Unsupervised Machine Learning: Clustering: Hierarchical Clustering

Alipio Jorge

January 2021

Clustering approaches

Top-down

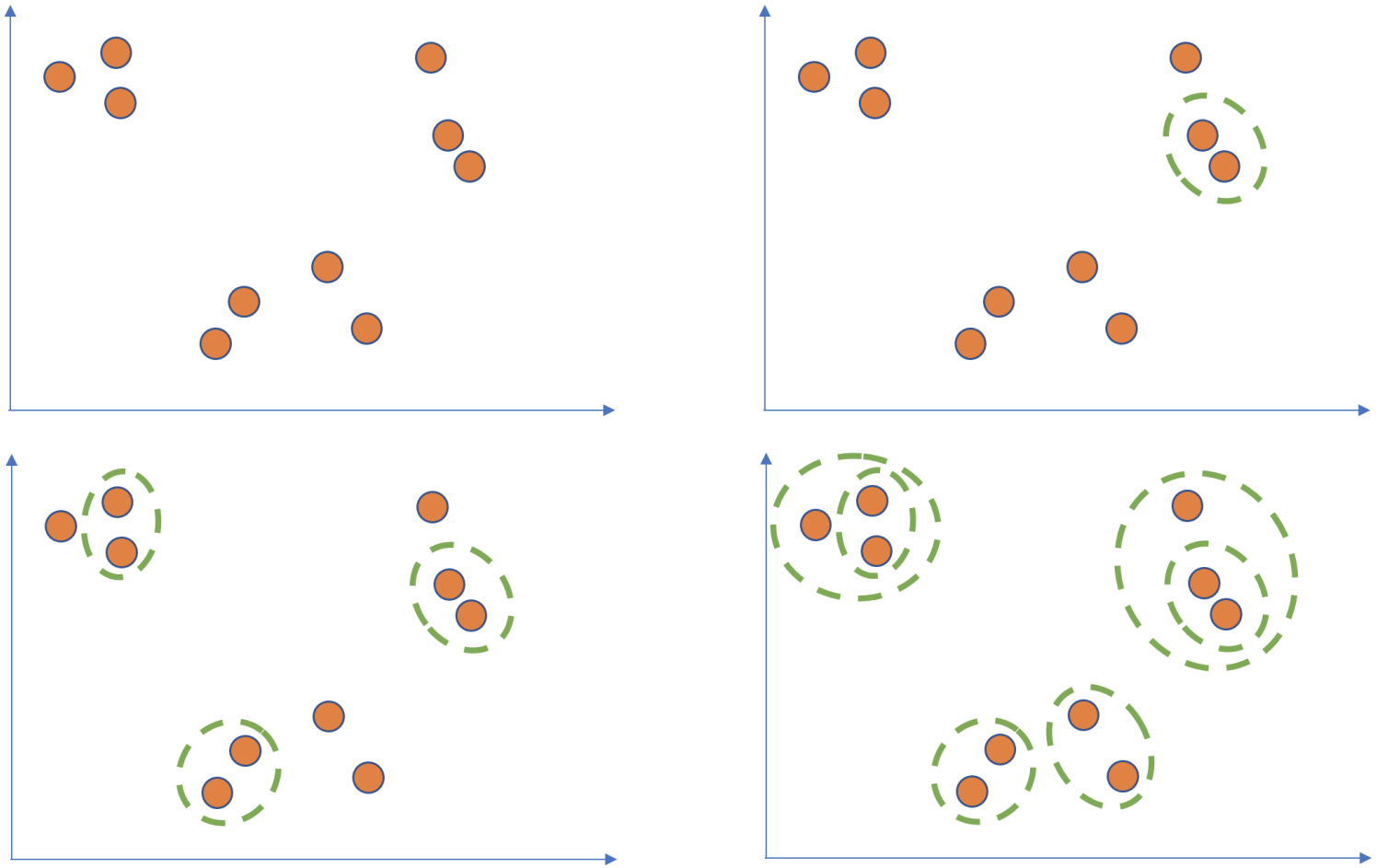
- Divisive approach
  - e.g. k-means algorithm

Figure 1: ' '

