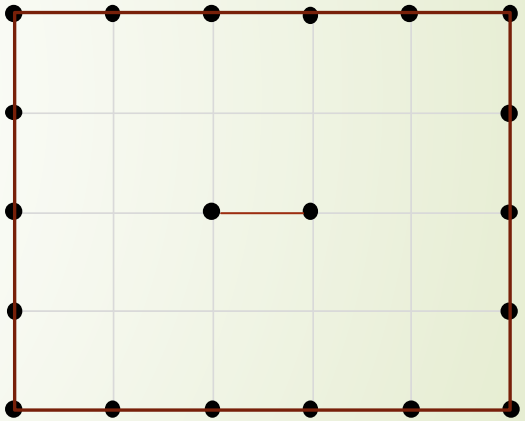
Cluster Analysis:
Hierarchical Methods

CS5483 Data Warehousing and Data Mining

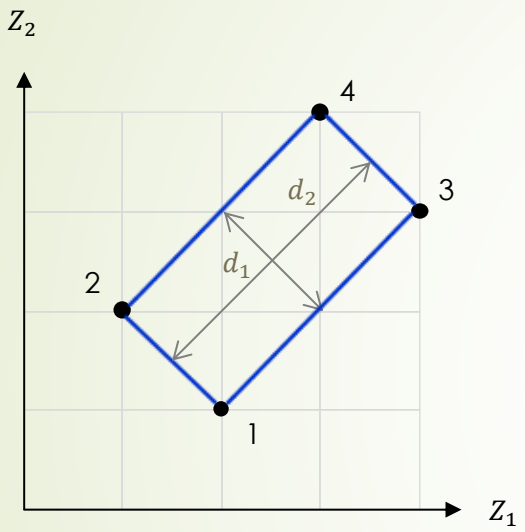
Main idea

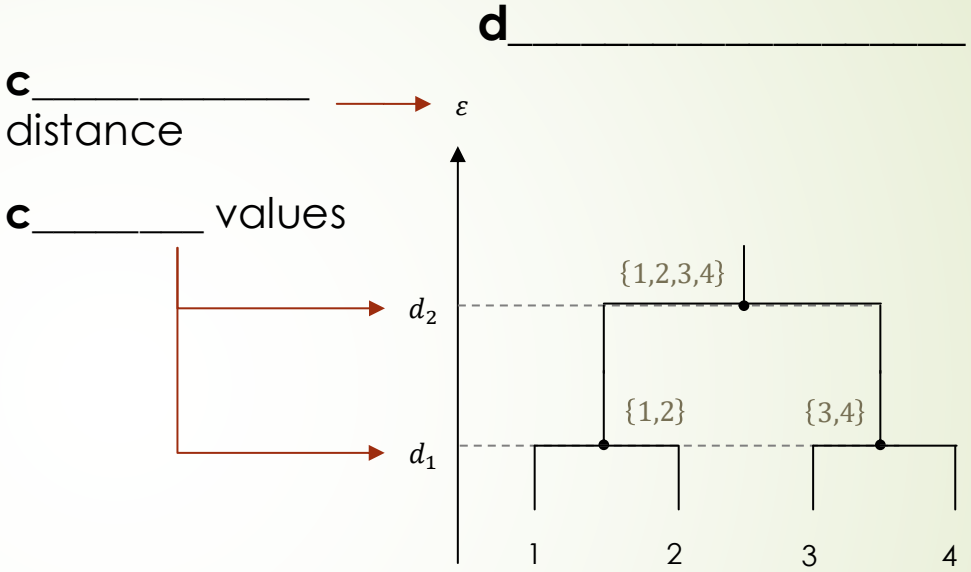
- Connect two points $\boldsymbol{p}', \boldsymbol{p}''$ whenever $\mathrm{dist}(\boldsymbol{p}', \boldsymbol{p}'') \leq \varepsilon$.
- Return **c**_____ **c**_____ as clusters.
- Benefits over k-means:
  - Can identify non-spherical clusters.
  - No need to choose _____
- How to choose ε?

