UNSUPERVISED MACHINE LEARNING (CLUSTERING)

Machine Learning – Yousef Elbaroudy

GUIDELINES

- Try to focus on the important information mentioned through the session
- Apply what you take on the practical section
- Do not try to memorize everything you got, just learn
- Don’t mind to ask about anything you want to know

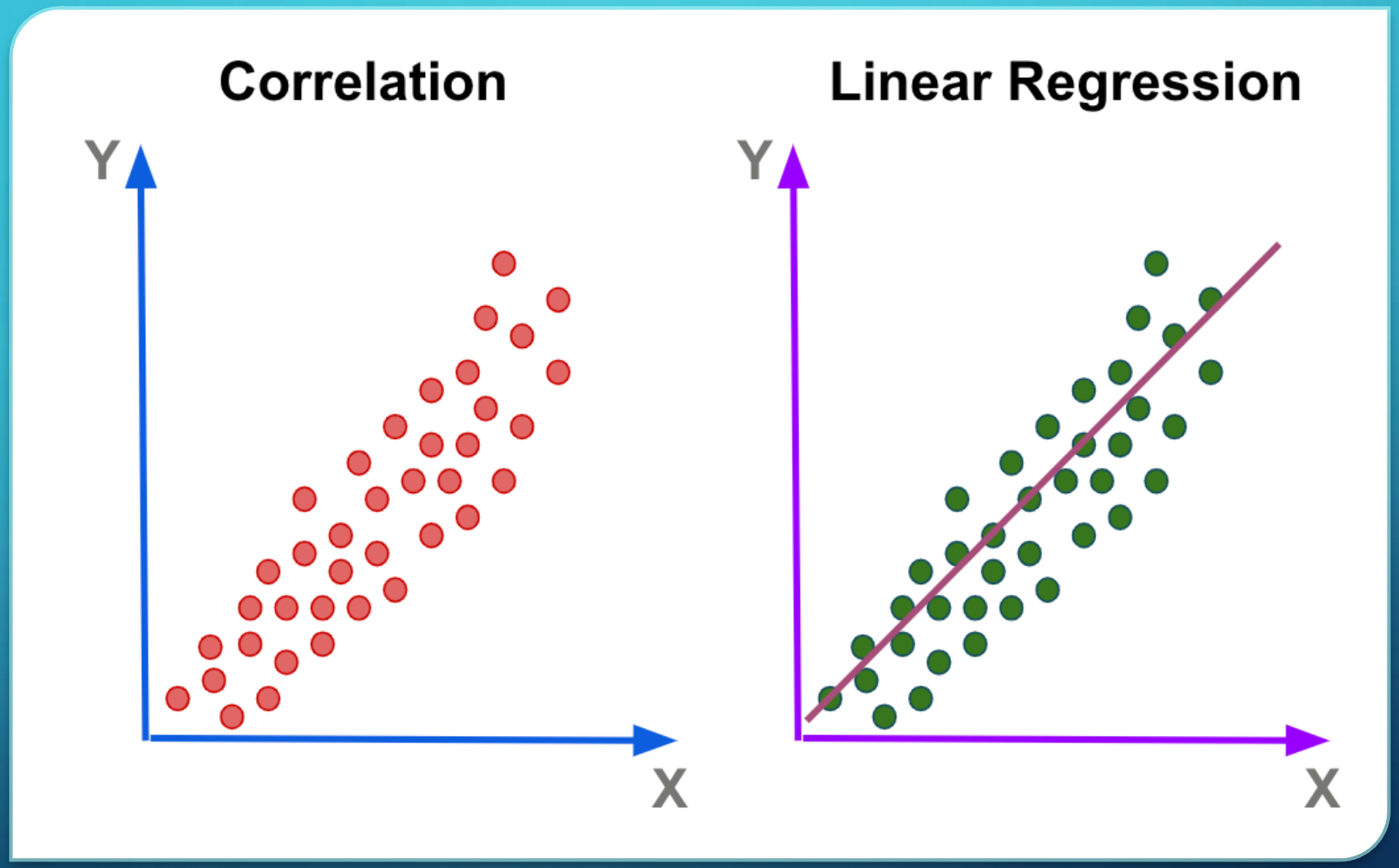
Enjoy the Session ☺

EVALUATION METRICS (REGRESSION)

COEFFICIENTS OF DETERMINATION

- The **coefficient of determination**, often denoted as R-squared (R2), is a statistical measure that represents the proportion of the variance in the dependent variable (the outcome) that can be explained by the independent variables (predictors) in a regression model.

- In other words, it quantifies the goodness of fit of a regression model to the data.


Correlation
Y
X
Linear Regression
Y
X

HOW TO CALCULATE R2 ?

