

Introduction to Machine Learning and Clustering

Mahdi KHODADADZADEH

Assistant Professor

Faculty of Geo-Information Science and Earth Observation (ITC)

Department of Geo-information Processing (GIP)

m.khodadadzadeh@utwente.nl

May 2024



Previous lesson

We discussed the fundamentals and importance of E(S)DA

We introduced some statistical and visualization methods

We discussed a real-world geospatial problem which can be addressed by using E(S)DA

We introduced some open-source Python tools

Question #1

Exploratory (spatial) data analysis is always recommended before modelling.

True

False

Question #2

For exploratory (spatial) data analysis is always recommended to use both statistical analysis and visualization tools.

True

False

Question #3

When a geospatial dataset is clustered distributed in space, the Moran's I value is positive.

True

False

THIS IS YOUR MACHINE LEARNING SYSTEM?

YUP! YOU POUR THE DATA INTO THIS BIG PILE OF LINEAR ALGEBRA, THEN COLLECT THE ANSWERS ON THE OTHER SIDE.

WHAT IF THE ANSWERS ARE WRONG?

JUST STIR THE PILE UNTIL THEY START LOOKING RIGHT.



From: <https://xkcd.com>

This lesson's learning objectives

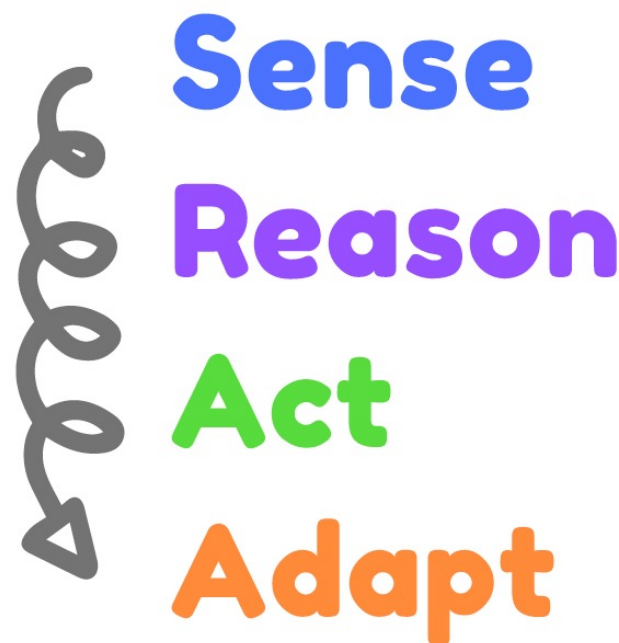
Explain to peers

- the fundamentals of **Artificial Intelligence**
- the principles of **Machine Learning** methods
- the basics of **clustering** methods

Apply some **open-source Python tools** to perform geodata clustering

Artificial Intelligence (AI)

- AI is the **science of creating intelligent machines** that can **simulate human** thinking capability and behavior.



AI takes raw data (images, sound, text) and processes it using image or text processing.

AI thinks about the information it has received and how it relates to what it recognises and has learned previously.

The AI performs a task or action based on the information it has processed.

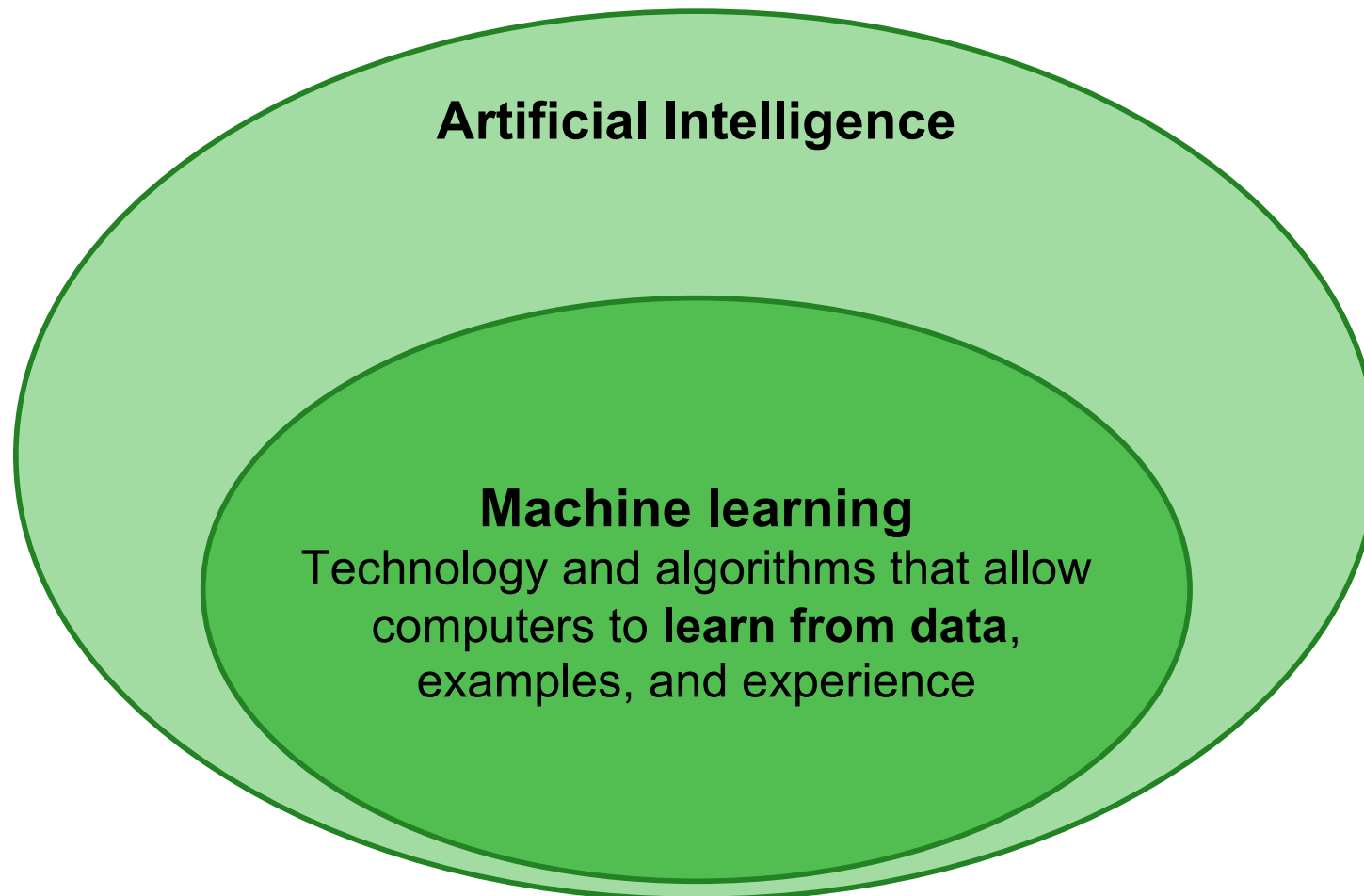
The AI uses the successful or unsuccessful outcome as feedback.

