

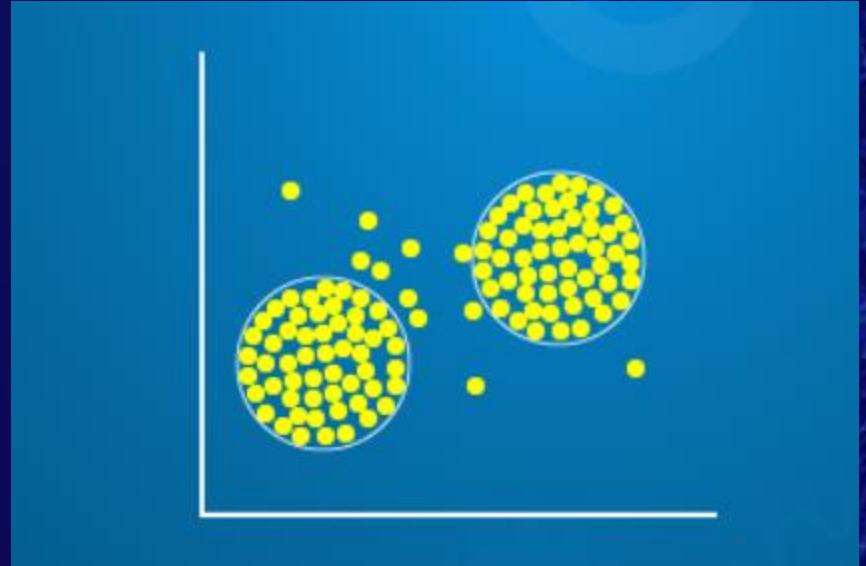


edunet  
foundation



## Unit 3.1

# Unsupervised Learning

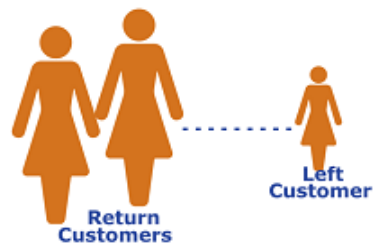


## Disclaimer

The content is curated from online/offline resources and used for educational purpose only



**Customer Segmentation  
and Target Marketing**



**ADDRESSING CUSTOMER CHURN**



**CHATBOTS**



**IDC survey says  
83% companies use  
predictive LEAD SCORING**



**Dynamic Pricing Models  
Buy 2 Get 3 – Buy 4 Get 7**



**Content Creation  
Optimization &  
Deployment**



**Sentiment Analysis**

**Personalised Customer  
Experience**

## Learning Objectives

- Supervised and Unsupervised
- Clustering
- Case Study
- Types of Clustering Methods
- Applications
- Distance Measures
- Different types of Distance Measures



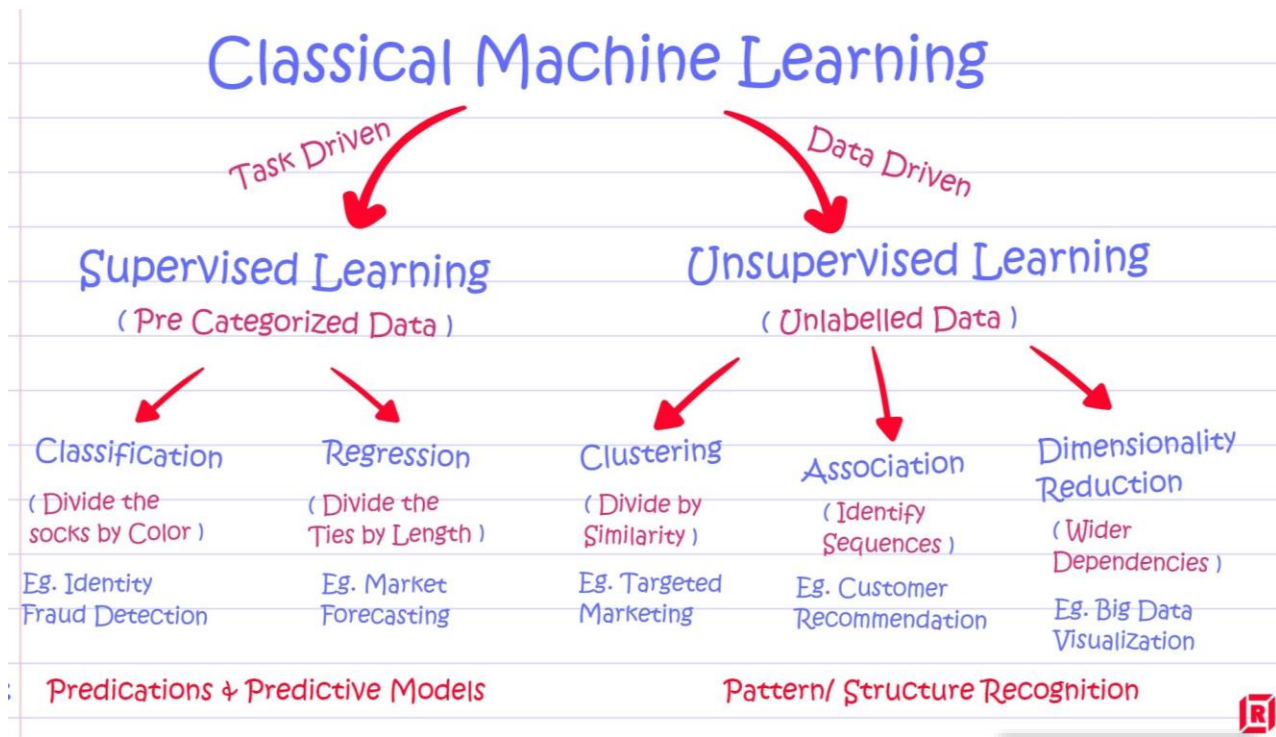
[Click here](#)

