MachineLearnAthon - Microlecture Clustering

Recorded by Lara Kuhlmann

Recap previous microlectures – Introduction to ML

This is not a complete overview of methods!

Supervised Learning

Input: Historical input data with a known result or output (label). Output (label) can have the form of a numerical value or category.
Goal: Predict the output of unseen or future data.
Approach: Learn a function (model) that describes the relationship (or mapping) between the historic input and the output variables and apply this function to the unseen or future data.

Classification

- Logistic regression
- Decision trees and random forests

Regression

- Linear regression
- Decision trees and random forests

Unsupervised Learning

Input: Historical input data without a known result.
Goal: Build meaningful subgroups (clusters), find associations or compress data (reduce dimensionality).
Approach: Find groups of objects sharing similarities, find hidden structures or associations, and remove noise from the data.

Clustering

- k-means
- Hierarchical clustering

Dimensionality reduction

- Principal component analysis (PCA)
- Nonlinear PCA

Association rule learning

- Aprior
- Eclat

Reinforcement Learning

Input: Environment which determines the state and the corresponding reward (e.g. win or lose) after an applied action.
Goal: (Sequentially) select good actions given the current state of the environment.
Approach: Agent learns a policy or a value function through interacting with the environment by "trying" actions (play); the policy or value functions maximize the expected rewards of actions depending on the current state.

Value based

- Deep Q Network

Policy based

- Policy Gradient

technische universität dortmund
TU Delft
Co-funded by the Erasmus+ programme of the European Union
Machine LearnAthon

Recap previous microlecture - Data Preparation

Tabular data may contain different data types. The most common data types are:

- Numerical data
  - Integer: Whole numbers (e.g. number of products)
  - Float: Decimal numbers (e.g. stock prices)
- Categorical data
  - Nominal: Categories without any order (e.g. countries)
  - Ordinal: Categories with an order (e.g. rating scales)
- Boolean data: Binary values (True/ False or 0/1)
- Text (string) data: Free-form text (e.g. names, addresses)
- Datetime data

Learning outcomes of today

After successfully completing this micro-lecture, you are able to….

- Define the term clustering and identify clustering use cases
- Explain the differences between the distance measures Euclidean distance, Manhattan distance, Hemming distance and Jaccard distance
- Analyze the similarity of data points based on the distance measures

Agenda for today

- Introduction to Clustering
- Hard vs. soft clustering
- Clustering applications
  - Clustering in Customer Analysis
  - Clustering in Medical Diagnostics
  - Clustering in Image Processing
  - Clustering in Text Analysis
- Distance Measures
  - Distance measures for numerical values
  - Distance measures for non-numerical values

Clustering

- Clustering literally means grouping
- It is an unsupervised learning process that helps grouping data points
- “Clustering analysis aims to group individuals into a number of classes or clusters using some measure such that the individuals within classes or clusters are similar in some characteristics, and the individuals in different classes or clusters are quite distinct in some features”[1]
- Use Cases include customer analysis, image segmentation, behavior pattern analysis,…

[1] Tang, N., & Wu, Y. (2022). Introductory Chapter: Development of Data Clustering. In *Data Clustering*. IntechOpen.
Xu, R., & Wunsch, D. (2008). *Clustering*. John Wiley & Sons.

technische universität dortmund
TU Delft
Co-funded by the Erasmus+ programme of the European Union
Machine LearnAthon

Hard vs. soft clustering

- Hard clustering (e.g., k-Means)
  - Each data point belongs to exactly one cluster.
  - There are clear boundaries between clusters.
  - Suitable when groups are well-separated.

- Soft clustering (e.g., Gaussian Mixture Models)
  - A data point can belong to multiple clusters with certain probabilities.
  - Allows overlapping and fuzzy assignments.
  - Useful when data distributions are complex.

Xu, R., & Wunsch, D. (2008). Clustering. John Wiley & Sons.
Tang, N., & Wu, Y. (2022). In Data Clustering. IntechOpen.

Co-funded by the Erasmus+ programme of the European Union

Clustering applications

- Clustering is widely used across industries. Common applications include:
  - Customer segmentation
  - Image processing and computer vision
  - Text and document analysis
  - Behavior pattern detection
  - Medical diagnosis
  - Anomaly detection (e.g. fraud)
  - Recommender systems

Xu, R., & Wunsch, D. (2008). *Clustering*. John Wiley & Sons..
Aggarwal, C. C., & Reddy, C. K. (2014). Data Clustering: Algorithms and Applications. Chapman & Hall/CRC.

technische universität dortmund
TU Delft
Co-funded by the Erasmus+ programme of the European Union
Machine LearnAthon

Clustering in Customer Analysis

- Segment customers based on behavior, demographics, or preferences
- Enables targeted marketing and personalized offers
- Example: Group customers into “Bargain Seekers”, “Brand Loyalists”, “Occasional Shoppers”
- Algorithms: k-Means, DBSCAN, Hierarchical Clustering

Aggarwal, C. C., & Reddy, C. K. (2014). Data Clustering: Algorithms and Applications. Chapman & Hall/CRC.