From: https://csermoocs.appspot.com/ai_secondary/unit?unit=2&lesson=4



Chat GPT

Artificial Intelligence (AI) vs. Machine Learning (ML)



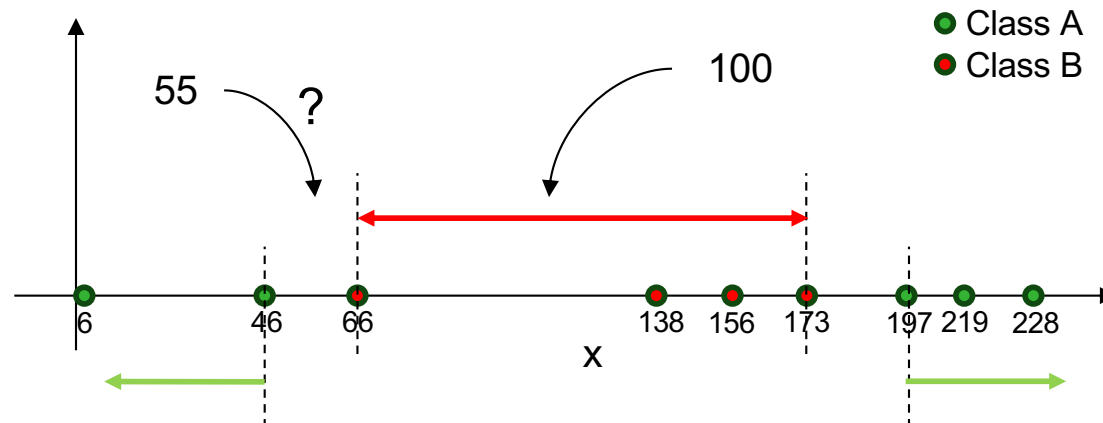
What is Machine Learning?

- Machine Learning = Learning from data
= Finding a rule that is consistent with the data

Dataset

X	Class
228	A
66	B
138	B
219	A
156	B
46	A
197	A
6	A
173	B
100	?

How can we improve the definition of the rule?



Rule = {A: $X \geq 197$ or $X \leq 46$ } and {B: $66 \leq X \leq 173$ }

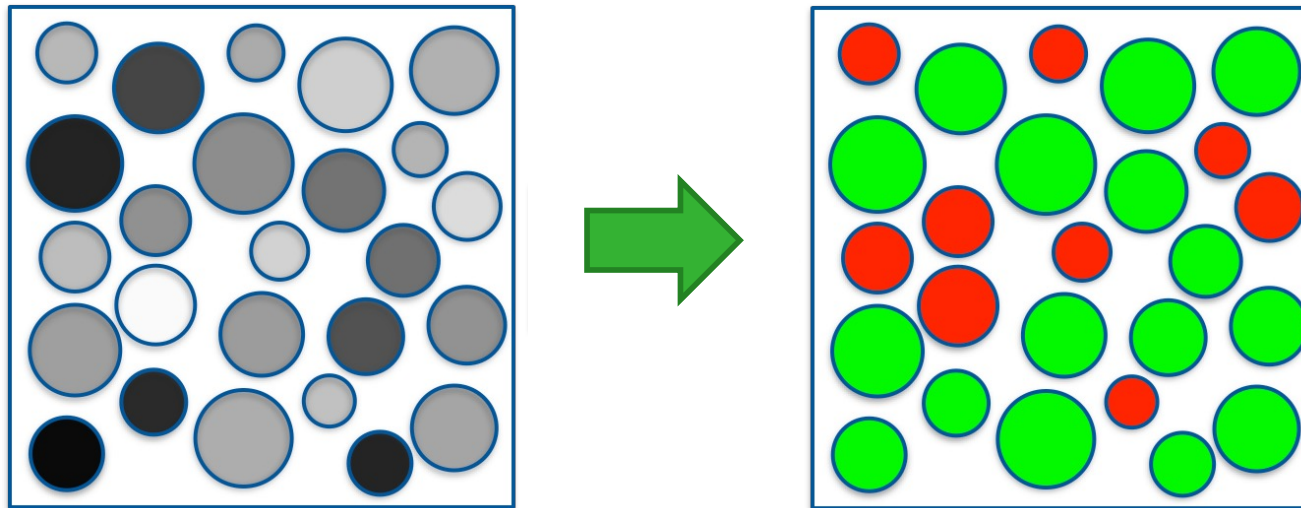
What is Machine Learning?