[Reference link](#)

## Supervised Learning vs. Unsupervised Learning

Supervised learning: Discover patterns in the data that relate data attributes with a target (class) attribute.

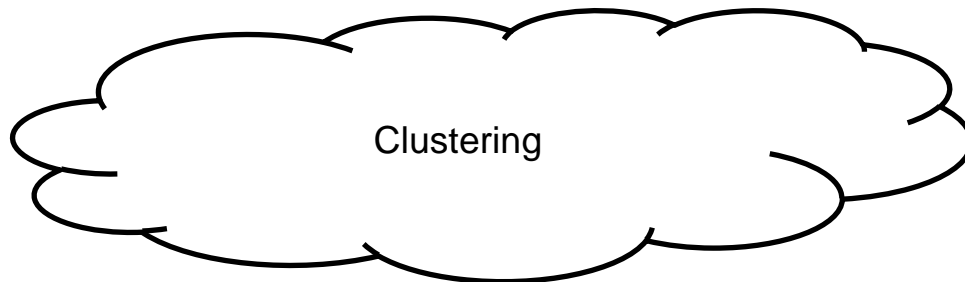
Unsupervised machine learning uses unlabelled data and learns on itself without any supervision. Aim is to group or categorize the unsorted dataset according to the similarities, patterns, and differences.



[Reference link](#)