With $\varepsilon = 1$,



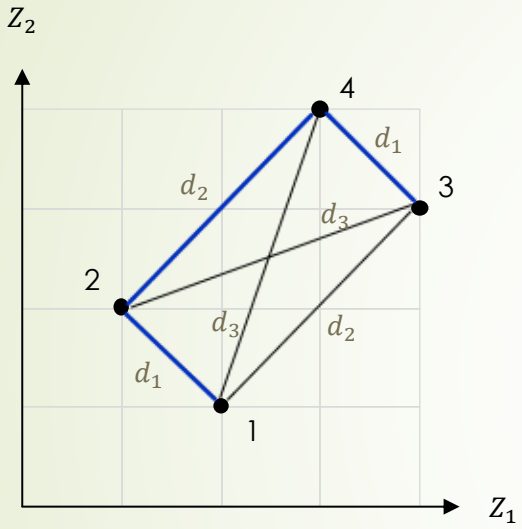
Hierarchical clustering

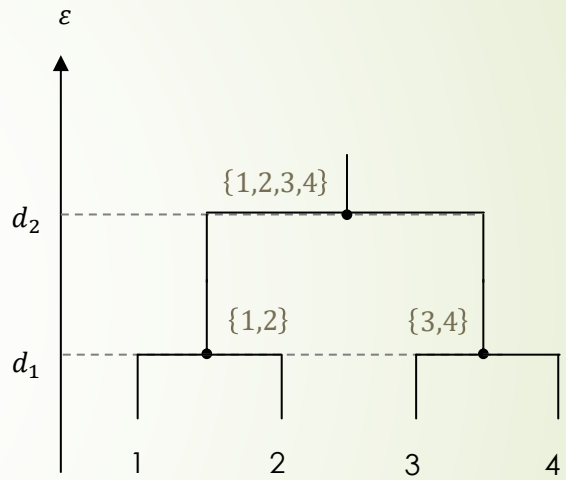




- Try all possible ε from 0.
- Space complexity: _____ (∵at most _____ internal nodes in the tree.)
- Time complexity?

Single-linkage method





1. Start with singletons as clusters.
2. Merge clusters with a single shortest link.

- The links form a **m**_____ **s**_____ **t**ree. See Kruskal's algorithm.

Minimum spanning tree (MST) algorithm

- **Kruskal's algorithm**:

input: weighted connected graph $G = (V, E, w)$

output: minimum spanning tree $T = (V, E')$,
i.e., a spanning tree of G with minimum sum weight $\sum_{e \in E'} w(e)$.

$E' \leftarrow \emptyset$

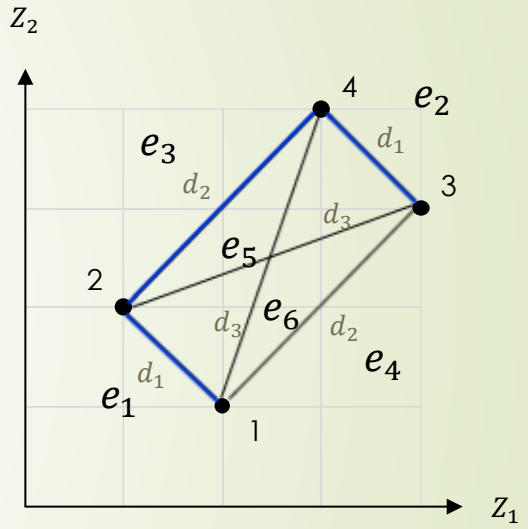
enumerate E as $e_1, e_2, \ldots, e_m$ such that $w(e_1) \leq w(e_2) \leq \cdots \leq w(e_m)$

for e from e_1 to e_m

    add e to E' if no cycle will be formed

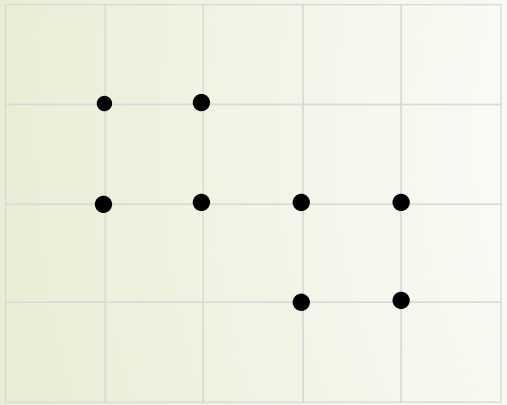
return $T \leftarrow (V, E')$

- Complexity: $O(|E| \log |V|)$ for sorting the edges.

