Applied AI & ML Industry Projects Lab (AAM-IPL)

Transforming Learning into Industry Solutions

Week-7

K-Means Clustering Recap and Project Implementation

V Semester - ML and AIUP, Aug-Oct 2024

Session Date and Time: 3rd Nov 2024 10:30 AM IST – 12:00 Noon IST

Venkateswar Reddy Melachervu

Visiting Faculty and CTO, Brillium Technologies

Department of Computer Science Engineering – AI and ML (CSM)

Email: venkat.reddy.gf@gprec.ac.in

G.Pulla Reddy Engineering College (Autonomous)

G.Pulla Reddy Nagar, Nandyal Road, Kurnool, AP 518007, India

Website: https://www.gprec.ac.in

Disclaimer and Confidentiality Notice

The content of this guest lecture, including all discussions, materials, and demonstrations, code is intended for educational purposes only and reflects the views and opinions of the speaker. While every effort has been made to ensure the accuracy and relevance of the information presented, it should not be considered as legal, financial, or professional advice.

Brillium Technologies retains unrestricted ownership of all information shared during this session. Participants must not record, reproduce, distribute, or disclose any part of the lecture or materials without prior written permission from Brillium Technologies. Unauthorized use or distribution of the content may result in legal action.

Additionally, all trademarks, service marks, and intellectual properties referenced or used in this presentation are the property of their respective owners. No ownership or rights over such third-party content are claimed or implied by the author or Brillium Technologies or by GPREC.

By attending this lecture, you agree to respect the confidentiality of the information shared and refrain from using it in any unauthorized manner. Failure to comply with these terms may result in legal action.

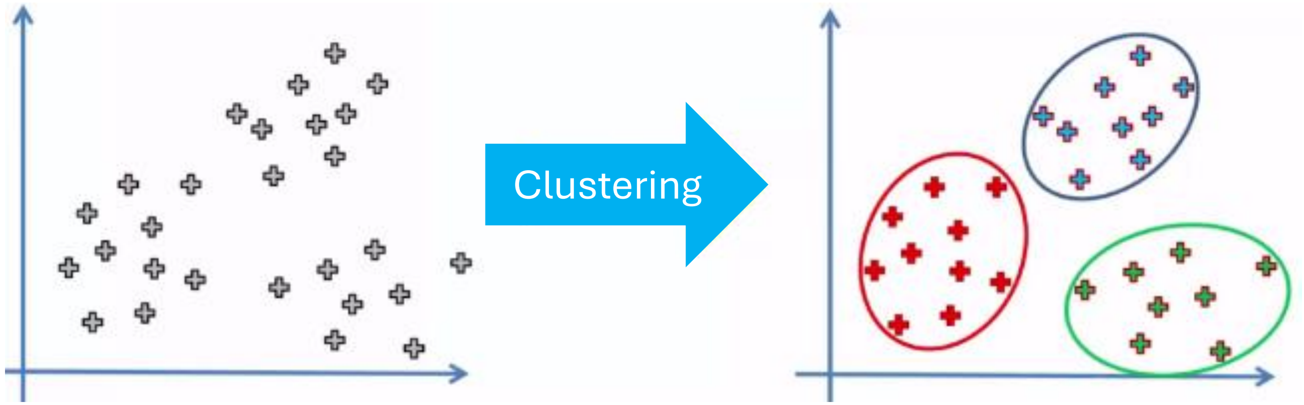
Thank you for your understanding and cooperation.