R-squared (R2) is calculated based on the variance of the data and the variance explained by the regression model. The formula for calculating R-squared is as follows:

$$R^2 = 1 - \frac{SS_{\text{res}}}{SS_{\text{total}}}$$

Where:

- SS_{res} is the sum of squared residuals (the sum of the squared differences between the actual observed values and the predicted values from the regression model).
- SS_{total} is the total sum of squares (the sum of squared differences between the actual observed values and the mean of the observed values).

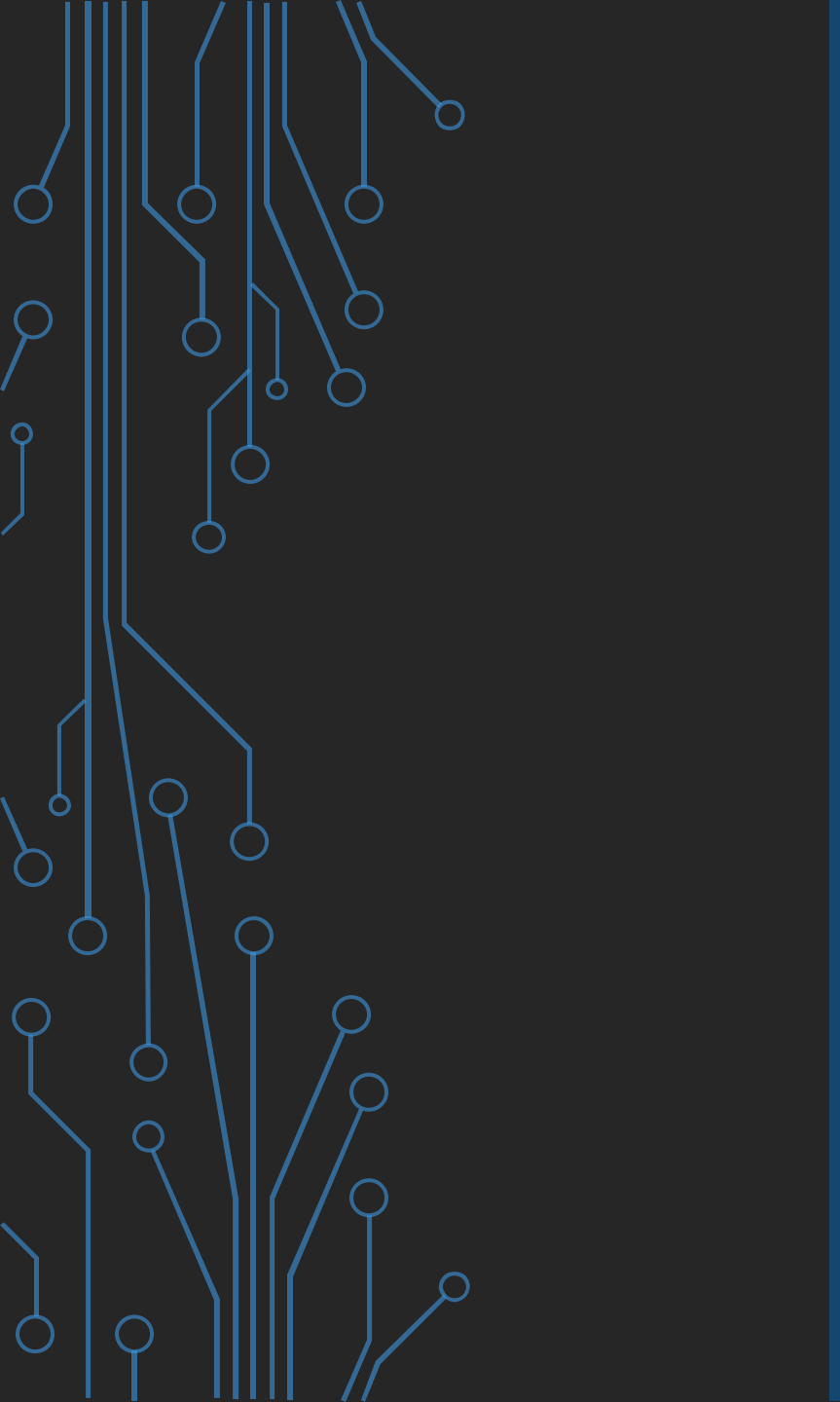
STEPS

1. Calculate the mean of the observed values: $\bar{y}$.
2. Calculate the total sum of squares (SS_{total}):
   $SS_{\text{total}} = \sum(y_i - \bar{y})^2$
3. Build and evaluate your regression model to obtain predicted values ($\hat{y}_i$) for each observation.
4. Calculate the sum of squared residuals (SS_{res}):
   $SS_{\text{res}} = \sum(y_i - \hat{y}_i)^2$
5. Plug the values of SS_{res} and SS_{total} into the R-squared formula:
   $R^2 = 1 - \frac{SS_{\text{res}}}{SS_{\text{total}}}$