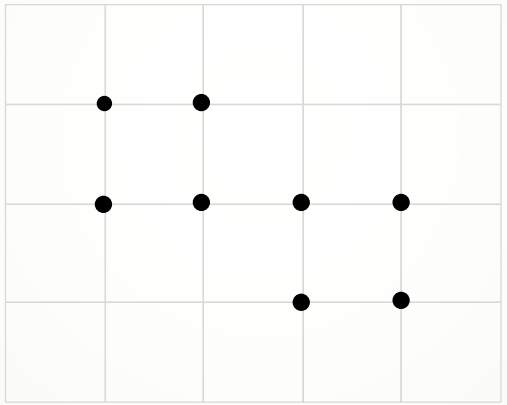


Limitation of single-linkage algorithm

Centroid-based with $k = 2$



Single-linkage with $\varepsilon = 1$

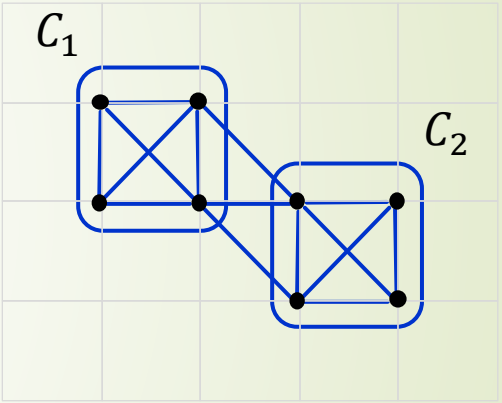


- **C_____ p_____**: which makes the clustering solutions sensitive to o_____.
- Remedy?

Complete-linkage

- Return *maximal* **c**_____ as clusters.
- Issues:
  - Clusters may o_____.
  - Computation is _____.
- How to return disjoint clusters efficiently?

Consider $\varepsilon = \sqrt{2}$



Agglomerative clustering
AGlomerative NESting (AGNES)

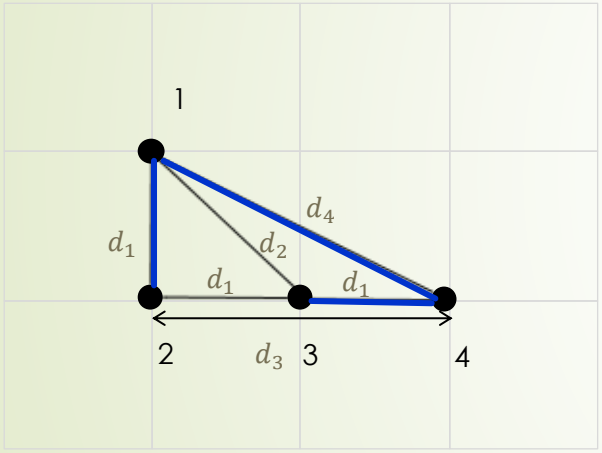
1. Start with singleton sets as clusters.
2. Repeatedly merge two clusters C, C' according to a linkage criteria such as minimizing certain cluster distance $\mathrm{dist}(C, C')$.

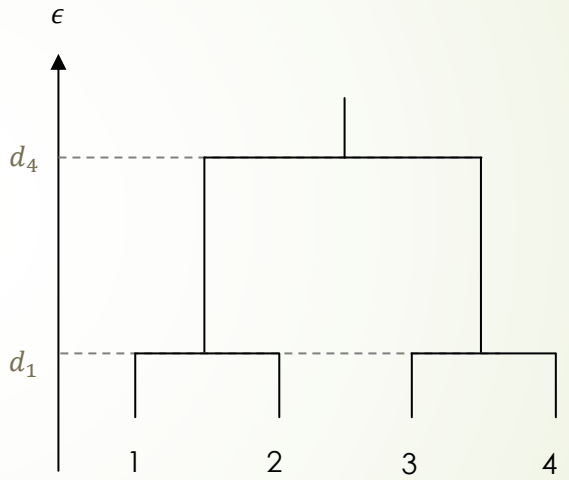
- May stop merging before reaching the trivial cluster if
  - clusters are dissimilar enough, or
  - the desired number of clusters is reached.
- [Optional] There is also a divisive approach called **DIvisive ANAlysis (DIANA).**