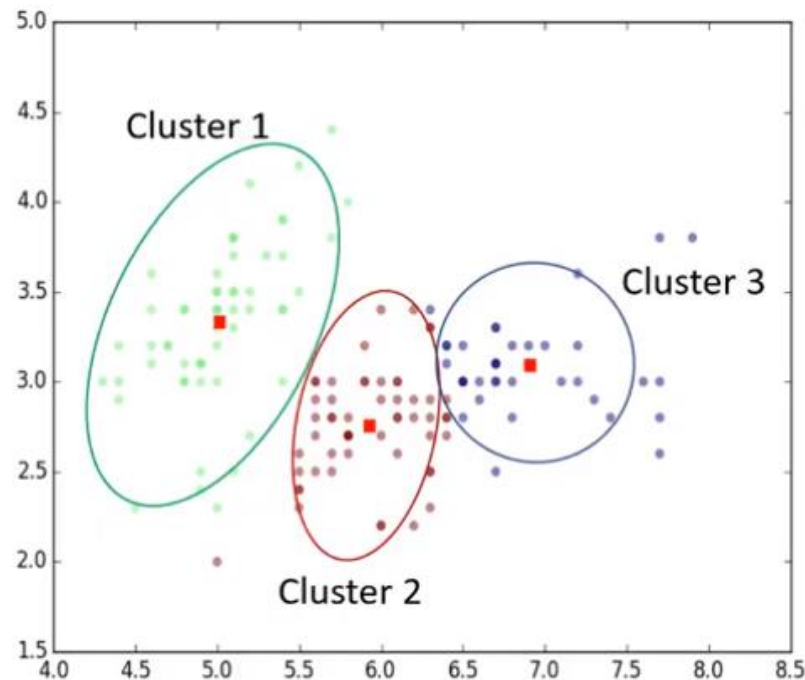
## Clustering

- Imagine that you are a Data Scientist working for a retail company and your boss requests for the customers' segmentation into the following groups: low, average, medium, or platinum customers based on spending behaviour for targeted marketing purposes and product recommendations.
- Knowing that there is no such historical label associated with those customers, how is it possible to categorize them?



## What is Clustering?

- Clustering is an unsupervised machine-learning technique used to find clusters in dataset.
- So, what is a cluster?
- A cluster is group of data points or objects in a dataset that are similar to other objects in the group, and dissimilar to data points in other clusters.



[Reference link](#)

## Case Study

- You have a customer dataset, and you need to apply customer segmentation on this historical data.
- Customer segmentation is the practice of partitioning a customer base into groups of individuals that have similar characteristics.
- Helps business to target specific groups of customers (ex: high profit & low risk customers or non-profile organization customers) by to more effectively allocating marketing resources.

Customer Id	Age	Edu	Years Employed	Income	Card Debt	Other Debt	Address	DebtIncomeRatio	Defaulted
1	41	2	6	19	0.124	1.073	NBA001	6.3	0
2	47	1	26	100	4.582	8.218	NBA021	12.8	0
3	33	2	10	57	6.111	5.802	NBA013	20.9	1
4	29	2	4	19	0.681	0.516	NBA009	6.3	0
5	47	1	31	253	9.308	8.908	NBA008	7.2	0
6	40	1	23	81	0.998	7.831	NBA016	10.9	1
7	38	2	4	56	0.442	0.454	NBA013	1.6	0
8	42	3	0	64	0.279	3.945	NBA009	6.6	0
9	26	1	5	18	0.575	2.215	NBA006	15.5	1

[Reference link](#)



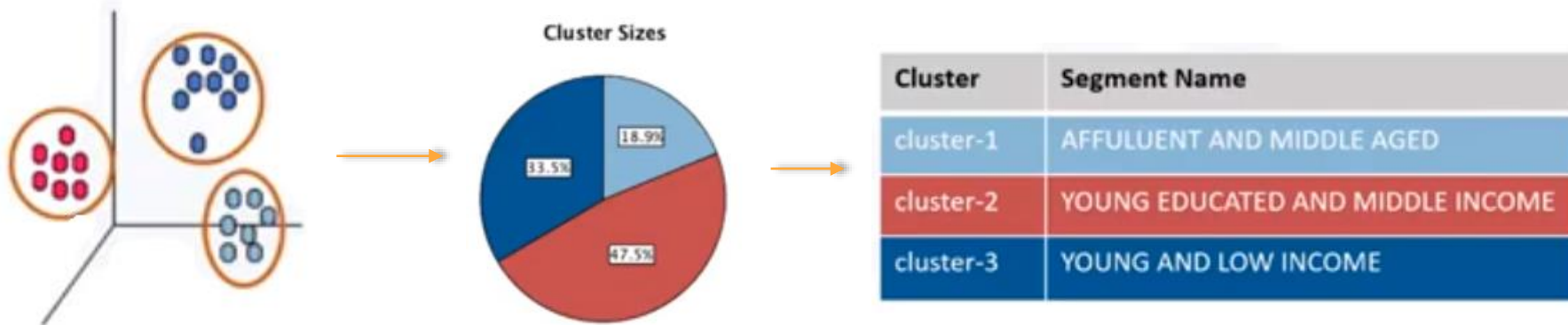
## Case Study...

- A general segmentation process is not usually feasible for large volumes of varied data.
- Therefore, you need an analytical approach to deriving segments and groups from large data sets.
- Customers can be grouped based on several factors: including age, gender, interests, spending habits, and so on.
- The important requirement is to use the available data to understand and identify how customers are similar to each other.
- One of the most adopted approaches that can be used for customer segmentation is clustering.