R-squared values range between 0 and 1, with higher values indicating a better fit of the model to the data. Here's what the R-squared value means:

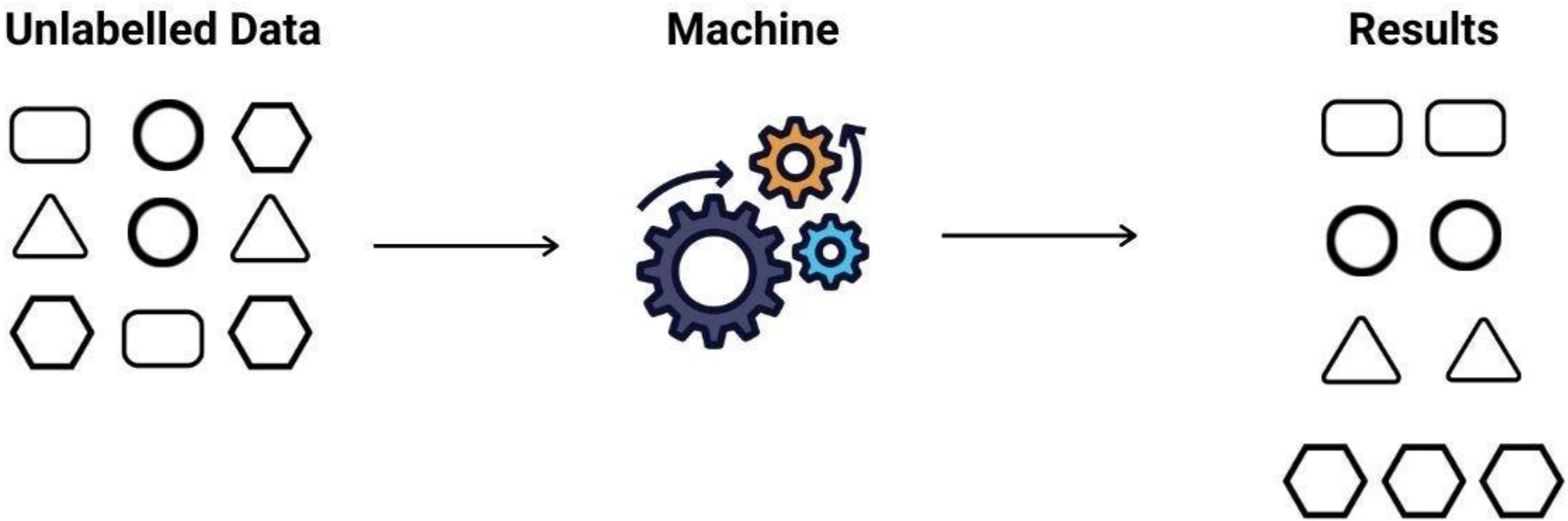
- R2 = 0: The model explains none of the variability in the dependent variable. It's essentially a poor fit.
- R2 = 1: The model explains all of the variability in the dependent variable. It's a perfect fit.

WHAT IT INDICATES TO ?



UNSUPERVISED MACHINE LEARNING

Unsupervised Learning

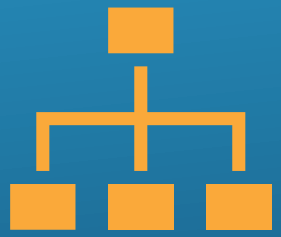


UNSUPERVISED MACHINE LEARNING

- Unsupervised machine learning is a branch of machine learning where the goal is to uncover patterns, relationships, or structures within a dataset without the presence of explicit target labels or predefined outcomes.

- Unlike supervised learning, which involves training a model to make predictions based on labeled data, unsupervised learning focuses on finding inherent structures or groupings within the data itself.

TYPES OF UNSUPERVISED MACHINE LEARNING



Clustering

K-Means
Hierarchical Clustering
DBSCAN

Dimensionality Reduction

Principle Component Analysis (PCA)
Singular Value Decomposition (SVD)
T-Distributed Stochastic Neighbor Embedding (t-SNE)

CLUSTERING

Clustering involves grouping similar data points together into clusters based on certain similarity criteria.

The Goal is to identify patterns or segments within the data.

DIMENSIONALITY REDUCTION

Dimensionality reduction techniques aim to reduce the number of features or variables in the dataset while preserving as much relevant information as possible.

This is useful for visualizing high-dimensional data, reducing noise, and speeding up subsequent analyses.

WHAT IS UNSUPERVISED LEARNING USED FOR ?

FIELDS OF UNSUPERVISED LEARNING

Data Exploration: Unsupervised techniques can help analysts understand the underlying structure of a dataset and discover relationships that might not be obvious at first glance.