- In machine learning, a feature is a **measurable property** of a phenomenon being observed.
- If we have **multiple features**, can you easily define a rule?

Dataset

X_1	X_2	X_3	X_4	X_5	X_6	X_7	Class
200	220	195	190	188	230	228	A
60	70	55	50	60	73	67	B
130	140	140	140	145	148	138	B
200	210	230	230	220	200	209	A
150	160	143	140	155	164	156	B
40	50	74	75	45	55	46	A
190	200	205	200	200	230	197	A
1	10	15	15	1	12	6	A
170	180	160	165	178	180	173	B

What is Machine Learning?



Can we define **a set of rules (specific procedure)** that a computer follows to perform this task?

Algorithm

When should we use Machine Learning?

- Problems where there are **no human experts**, so data cannot be labelled or categorized
- Problems where there are human experts, but it is **very hard (or impossible) to define rules**
- Problems where there are human experts and the rules can be defined, but where it is **not cost effective to implement**

- When we want to
 - ✓ **Detect** patterns/structures/themes/trends etc. in the data
 - ✓ **Make predictions** about future data and **make decisions**

Categories of Machine Learning

Unsupervised learning (clustering)

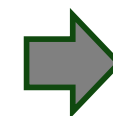
Class labels unknown
Checking for groups/clusters in the data

Dataset

X_1	X_2	X_3	X_4	X_5	X_6	X_7
200	220	195	190	188	230	228
60	70	55	50	60	73	67
130	140	140	140	145	148	138
200	210	230	230	220	200	209
150	160	143	140	155	164	156
40	50	74	75	45	55	46
190	200	205	200	200	230	197
195	205	200	190	180	210	188
170	180	160	165	178	180	173



Clustering
Algorithm



Class
A
B
B
A
B
A
A
A
B

Categories of Machine Learning

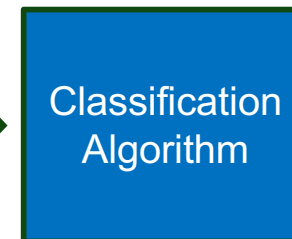
Supervised learning

Training data with known **target variable**
Making predictions on unseen/new data

Discrete labels
or
Continuous

Dataset

X ₁	X ₂	X ₃	X ₄	X ₅	X ₆	X ₇	Class
200	220	195	190	188	230	228	A
60	70	55	50	60	73	67	B
130	140	140	140	145	148	138	B
200	210	230	230	220	200	209	A
150	160	143	140	155	164	156	B
40	50	74	75	45	55	46	A



Class
A
A
B



190	200	205	200	200	230	197
195	205	200	190	180	210	188
170	180	160	165	178	180	173

Categories of Machine Learning

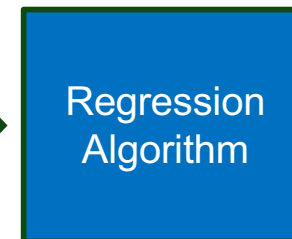
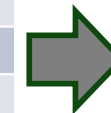
Supervised learning

Training data with known **target variable**
Making predictions on unseen/new data

Discrete labels
or
Continuous

Dataset

X ₁	X ₂	X ₃	X ₄	X ₅	X ₆	X ₇	Target
200	220	195	190	188	230	228	10
60	70	55	50	60	73	67	20
130	140	140	140	145	148	138	22
200	210	230	230	220	200	209	5
150	160	143	140	155	164	156	15
40	50	74	75	45	55	46	12



Target
12
18
22



190	200	205	200	200	230	197
195	205	200	190	180	210	188
170	180	160	165	178	180	173

