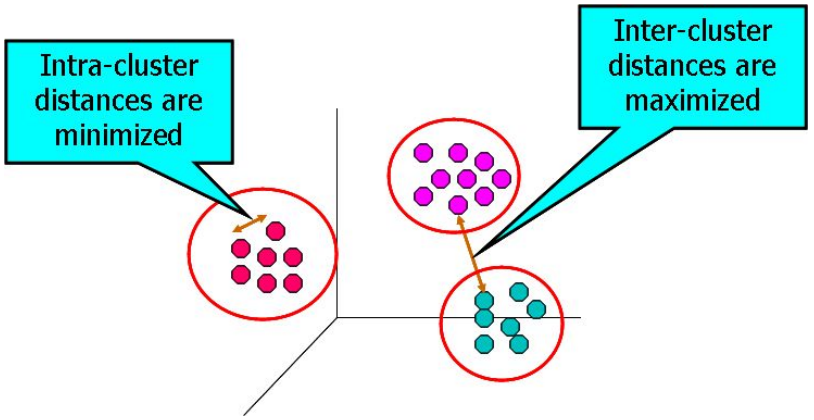
Clustering problem

Python for AI

What is Cluster Analysis?

- Cluster: a collection of data objects
  - Similar to one another within the same cluster
  - Dissimilar to the objects in other clusters
- Cluster analysis
  - Grouping a set of data objects into clusters
- Clustering is unsupervised classification: no predefined classes
- Typical applications
  - to get insight into data
  - as a preprocessing step



Examples of Clustering Applications

- Marketing: Help marketers discover distinct groups in their customer bases, and then use this knowledge to develop targeted marketing programs
- Land use: Identification of areas of similar land use in an earth observation database
- Insurance: Identifying groups of motor insurance policy holders with a high average claim cost
- City-planning: Identifying groups of houses according to their house type, value, and geographical location
- Earth-quake studies: Observed earth quake epicenters should be clustered along continent faults

Notion of a Cluster can be Ambiguous

How many clusters?

Six Clusters

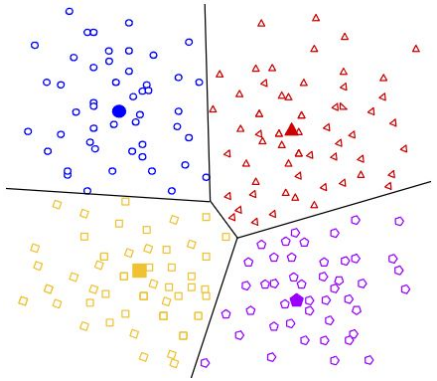
Two Clusters

Four Clusters

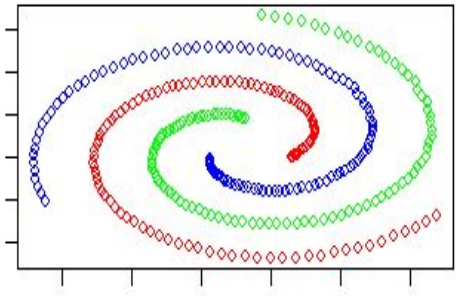
What Is Good Clustering?

- A good clustering method will produce high quality clusters with
  - high intra-class similarity
  - low inter-class similarity
- The quality of a clustering result depends on both the similarity measure used by the method and its implementation.
- The quality of a clustering method is also measured by its ability to discover some or all of the hidden patterns.

