Lecture 16: Clustering

Machine Learning, Summer Term 2019

Michael Tangermann Frank Hutter Marius Lindauer

University of Freiburg

Lecture Overview

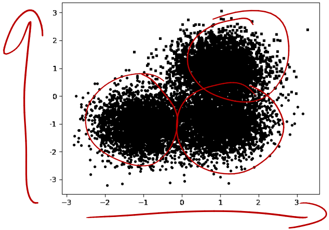
1. Motivation
2. Criteria for Clustering
3. K-Means
4. DBSCAN

What is Cluster Analysis?

Also called: clustering, segmentation analysis, taxonomy analysis, automatic classification, numerical taxonomy, botryology, typological analysis, community detection, ...

(Plot modified from scikit-learn clustering tutorial)

data points



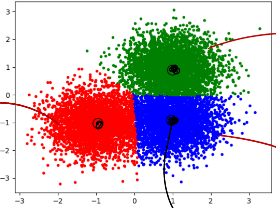
Unsupervised task:

Group objects such that objects within a group are more similiar/related to each other *(in some sense)* than to objects of another group.

What is Cluster Analysis?

Also called: clustering, segmentation analysis, taxonomy analysis, automatic classification, numerical taxonomy, botryology, typological analysis, community detection, ...

(Plot modified from scikit-learn clustering tutorial)

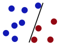
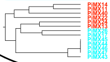


Unsupervised task:

Group objects such that objects within a group are more similiar/related to each other *(in some sense)* than to objects of another group.

ML Design Cycle

pre-processing → feature extraction / encoding → feature selection → **machine learning** → evaluation & model selection → post-processing

supervised learning	unsupervised learning	reinforcement learning
		

Cluster analysis is an unsupervised learning task

- Ground truth about clusters is not provided → evaluation is tricky!

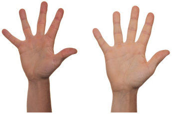
Given:

- N high dimensional data points $\mathbf{x}_i \in \mathbb{R}^D$ with $i = 1 \ldots N$.
- Data is collected in matrix $\mathbf{X} \in \mathbb{R}^{N \times D}$ no labels

Applications

recommender systems

preprocessing
cont → discrete



outlier detection

object detection, segmentation

Spike sorting

Example Applications (I)

- Medical imaging (fMRI, CT, PET): differentiate between different types of tissues, find tissue boundaries
- Biology: determine communities of organisms in space and time, compute data-driven phylogenetic trees
- Genetics: group DNA sequences into gene families
- Biochemistry / chemistry / pharmacology: group compounds according to their reaction mechanism
- Market research: detect clusters of customers with similar behavior, find market segments
- Social networks: recognize communities
- Search engines: Post-processing of search results into groups of hits that refer to vastly different topics

Example Applications (II)

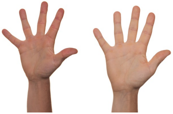
- Image segmentation: border detection, track objects
- Anomaly detection: identify outliers in data streams, network attacks, misbehaving software, sensor failures (robotics, production lines), predictive maintenance
- Finance: find stock clusters of similar behaviour
- Text analysis: clustering of documents into topics
- ...

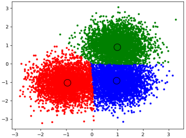
Lecture Overview

1. Motivation
2. Criteria for Clustering
3. K-Means
4. DBSCAN

How could Clusters be Determined?

Which metrics might be used to define clusters?



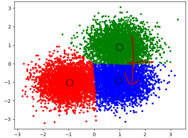


Mahalanobis

cluster means ? distance

How could Clusters be Determined?

Which metrics might be used to define clusters?



Zoo of clustering methods available, that exploit e.g.:

- distance/similarity function (between cluster members, between members of different clusters)
- connectivity structure using distances → single/avg/max linkage clustering, graph-based → clique
- centroid + neighborhood
- densities
- expected distributions

Lecture Overview

1 Motivation

2 Criteria for Clustering

3 K-Means

4 DBSCAN

K-Means Clustering (Steinhaus, 1957)

Find set $C = \{C_1, ..., C_k\}$ of k clusters represented by cluster centroids μ_k such, that the clusters have equal variance.