Question #1

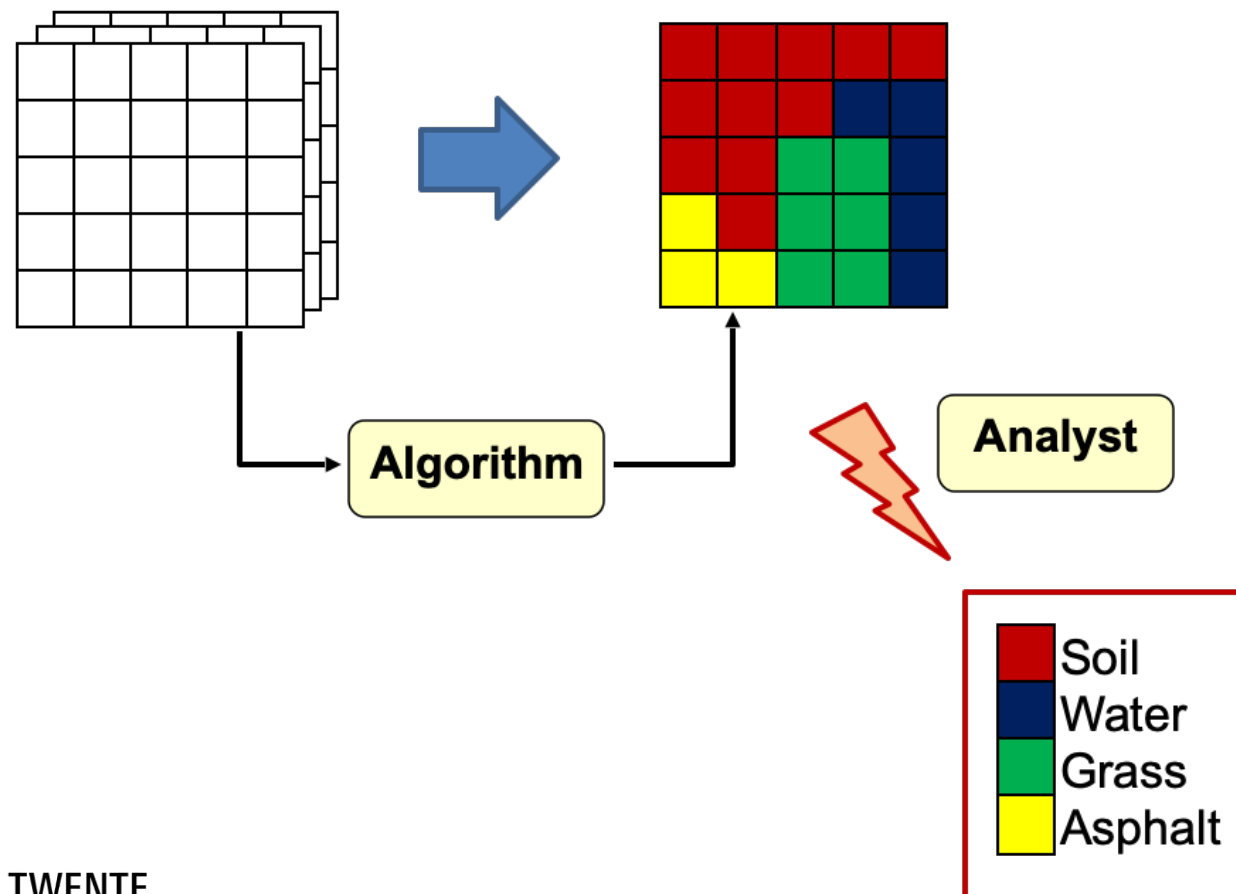
Identifying various land use categories using remote sensing images is a classification problem.