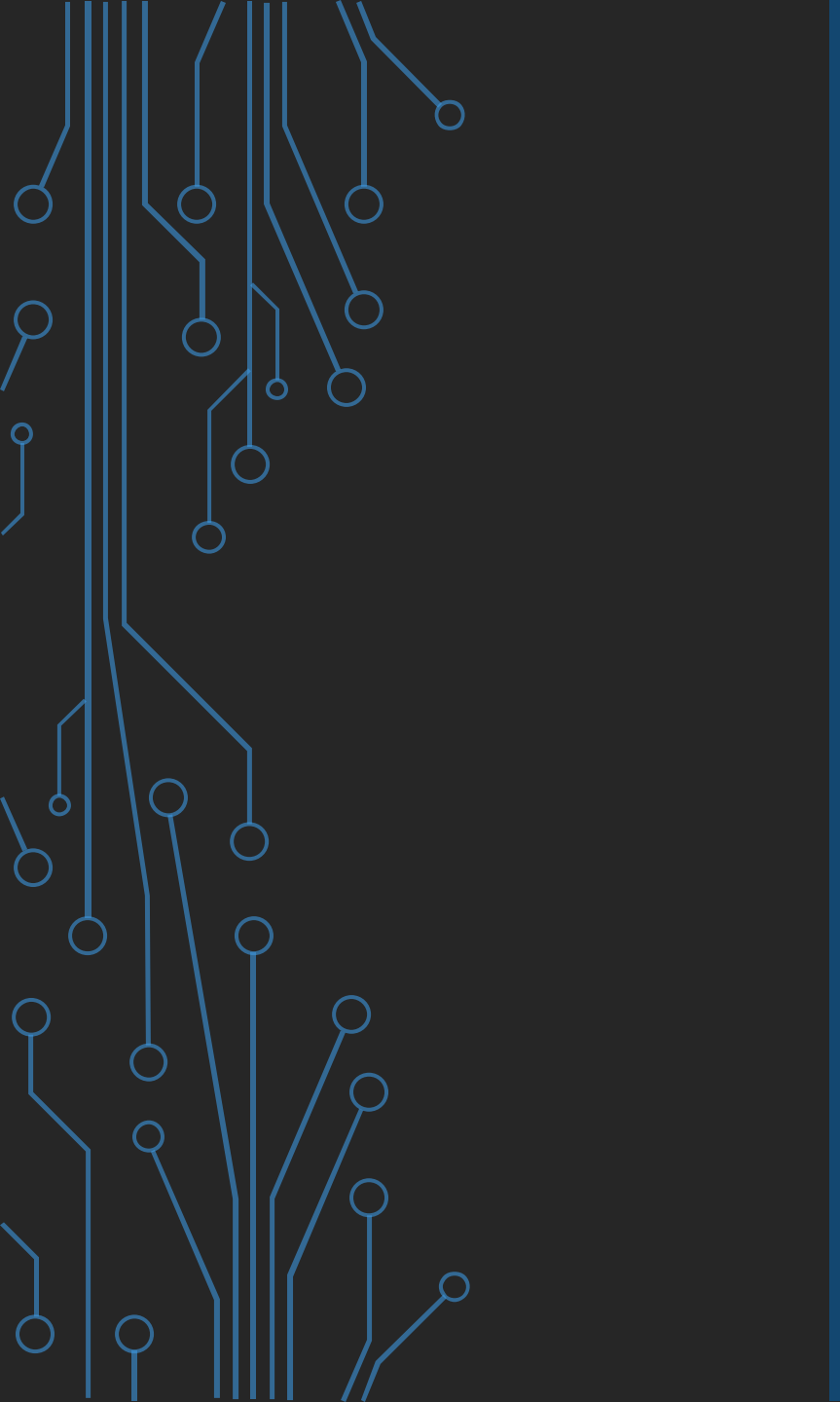
Anomaly Detection: By learning the normal patterns of the data, unsupervised methods can identify unusual or anomalous observations that deviate from those patterns.

Recommendation Systems: Unsupervised techniques can be used to group similar users or items, allowing for the creation of recommendation systems that suggest products.

Feature Engineering: Dimensionality reduction techniques can assist in reducing the complexity of a dataset, which can lead to simpler and more efficient models.

Text Analysis: Unsupervised learning is often used in natural language processing tasks such as topic modeling, where documents are grouped based on the topics they discuss.