## Clustering

- Clustering can group data only “unsupervised,” based on the similarity of customers to each other.
- It will partition your customers into mutually exclusive groups, ex: into 3 clusters.
- The customers in each cluster are similar to each other demographically.
- Now we can create a profile for each group, considering the common characteristics of each cluster.



[Reference link](#)

## Similarity Assessment

- Finally, we can assign each individual in our dataset to one of these groups or segments of customers.

Customer Id	Age	Edu	Years Employed	Income	Card Debt	Other Debt	Address	DebtIncomeRatio	Defaulted
1	41	2	6	19	0.124	1.073	NBA001	6.3	0
2	47	1	26	100	4.582	8.218	NBA021	12.8	0
3	33	2	10	57	6.111	5.802	NBA013	20.9	1
4	29	2	4	19	0.681	0.516	NBA009	6.3	0
5	47	1	31	253	9.308	8.908	NBA008	7.2	0
6	40	1	23	81	0.998	7.831	NBA016	10.9	1
7	38	2	4	56	0.442	0.454	NBA013	1.6	0
8	42	3	0	64	0.279	3.945	NBA009	6.6	0
9	26	1	5	18	0.575	2.215	NBA006	15.5	1

Customer ID	Segment
1	YOUNG AND LOW INCOME
2	AFFLUENT AND MIDDLE AGED
3	AFFLUENT AND MIDDLE AGED
4	YOUNG AND LOW INCOME
5	AFFLUENT AND MIDDLE AGED
6	AFFLUENT AND MIDDLE AGED
7	YOUNG AND LOW INCOME
8	YOUNG AND LOW INCOME
9	AFFLUENT AND MIDDLE AGED

[Reference link](#)

## Inference

- When we cross-join this segmented dataset, with the dataset of the product or services that customers purchase from our company.
- Would really help us to understand and predict the differences in individual customers' preferences and their buying behaviors across various products, allowing your company to provide highly personalized experiences for each segment.

Customer Id	Age	Edu	Years Employed	Income	Card Debt	Other Debt	Address	DebtIncomeRatio	Defaulted
1	41	2	6	19	0.124	1.073	NBA001	6.3	0
2	47	1	26	100	4.582	8.218	NBA021	12.8	0
3	33	2	10	57	6.111	5.802	NBA013	20.9	1
4	29	2	4	19	0.681	0.516	NBA009	6.3	0
5	47	1	31	253	9.308	8.908	NBA008	7.2	0
6	40	1	23	81	0.998	7.831	NBA016	10.9	1
7	38	2	4	56	0.442	0.454	NBA013	1.6	0
8	42	3	0	64	0.279	3.945	NBA009	6.6	0
9	26	1	5	18	0.575	2.215	NBA006	15.5	1

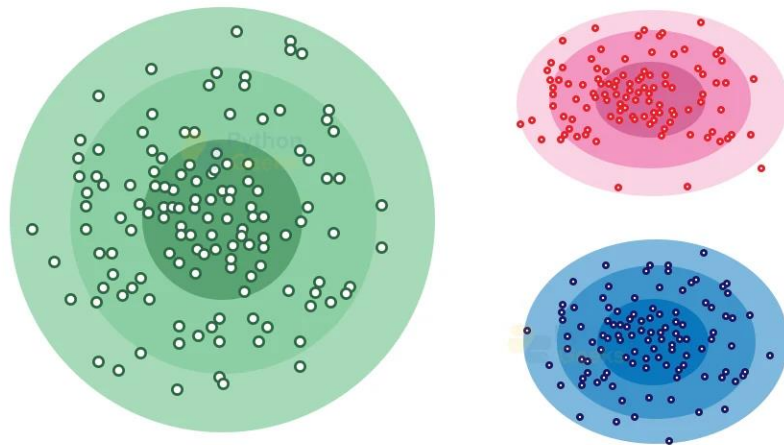
Customer ID	Segment
1	YOUNG AND LOW INCOME
2	AFFLUENT AND MIDDLE AGED
3	AFFLUENT AND MIDDLE AGED
4	YOUNG AND LOW INCOME
5	AFFLUENT AND MIDDLE AGED
6	AFFLUENT AND MIDDLE AGED
7	YOUNG AND LOW INCOME
8	YOUNG AND LOW INCOME
9	AFFLUENT AND MIDDLE AGED

[Reference link](#)

## Types of Clustering Methods

Different types of clustering methods are:

