Karlsruhe Institute of Technology

Clustering Algorithms

Doctoral Research Meeting 2019

Benjamin Ertl

KARLSRUHE INSTITUTE OF TECHNOLOGY (KIT)

Introduction

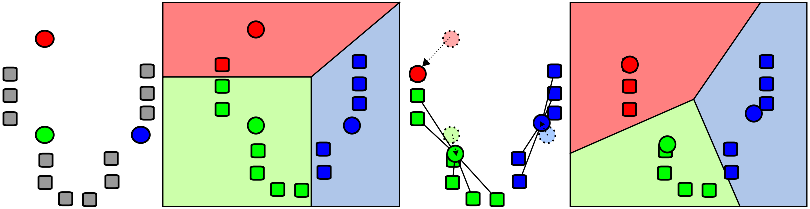
Clustering

Given a set of points in multidimensional space, find a partition of the points into *clusters* so that the points within each cluster are similar to one another.[AY00]

- Early history
  - 1932, Driver and Kroeber, Anthropology
  - 1938, Joseph Zubin, Psychology
  - 1939, Robert Tryon, Psychology
  - 1943, Raymond Cattell, Psychology
- More recent history (very brief)
  - 1957/65, Stuart Lloyd & Edward W. Forgy, k-means
  - 1975, Fukunaga and Hostetler, mean shift
  - 1996, Ester et al., DBSCAN

k-means

Link



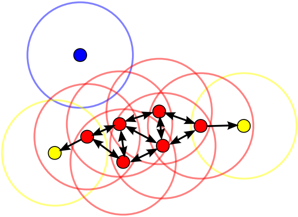
1. k initial "means" (in this case k=3) are randomly generated within the data domain (shown in color).
2. k clusters are created by associating every observation with the nearest mean.
3. The centroid of each of the k clusters becomes the new mean.
4. Steps 2 and 3 are repeated until convergence has been reached. [1]

[1] https://en.wikipedia.org/wiki/K-means_clustering

k-means family

DBSCAN

Link



1. Find the points in the ϵ (eps) neighborhood of every point, and identify the core points with more than minPts neighbors.
2. Find the connected components of core points on the neighbor graph, ignoring all non-core points.
3. Assign each non-core point to a nearby cluster if the cluster is an ϵ (eps) neighbor, otherwise assign it to noise. [2]

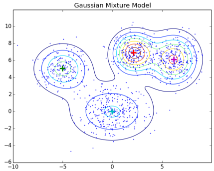
[2] `https://en.wikipedia.org/wiki/DBSCAN`

Density-based clustering

Karlsruhe Institute of Technology

Expectation–Maximization

Link



1. Randomly initialize Gaussian distributions for a given number of clusters.
2. Compute the probability that each data point belongs to a particular cluster.
3. Based on the probabilities, compute new set of parameters for the Gaussian distributions such that the probabilities of data points within the clusters are maximized. [3]

[3] `https://tinyurl.com/ybmwltyv`

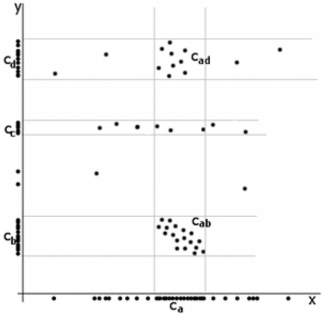
Hierarchical clustering family

High-dimensional data

Curse of dimensionality

In high dimensional space, the distance between every pair of points is almost the same for a wide variety of data distributions and distance functions. [BGRS99]

- **Solution** Feature selection
- **Problem** Correlations specific to locality
- **Solution** Subspace clustering[4]
- **Problem** Loss of information
- **Solution** Correlation clustering

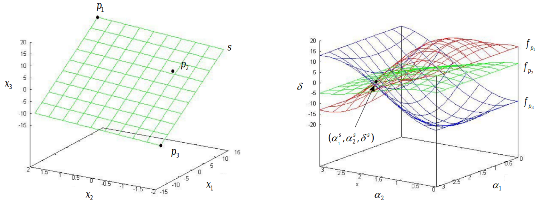


---

[4] `https://tinyurl.com/y3whbwhl`

Subspace clustering algorithms

CASH correlation clustering



- Robust clustering in arbitrarily oriented subspaces [ABD$^+$]
- Map points to functions in d-dimensional parameter space
  - Hough transform
  - Spherical coordinates
  - Parametrization function
- Find dense regions in parameter space

Correlation clustering algorithms

Conclusion

Clustering

Given a set of points in multidimensional space, find a partition of the points into *clusters* so that the points within each cluster are similar to one another.[AY00]

- Full-space clustering:
  - k-means, DBSCAN, EM
  - work best for low dimensional data
- Subspace/correlation clustering:
  - PROCLUS, ORCLUS, CASH
  - work also for high dimensional data
- Tools:
  - ELKI
  - scikit-learn

Literatur

Elke Achtert, Christian Böhm, Jörn David, Peer Kröger, and Arthur Zimek.
Robust Clustering in Arbitrarily Oriented Subspaces, pages 763–774.

Charu C. Aggarwal and Philip S. Yu.
Finding generalized projected clusters in high dimensional spaces.
SIGMOD Rec., 29(2):70–81, May 2000.

Kevin S. Beyer, Jonathan Goldstein, Raghu Ramakrishnan, and Uri Shaft.
When is ”nearest neighbor” meaningful?
In *ICDT*, 1999.