Different choices of cluster distances

1. Minimum distance/s_____: $\min_{\boldsymbol{p}\in C,\boldsymbol{p}'\in C'} \mathrm{dist}(\boldsymbol{p},\boldsymbol{p}')$
2. Maximum distance/complete-linkage: $\max_{\boldsymbol{p}\in C,\boldsymbol{p}'\in C'} \mathrm{dist}(\boldsymbol{p},\boldsymbol{p}')$
3. Centroid: $\mathrm{dist}(\boldsymbol{c},\boldsymbol{c}')$ where $\boldsymbol{c}$ and $\boldsymbol{c}'$ are the centroids of C and C' resp.
4. Ward: $\min_{\boldsymbol{c}} \sum_{\boldsymbol{p}\in C\cup C'} \mathrm{dist}(\boldsymbol{p},\boldsymbol{c})^2$
5. Average: $\frac{1}{|C||C'|}\sum_{\boldsymbol{p}\in C,\boldsymbol{p}'\in C'} \mathrm{dist}(\boldsymbol{p},\boldsymbol{p}')$
6. Group-average: $\frac{1}{\binom{|C\cup C'|}{2}}\sum_{\boldsymbol{p}\neq\boldsymbol{p}'\in C\cup C'} \mathrm{dist}(\boldsymbol{p},\boldsymbol{p}')$

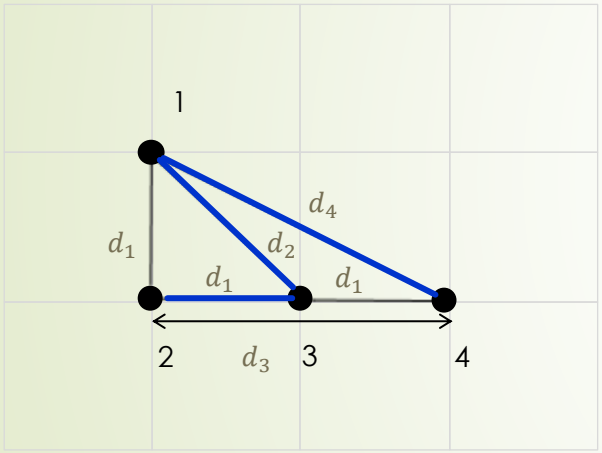
Complete-linkage method

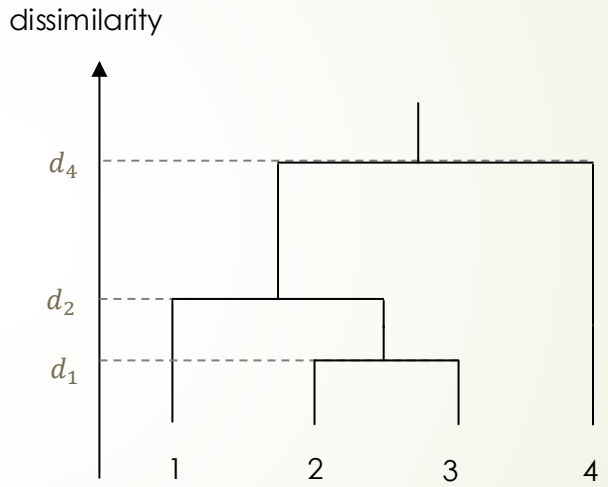




- Run AGNES with $\mathrm{dist}(C, C') := \max_{\boldsymbol{p} \in C, \boldsymbol{p}' \in C'} \mathrm{dist}(\boldsymbol{p}, \boldsymbol{p}')$.
- C, C' are merged only if $C \cup C'$ is a clique. Why? _____

