MACHINE LEARNING AND OPTIMIZATION FOR TESTING

DI Dr Branka Stojanovic

ORGANISATION

Unit	Date	Type/Scope	Content
1	06.10.2023	2 LE	Introduction – AI/ML and testing
2	13.10.2023	2 EXE	Introduction – data science and Python
3	20.10.2023	2 LE	Machine learning, testing and data preparation
4	20.10.2023	2 EXE	Data preparation and Python; Homework assignments
5	10.11.2023	2 LE	Supervised Machine Learning
6	10.11.2023	2 EXE	Supervised Machine Learning
7	17.11.2023	2 LE	Unsupervised Machine Learning
8	17.11.2023	2 EXE	Unsupervised Machine Learning
9	01.12.2023	2 LE	Neural Networks and Deep Learning
10	01.12.2023	2 EXE	Neural Networks and Deep Learning
11	15.12.2023	2 LE	Final project introduction and assignments
12	12.01.2024	2 EXE	Hands-on - Final project consultation; Homework discussion
13	19.01.2024	2 LE	Final project tutorial and results presentations and discussion
14	19.01.2024	2 EXE	Final project results demonstrations
15	26.01.2024	2 LE	Recap and Q&A
16	09.02.2024	1 EXM	Exam

4. UNSUPERVISED MACHINE LEARNING

OUTLINE

4.1 Unsupervised Machine Learning

4.2 Clustering

4.3 Dimensionality reduction

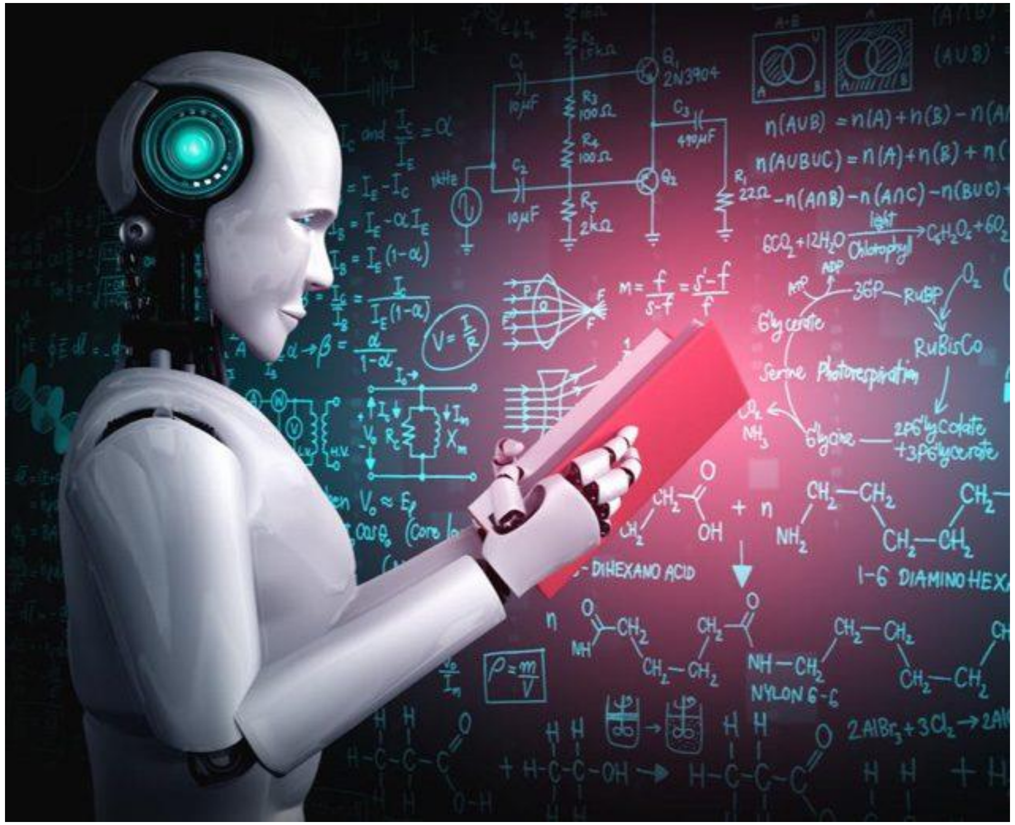


Image source: https://www.eweek.com/enterprise-apps/what-is-artificial-intelligence

OUTLINE

4.1 Unsupervised Machine Learning

- Overview

4.2 Clustering

- K-Means Clustering

4.3 Dimensionality reduction

- Principle Component Analysis (PCA)
- T - Distributed Stochastic Neighbour Embedding (T - SNE)

