Institut Informatika & Bisnis DARMAJAYA
Yayasan Alfian Husin

SAINS DATA
Institut Informatika & Bisnis Darmajaya

Kampus Merdeka INDONESIA JAYA

MERDEKA BELAJAR

DATA SCIENCE DARMAJAYA
"YOUR BEST FUTURE IN DATA"

MEETING: [13]

UNSUPERVISED LEARNING

BY: HENDRA KURNIAWAN

[12-09-2023] | [KODE MK: BDG23407 | MATAKULIAH: DATA SCIENCE AND BUSINESS ANALYTICS | SKS: 2/2 | VERSI: 01

UNSUPERVISED LEARNING

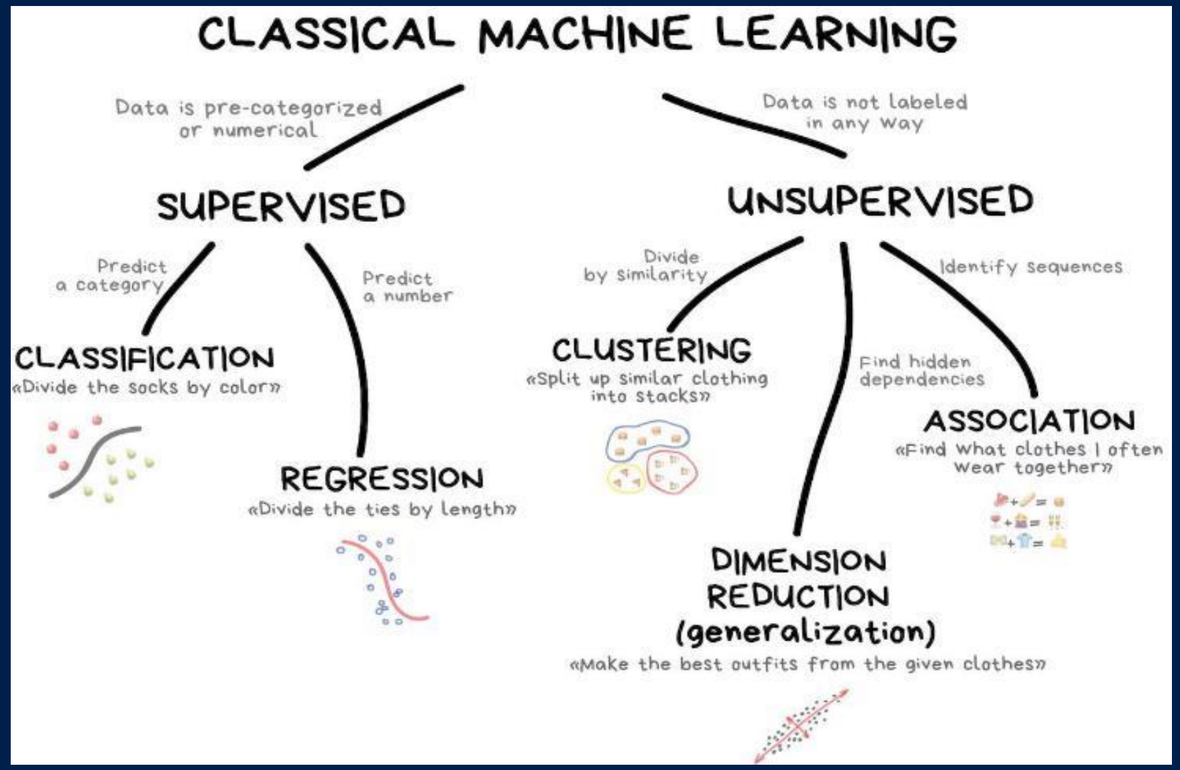
1. Basic Concept
2. Supervised vs Unsupervised Learning
3. Clustering?
4. What is Clustering for?
5. Aspects of Clustering

Supervised learning vs. unsupervised learning

- Supervised learning: discover patterns in the data that relate data attributes with a target (class) attribute.
  - These patterns are then utilized to predict the values of the target attribute in future data instances.
- Unsupervised learning: The data have no target attribute.
  - We want to explore the data to find some intrinsic structures in them.

Supervised learning vs. unsupervised learning

- Supervised
  - Classification
  - Regression
- Unsupervised
  - Clustering
  - Association
  - Dimension Reduction