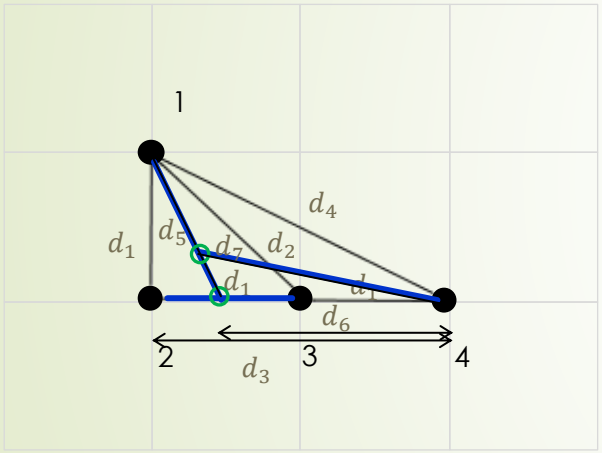
With a different order of agglomeration

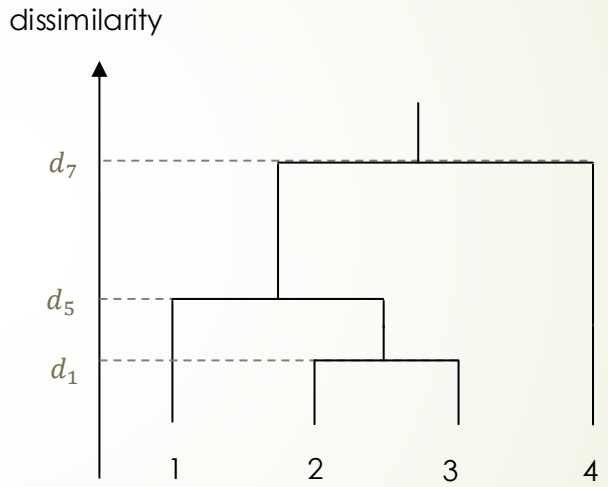




- Is the clustering solution unique? Yes/No

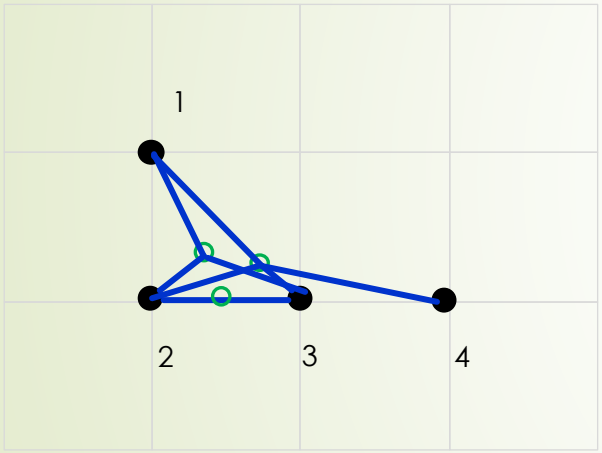
Centroid method

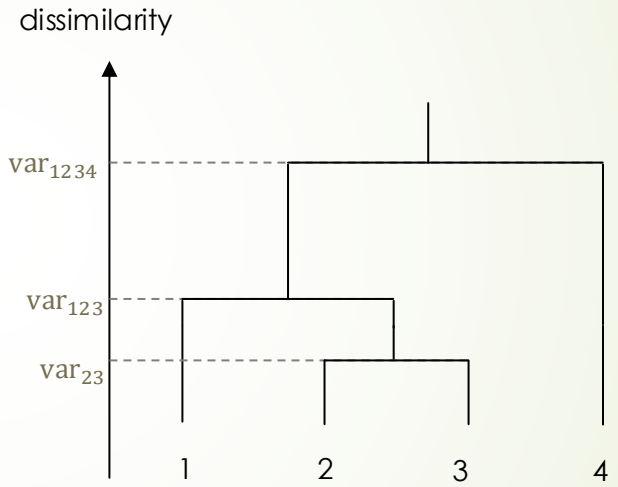




- Is the clustering solution unique? Yes/No

Ward's method





- Minimizes the variance of the merged cluster.
- Is the clustering solution unique? Yes/No

References

- 10.3 Hierarchical Methods
  (up to an including 10.3.2 Distance Measures in Algorithmic Methods)
- Supplementary readings (Optional):
  - R. Sibson (1973). "SLINK: an optimally efficient algorithm for the single-link cluster method". *The Computer Journal*. British Computer Society. **16** (1): 30–34. (Wikipedia page)
  - D. Defays (1977). "An efficient algorithm for a complete link method". *The Computer Journal*. British Computer Society. **20** (4): 364–366. (Wikipedia page)
  - Ward, Joe H. "Hierarchical Grouping to Optimize an Objective Function." *Journal of the American Statistical Association* 58, no. 301 (1963): 236–44. (Wikipedia page)
  - Kaufman, Leonard, and Peter J. Rousseeuw. "*Finding groups in data: an introduction to cluster analysis.*" John Wiley & Sons, 2009. (Wikipedia page)