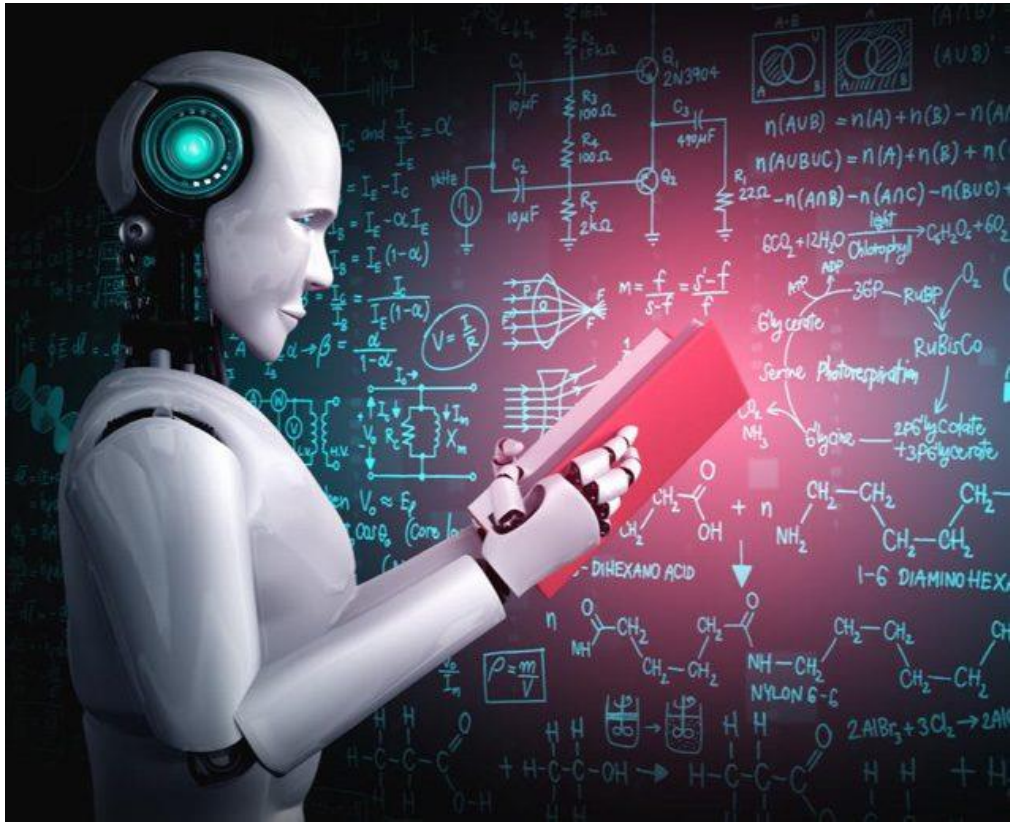


Image source: https://www.eweek.com/enterprise-apps/what-is-artificial-intelligence

4.1 UNSUPERVISED MACHINE LEARNING

- In Unsupervised Learning, the machine uses unlabeled data and learns on itself without any supervision
- The machine tries to find a pattern in the unlabeled data and gives a response
- Let's take a similar example as before, but this time we do not tell the machine whether it's a spoon or a knife
- The machine identifies patterns from the given set and groups them based on their patterns, similarities, etc.