- Centroid-based or Partition Clustering
- Connectivity-based Clustering (Hierarchical Clustering)
- Density-based Clustering (Model-based Methods)
- Distribution-Based Clustering

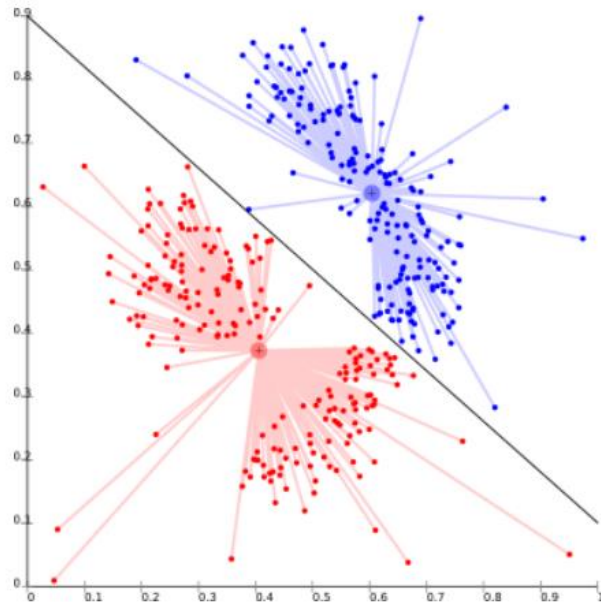


[Reference link](#)

## Types of Clustering Methods

### Centroid-based or Partition Clustering:

- Organizes the data into non-hierarchical clusters.
- Centroid-based clustering methods partition the data points/objects into 'k' number of clusters.
- These clustering methods iteratively measure the distance between each data point and its nearest cluster's centroid using various distance metrics.
- Performance is affected when there is more noise in the data.
- K means clustering lies in this category.



[Reference link](#)

## Types of Clustering Methods

### Connectivity-based Clustering (Hierarchical Clustering)

- This category of model is based on the idea that objects are more related to nearby objects than those further away. Clusters are thus developed based on distance between objects in the data space.
- These model do not scale well to big datasets.

### Density-based Clustering (Model-based Methods)

- Density-Based Clustering refers to unsupervised machine learning methods that identify distinctive clusters in the data, based on the idea that a cluster/group in a data space is a contiguous region of high point density, separated from other clusters by sparse regions.
- DBSCAN method is an example of Density based clustering.

## Types of Clustering Methods

### Distribution Based Clustering

- Distribution based clustering is a type of hierarchical clustering that is used when the distribution of data points is known such as Gaussian or normal distribution.
- As distance from the distribution's centre increases, the probability that a point belongs to the distribution decreases
- The following picture shows three different distribution-based clusters.

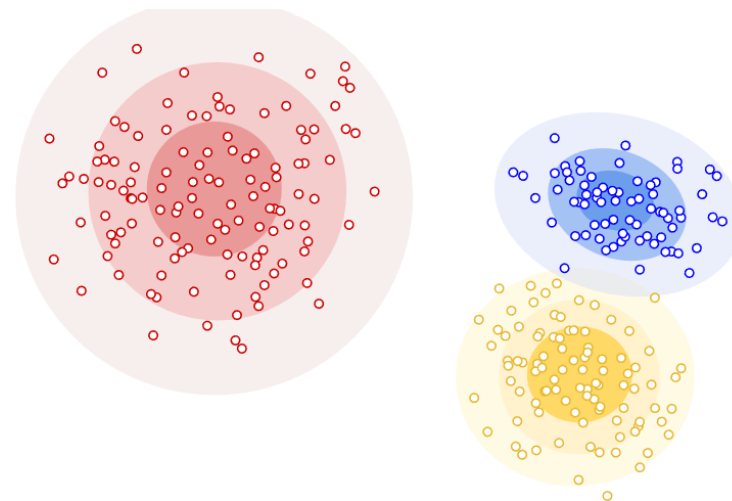


Image: Density Based Clustering

Click here

[Reference link](#)



## Applications

- Customer segmentation is one of the popular usages of clustering.
- Cluster analysis also has many other applications in different domains.
- Retail/Marketing
  - Based on customer's demographic characteristics, identifying buying patterns of customer groups
  - Recommending new books or movies to customers
- Banking
  - Find clusters of normal transactions to find pattern of fraudulent credit card usage i.e. anomaly detection.
  - Identify clusters of customers to find loyal customers versus churn customers.