BREAK (20 MINUTES)



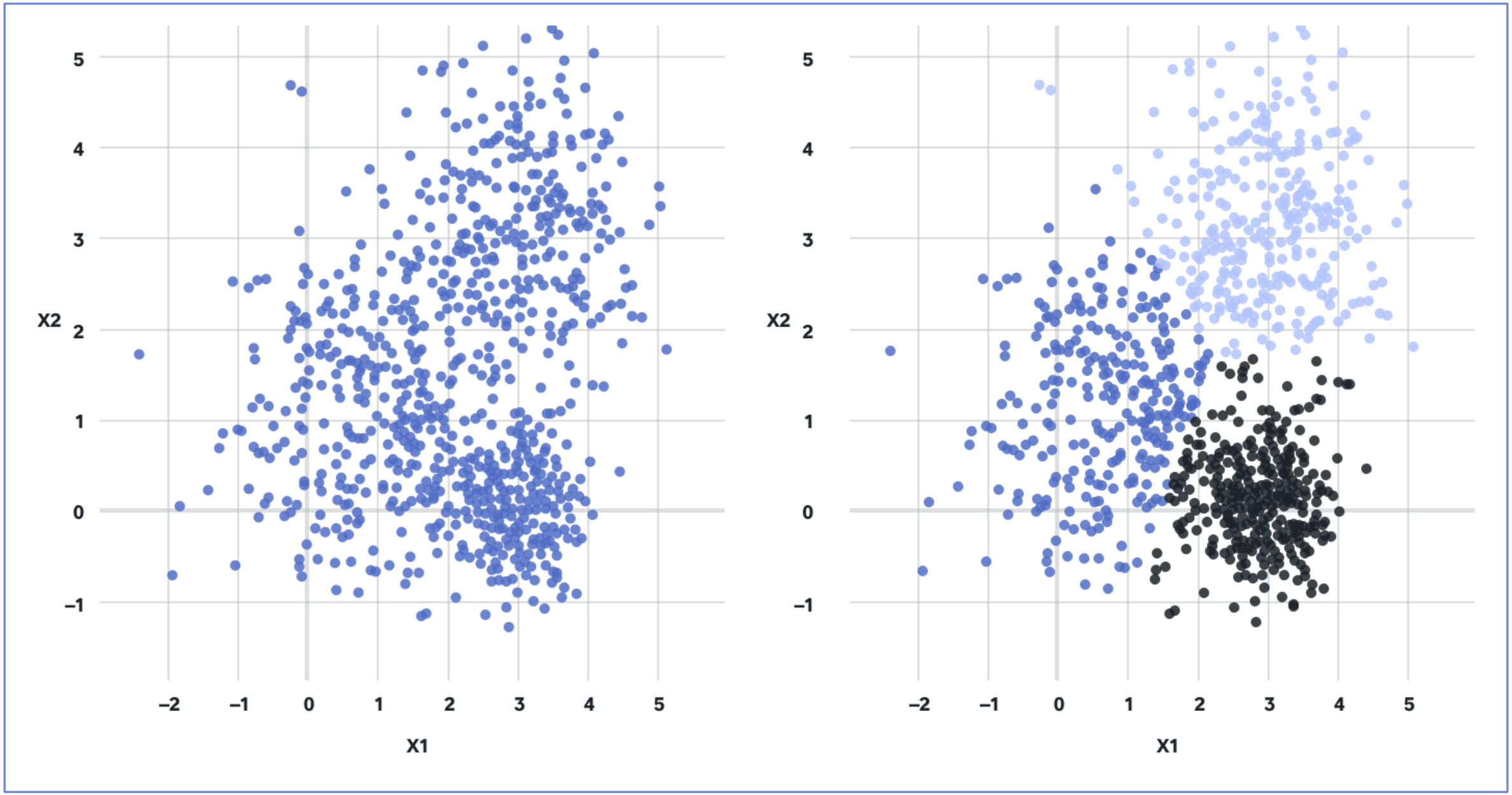
UNSUPERVISED LEARNING (CLUSTERING)

CLUSTERING

- **Clustering** is an unsupervised machine learning technique that involves grouping similar data points together based on certain criteria.

- **The goal of clustering** is to find natural groupings within a dataset without any predefined labels.

- Each group of data points is referred to as a "cluster," and the process of creating these clusters is known as "clustering."