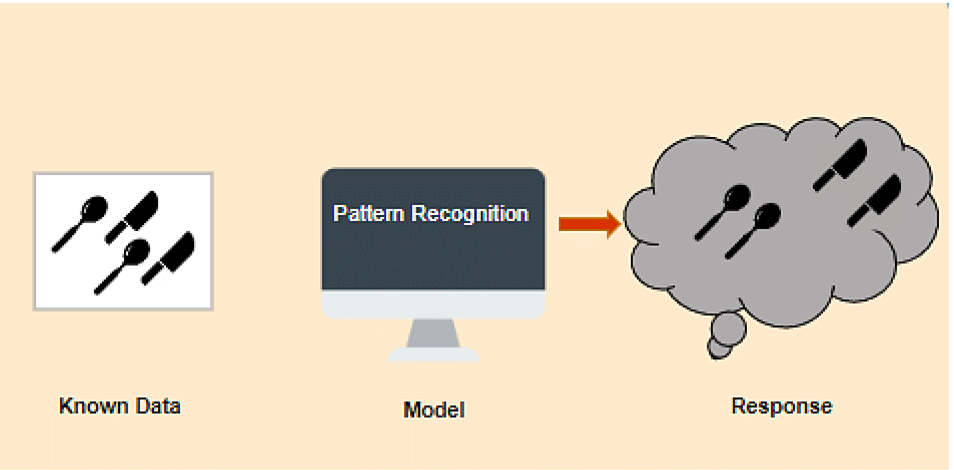


Image source: https://www.simplilearn.com/tutorials/machine-learning-tutorial

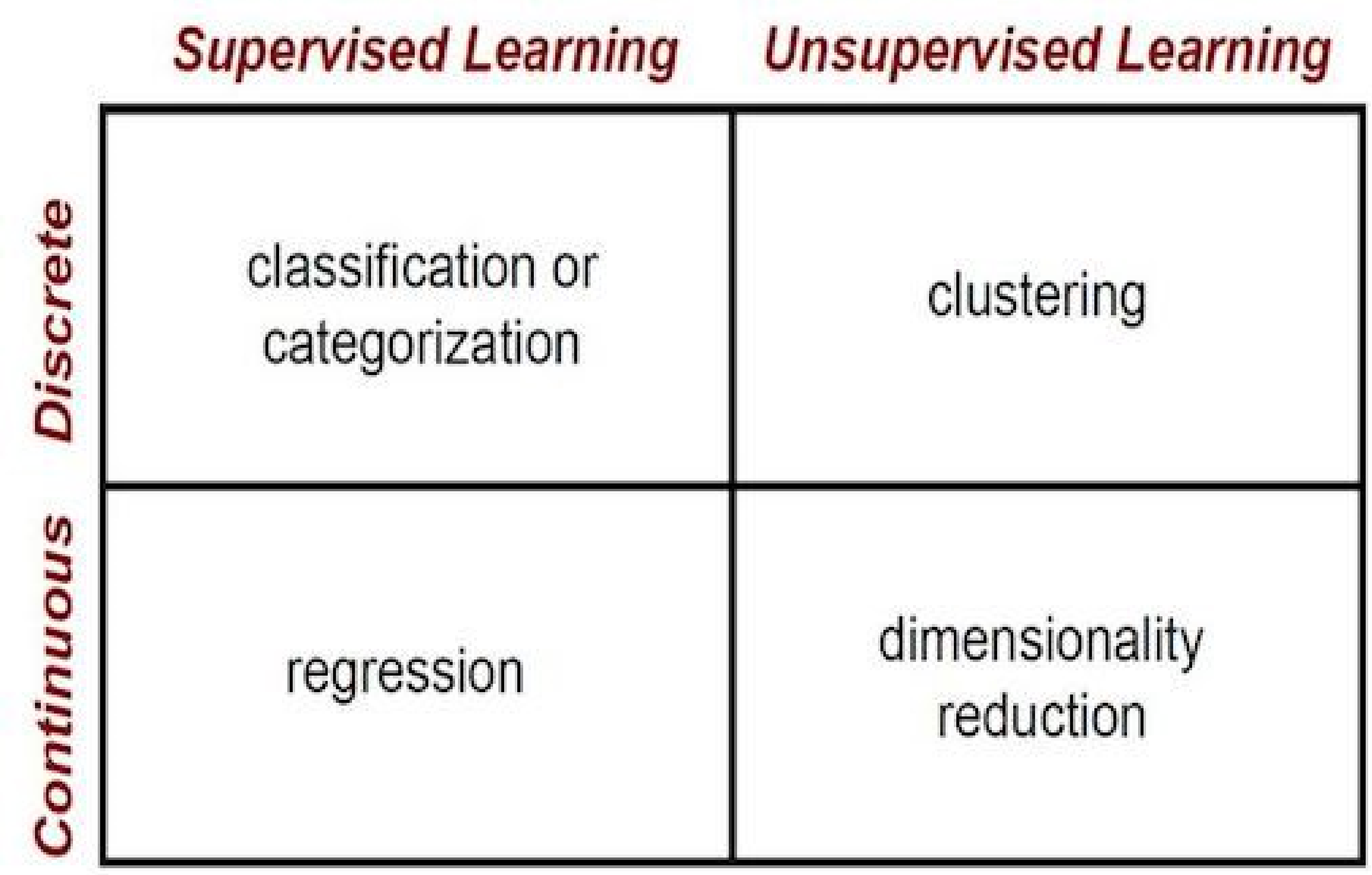
Overview

- The most common **tasks** within unsupervised learning are
  - clustering
  - representation learning
  - density estimation
- In all of these cases, we wish to learn the inherent structure of our data **without** using explicitly-provided **labels**
- Some common algorithms include **k-means** clustering, **principal component analysis**, and **autoencoders**
- Since no labels are provided, there is no specific way to compare model performance in most unsupervised learning methods
- Two common **use-cases** for unsupervised learning are
  - exploratory analysis
  - dimensionality reduction

Use-cases

- Unsupervised learning is very useful in **exploratory analysis** because it can automatically identify structure in data, e.g.
  - If an analyst were trying to segment consumers, unsupervised clustering methods would be a great starting point for their analysis
  - In situations where it is either impossible or impractical for a human to propose trends in the data, unsupervised learning can provide initial insights that can then be used to test individual hypotheses
- **Dimensionality reduction** - methods used to represent data using less columns or features
  - In **representation learning**, we wish to learn **relationships** between individual features, allowing us to represent our data using the **latent features** that interrelate our initial features
  - This sparse latent structure is often represented using far fewer features than we started with, so it can make further data processing much **less intensive**, and can **eliminate redundant** features

Recap



	Supervised Learning	Unsupervised Learning
Discrete	classification or categorization	clustering
Continuous	regression	dimensionality reduction

Image source: https://towardsdatascience.com/supervised-vs-unsupervised-learning-14f68e32ea8d

OUTLINE

4.1 Unsupervised Machine Learning

- Overview

4.2 Clustering

- K-Means Clustering

4.3 Dimensionality reduction

- Principle Component Analysis (PCA)
- T - Distributed Stochastic Neighbour Embedding (T - SNE)

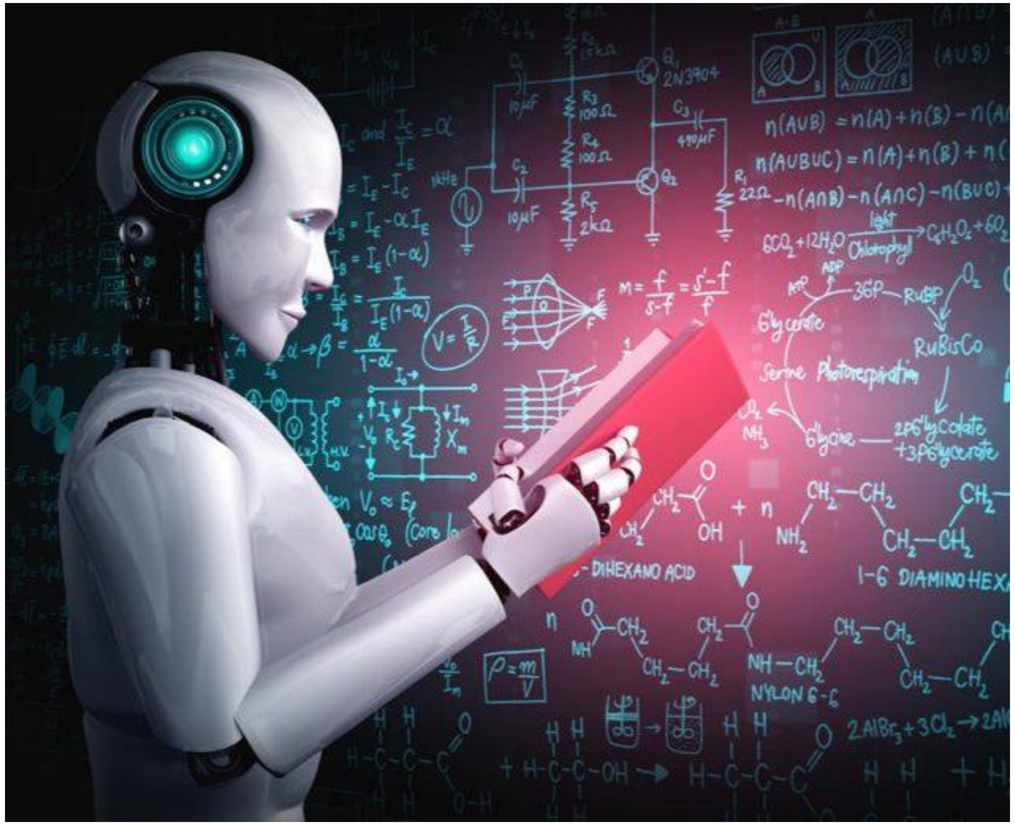


Image source: https://www.eweek.com/enterprise-apps/what-is-artificial-intelligence