$\rightarrow$ Minimize the *inertia* or *within-cluster sum-of-squares* criterion:

$$\underset{C}{\operatorname{argmin}} \sum_{i=1}^{k} \sum_{\mathbf{x} \in C_i} ||\mathbf{x} - \mu_j||^2$$

Observations:

- Cluster centroids μ_j do not need to be points of the training data sets
- Unfortunately NP-hard problem!
- Clustering can be represented by Voronoi tesselation

K-Means Clustering

Practical solution by heuristic approximation
(e.g. Lloyd's algorithm (1957, 1982), similar to expectation-maximization):

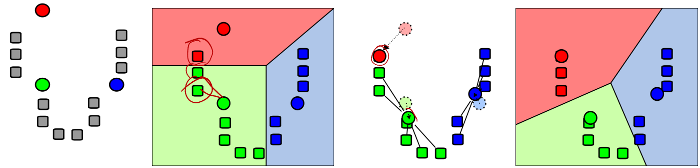
Initialize k data points as cluster centroids.
Then iterate these two steps until convergence of the centroids:

~~1~~ 2 Assign each data point to its nearest centroid
(→ approximations necessary for high dimensions!)

~~2~~ 3 Create k new centroids by taking the mean value of all of the data points assigned to each novel centroid
(→ approximations necessary for many data points)

1 2 3

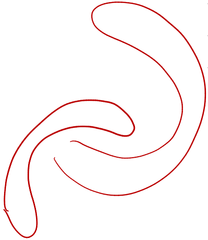
k=3

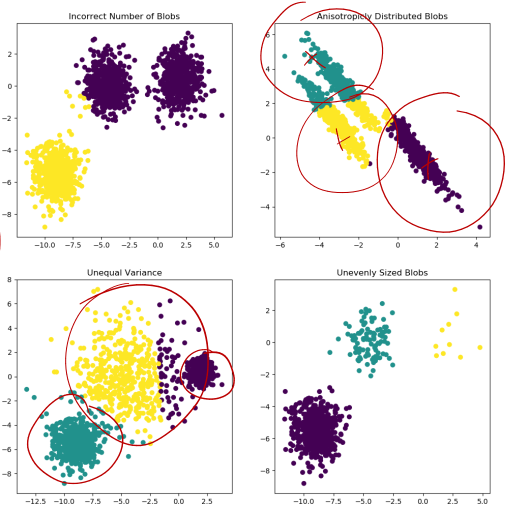


K-Means Clustering

Problematic data sets for k-means:

k=2





K-Means Clustering

Pros:

- conceptually simple algorithm
- mini-batches and different kind of initialization strategies are available (→ k-means++)
- scales to many data points (if approximations are utilized)

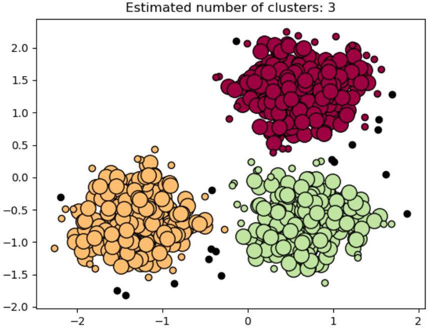
Cons:

- sensitive to initialization of centroids
- can not model noise or outliers
- concave cluster shapes are problematic
- can not deal with uneven variance between clusters
- number k of clusters needs to be provided

Lecture Overview

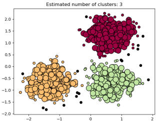
1. Motivation
2. Criteria for Clustering
3. K-Means
4. DBSCAN

DBSCAN (Ester et al., 1996)



Key idea: DBSCAN assumes that clusters are areas of high density, which are separated by areas of lower density.

DBSCAN (Ester et al., 1996)



Key idea: DBSCAN assumes that clusters are areas of high density, which are separated by areas of lower density.

A cluster is formed by two types of data points:

- "core samples" are data points in areas of high density (defined by at least `min samples` within an `eps`-neighborhood)
- "non-core samples" are data points which are close to a core sample but that are not core samples themselves (e.g. at the fringes of a cluster)

Samples which have a distance of more than eps to a core sample are considered outliers.