Co-funded by the Erasmus+ programme of the European Union

Clustering in Medical Diagnostics

- Group patients based on symptoms, gene expression, or imaging data
- Goals: Diagnosis support, treatment personalization
- Example: Identify cancer subtypes from gene profiles
- Algorithms: Hierarchical Clustering, k-Means, Gaussian Mixture Models
- Data: high-dimensional genetic or image data

Xu, R., & Wunsch, D. (2008). *Clustering*. John Wiley & Sons..

technische universität dortmund
TU Delft
Co-funded by the Erasmus+ programme of the European Union
Machine LearnAthon

Clustering in Image Processing

- Segment pixels into regions (e.g., object vs. background)
- Applications: face recognition, object tracking
- Example: Extract foreground object from an image
- Algorithms: k-Means (color space), DBSCAN
- Data: RGB values, spatial features

Igual, L., & Seguí, S. (2024). *Introduction to Data Science*. Springer.

Co-funded by the Erasmus+ programme of the European Union

Clustering in Text Analysis

- Group similar documents (e.g., news, tweets, reviews)
- Useful for topic detection without labels
- Example: Cluster customer reviews by topic (e.g., delivery, product quality)
- Algorithms: k-Means, Agglomerative Clustering
- Data: TF-IDF vectors, word embeddings (e.g., BERT)

Xu, R., & Wunsch, D. (2008). *Clustering*. John Wiley & Sons..

technische universität dortmund
TU Delft
Co-funded by the Erasmus+ programme of the European Union
Machine LearnAthon

Distance measures

- Distance measures quantify the similarity or dissimilarity between two data points
- They are crucial to define relationships between points
- Distance measures depend on the data type

Igual, L., & Seguí, S. (2024). Introduction to data science. In *Introduction to Data Science: A Python Approach to Concepts, Techniques and Applications* (pp. 1-4). Cham: Springer International Publishing.

technische universität dortmund
TU Delft
Co-funded by the Erasmus+ programme of the European Union
Machine LearnAthon

Distance measures for numerical values

- To measure the distance between numerical values, there are two popular measures, the Euclidean and the Manhattan distance
- The Euclidean distance is best for continuous data with normally distributed features
- The Manhattan distance is suitable for high-dimensional spaces or where the axes are of primary importance

For two points $p = (p_1, p_2, \ldots, p_n)$ and $q = (q_1, q_2, \ldots, q_n)$

the Euclidean distance is defined as:

$$d(p,q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_n - q_n)^2}$$

And the Manhattan distance as:

$$d(p,q) = |p_1 - q_1| + |p_2 - q_2| + \ldots + |p_n - q_n|$$

Igual, L., & Seguí, S. (2024). Introduction to data science. In *Introduction to Data Science: A Python Approach to Concepts, Techniques and Applications* (pp. 1-4). Cham: Springer International Publishing.

technische universität dortmund
TU Delft
Co-funded by the Erasmus+ programme of the European Union
Machine LearnAthon

Distance measures for non-numerical values (I)

- When dealing with non-numerical values (categorical or textual data), distance measures like Euclidean or Manhattan distance are not suitable. Instead, specific distance measures are used to calculate the similarity or dissimilarity between categorical or textual data points.

For two strings s_1 and s_2 of equal length

the Hamming distance is defined as:

$$d(s_1, s_2) = number\ of\ positions\ where\ s_1[i] \neq s_2[i]$$

Aggarwal, C. C., & Reddy, C. K. (2014). Data clustering. *Algorithms and applications. Chapman&Hall/CRC Data mining and Knowledge Discovery series, Londra*.

technische universität dortmund
TU Delft
Co-funded by the Erasmus+ programme of the European Union
Machine LearnAthon

Distance measures for non-numerical values (II)

- Two sets can be compared using the Jaccard Distance, which uses their intersection and union. The measure compares the presence or absence of attributes.

The Jaccard distance is defined as:

$$1 - \frac{|A \cap B|}{|A \cup B|}$$

The distance between two shopping lists can be calculated in the following way:

$$List_1 = \{Apples, pasta, milk\}, \text{List}_2 = \{Apples, milk, ham\}$$

$$1 - \frac{|2|}{|4|} = 0.5$$

Aggarwal, C. C., & Reddy, C. K. (2014). Data clustering. *Algorithms and applications. Chapman&Hall/CRC Data mining and Knowledge Discovery series, Londra*.

technische universität dortmund
TU Delft
Co-funded by the Erasmus+ programme of the European Union
Machine LearnAthon

Recap this lecture

After successfully completing this micro-lecture, you are able to….

- Define the term clustering and identify clustering use cases
- Explain the differences between the distance measures Euclidean distance, Manhattan distance, Hemming distance and Jaccard distance
- Analyze the similarity of data points based on the distance measures

The European Commission support for the production of this publication does not constitute an endorsement of the contents which reflects the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

This material was created with the assistance of an AI language model.

technische universität dortmund
TU Delft
Co-funded by the Erasmus+ programme of the European Union
Machine LearnAthon