True

False

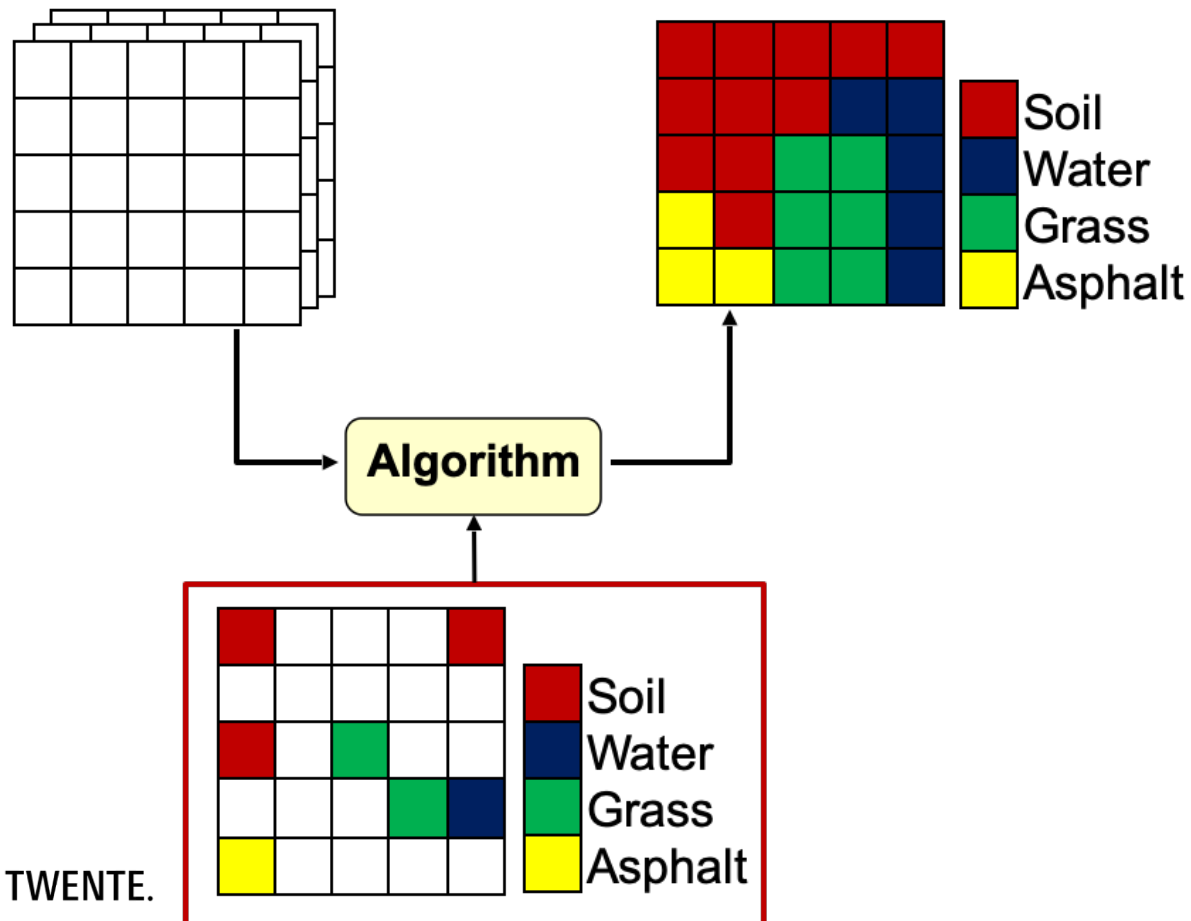
Clustering

- Machine learning algorithm finds hidden patterns or intrinsic structures in input data



Classification

- Machine learning algorithm trains a model on known input and output data so that it can predict future outputs (labels).



Question #2

Predicting house prices based on various spatial and non-spatial factors is a regression problem.

True

False

Question #3

We would like to characterize tick bites risk in a study area using some observational and geospatial datasets. We can formulate this problem both as a regression or classification.

True

False

Target variable	
Categorical	Continuous
1: Very low risk 2: Low risk 3: Moderate risk 4: High risk 5: Very high risk	A probability or a score, the range could be: 0.0 to 1.0

Unsupervised Learning (Clustering)

TRUE
TRUE
TRUE

AI: machine that can simulate human (behaviour, and thinking)

Machine learning is the brain which learns from data, it's part of AI itself

How can we improve the definition of a rule?

- Adding a new class
- Adding more data

Regression: similar to classification but the value is continuous. It can predict new data

Classification: target is categorized

Land use is categorized so it is clustering

House prices consider as regression because their continuous values.

Ticks: continuous variable: risk (1, 2, 3) -> classification

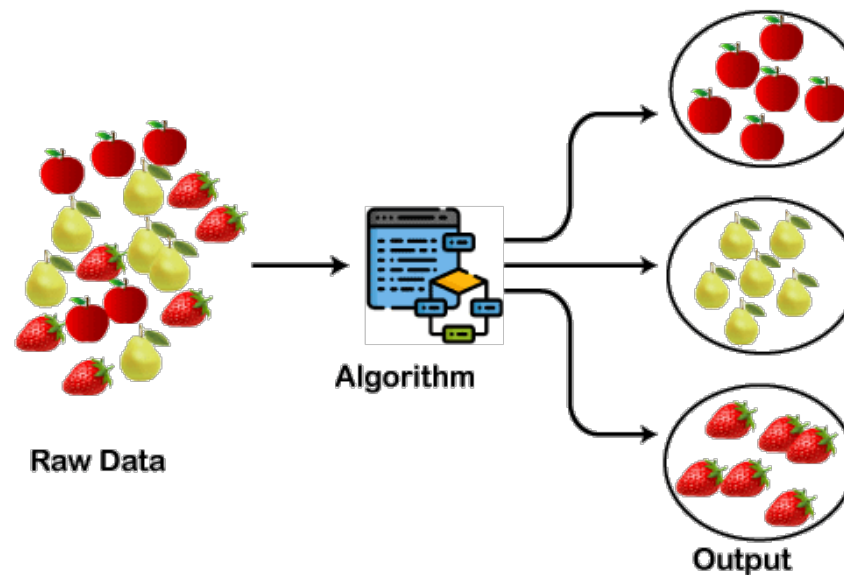
Categorical variables : very low risk, very high risk -> regression

K-means result can be different because of :

- How many clusters that we need

Clustering

- *Clustering* is an important task in data mining that aims at **identifying groups of elements** that are similar among themselves but dissimilar to the elements in other groups.
- It provides a **high-level abstraction** of the data, which facilitates the **extraction of useful information**.



<https://www.javatpoint.com/clustering-in-machine-learning>

Unsupervised Learning - clustering

- Contrarily to classification/regression task there is **NOT a target**
- It is suitable in problems where we have **unlabeled objects** or where the process of labelling is expensive (time / money).
- It is often done in **combination with the E(S)DA** to get a better grip on the objects at hand

K-means

- One of the most **popular** clustering algorithms
- It is an **iterative** algorithm that first assigns points to clusters (**k classes**) and then recomputes the centers of those clusters (means) until finding an optimum solution.
- Each point belong to one and only one cluster
- Each cluster is represented by the cluster mean (centroid)