Lecture Outline

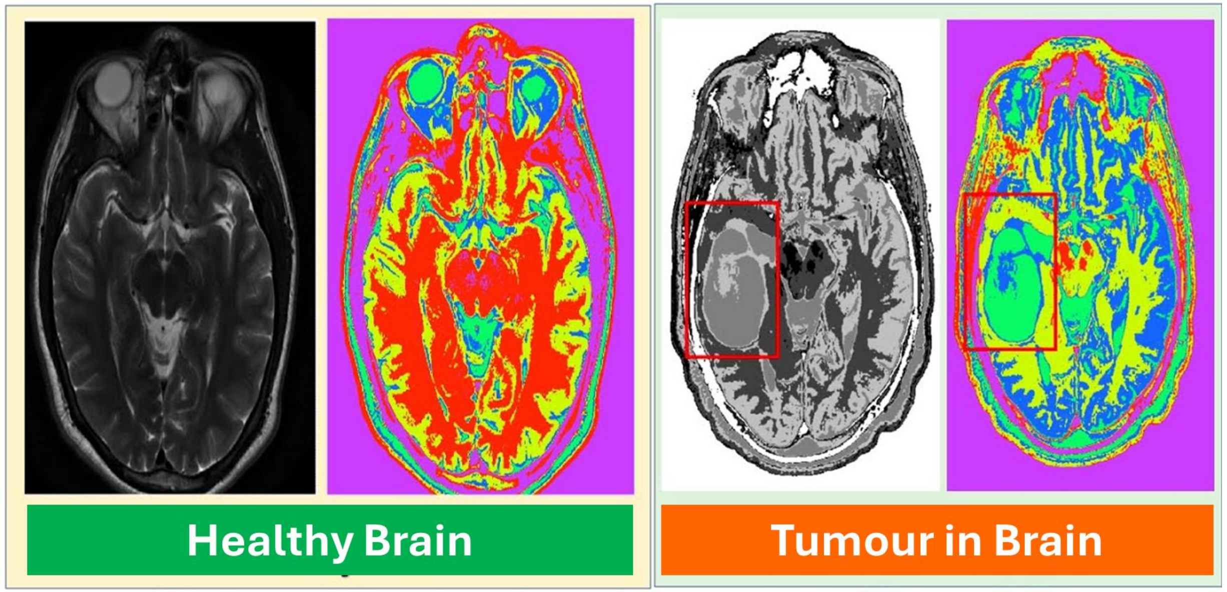
- What is Clustering?
- Real-life Clustering Scenarios
- Clustering Business Use Cases
- K-Means Clustering Overview
- Clustering Assignment
- K-Means Clustering Algorithm
- Elbow Method
- K-Nearest Neighbhours Distance Metric and Norms
- Centroid Convergence
- Quality of Clustering – Silhouette Score
- K-Means Clustering Vs K-Nearest Neighbhours
- K-Means Clustering - Strengthens and Limitations
- Project Announcement
- Project Implementation Steps
- Project Demo
- Q&A

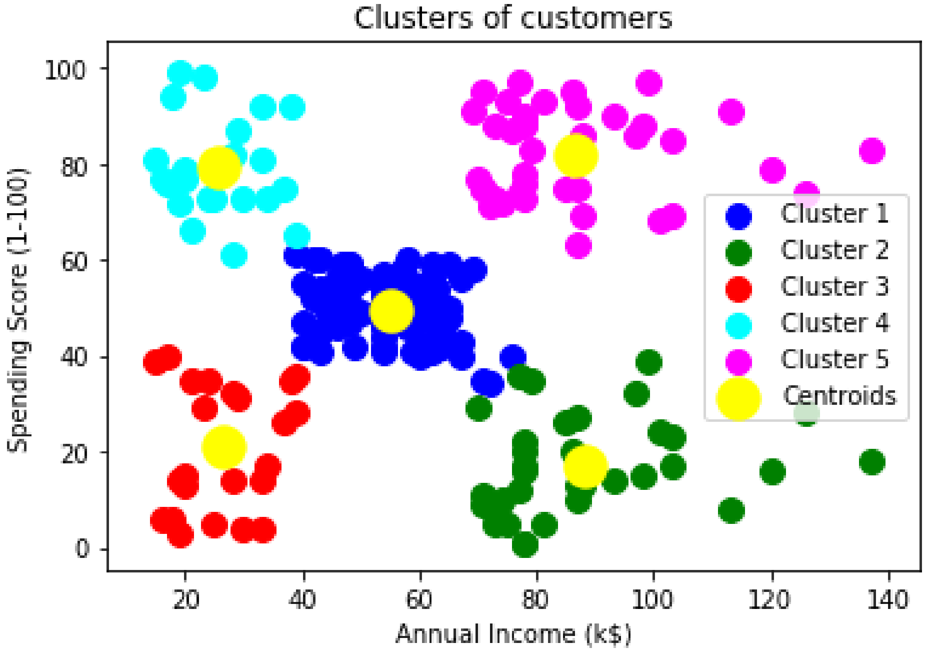
What is Clustering?