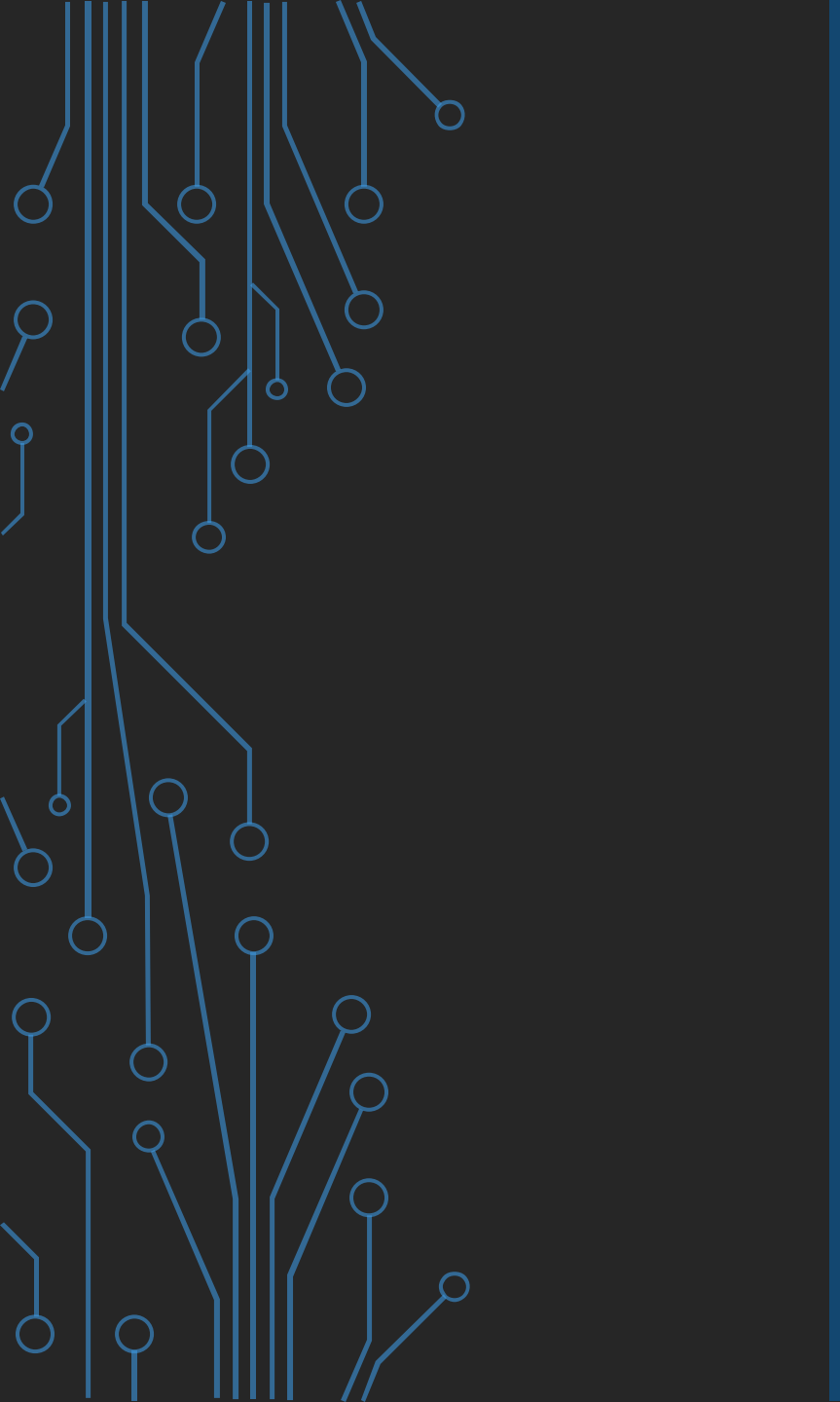

X2
X1
5
4
3
2
1
0
−1
−2
−1
0
1
2
3
4
5

ALGORITHMS FOR CLUSTERING

K-Means Clustering: This algorithm partitions the data into a specified number (k) of clusters by iteratively assigning data points to the nearest cluster center (centroid).

Hierarchical Clustering: Hierarchical clustering builds a hierarchy of clusters by either starting with each data point as its own cluster and merging them or by starting with all data points in a single cluster and recursively dividing them into smaller clusters.

DBSCAN (Density-Based Spatial Clustering of Applications with Noise): DBSCAN groups together data points that are closely packed together while considering regions with lower data density as noise.



K-MEANS CLUSTERING

K-MEANS ALGORITHM

- **K-Means** is one of the most widely used clustering algorithms.
- It's a partitioning method that divides a dataset into a specified number of clusters (k) based on the similarity of data points.
- The algorithm iteratively assigns data points to clusters and updates the cluster centers until convergence.

CENTROID

- A **centroid** is a ***CENTRAL POINT*** that represents a group of data points within a cluster in a clustering algorithm
- the centroid serves as a reference point around which the data points in a cluster are grouped.
- It's used to calculate the "center" of the cluster and is updated iteratively as the clustering algorithm progresses.

HOW IT WORKS ?