K-means

- K-means minimizes the total squared distance between each input point (x_i) and its cluster center (c_j)

number of clusters number of cases centroid for cluster j

case i

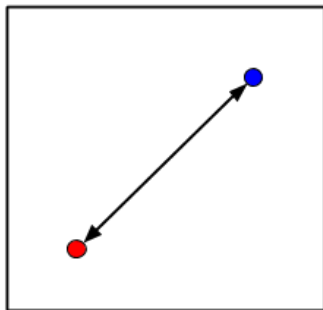
objective function $\leftarrow J = \sum_{j=1}^k \sum_{i=1}^n \underbrace{\|x_i^{(j)} - c_j\|^2}_{\text{Distance function}}$

The diagram shows the objective function $J = \sum_{j=1}^k \sum_{i=1}^n \|x_i^{(j)} - c_j\|^2$. Annotations include: 'number of clusters' pointing to k , 'number of cases' pointing to n , 'case i ' pointing to $x_i^{(j)}$, 'centroid for cluster j ' pointing to c_j , and 'Distance function' pointing to the term $\|x_i^{(j)} - c_j\|^2$. The entire expression is labeled 'objective function' with an arrow pointing to J .

Common Distance Measures in Data Science

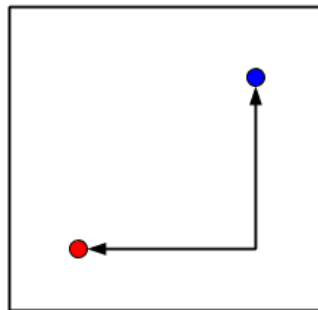
- Euclidean distance
 - The most common
- Manhattan distance
 - Approximation to Euclidean distance and cheaper to compute
 - Sum of the absolute differences of their Cartesian coordinates
- Minkowski distance
 - A generalization of both the Euclidean & Manhattan distance

Euclidean



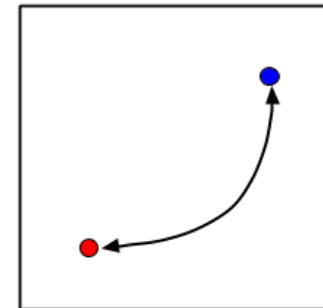
$$d(x, y) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

Manhattan



$$d(x, y) = \sum |x_i - y_i|$$

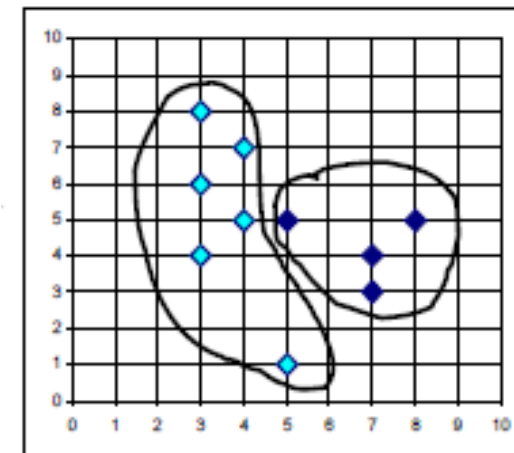
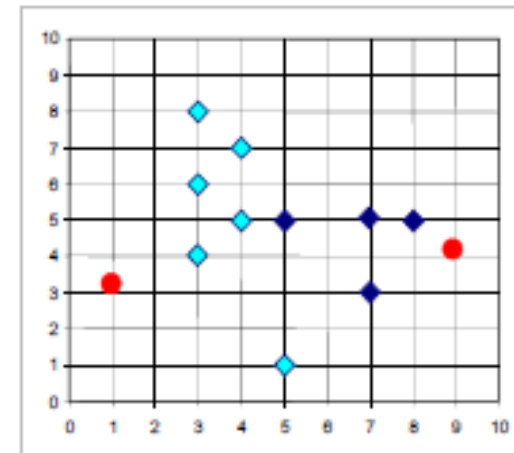
Minkowski



$$d(x, y) = \sqrt[h]{|x_1 - y_1|^h + |x_2 - y_2|^h + \dots + |x_p - y_p|^h}$$

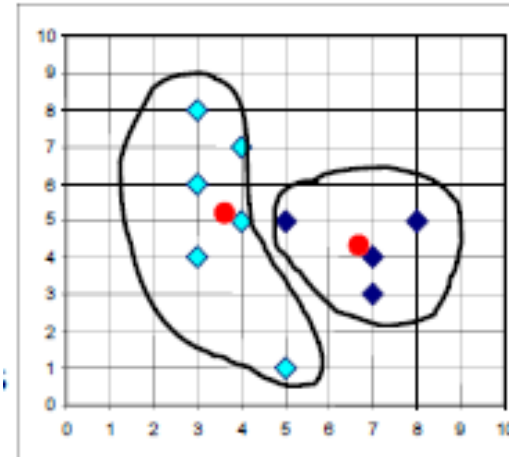
K-means

1. **Initialization:** arbitrary initialization of the centers
2. **Data assignment:** each point is assigned to its closest cluster (center). Ties are broken by randomly assigning the point to one of the clusters. This yields a partitioning of the data.

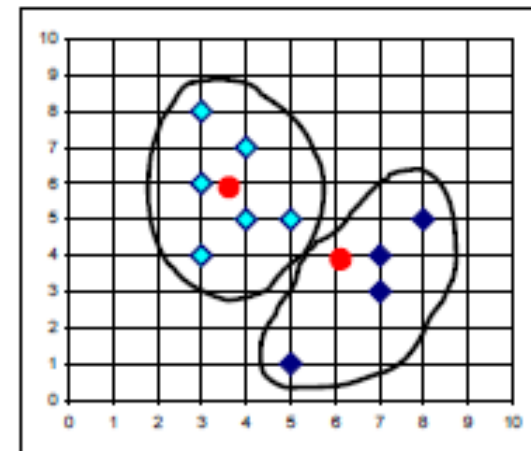


K-means

3. **Relocating “centers”**: each cluster representative is moved to the center of the points assigned to it.

