Hierarchical clustering

Sanjoy Dasgupta

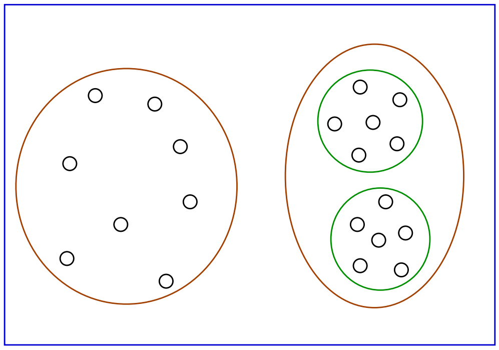
University of California, San Diego

Topics we'll cover

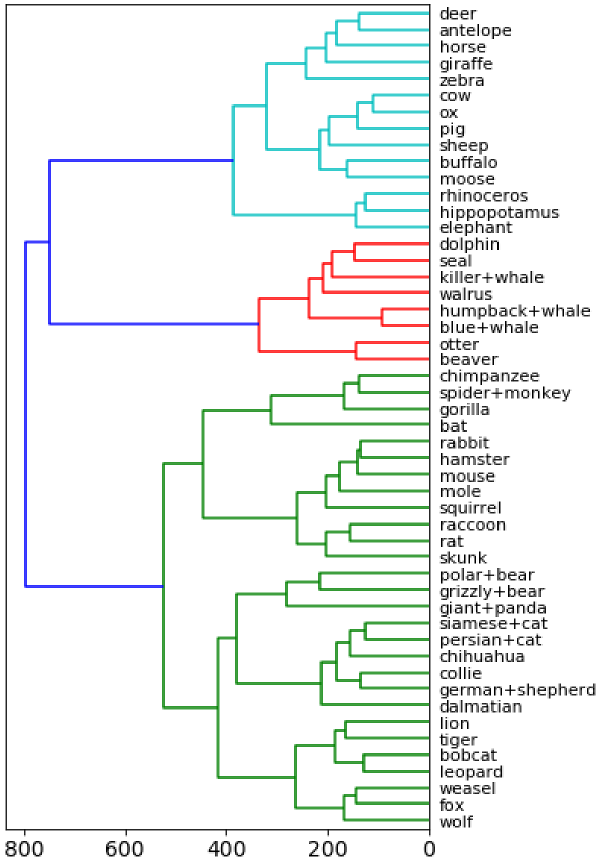
1. What is hierarchical clustering?
2. Single linkage
3. The other linkage schemes

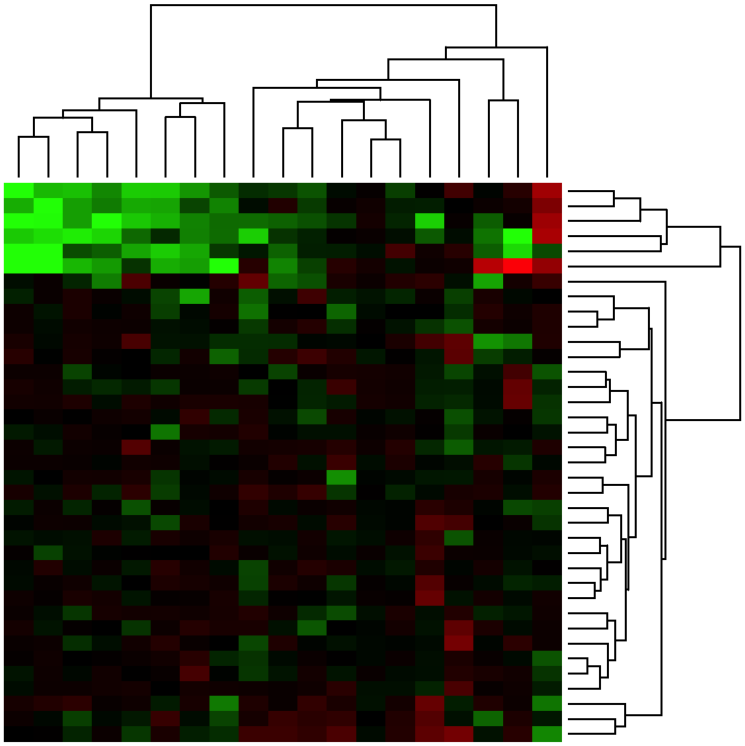
Hierarchical clustering

Choosing the number of clusters (k) is difficult.

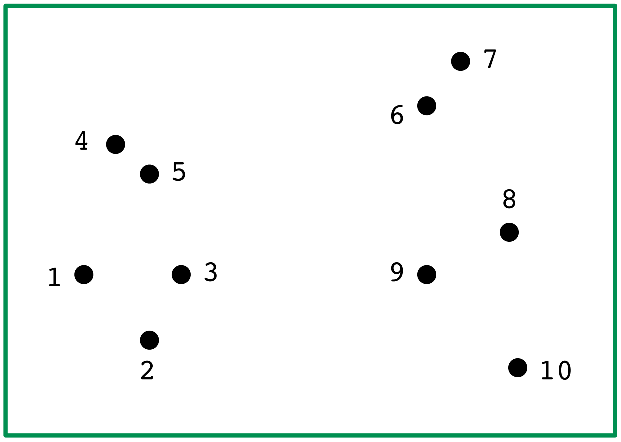


Often there is no single right answer, because of multiscale structure.





The single linkage algorithm

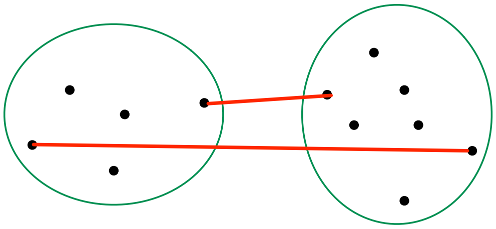


- Start with each point in its own, singleton, cluster
- Repeat until there is just one cluster:
  - Merge the two clusters with the closest pair of points

Linkage methods

- Start with each point in its own cluster
- Repeat until there is just one cluster:
  - Merge the two "closest" clusters

How to measure the distance between two clusters C, C'?



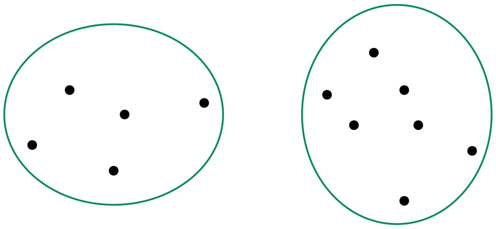
- Single linkage

$$\text{dist}(C, C') = \min_{x \in C, x' \in C'} \|x - x'\|$$

- Complete linkage

$$\text{dist}(C, C') = \max_{x \in C, x' \in C'} \|x - x'\|$$

Average linkage



❶ Average pairwise distance between points in the two clusters

$$\text{dist}(C, C') = \frac{1}{|C| \cdot |C'|} \sum_{x \in C} \sum_{x' \in C'} \|x - x'\|$$

❷ Distance between cluster centers

$$\text{dist}(C, C') = \|\text{mean}(C) - \text{mean}(C')\|$$

❸ Ward's method: increase in k-means cost from merging the clusters

$$\text{dist}(C, C') = \frac{|C| \cdot |C'|}{|C| + |C'|} \|\text{mean}(C) - \text{mean}(C')\|^2$$