## More Applications

- Insurance
  - Analyse genuine claim clusters to identify false claim fraud.
  - Evaluate insurance risk of customers based on their segments
- Publication
  - Clustering is used to auto-categorize news based on its content,
  - Tag news, then cluster it, so as to recommend similar news articles to readers
- Medicine
  - Characterize patient behaviour, based on their similar characteristics, so as to identify successful medical therapies for different illnesses.

## Where Clustering can be used ?

Clustering can be used for one of the following purposes:

- Exploratory data analysis
- Summary generation or reducing the scale
- Outlier detection, especially to be used for fraud detection
- Noise removal
- Finding duplicates in datasets
- Pre-processing step for either prediction, other data mining tasks
- Part of a complex system.

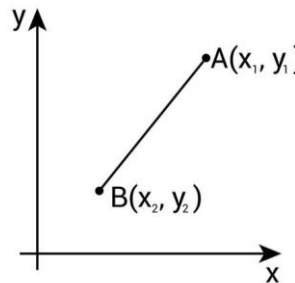


## Distance Measures

- Distance measures play an important role in machine learning.
- Determines the similarity between two elements and it influences the shape of the clusters.
- Provide the foundation for many popular and effective machine learning algorithms like k-nearest neighbours for supervised learning and k-means clustering for unsupervised learning.

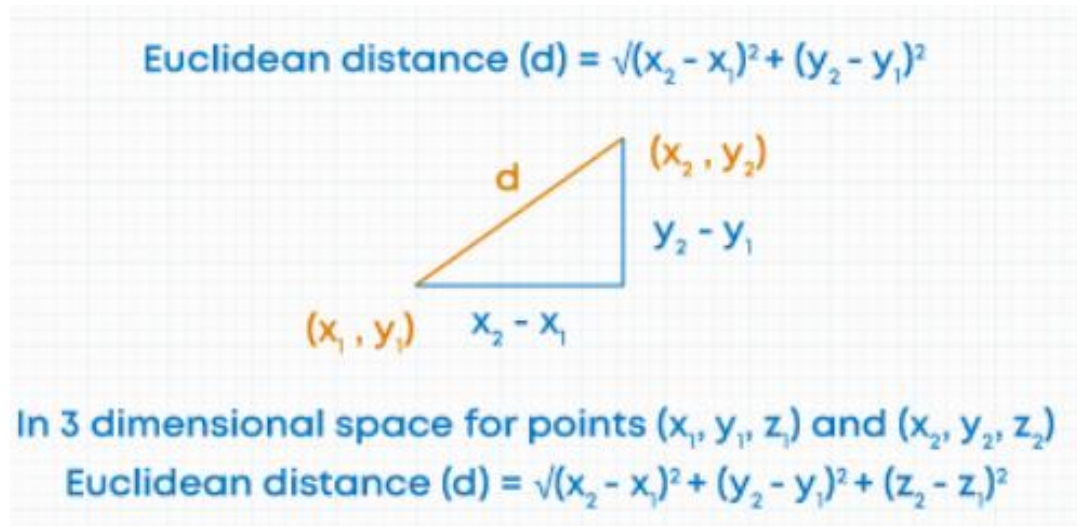
Different distance measures are given below:

- Euclidean Distance
- Manhattan Distance
- Minkowski Distance
- Mahalanobis Distance
- Hamming Distance
- Cosine Similarity
- Chebyshev distance


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

## Distance Measures

- **Euclidean distance:** Most widely used distance measure when the variables are continuous.
- The Euclidean distance between two points calculates the length of a segment connecting the two points.

