MADS-MMS – Mathematics and Multivariate Statistics

Clustering – Overview

Prof. Dr. Stephan Doerfel

FACHHOCHSCHULE KIEL
University of Applied Sciences

SCAN ME

Moodle (SoSe 2025)

Agenda

Motivation

Goals of Clustering

What is Clustering?

Clustering Methods

Ingredients

Outline

Motivation

Goals of Clustering

What is Clustering?

Clustering Methods

Ingredients

Chapter Goals

- overview on the topic of clustering
- categorization of methodology
- understand motivation and application of clustering

Outline

Motivation

Goals of Clustering

What is Clustering?

Clustering Methods

Ingredients

Goals of Clustering

Goals:

- identify clusters (categories / subsets / groups) in datasets
- instances in the same cluster should be as similar as possible
- instances in different clusters should have low similarity
- (identify instances that belong to no group: outliers, noise)
- NOT: assign instances to known classes

Context:

- detect sets of “comparable/similar/close elements”
- explore and analyze unknown data
- engineer classes / features
- semi-automatic – often data scientist has to “judge” and interprete clusterings
- requires a useful and meaningful distance/similarity function (often individually chosen or designed)

Applications for Clustering

Applications:

- Market segmentation / customer base segmentation
- Pattern recognition
- Community discovery in social networks
- Topic detection in text corpora
- Tracking of evolutional steps
- Identifying common behavior or common interests (e.g. for pdf recommender systems)
- Identifying common physical properties in sensor data
- ...

Outline

Motivation

Goals of Clustering

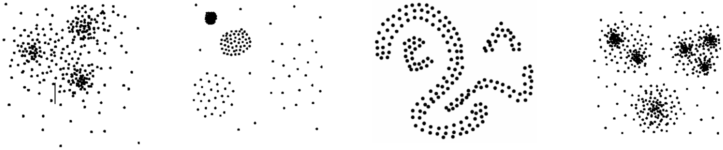
What is Clustering?

Clustering Methods

Ingredients

Examples

Clusters of different size, form, density, and hierarchical structure


[1]

[1] Source: [1], Abb. 3-1

Clustering Formally

There is no hard mathematical definition of clustering in general.

Definition 1 (Clustering)

Clustering comprises (machine learning) methods of **unsupervised learning** to collect data instances into **groups, categories, or classes**, called **clusters**. The set of all clusters is called a **clustering**.

Criteria for the grouping can be

intra-class similarity: similarity within a cluster

inter-class dissimilarity: dissimilarity between different clusters

Machine Learning Disciplines

Supervised

- labelled data
- goal: class/prediction of unknown/future data
- idea: Learn by deriving a model from looking at examples
- correctness of the training can be assessed (supervised)
- examples: classification, regression

Unsupervised

- unlabelled data
- goal: Detect patterns (groups, structure) in the data
- learning is unsupervised, no "correct" result that we can compare to
- examples: **clustering**, dimensionality reduction

Clustering Process

Definition 2 (Clustering Process)

A clustering process comprises the following steps:

- representation of the data
- definition of a similarity measure (domain-specific)
- creating the clusters
- optionally abstraction of knowledge
- optionally evaluation of the output

Outline

Motivation

Goals of Clustering

What is Clustering?

Clustering Methods

Ingredients

Classification of Clustering Methods

Partitioning Methods (Hard Clustering)

- determines a partition into disjoint subsets, minimizing a cost function
- typical parameters: number of clusters k, distance function

Density-based Methods

- adds neighbors to clusters, as long as density does not fall below threshold
- distinguishes between the cores of clusters, its borders, and noise
- parameters: minimal acceptable density in a cluster, distance function

Hierarchical Methods

- determine a hierarchy of clusters, fuses most similar clusters
- parameters: distance function

Other Methods (incomplete)

- Soft Clustering (Fuzzy Clustering, Overlapping Clustering)
- Graph-based Methods

Outline

Motivation

Goals of Clustering

What is Clustering?

Clustering Methods

Ingredients

Ingredients

We need:

- a way to express different kinds of data mathematically, and
- a way to measure distance/similarity between points
- all that in many-dimensional realms
- 2D-visualizations despite multidimensional data
- algorithms that cluster
- basic mathematics like logarithms, vector geometry, matrices

📎 Exercises 1

References

M. Ester and J. Sander.
Knowledge Discovery in Databases.
Springer-Verlag, Berlin/Heidelberg, 2000.