4.2 CLUSTERING

- Clustering is a Machine Learning technique that involves the grouping of data points

- Given a set of data points, we can use a clustering algorithm to classify each data point into a specific group
  - Example: finding out which customers made similar product purchases

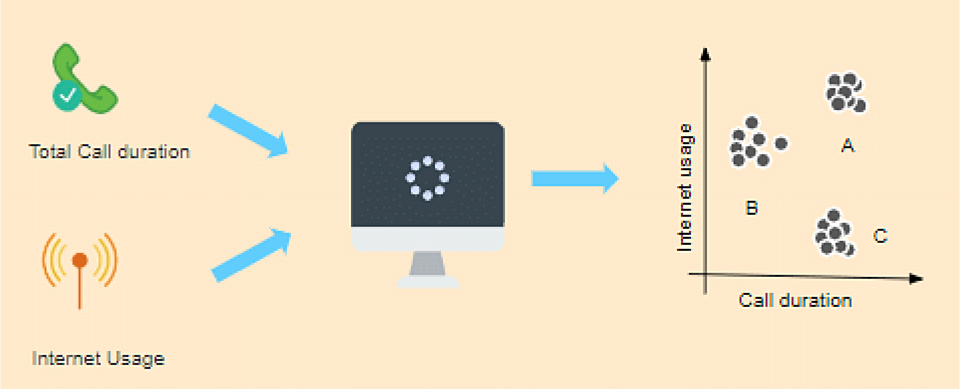


Image source: https://www.simplilearn.com/tutorials/machine-learning-tutorial

Overview

- In theory, data points that are in the same group should have similar properties and/or features, while data points in different groups should have highly dissimilar properties and/or features

- Clustering is a method of unsupervised learning and is a common technique for statistical data analysis used in many fields

- In Data Science, we can use clustering analysis to gain some valuable insights from our data by seeing what groups the data points fall into when we apply a clustering algorithm

Real-Life Applications of Unsupervised Learning

- **Market Basket Analysis**
  - It is a machine learning model that determines if you buy a certain group of items, you are less or more likely to buy another group of items
- **Semantic Clustering**
  - Semantically similar words share a similar context. People post their queries on websites in their own ways. Semantic clustering groups all these responses with the same meaning in a cluster to ensure that the customer finds the information they want quickly and easily. It plays an important role in information retrieval and good browsing experience.
- **Delivery Store Optimization**
  - Machine learning models are used to predict the demand and keep up with supply. They are also used to open stores where the demand is higher and optimizing roots for more efficient deliveries according to past data and behavior.
- **Identifying Accident Prone Areas**
  - Unsupervised machine learning models can be used to identify accident-prone areas and introduce safety measures based on the intensity of those accidents.

Supervised vs. Unsupervised Learning

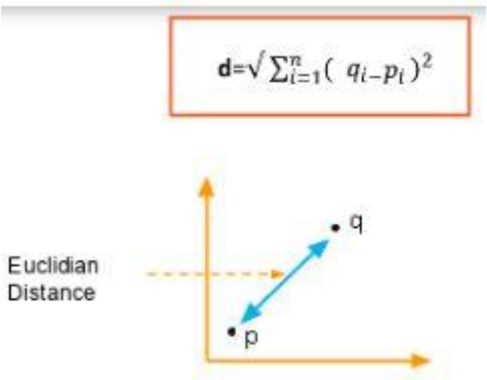
Supervised Learning	Unsupervised Learning
It uses known and labeled data as input	It uses unlabeled data as input
It has a feedback mechanism	It has no feedback mechanism
The most used supervised learning algorithms are: •KNN •Logistic regression •Support vector machine	The most used unsupervised learning algorithms are: •K-means clustering •PCA

K-Means Clustering

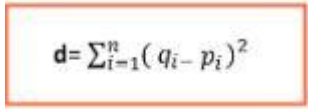
- The term ‘**K**’ is a number
  - You need to tell the system how many clusters you need to create
  - For example, K = 2 refers to two clusters
  - There is a way of finding out what is the best or optimum value of K for a given data
- **Applications** of K-Means Clustering
  - Academic Performance
    - Based on the scores, students are categorized into grades like A, B, or C
  - Diagnostic systems
    - The medical profession uses k-means in creating smarter medical decision support systems
  - Search engines
    - Clustering forms a backbone of search engines. When a search is performed, the search results need to be grouped, and the search engines very often use clustering to do this.
  - Wireless sensor networks
    - The clustering algorithm plays the role of finding the cluster heads, which collect all the data in its respective cluster

Distance Measure

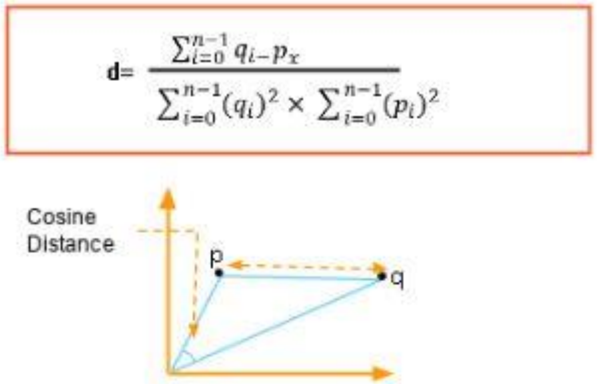
- Distance measure determines the similarity between two elements and influences the shape of clusters
- K-Means clustering supports various kinds of distance measures:
    - Euclidean distance measure
    - Squared Euclidean distance measure
    - Manhattan distance measure
    - Cosine distance measure