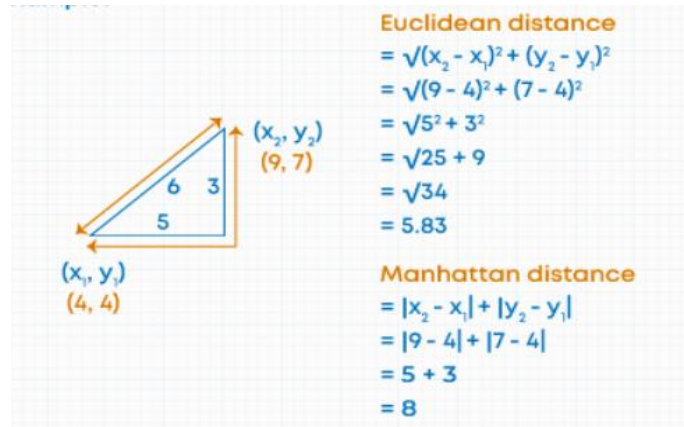
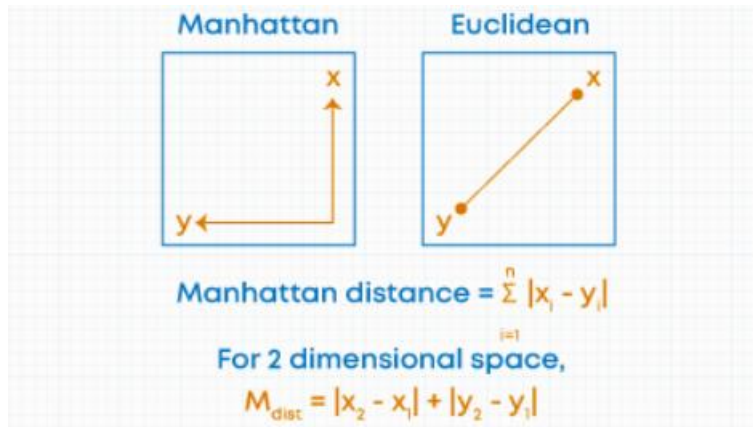


Click here

[Reference link](#)

## Distance Measures

- **Manhattan Distance:** is the sum of absolute differences between points across all the dimensions.
- If we wanted to measure a distance between two retail stores in a city, then Manhattan distance will be more suitable to use, instead of Euclidean distance.

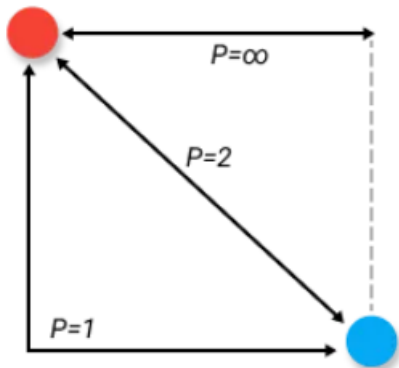


[Click here](#)

[Reference link](#)

## Distance Measures

- **Minkowski Distance** : It is a generalization of the Euclidean and Manhattan distance measures and adds a parameter, called the “order” or “p”, that allows different distance measures to be calculated.
- It determines the similarity of distances between two or more vectors in space.



$$D(x, y) = \left( \sum_{i=1}^n |x_i - y_i|^p \right)^{\frac{1}{p}}$$

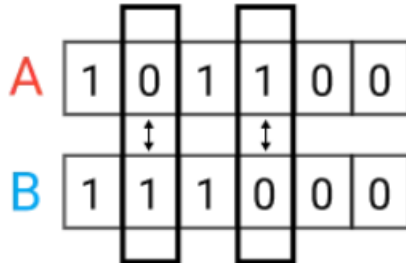
Image: Minkowski distance (left) , Minkowski distance formula (right)

[Click here](#)

[Reference link](#)

## Distance Measures

- **Hamming Distance** : Hamming distance is the number of values that are different between two vectors.
- It is typically used to compare two binary strings of equal length. It can also be used for strings to compare how similar they are to each other by calculating the number of characters that are different from each other.



[Click here](#)

[Reference link](#)



## Distance Measures

- **Mahalanobis Distance:** Mahalanobis distance is an effective multivariate distance metric that measures the distance between a point (vector) and a distribution.
- It has excellent applications in multivariate anomaly detection, classification on highly imbalanced datasets and one-class classification.
- The formula to compute Mahalanobis distance is as follows:

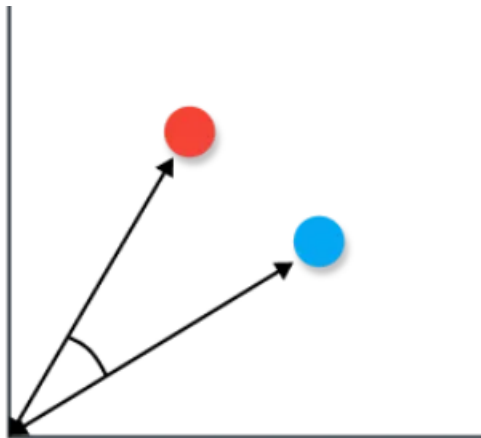
Where:

- $x$  is the vector of observation.
- $m$  is a vector of mean values of each column.
- $C$  is a Covariance Matrix

$$D^2 = (x - m)^T \cdot C^{-1} \cdot (x - m)$$

## Distance Measures

- **Cosine similarity:** Cosine similarity has often been used as a way to counteract Euclidean distance's problem with high dimensionality. The cosine similarity is simply the cosine of the angle between two vectors.