Clustering approaches

Bottom-up

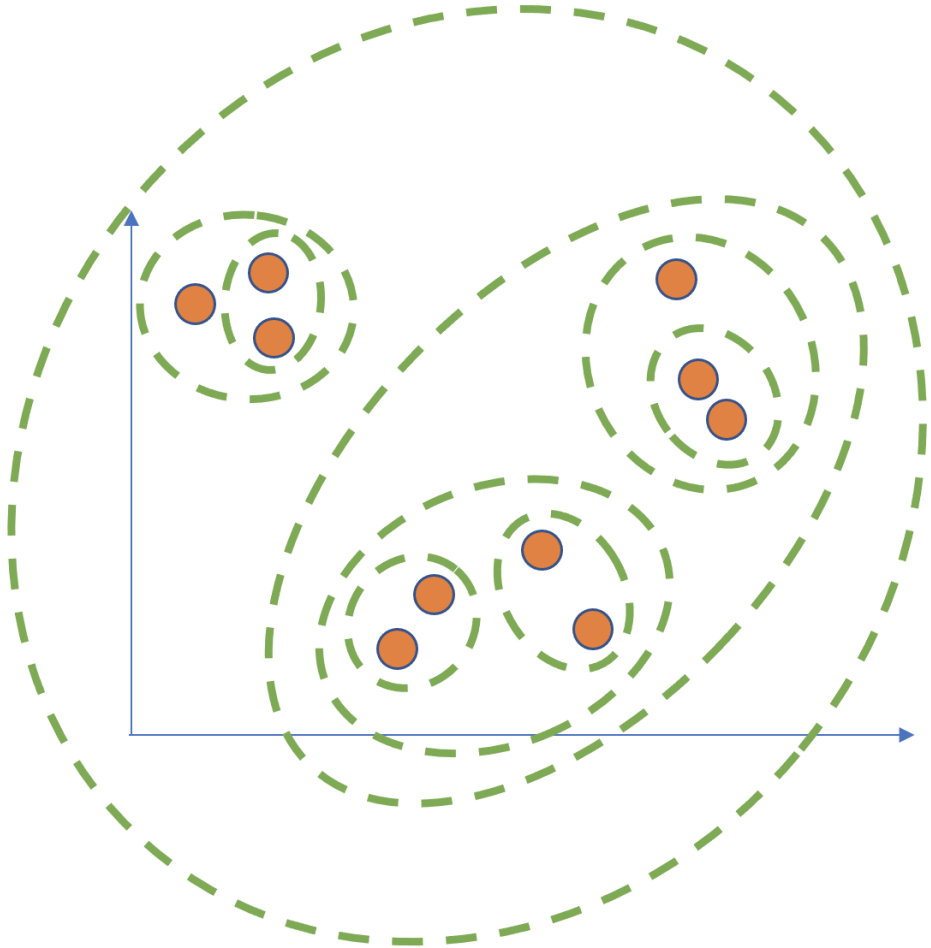
- Agglomerative appraoches
  - e.g. Hierarchical Clustering



Clustering approaches

Bottom-up

- Agglomerative appraoches
  - e.g. Hierarchical Clustering



Hierarchical CLustering

Algorithm

- **Given**: data D, object distance d, cluster distance d_C
- **Output**: a clustering hierarchy H
- **DO**:
  - Each case is a cluster
  - **Repeat**
    - join the **two nearest** clusters A and B into C
    - A and B are the two **branches** of C
  - **Until** there is only one cluster H

