Clustering

scikit
learn

Examples in Economics

- Marko Terviö (2011): "Divisions within Academia: Evidence from Faculty Hiring and Placement," The Review of Economics and Statistics 93(3), 1053-1062.

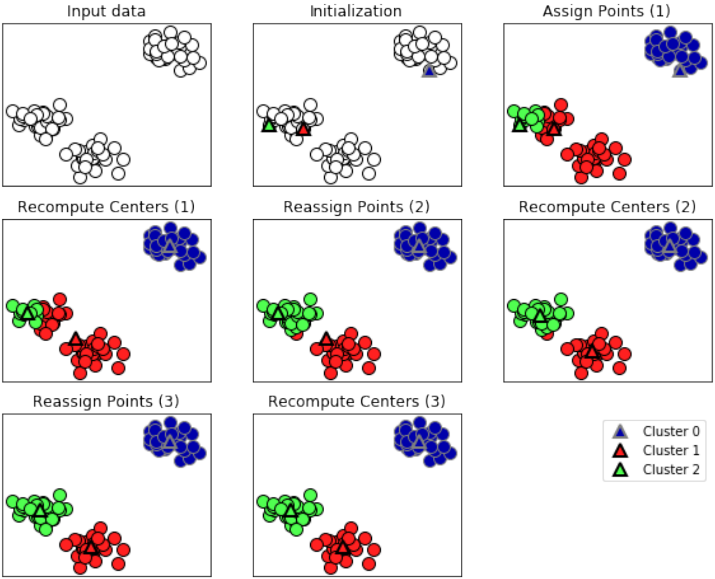
k-Means Clustering

- Partitional Clustering
- Find Cluster Centers representative of regions of data
- Algorithm
  1. Initialize k points as cluster means randomly
  2. Assign each point to one cluster center
  3. Reset cluster center as mean of points assigned to it
  4. Repeat 2 and 3 until convergence
- 1 parameter
  1. How many clusters?

+ Fast and transparent

- Performs badly for non-simple shapes (e.g. where clusters don't have same diameter)

k-Means Clustering, cont.



from: Andreas Müller and Sarah Guido (2016): Introduction to Machine Learning with Python, O'Reilly

Optimal k: Elbow plot

- Find clustering step where the acceleration of distance growth is the biggest
- Looks like an elbow when plotting SSE for increasing k
- For partitional clusterings only

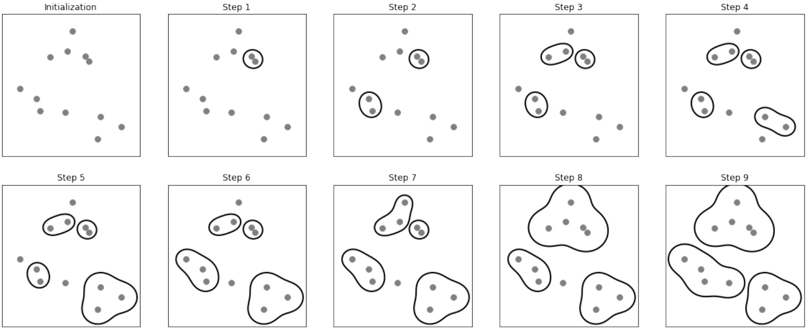
Agglomerative Clustering

- Hierarchical Clustering
- Algorithm
  - Make each point its own cluster
  - Iteratively merge two closest clusters
  - Stop when k clusters are left
- 2 Parameters
  1. Which number of clusters?
  2. Which clustering method (`ward`, `average` or `complete`)?

+ Good for hierarchical data

- No prediction, performs badly for non-simple shapes

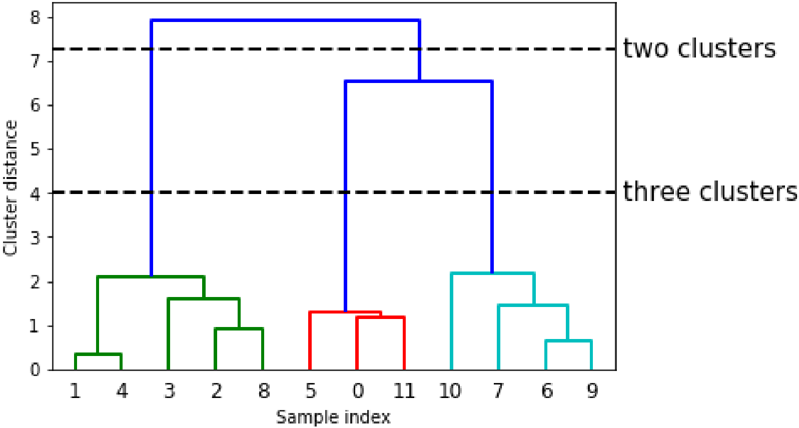
Agglomerative Clustering, cont.



from: Andreas Müller and Sarah Guido (2016): Introduction to Machine Learning with Python, O'Reilly

Optimal k: Dendrogram

- Visualizes a linkage array, depicting distances between clusters



from: Andreas Müller and Sarah Guido (2016): Introduction to Machine Learning with Python, O'Reilly

DBSCAN

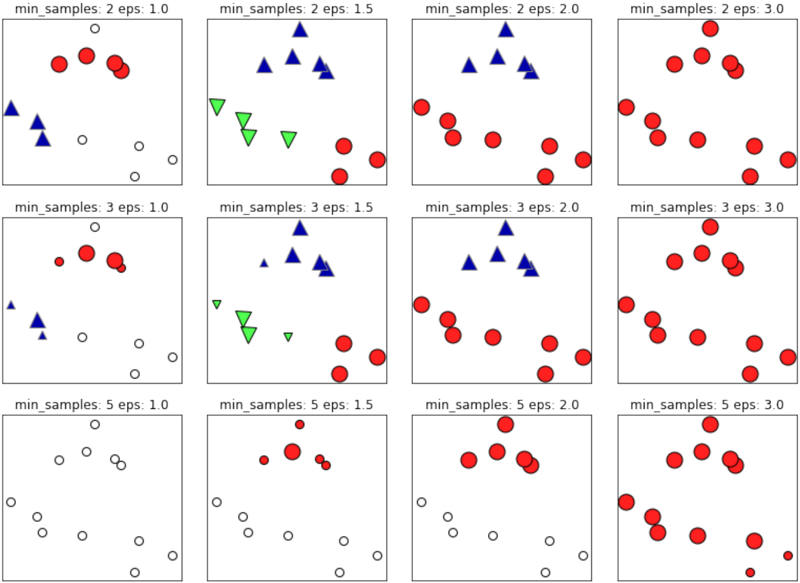
- Algorithm
  1. Pick an arbitrary observation
  2. If parametric conditions are met, point and neighbors become core cluster, otherwise noise
  3. Repeat for neighbors
  4. Repeat until all observations have been visited

  2 Parameters
  1. How many observations in a cluster at least?
  2. How close at least?

+ No a priori number of clusters needed, captures complex shapes

- Slower than the others

DBSCAN, cont.



from: Andreas Müller and Sarah Guido (2016): Introduction to Machine Learning with Python, O'Reilly