Statistically Speaking: Agglomerative and divisive clustering

DAVID L. TABB, PH.D.

NOVEMBER 16, 2017

Overview

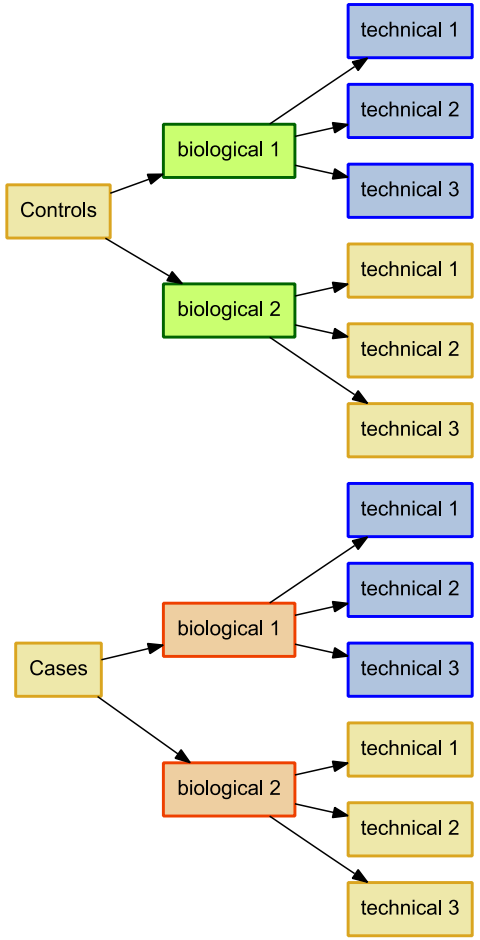
- Discovering sample relationships
- Distance metrics
- Agglomerative versus Divisive
- Interpreting dendrograms
- Tore Dalenius, statistician

Why cluster?

- A dendrogram visualizes nested sample relationships derived from data.
- Sample structure can be inferred from sample data; is this story different than our preconceptions of the samples?
- Can we find samples that may have been misclassified before we acquired data?

Hierarchy of relationships

- In this experiment, which samples should be most similar?
- Multiple levels of relationship constitute a hierarchy.
- An inferred hierarchy should match the design hierarchy.
- What if Control: Bio2 is actually a case rather than a control?



Distance metrics

- Manhattan $d = \sum |P_i - Q_i|$
  - Sum of measurement differences
- Euclidean $d = \sqrt{\sum |P_i - Q_i|^2}$
  - Pythagorean distance
- Chebyshev $d = \max_i |P_i - Q_i|$
  - Maximum of measurement differences
- Kullback-Leibler $d = \sum P_i \ln \frac{P_i}{Q_i}$
  - Relative entropy

Lumpers versus splitters

HIERARCHICAL AGGLOMERATIVE CLUSTERING

- Start each datum in its own cluster.
- Find closest clusters and join them.
- Repeat above until all clusters are joined.

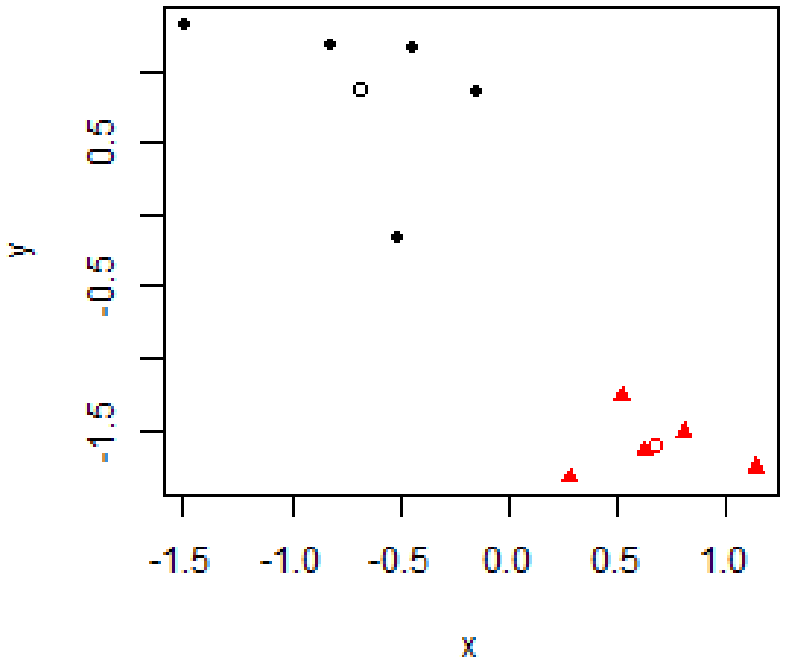
DIVISIVE: TOP DOWN CLUSTERING

- Start cluster contains all data points.
- Cluster is split to two clusters by rule.
- Repeat until all points are clusters.