Hierarchical Clustering

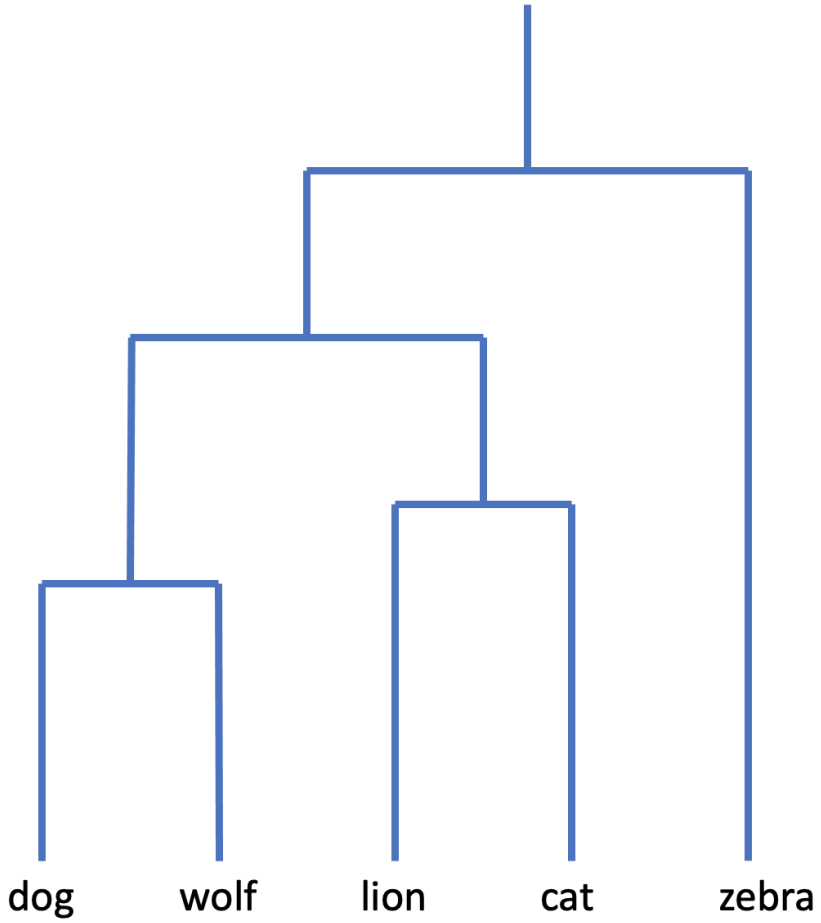
Simple Example

- Objects are: dog, wolf, lion, cat, zebra
- Step 1: join dog and wolf into dw
- Step 2: join lion and cat into lc
- Step 3: join dog-wolf and lion-cat into dwlc
- Step 4: join zebra and dwlc into H

Hierarchical Clustering

Dendrogram

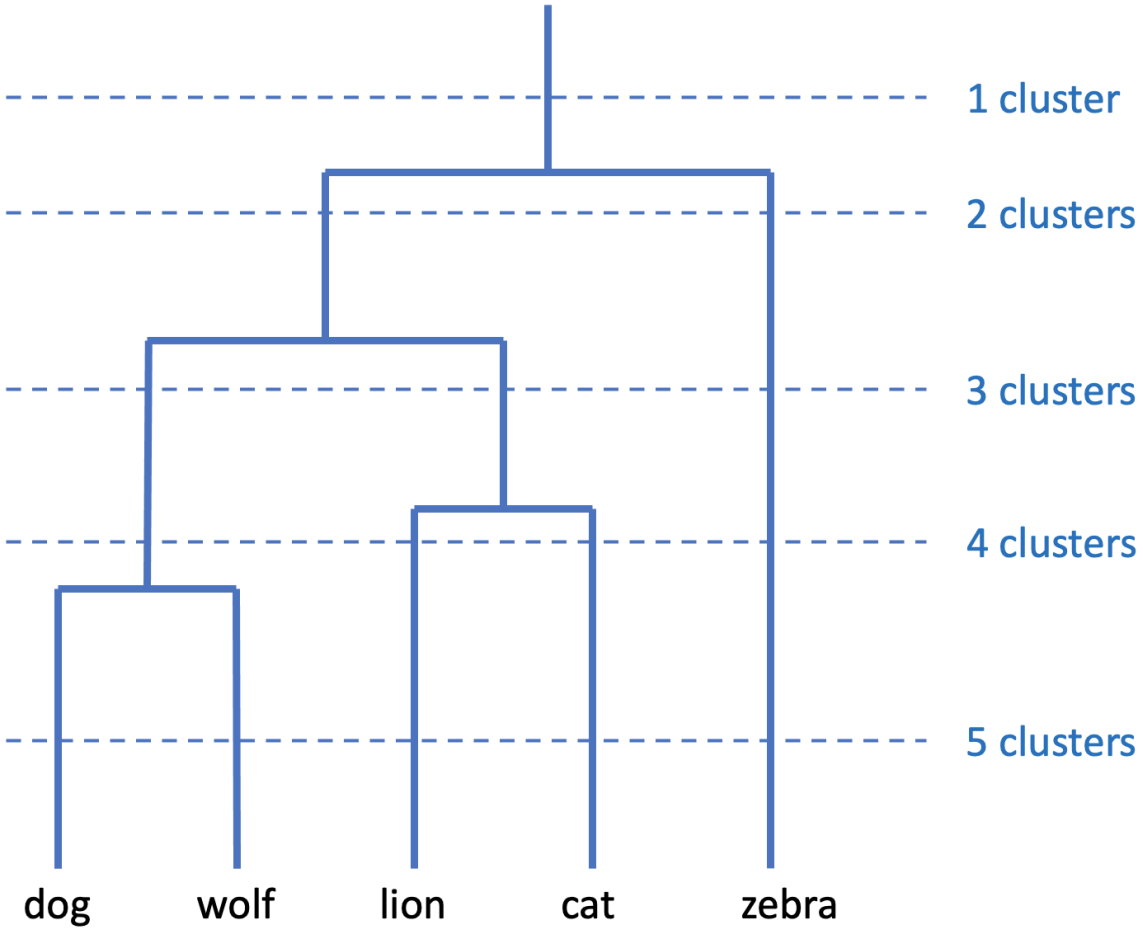
- The resulting hierarchy can be visualized as a tree
  - that chart is called a **dendrogram**