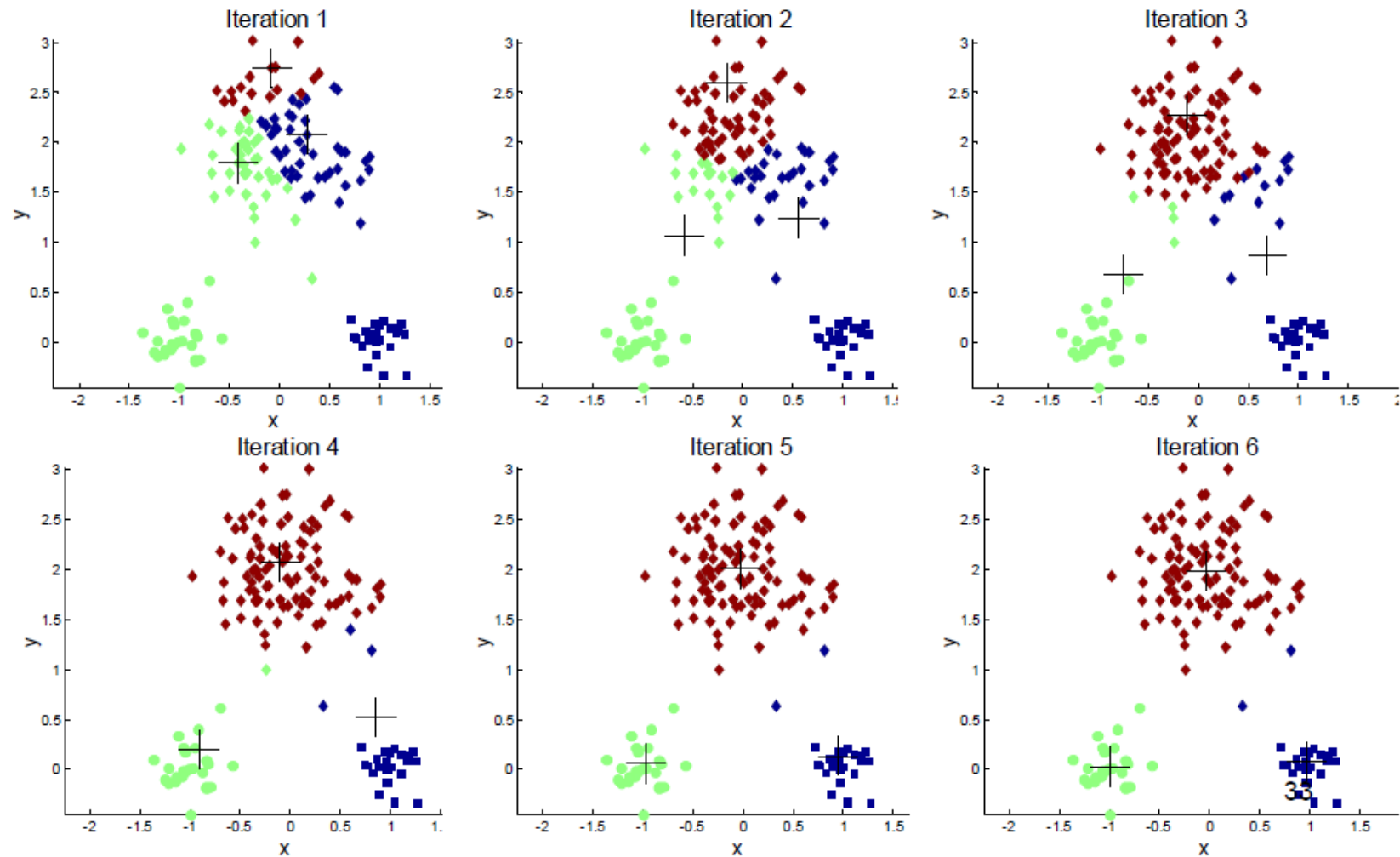


4. **Repeat 2 and 3** until the centers do not longer change.



Each iteration requires $N \cdot k$ comparisons. It takes long time for large datasets