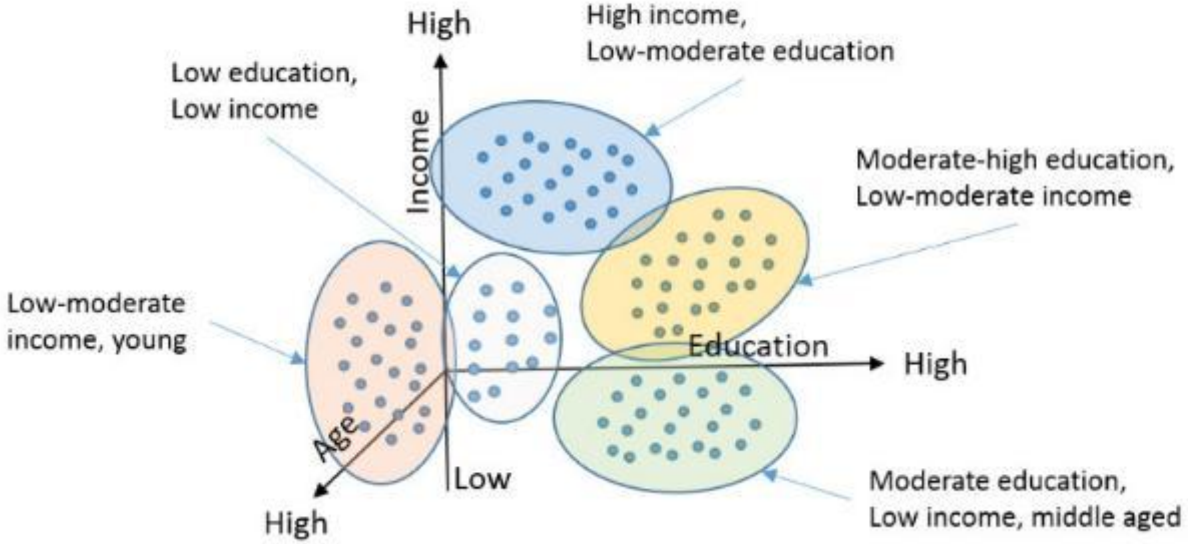
- Grouping similar data points together
- A clustering algorithm makes **birds of a feather flock together**
- Uses data features
- No need of labelled data
- Unsupervised learning
- Examples
  - Customer segmentation
  - Image compression
  - Anomaly detection
- Purpose - Discover patterns/grouping in data for taking related decisions



Real life Clustering Scenarios







High income, Low-moderate education

Low education, Low income

Moderate-high education, Low-moderate income

Low-moderate income, young

Moderate education, Low income, middle aged

Clustering Business Use Cases

Customer segmentation – Retail and e-Commerce

- Objective
  - Group customers based on purchasing behavior, demographics, or browsing habits
- Benefits
  - Enables targeted marketing, personalized product recommendations, and optimized inventory management
- Example
  - Segmenting customers into groups such as frequent buyers, seasonal buyers, or one-time shoppers for tailored promotional strategies