Hierarchical Clustering

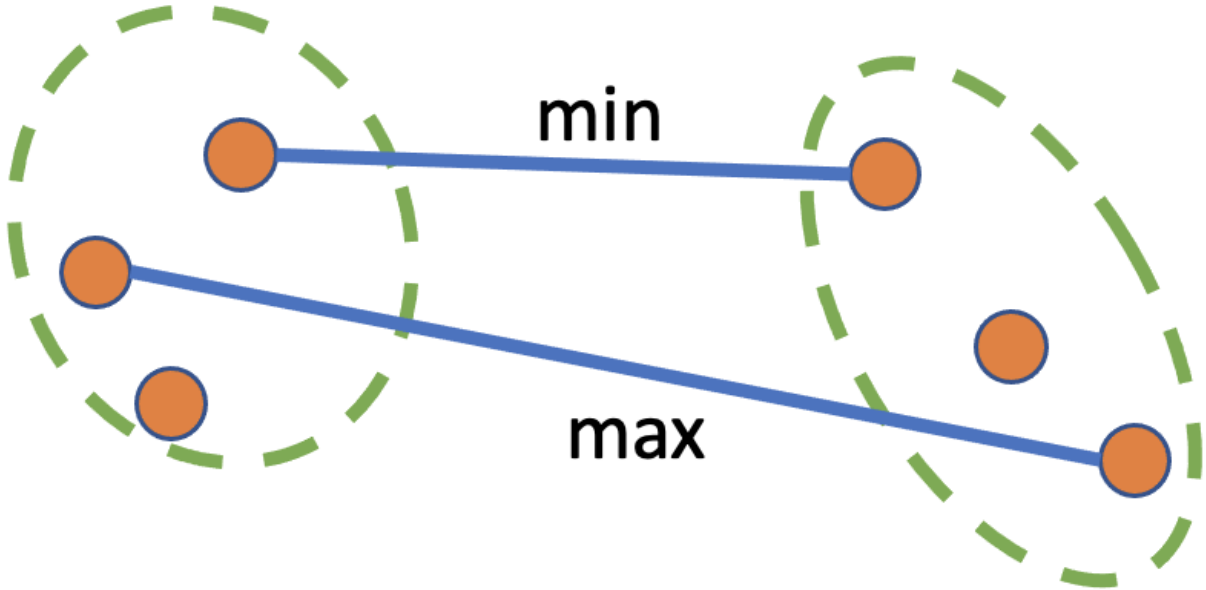
Obtaining clusters

- We can obtain clusters byt cutting the tree
  - different **height cuts** give different numbers of clusters



