Hierarchical Clustering

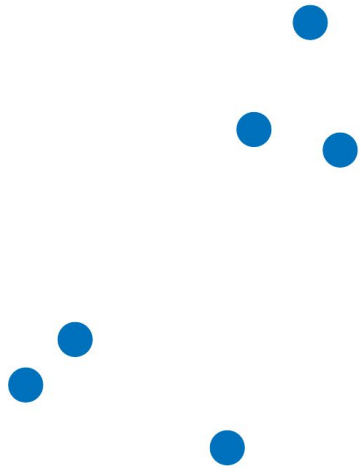
Hierarchical clustering

Building a hierarchy of clusters sequentially.

- Agglomerative (bottom-up): Start considering each point as a cluster then merge the closest ones and repeat
- Divisive (top-down): Start with one single cluster and divide to have groups with reduced variance

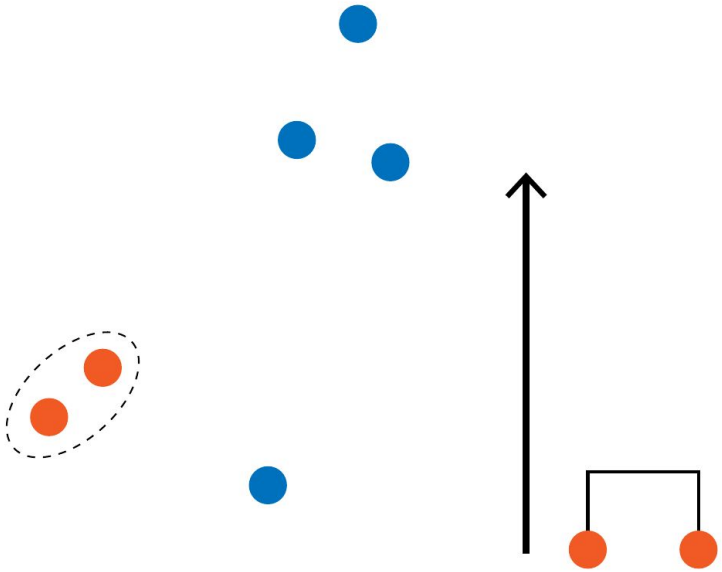
Let’s take a look at agglomerative hierarchical clustering (a.k.a. linkage)