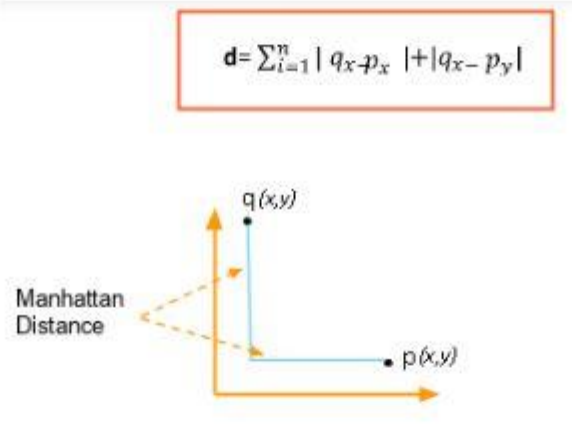
$$d=\sqrt{\sum_{i=1}^{n}(q_i - p_i)^2}$$



$$d=\sum_{i=1}^{n}(q_i - p_i)^2$$



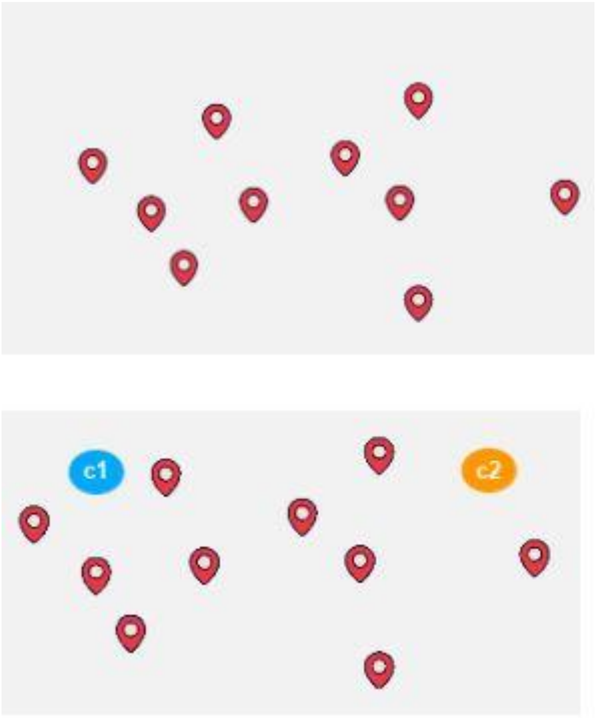
$$d=\frac{\sum_{i=0}^{n-1} q_i - p_x}{\sum_{i=0}^{n-1}(q_i)^2 \times \sum_{i=0}^{n-1}(p_i)^2}$$



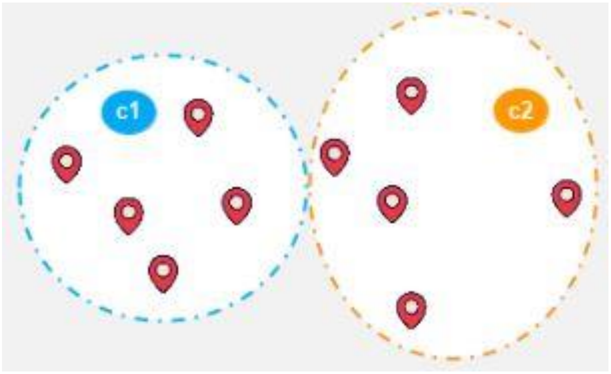
$$d=\sum_{i=1}^{n} | q_x - p_x | + | q_x - p_y |$$

Image source: https://www.simplilearn.com/tutorials/machine-learning-tutorial

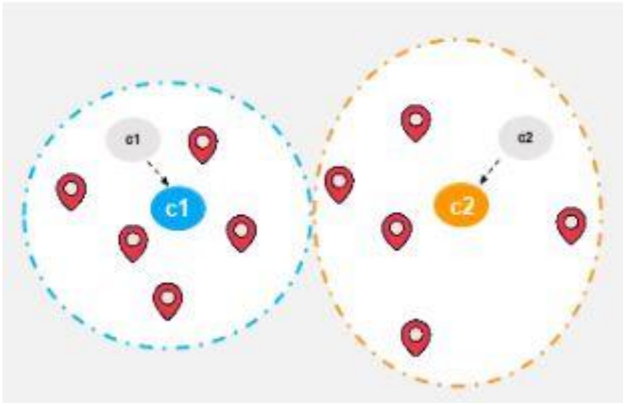
Workflow



Randomly assign centroids



Measure distance and assign initial groups



Compute actual centroids

Image source: https://www.simplilearn.com/tutorials/machine-learning-tutorial

How does it work?