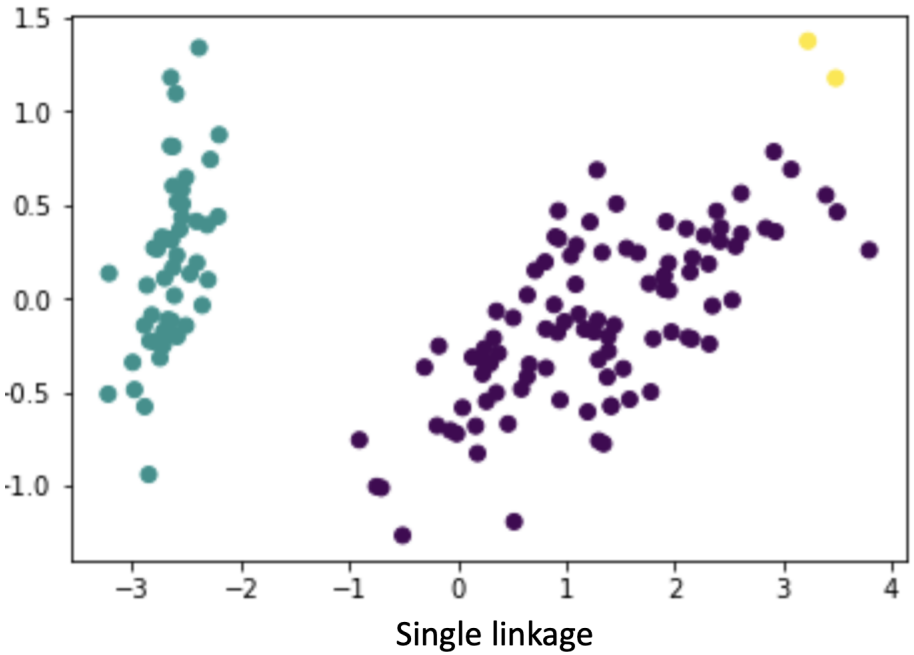
Distance between clusters

- In hierarchical clustering we need a **cluster distance measure**
- Different measures give **different results**
  - minimum distance (single **linkage**)
  - maximum distance (complete linkage)
  - average distance
  - Ward's distance

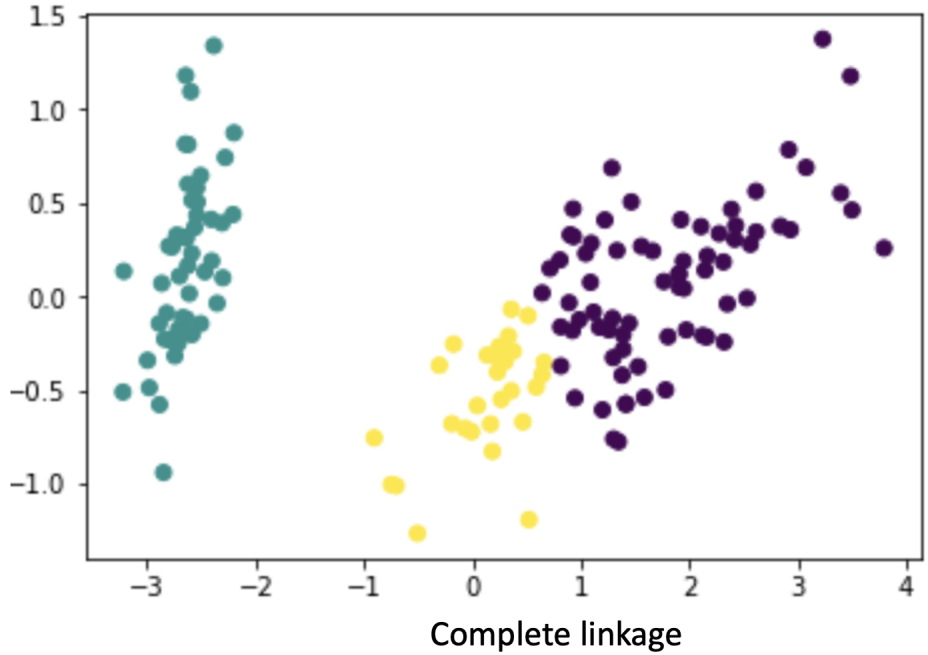


Distance between clusters

- Different distances correspond to different algorithms
  - and different strategies