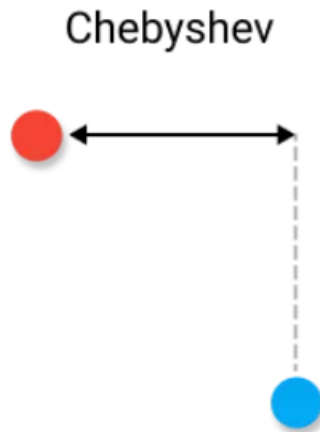


Click here

[Reference link](#)

## Distance Measures

- **Chebyshev distance:** Chebyshev distance is defined as the greatest of difference between two vectors along any coordinate dimension.
- In other words, it is simply the maximum distance along one axis. Due to its nature, it is often referred to as Chessboard distance since the minimum number of moves needed by a king to go from one square to another is equal to Chebyshev distance.



$$D(x, y) = \max_i (|x_i - y_i|)$$

Click here

[Reference link](#)

## Summary

**1.Clustering Unveiled:** Clustering is a powerful unsupervised learning technique that groups similar data points, enabling insights, personalization, and decision-making across various domains.

**2.Distance Measures at the Core:** Distance measures serve as the foundation for assessing similarity between data points. They influence clustering outcomes, shape algorithms, and play roles in anomaly detection, classification, and more.

**3.Distance Variety:** Our exploration covered a range of distance measures including Euclidean, Manhattan, Minkowski, Mahalanobis, Hamming, Cosine similarity, and Chebyshev. Each measure caters to specific scenarios and data types.

**4.Practical Applications:** Clustering finds applications in customer segmentation, fraud detection, healthcare, and more. Distance measures empower diverse domains, from urban navigation to image processing, with insights and informed decisions.

## Quiz

1. What is the primary objective of unsupervised learning?
- A. To classify data into predefined categories
  - B. To predict a target variable based on input features
  - C. To discover hidden patterns or structures in data
  - D. To minimize errors in model predictions

**Answer: C**

## Quiz

2. Which of the following is an example of an unsupervised learning algorithm?

- A. Linear Regression
- B. K-Means Clustering
- C. Decision Trees
- D. Support Vector Machines

**Answer: B**

## Quiz

3. In unsupervised learning, when data points are grouped together based on similarity, it is called:

- A. Regression
- B. Classification
- C. Clustering
- D. Feature extraction

**Answer: C**

## Quiz

4. Which type of learning is most suitable for anomaly detection?

- A. Supervised learning
- B. Unsupervised learning
- C. Semi-supervised learning
- D. Reinforcement learning

**Answer: B**



## Quiz

5. What is the main difference between supervised and unsupervised learning?
- A. Supervised learning requires labeled data, while unsupervised learning does not.
  - B. Supervised learning always involves classification, while unsupervised learning involves regression.
  - C. Supervised learning can handle larger datasets than unsupervised learning.
  - D. Supervised learning is faster and more accurate than unsupervised learning.

**Answer: A**

## References

- <https://en.wikipedia.org/wiki/Git>
- [https://en.wikipedia.org/wiki/Unsupervised\\_learning](https://en.wikipedia.org/wiki/Unsupervised_learning)
- [https://scikit-learn.org/stable/unsupervised\\_learning.html](https://scikit-learn.org/stable/unsupervised_learning.html)
- <https://www.coursera.org/learn/machine-learning>
- <http://cs229.stanford.edu/notes2020spring/cs229-notes8.pdf>
- <https://developers.google.com/machine-learning/clustering>
- <https://archive.ics.uci.edu/ml/index.php>
- <https://towardsdatascience.com/tagged/unsupervised-learning>
- <https://www.kaggle.com/kernels>

# Thank you...!