DBSCAN

DBSCAN creates clusters by sequentially considering the training data points.

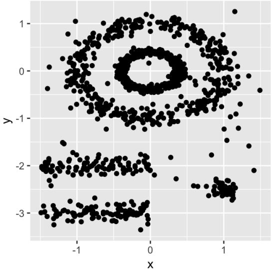
Pros:

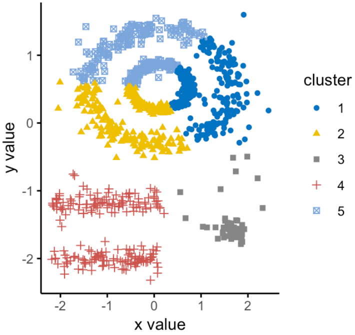
- DBSCAN is fast and deterministic for a fixed sequence of data points.
- Number of clusters is determined automatically.
- Hyperparameter `min samples` can express prior knowledge about noise.
- Different distance metrics can be utilized
- Hierarchical variant HDBSCAN available

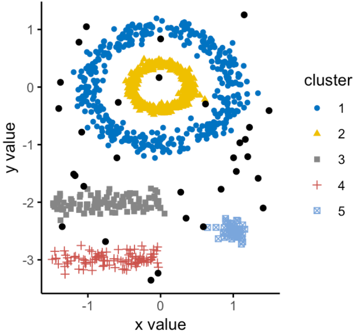
Cons:

- Varying the sequence of data points processed can lead to different clusterings.
- Hyperparameter `eps` is critical, no good default!

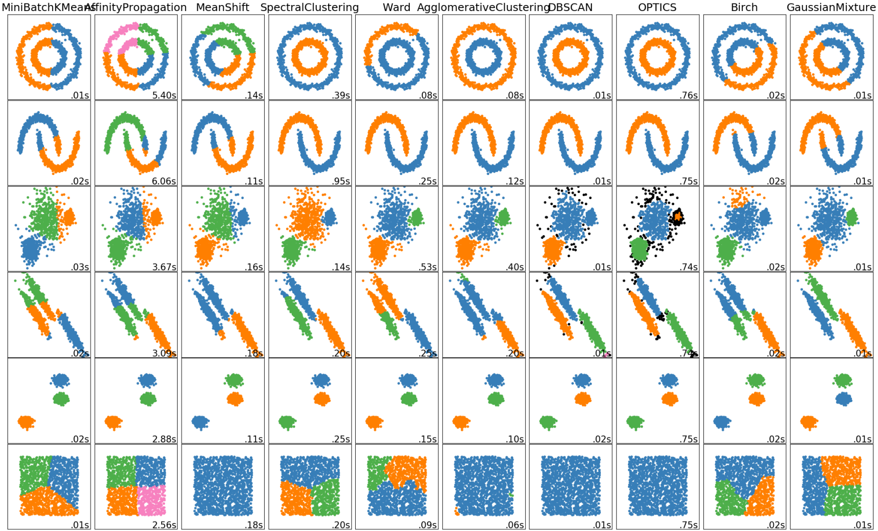
Comparison of K-Means and DBSCAN





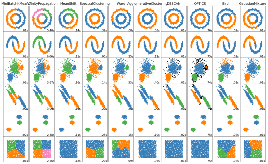


A few Clustering Algorithms on Toy Data



https://scikit-learn.org/stable/modules/clustering.html

Criteria for the Choice of Clustering Algorithms



Does the algorithm...

- expects each cluster to follow a specific distribution? (e.g. Gaussian)?
- considers density of data points?
- deal well with noisy data / high-dimensional data / redundant dimensions / irrelevant dimensions?
- deliver hard / soft clustering?
- deliver a strict partitioning (i.e. each object belongs to exactly one cluster)?
- deliver a hierarchical clustering?

Wrap-Up: Summary by Learning Goals

Having heard this lecture and doing the assigment on clustering, you will be able to:

- Explain, which metrics can be used to create a clustering from unlabeled data
- formulate the optimization problem for k-means clustering and implement an iterative heuristic
- Describe pros and cons of k-means and DBSCAN
- Derive a metric for the quality of a given clustering (e.g. via the "**silhouette score**", see assignment)