Single linkage



Complete linkage

Ward's distance

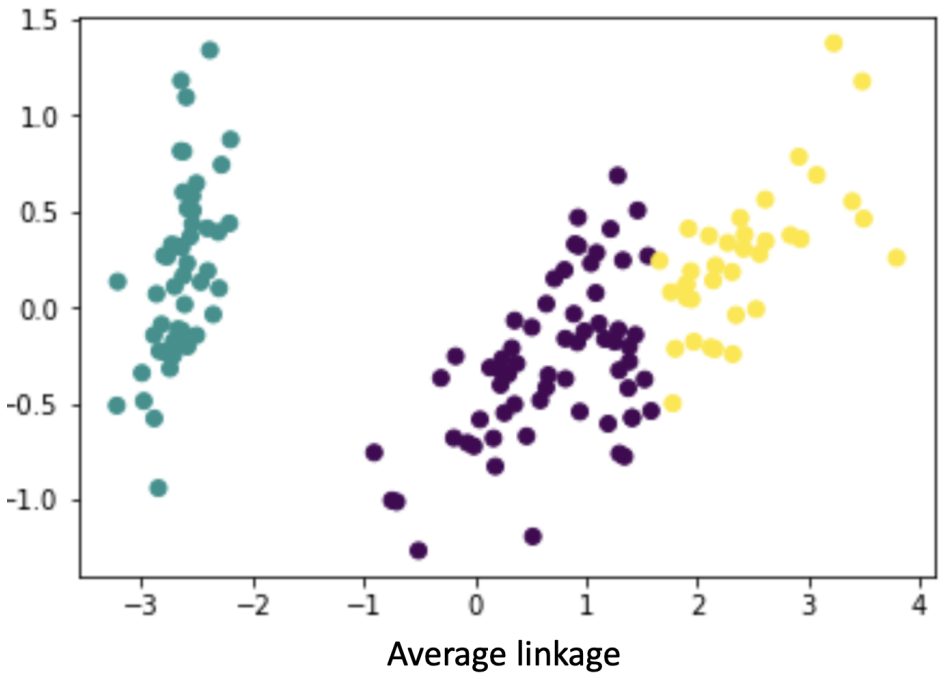
- Ward's distance measures the **increase** in within-cluster divergence
  - minimizes the square distances in clusters
- Suppose we have clusters A, B and C and $m_{cluster}$ is the center of $m_{cluster}$

$$d(A,B) = \sum_{i \in A \cup B} \|x_i - m_{A \cup B}\|^2 - \sum_{i \in A} \|x_i - m_A\|^2 - \sum_{i \in B} \|x_i - m_B\|^2$$

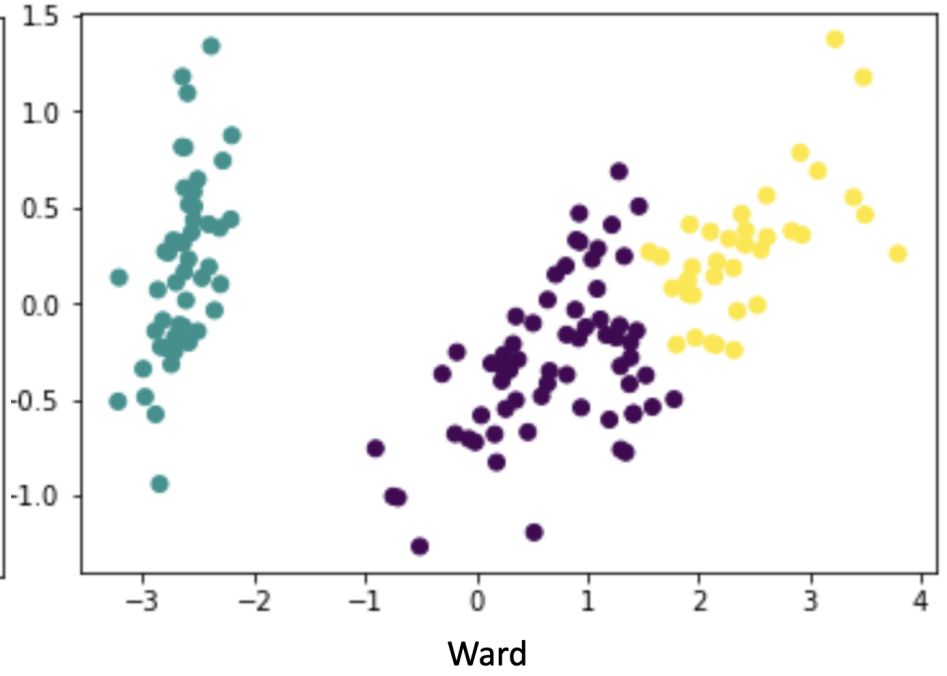
- Computationally
  - Ward's distance is computed recursively (end efficiently)

Average linkage and Ward's distance

- Average is a good comprmise between single and complete
- Ward's distance is the standard "default" choice



Average linkage



Ward

Other clustering approaches

- Density based
  - find points with dense neibourhoods
  - method DBSCAN
- Artificial Neural Networks
  - Self organizing maps (Kohonen Nets)

References

- Books
  - Han, Kamber & Pei, Data Mining Concepts and Techniques, Morgan Kaufman.
- Scikit docs
  - https://scikit-learn.org/stable/modules/generated/sklearn.cluster.AgglomerativeClustering.html