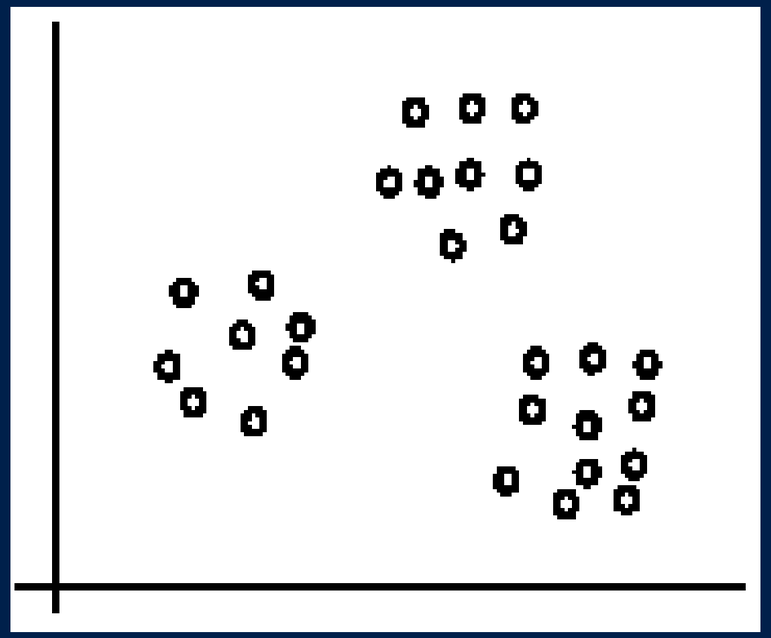


Clustering

- Clustering is a technique for finding similarity groups in data, called **clusters**. I.e.,
  - it groups data instances that are similar to (near) each other in one cluster and data instances that are very different (far away) from each other into different clusters.
- Clustering is often called an **unsupervised learning** task as no class values denoting an *a priori* grouping of the data instances are given, which is the case in supervised learning.
- Due to historical reasons, clustering is often considered synonymous with unsupervised learning.
  - In fact, association rule mining is also unsupervised.

An Illustration

- The data set has three natural groups of data points, i.e., 3 natural clusters.



Example of Using Clustering?

Given customer profile data, how to select potential customer data to offer certain products

	Customer Id	Age	Edu	Years Employed	Income	Card Debt	Other Debt	Defaulted	Address	DebtIncomeRatio
0	1	41	2	6	19	0.124	1.073	0.0	NBA001	6.3
1	2	47	1	26	100	4.582	8.218	0.0	NBA021	12.8
2	3	33	2	10	57	6.111	5.802	1.0	NBA013	20.9
3	4	29	2	4	19	0.681	0.516	0.0	NBA009	6.3
4	5	47	1	31	253	9.308	8.908	0.0	NBA008	7.2
...	...	...	...	...	...	...	...	...	...	...
845	846	27	1	5	26	0.548	1.220	NaN	NBA007	6.8
846	847	28	2	7	34	0.359	2.021	0.0	NBA002	7.0
847	848	25	4	0	18	2.802	3.210	1.0	NBA001	33.4
848	849	32	1	12	28	0.116	0.696	0.0	NBA012	2.9
849	850	52	1	16	64	1.866	3.638	0.0	NBA025	8.6

850 rows × 10 columns

We are asked to group customer data side by side, based on similarities in customer profiles

↓

Customer Segmentation

↓

Clustering

Example of Using Clustering? RESULT

	Customer Id	Age	Edu	Years Employed	Income	Card Debt	Other Debt	Defaulted	DebtIncomeRatio	Cluster
0	1	41	2	6	19	0.124	1.073	0.0	6.3	2
1	2	47	1	26	100	4.582	8.218	0.0	12.8	0
2	3	33	2	10	57	6.111	5.802	1.0	20.9	2
3	4	29	2	4	19	0.681	0.516	0.0	6.3	2
4	5	47	1	31	253	9.308	8.908	0.0	7.2	1
...	...	...	...	...	...	...	...	...	...	...
845	846	27	1	5	26	0.548	1.220	NaN	6.8	2
846	847	28	2	7	34	0.359	2.021	0.0	7.0	2
847	848	25	4	0	18	2.802	3.210	1.0	33.4	2
848	849	32	1	12	28	0.116	0.696	0.0	2.9	2
849	850	52	1	16	64	1.866	3.638	0.0	8.6	0

Each customer is successfully categorized

What is Clustering for?

- Let us see some real-life examples
- Example 1: groups people of similar sizes together to make “small”, “medium” and “large” T-Shirts.
  - Tailor-made for each person: too expensive
  - One-size-fits-all: does not fit all.
- Example 2: In marketing, segment customers according to their similarities
  - To do targeted marketing.

What is Clustering for?

- Example 3: Given a collection of text documents, we want to organize them according to their content similarities,
  - To produce a topic hierarchy
- In fact, clustering is one of the most utilized data mining techniques.
  - It has a long history, and used in almost every field, e.g., medicine, psychology, botany, sociology, biology, archeology, marketing, insurance, libraries, etc.
  - In recent years, due to the rapid increase of online documents, text clustering becomes important.

Aspects of Clustering?

- A clustering algorithm
  - Partitional clustering (K-Means, K-Medoid, K-Median, Fuzzy C-Means, dll
  - Hierarchical clustering (Agglomerative, Divisive, dll
  - Density Based Clustering (DBSCAN, dll)
- A distance (similarity, or dissimilarity) function
- Clustering quality
  - Inter-clusters distance $\Rightarrow$ maximized
  - Intra-clusters distance $\Rightarrow$ minimized
- The quality of a clustering result depends on the algorithm, the distance function, and the application.

Institut Informatika & Bisnis DARMAJAYA
Yayasan Alfian Husin

SAINS DATA
Institut Informatika & Bisnis Darmajaya

Kampus Merdeka INDONESIA JAYA

MERDEKA BELAJAR

THANK YOU!!

DATA SCIENCE DARMAJAYA “YOUR BEST FUTURE IN DATA”