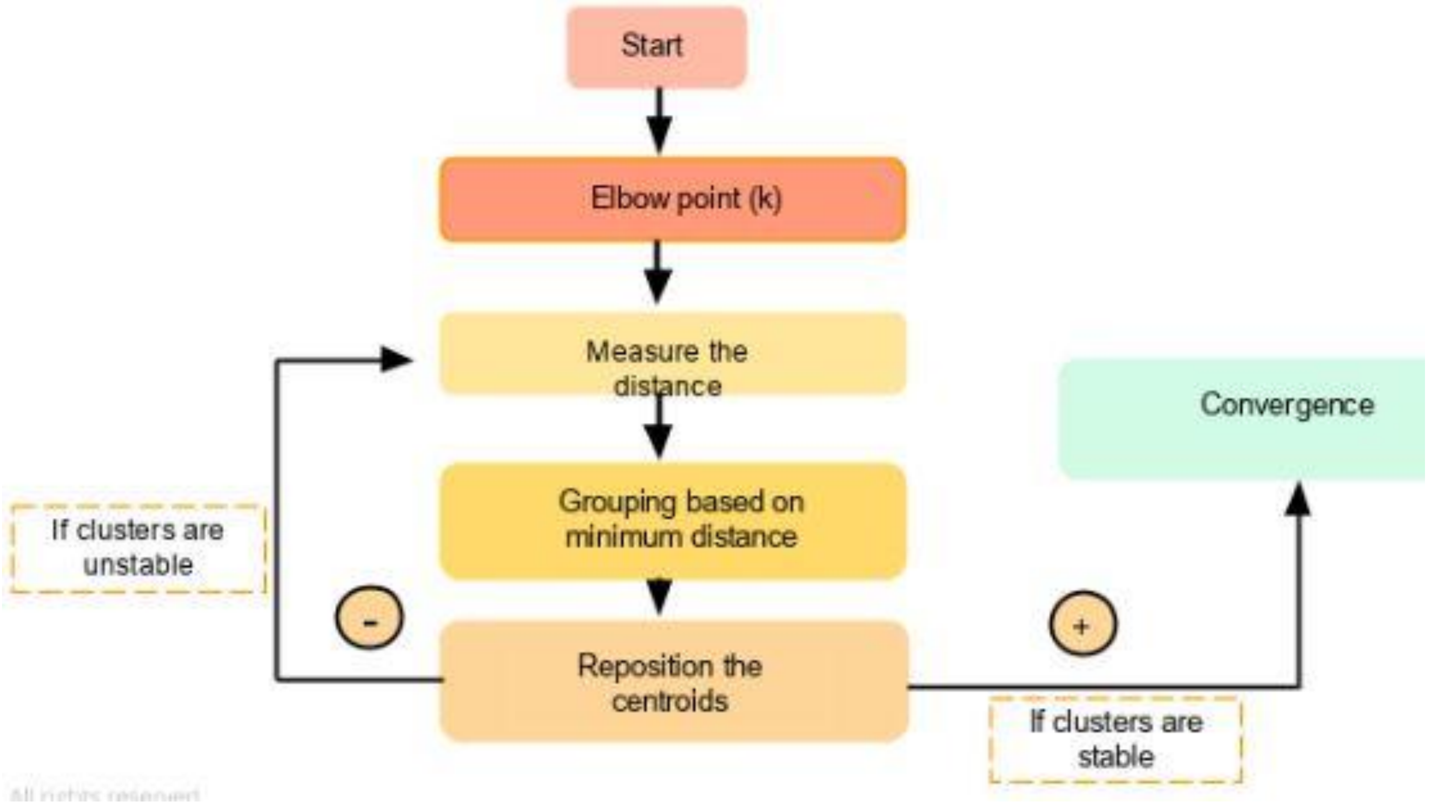
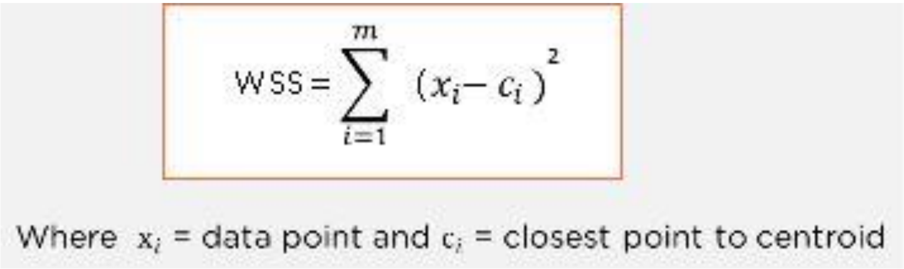


Image source: https://www.simplilearn.com/tutorials/machine-learning-tutorial

Elbow method

- The **Elbow** method is the best way to find the number of clusters
- We run k-means and use **within-sum-of-squares (WSS)** as a measure to find the optimum number of clusters
- WSS (also called **Inertia**) is defined as the sum of the squared distance between each member of the cluster and its centroid



$$\text{WSS} = \sum_{i=1}^{m} (x_i - c_i)^2$$

Where x_i = data point and c_i = closest point to centroid

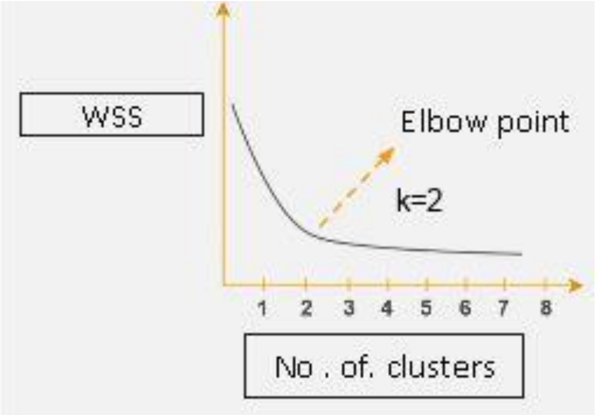


Image source: https://www.simplilearn.com/tutorials/machine-learning-tutorial

OUTLINE

4.1 Unsupervised Machine Learning

- Overview

4.2 Clustering

- K-Means Clustering

4.3 Dimensionality reduction

- Principle Component Analysis (PCA)
- T - Distributed Stochastic Neighbour Embedding (T - SNE)