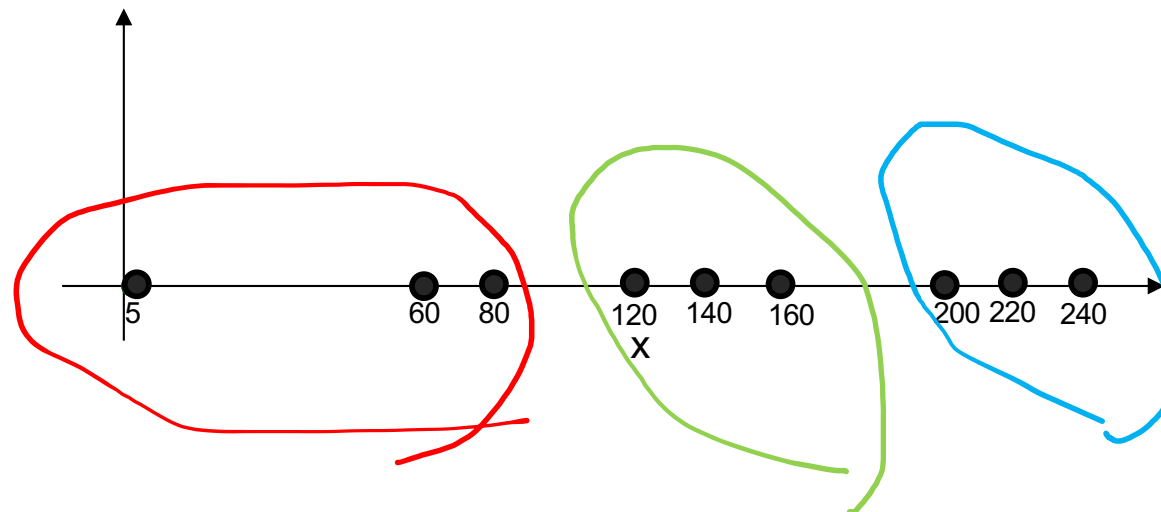
A K-means example of 3 clusters



K-means Limitations

Dataset

X
240
80
140
220
160
60
200
5
120

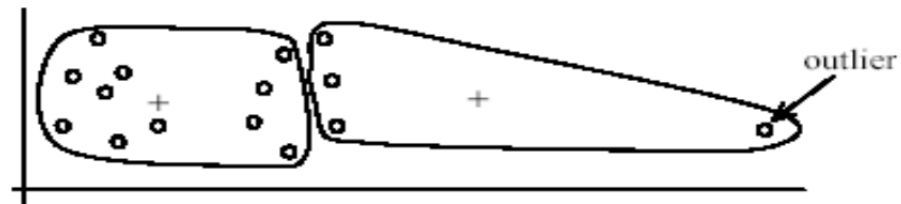


K-means Limitations

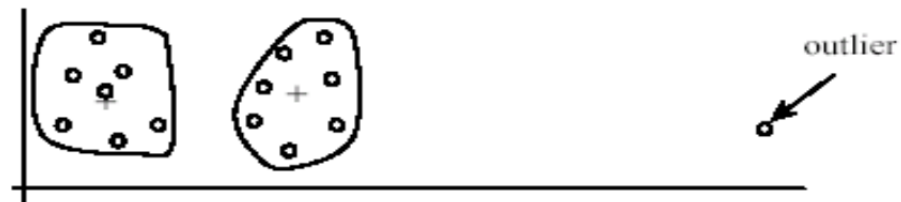
The algorithm is sensitive to outliers

Clusters centers are computed using the “mean” function which is sensitive to outliers

Outliers



(A): Undesirable clusters



(B): Ideal clusters

Kmeans limitations

It is quite sensitive to the initial set-up (location of cluster centers) which might lead to finding local minima instead of the absolute minimum.

Better to run multiple times (different initializations) or look for a robust way of initializing the algorithm.

Choosing the value of k is difficult (unless you have a priori knowledge about the “natural” number of groups present in the data).

Silhouette method

Elbow method

EDA?

Team based learning

Ghelgheli decided to improve his customers' satisfaction. He heard a lot about a magic tool that could learn from data and help him tailor marketing strategies and promotions to different customer groups.

Intrigued, Ghelgheli began exploring this tool. He applied it to a set of data and discovered that his customers fell into three main categories: "Daily Commuters", "Weekend Relaxers" and "Tea Enthusiasts". The "Daily Commuters" were regular visitors who stopped by for a quick tea on their way to work. The "Weekend Relaxers" visited the teahouse mainly on weekends. The "Tea Enthusiasts" were passionate about trying different varieties and learning about tea.

After the analysis, Ghelgheli introduced a program offering discounts on morning tea purchases targeting "Daily Commuters". For the "Weekend Relaxers", he organized weekend events and introduced special weekend-only blends. To satisfy "Tea Enthusiasts", he started hosting monthly tea-tasting events and workshops. As a result, Ghelgheli's teahouse became even more popular.

- **Which data and methods do you think Ghelgheli utilized for his analysis?**
- **Can you list the potential steps that Ghelgheli took for such an analysis?**
- **Can you provide some real-life examples similar to Ghelgheli's experience?**

Conclusion

- Data: Customer purchase history (e.g., items bought, frequency of purchases), demographics (e.g., age, occupation, location), customer feedback and reviews, and visit times and days.
- Methods: Data cleaning and preparation, EDA (descriptive statistics and visualizations), Clustering algorithms (e.g., K-means clustering) to identify distinct customer groups
- Steps: Data collection, data cleaning, EDA, data standardization (if necessary), choosing a clustering algorithm (K-means), determining the number of clusters, applying the clustering algorithm, analyzing and interpreting the resulting clusters

Conclusion

- Please reply to the questions and write your answers on the yellow post-it
- What is the most important idea/insight you will remember from today's lesson?
- What questions do you still have?

Conclusion

I can **Explain** to peers

- the fundamentals of **Artificial Intelligence**
- the principles of **Machine Learning** methods
- the basics of **clustering** methods

not yet ☹️

very well 😊