INITIALIZATION

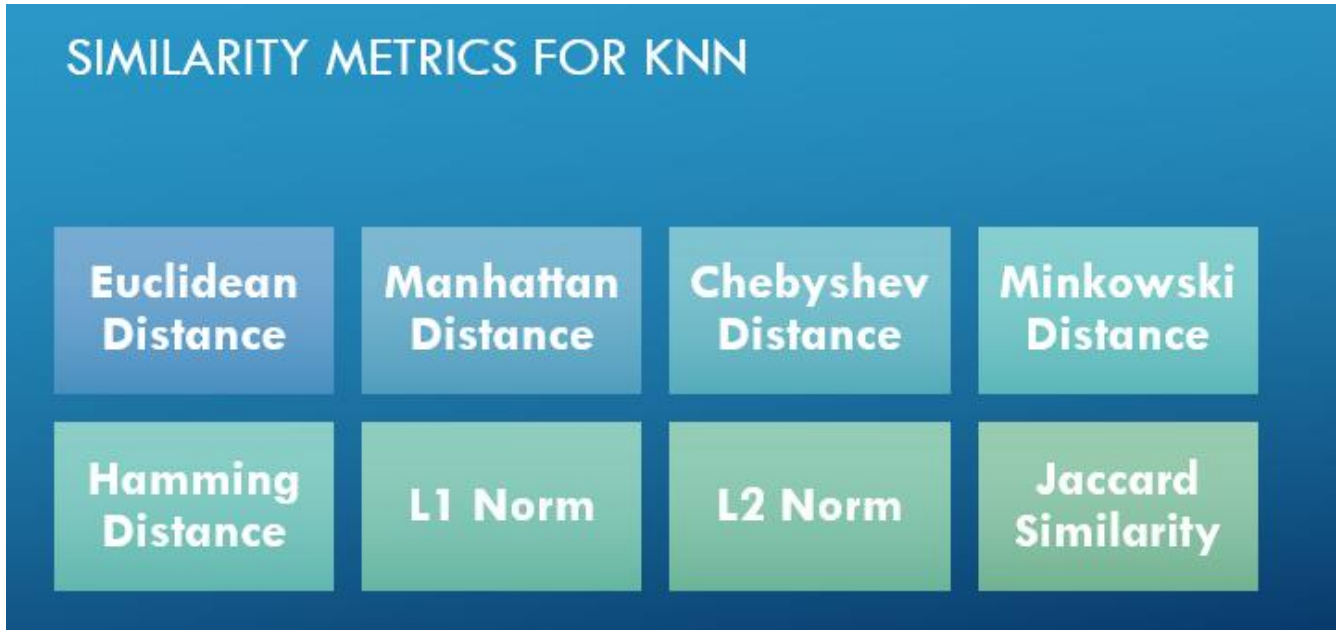
Choose the number of clusters (k) you want to create.

Initialize k cluster centroids randomly by selecting k data points from the dataset. These centroids will serve as the initial cluster centers.

ASSIGNMENT STEP

For each data point in the dataset, calculate the distance (**USING ANY DISTANCE METRIC**) to each of the k cluster centroids.

Assign the data point to the cluster whose centroid is the closest (i.e., the cluster with the minimum distance).



REVISE: DISTANCE METRICS

UPDATE STEP

After all data points are assigned to clusters, calculate the mean (average) of the data points in each cluster. This mean becomes the new centroid for that cluster.

Update the cluster centroids to be the newly calculated means.

ITERATION

Repeat the Assignment and Update steps for a predefined number of iterations or until a **stopping criterion** is met.

The algorithm iteratively refines the cluster assignments and updates centroids to improve the fit.

CONVERGENCE

- The algorithm will converge when the cluster assignments and centroids no longer change significantly between iterations.

HOW TO CHOOSE APPROPRIATE NUMBER OF K ?

COST FUNCTION OF K-MEANS

- The cost function of K-Means, also known as the objective function or distortion function, quantifies how well the data points in a cluster are grouped around their respective centroids.