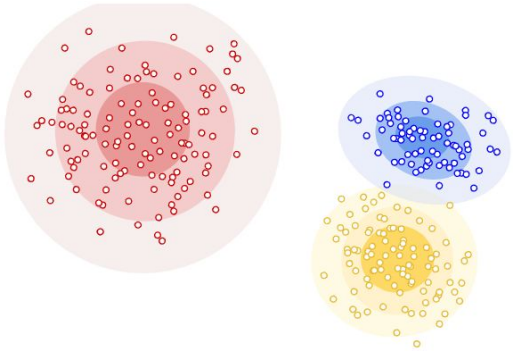
Clustering algorithms



Center-based clustering

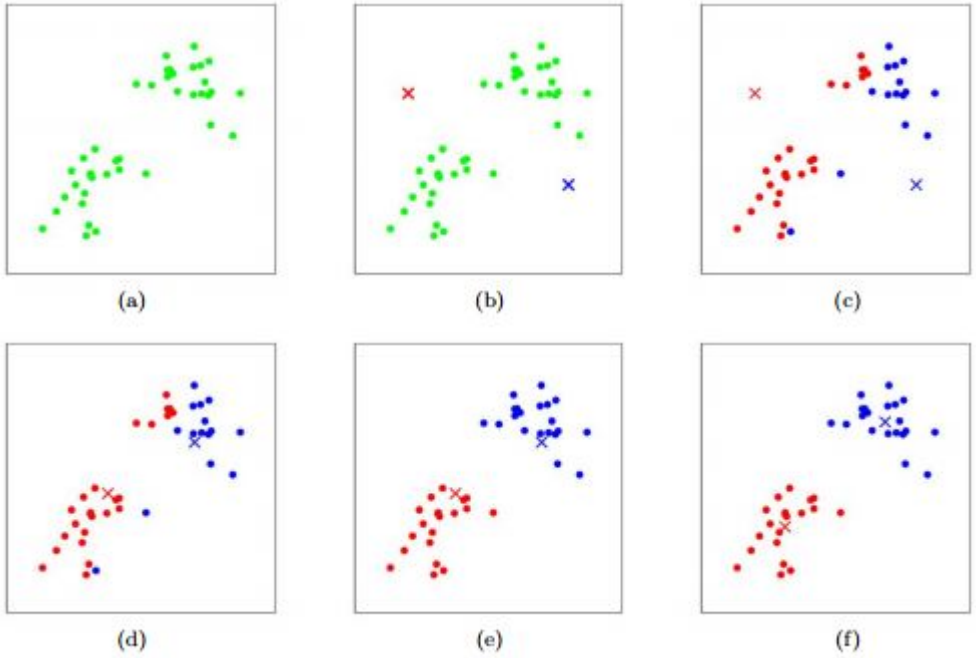


Density-based clustering

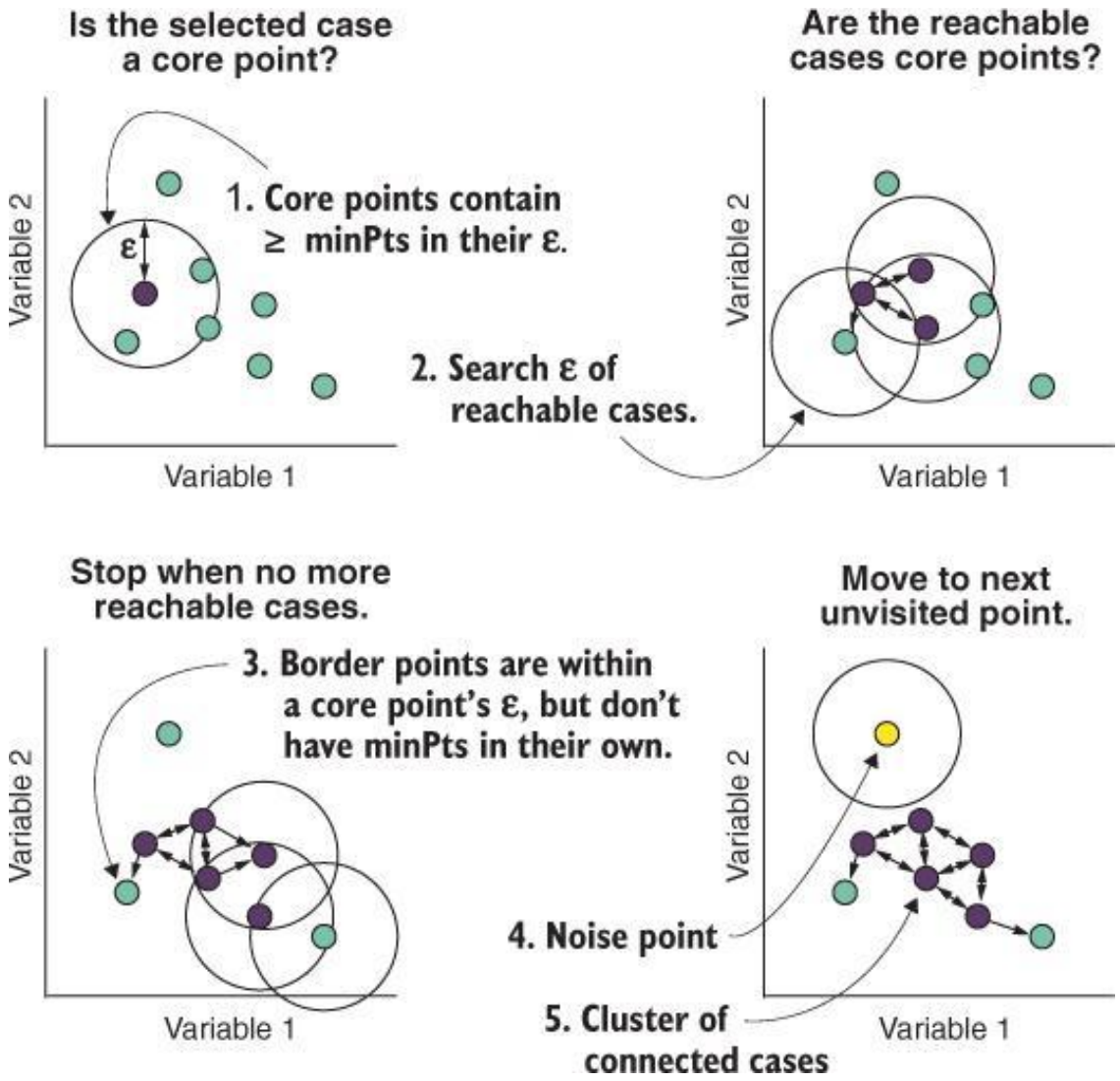


Probability distribution-based clustering

K-means

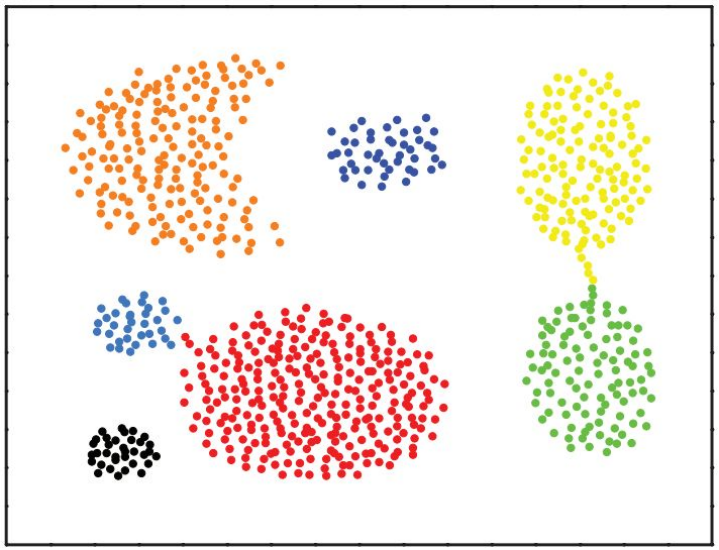


DBSCAN


Is the selected case a core point?
1. Core points contain ≥ minPts in their ε.
ε
Variable 2
Variable 1
2. Search ε of reachable cases.
Are the reachable cases core points?
Variable 2
Variable 1
Stop when no more reachable cases.
3. Border points are within a core point's ε, but don't have minPts in their own.
Variable 2
Variable 1
Move to next unvisited point.
4. Noise point
5. Cluster of connected cases
Variable 2
Variable 1

Density peak clustering

- Original paper: Clustering by fast search and find of density peaks, Science, 2014
- Google scholar citations: 2528
- A clustering algorithm based on the idea that cluster centers are characterized by a higher density than their neighbors and by a relatively large distance from points with higher densities



Example and code

- Download code in the classroom
- On class: follow a step by step tutorial