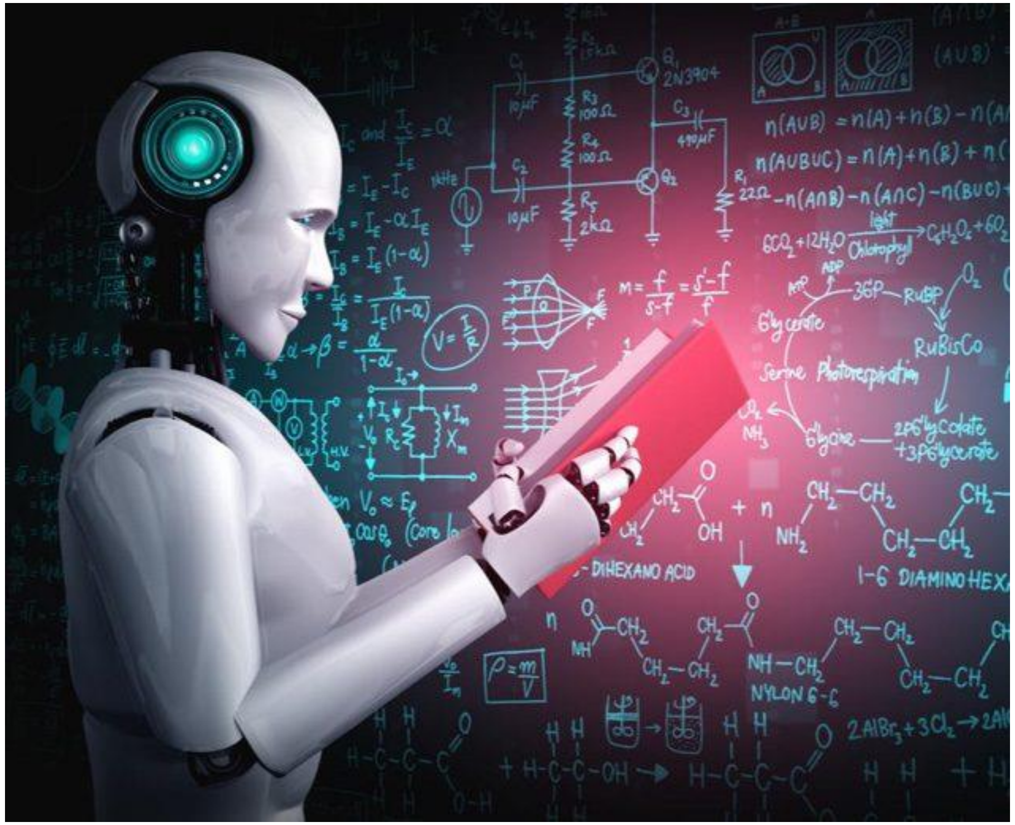


Image source: https://www.eweek.com/enterprise-apps/what-is-artificial-intelligence

4.4 DIMENSIONALITY REDUCTION

- While working with high-dimensional data, machine learning models often seem to overfit, and this reduces the ability to generalize past the training set examples

- Hence, it is important to perform dimensionality reduction techniques before creating a model

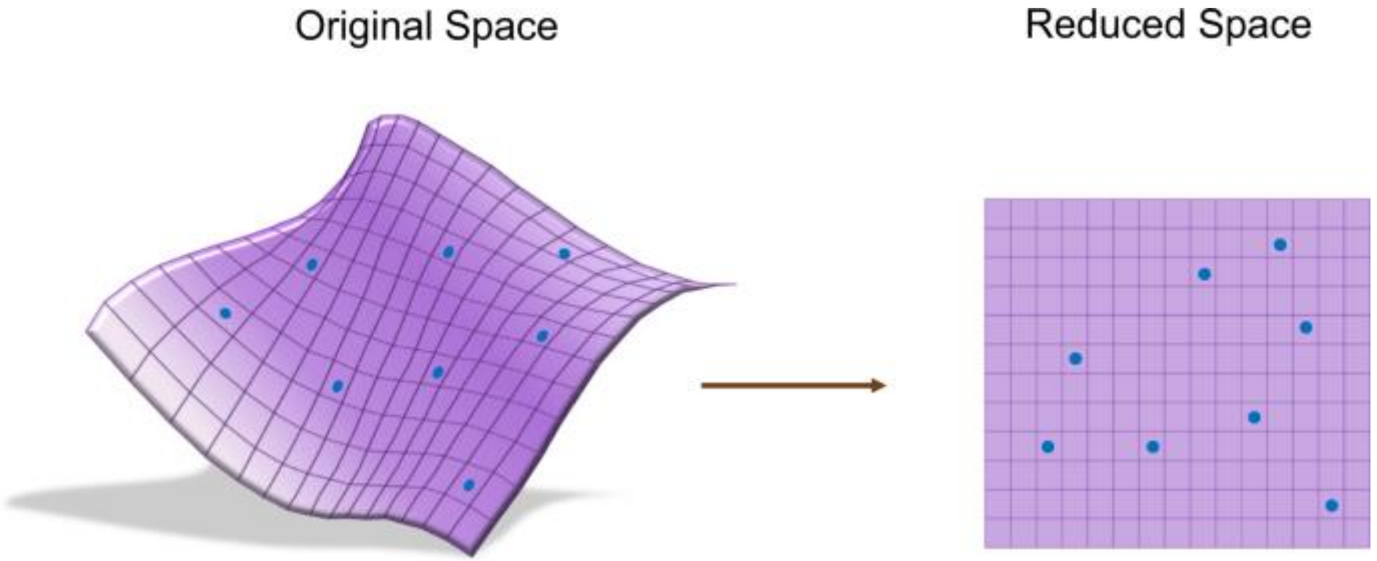


Image source: https://www.nature.com/articles/s41524-020-0276-y

Principal Component Analysis (PCA)

- Principal Component Analysis is a popular unsupervised learning technique for reducing the dimensionality of data
- It increases interpretability yet, at the same time, it minimizes information loss
- It helps to find the most significant features in a dataset and makes the data easy for plotting in 2D and 3D

- Applications:
  - PCA is used to visualize multidimensional data
  - It is used to reduce the number of dimensions in healthcare data
  - PCA can help resize an image
  - It can be used in finance to analyze stock data and forecast returns
  - PCA helps to find patterns in the high-dimensional datasets

What is a Principal Component?

- The Principal Components are straight lines that capture most of the variance of the data
- They have a direction and magnitude
- Principal components are orthogonal projections (perpendicular) of data onto lower-dimensional space

Example:

PC1 is the primary principal component that explains the maximum variance in the data. PC2 is another principal component that is orthogonal to PC1.