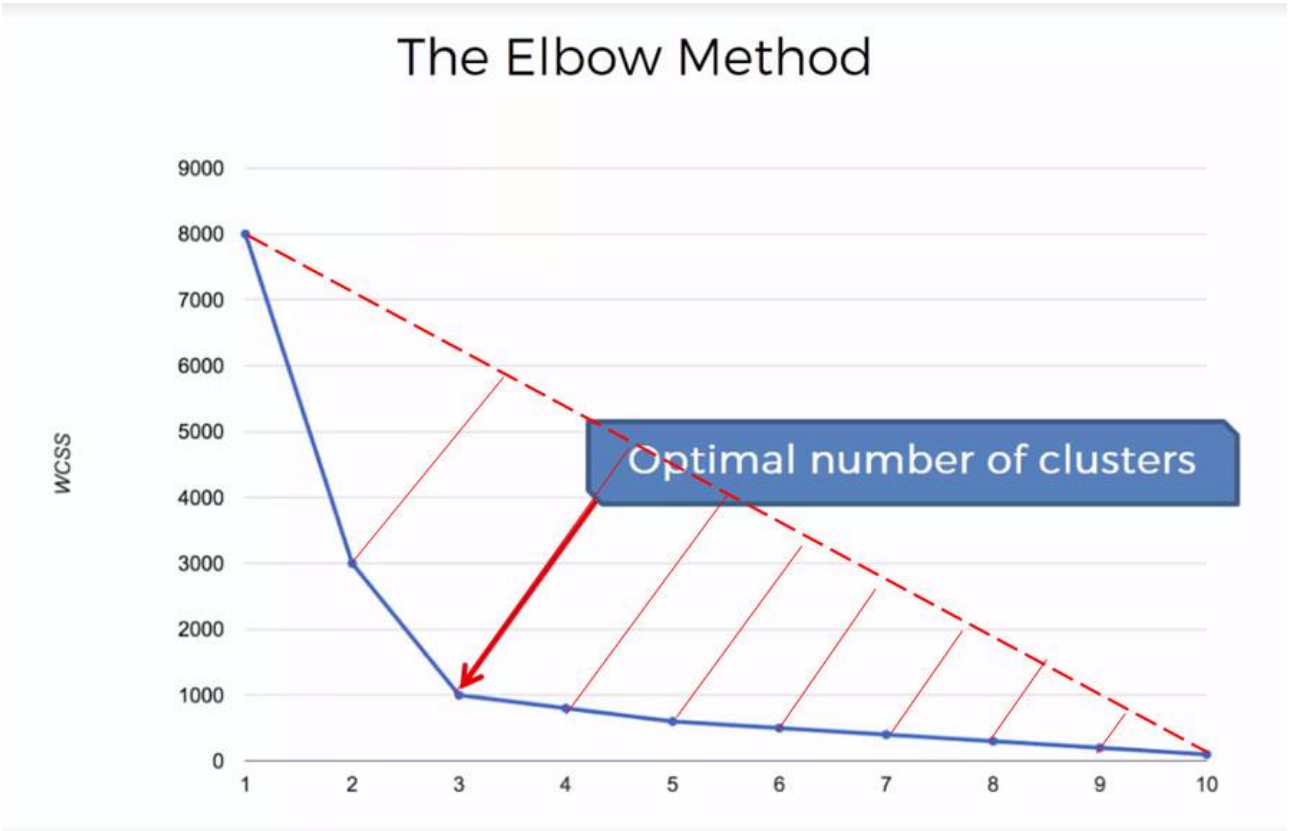
- The goal of the K-Means algorithm is to minimize this cost function, which essentially means **MINIMIZING** **the sum of squared distances between data points and their assigned cluster centroids**.

$$J = \sum_{i=1}^{k} \sum_{x \in C_i} ||x - \mu_i||^2$$

Where:

- k is the number of clusters.
- C_i represents the i-th cluster.
- x is a data point in cluster C_i.
- μ_i is the centroid of cluster C_i.
- $||x - \mu_i||^2$ is the squared Euclidean distance between data point x and the centroid μ_i.

COST FUNCTION FOR DIFFERENT NUMBER OF K (ELBOW METHOD)



ADVANTAGES OF K-MEANS

1. **Simple and Efficient:** K-Means is relatively easy to understand and implement. It's computationally efficient and can handle large datasets.
2. **Scalability:** K-Means can be applied to datasets with a large number of data points, making it suitable for big data scenarios.
3. **Parallelizability:** The algorithm's nature allows it to be parallelized, leading to faster execution on multi-core systems.
4. **Interpretability:** The resulting clusters and centroids are easy to interpret and can provide insights into the data structure.
5. **Well-Suited for Spherical Clusters:** K-Means works well when clusters are approximately spherical and of equal size.
6. **Useful for Preprocessing:** K-Means can be used as a preprocessing step for other algorithms, like anomaly detection or feature engineering.

DISADVANTAGES OF K-MEANS

1. **Sensitive to Initialization:** K-Means is sensitive to the initial placement of centroids, which can result in different final clusterings. Using K-Means++ initialization can mitigate this to some extent.
2. **Assumption of Equal Sized Clusters:** K-Means assumes that clusters are of equal size, which may not be true in some cases.
3. **Assumption of Spherical Clusters:** K-Means assumes that clusters are spherical in shape and can struggle with elongated or irregular clusters.
4. **Number of Clusters (k) Selection:** Determining the optimal number of clusters (k) can be subjective and challenging. Various techniques, like the elbow method and silhouette analysis, can help, but there's no definitive solution.
5. **Sensitive to Outliers:** Outliers can significantly affect the centroid calculation and thus the clustering results.
6. **Convergence to Local Optima:** K-Means can converge to local optima, meaning it might not find the best solution. Running the algorithm with different initializations or using more robust clustering methods might help.
7. **Cannot Handle Non-Numeric Data:** K-Means requires a numeric distance metric, making it unsuitable for datasets with non-numeric attributes or categorical data.
8. **Need to Specify k:** The algorithm requires you to specify the number of clusters (k) beforehand, which might not always be straightforward.

TIME FOR PRACTICALITY (15 MINUTES)

QUESTIONS


THANK YOU