Inventory and Stock Optimization – Retail, Q-Commerce, Supply Chain

- Objective
  - Categorize products based on sales velocity, demand patterns, and seasonality
- Benefits
  - Assists in managing stock levels, reducing overstock/understock situations, and optimizing warehousing
- Example
  - Grouping products into fast-moving, slow-moving, and seasonal categories for better inventory control

Image Compression

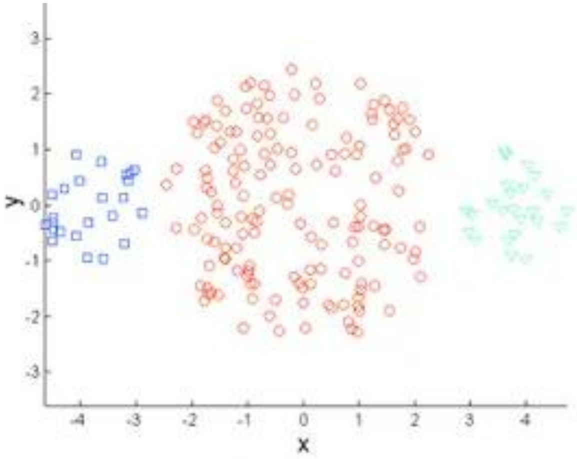
- Objective
  - Reduce image file sizes by grouping pixels with similar colors and replacing them with a single color (centroid)
- Benefits
  - Decreases storage requirements and speeds up image processing
- Example
  - Used in digital media platforms to compress images for faster loading while maintaining visual quality

Recommendation Systems (Streaming Services)

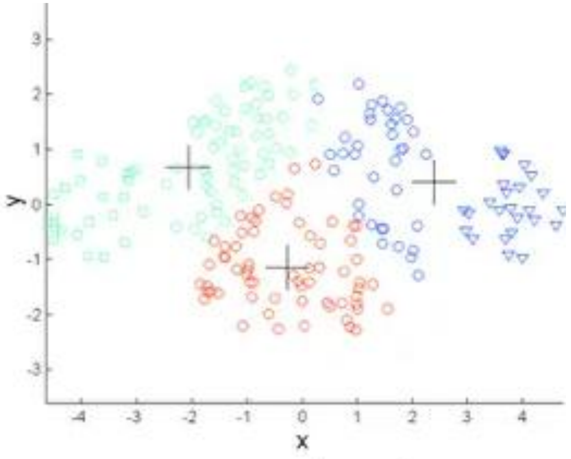
- Objective
  - Group content or users based on viewing patterns to improve recommendation accuracy
- Benefits
  - Personalizes content recommendations, enhancing user engagement and satisfaction
- Example
  - Streaming platforms like Netflix cluster users based on viewing history, genres, and ratings to recommend content that similar viewers enjoy