Jain et al. *ACM Computing Surveys* (1999) 31: 264.

Linkage: determine inter-group distance

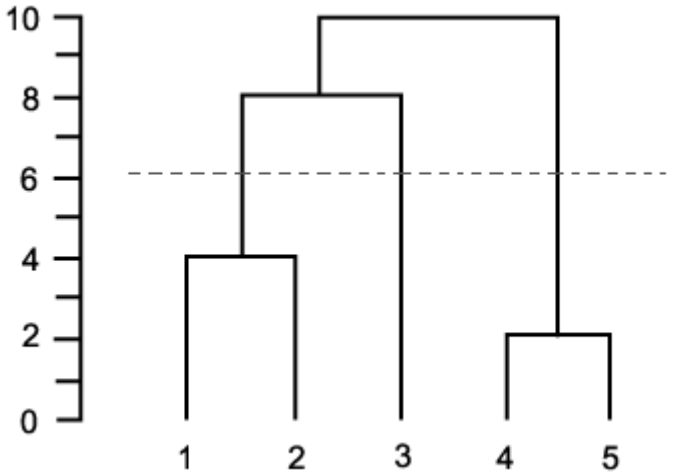
- *Complete*: maximum distance between: 4.05
- *Single*: minimum distance between: 1.51
- *Mean*: average of all distances between: 2.86
- *Centroid*: distance between centroids: 2.82



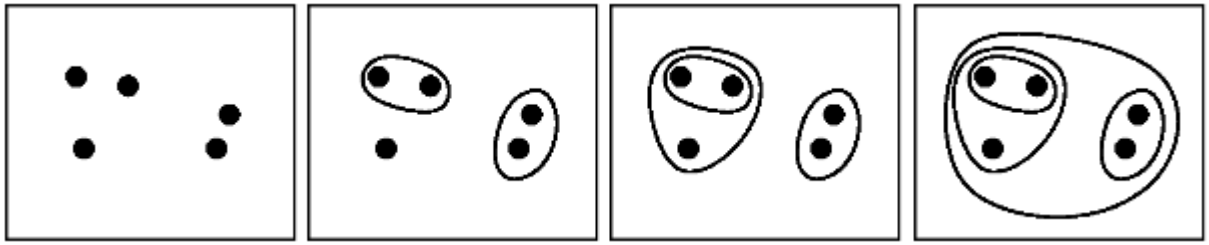
- Filled: cluster members
- Empty: cluster centroids

Agglomerative clusters form step by step

- At every step, software seeks the closest points/groups to join.
- Sometimes, this approach is called a "greedy" algorithm.



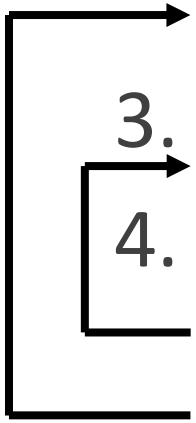
How many clusters do I claim?



https://stackoverflow.com/questions/8016313/agglomerative-clustering-in-matlab

K-means clustering

1. User specifies *k*, the number of groups anticipated from the data, and *n*, the number of starting positions.
2. From *n* different starts, software labels each member to a random group.
3. The centroid for each group is computed.
4. Each member is re-labelled to the group with the closest centroid.



UNIVERSITEIT STELLENBOSCH UNIVERSITY

Questions to ask your dendrogram

- What happens if I rotate a branch to the other orientation?
  - *These two views are both correct!*
- Are any of the branch points very close together?
  - *Ordering of these splits may be dubious.*
- What does the length of the arms mean?
  - *Genetic diversity is hard to map to years!*

Tore E. Dalenius 1917-2002

DOI: 10.1007/
978-3-0348-5513-6_5

- Swedish statistician at University of Stockholm and Brown University with key contributions in survey design and in data privacy.
- 1950 developed SSQ Clustering / Stratified Sampling, a predecessor of *K*-means clustering.
- 1953 named fellow of American Statistical Association
- 1957 defended Ph.D.: *Sampling in Sweden: Contributions to the Methods and Theories of Sample Survey Practice*
- 1977 introduced Statistical Disclosure Control, and later Data-swapping, to guard privacy of study participants

Takeaways

- Clustering almost always involves many arbitrary decisions. No cluster should be considered perfect, inarguable truth.
- When you know the number of different types in your data, K-means is a nice option.
- Always keep the rotation principle in mind when reading a dendrogram.