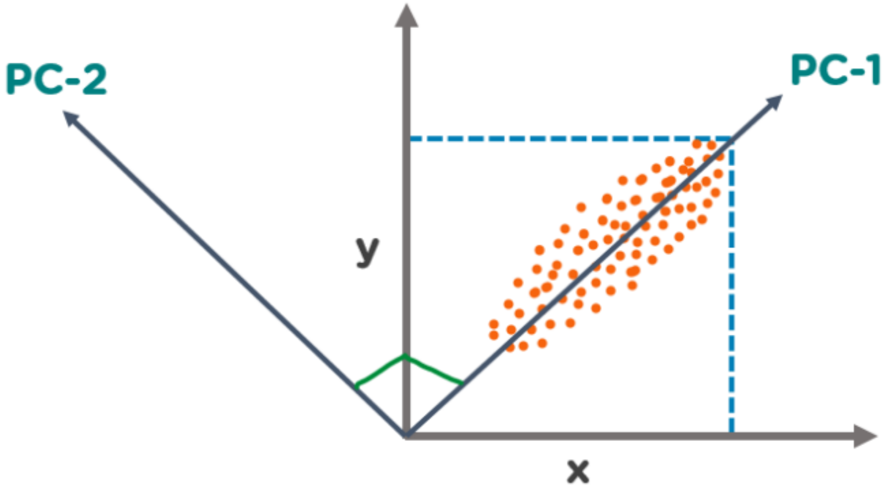


Image source: https://www.simplilearn.com/tutorials/machine-learning-tutorial

How does it work?

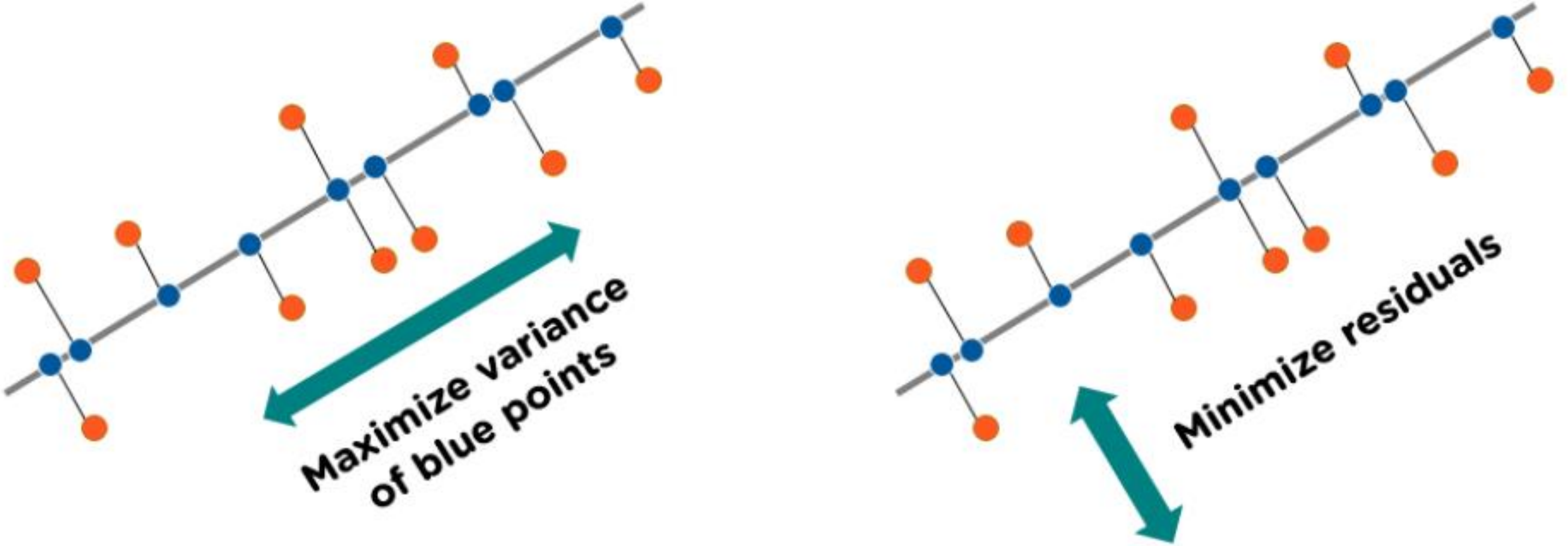
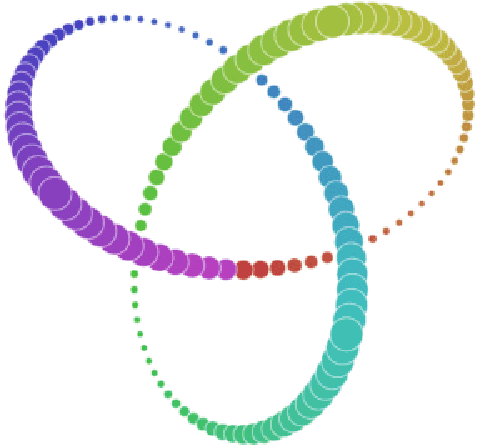


Image source: https://www.simplilearn.com/tutorials/machine-learning-tutorial

T - Distributed Stochastic Neighbor Embedding (T - SNE)

- **t-SNE** was developed by Laurens van der Maaten and Geoffrey Hinton in 2008
- t-SNE is something called **nonlinear dimensionality reduction**
- What that means is this algorithm allows us to separate data that cannot be separated by any straight line
- Examples below won’t return any reasonable results when parsed through PCA, and therefore a nonlinear dimensionality reduction is very important



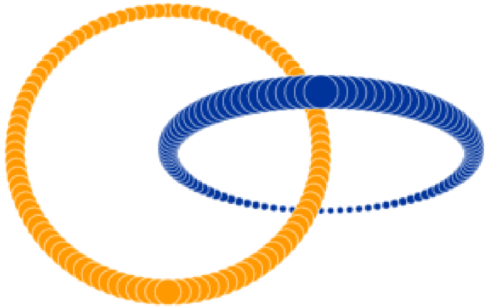


Image source: https://distill.pub/2016/misread-tsne/

ACKNOWLEDGEMENTS

- Materials based on:
  - https://scikit-learn.org/stable/model_selection.html
  - https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47
  - https://towardsdatascience.com/supervised-vs-unsupervised-learning-14f68e32ea8d
  - https://towardsdatascience.com/t-sne-clearly-explained-d84c537f53a
  - https://www.simplilearn.com/tutorials/machine-learning-tutorial