Linkage algorithm



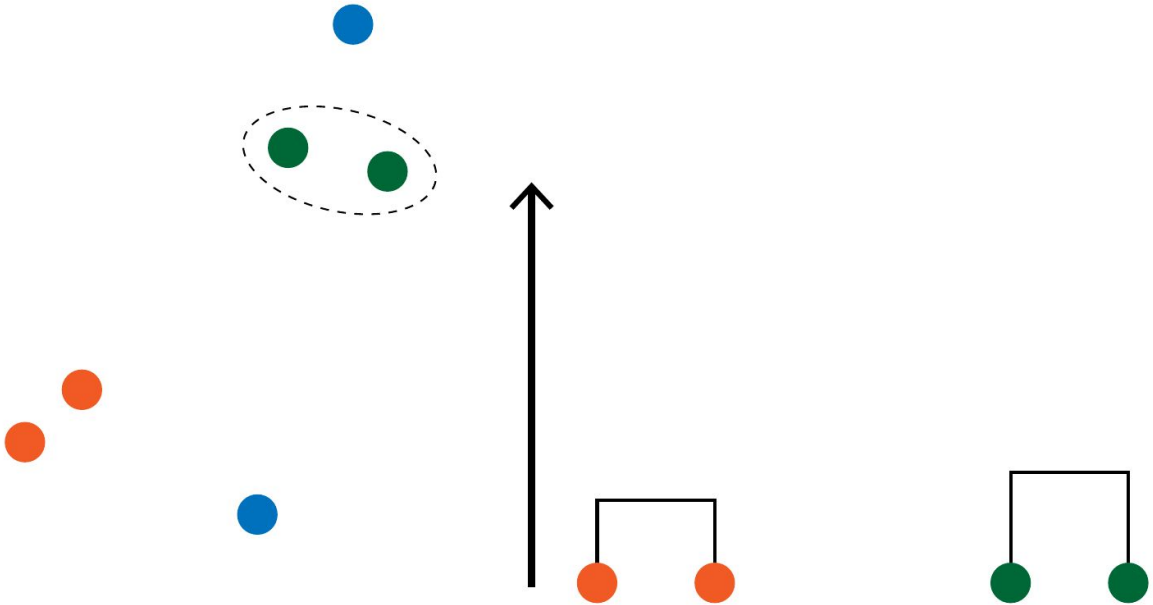
CAMBRIDGE SPARK

Linkage algorithm



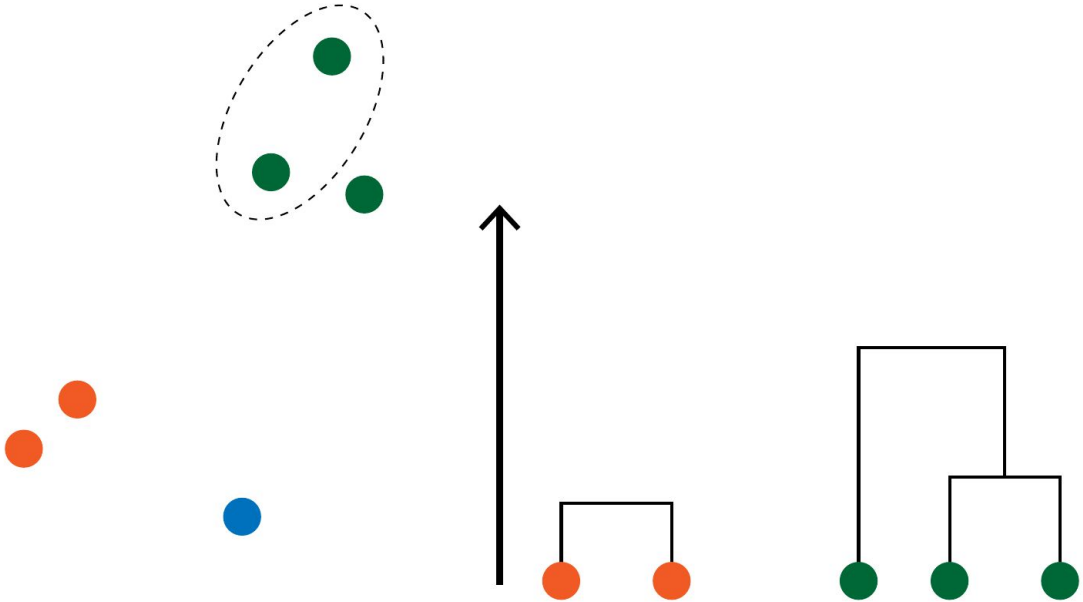
CAMBRIDGE SPARK

Linkage algorithm



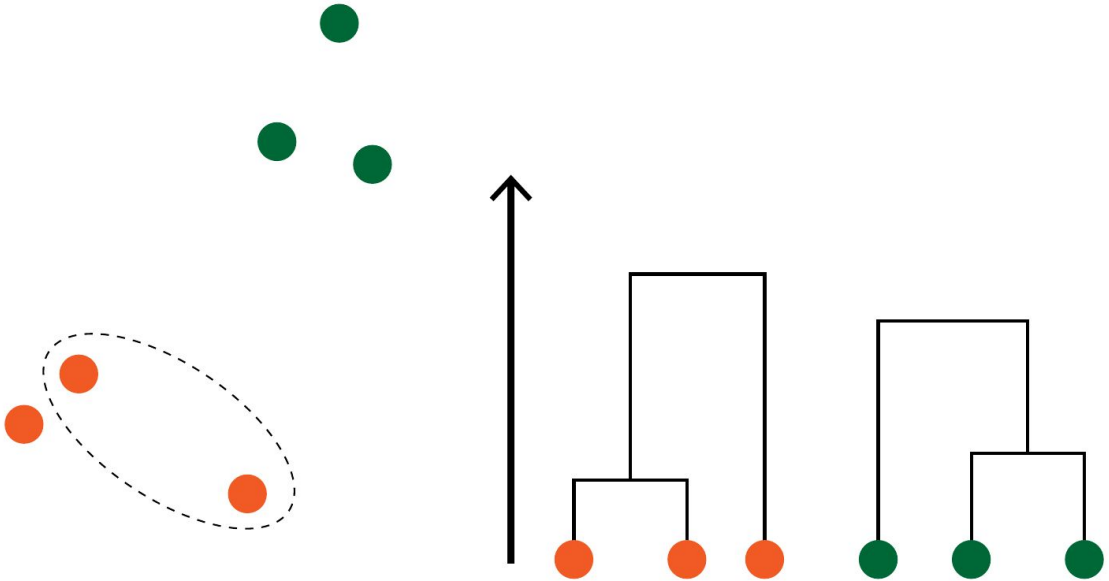
CAMBRIDGE SPARK

Linkage algorithm



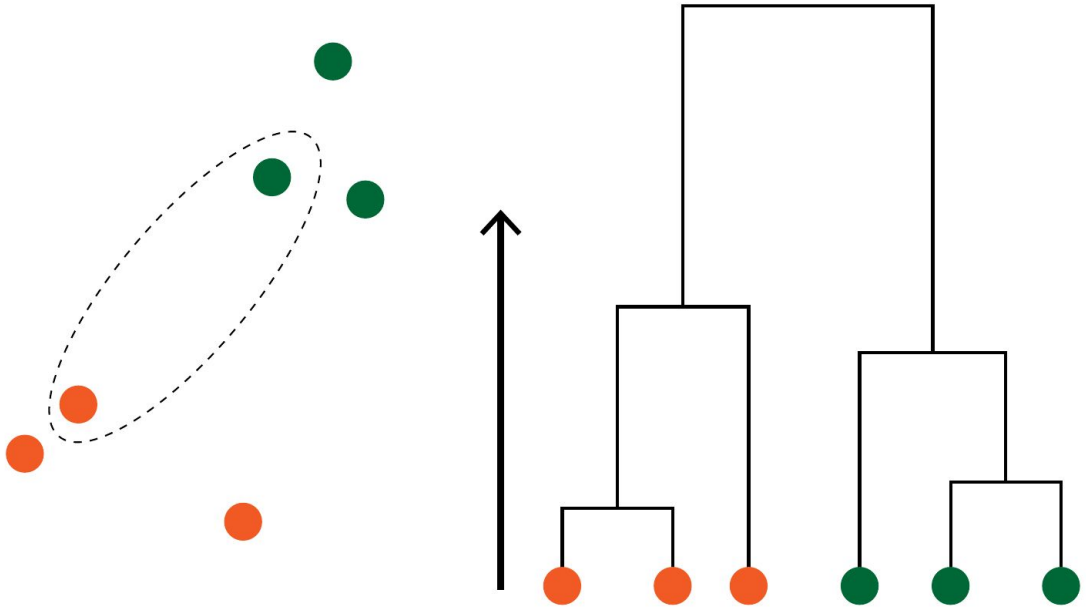
CAMBRIDGE SPARK

Linkage algorithm



CAMBRIDGE SPARK

Linkage algorithm

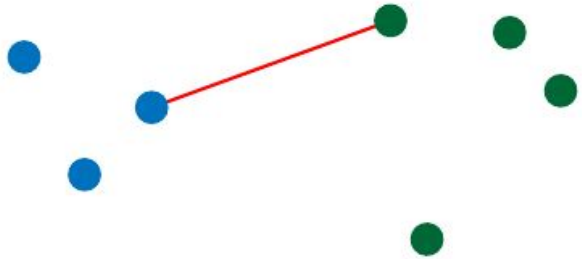


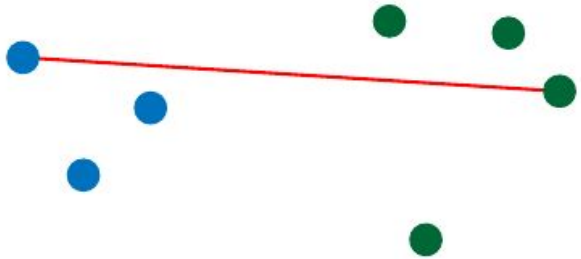
CAMBRIDGE SPARK

Linkage algorithm

Two strategies to merge clusters:

- Single linkage: closest point distance
- Complete linkage: furthest point distance





CAMBRIDGE SPARK

Linkage: pros and cons

Pros

- Clusters are not necessarily globular
- No dependance upon initialisation
- Dendogram shows a good summary

Cons

- Slower than K-means
- Still need to pick a number of clusters
- Assigning a new point is not straightforward
- Sensitive to noise

Hands-on session

hierarchical_clustering.ipynb