K-Means Clustering Overview

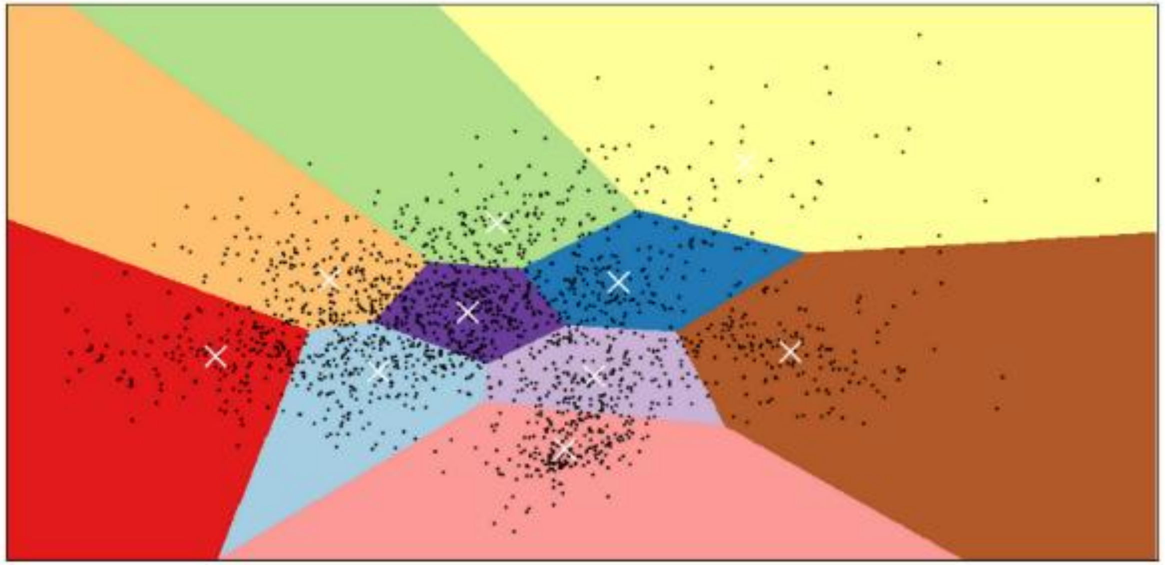
- Unsupervised Learning
  - No labeled data required
- Goal
  - Group similar data points into clusters
- Process
  - Randomly initialize K cluster centroids
  - Assign data points to the nearest centroid
  - Recalculate centroids based on assigned points
  - Repeat until convergence
- Key Considerations
  - Choice of number of clusters k
  - Distance metric - e.g., Euclidean distance
  - Initial centroid selection
- Applications
  - Customer segmentation
  - Image compression
  - Anomaly detection
  - Document clustering



Original Points

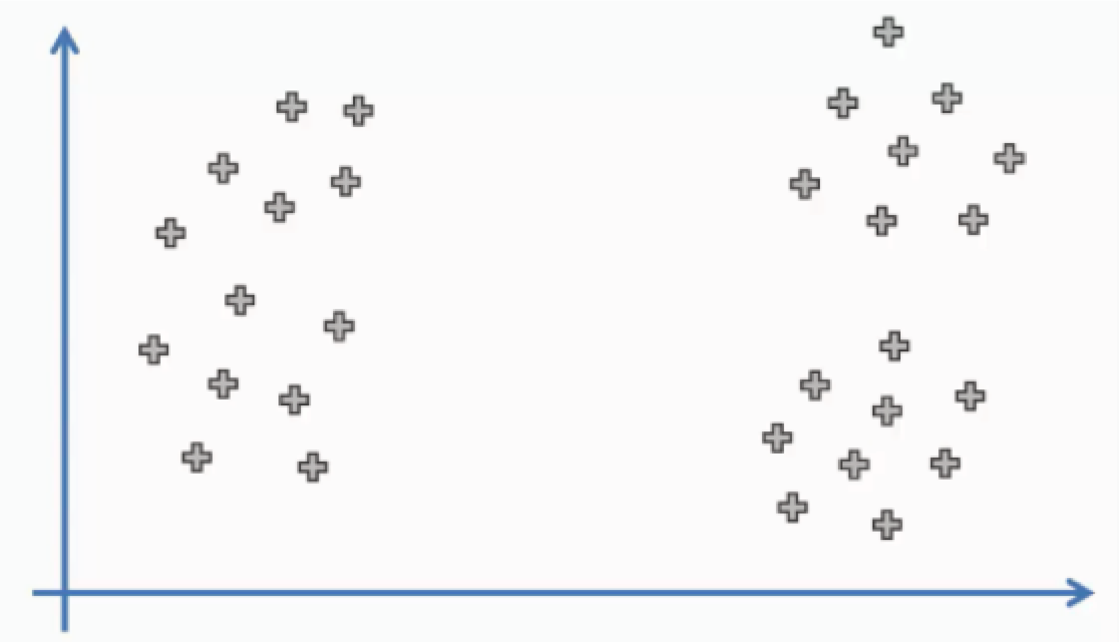


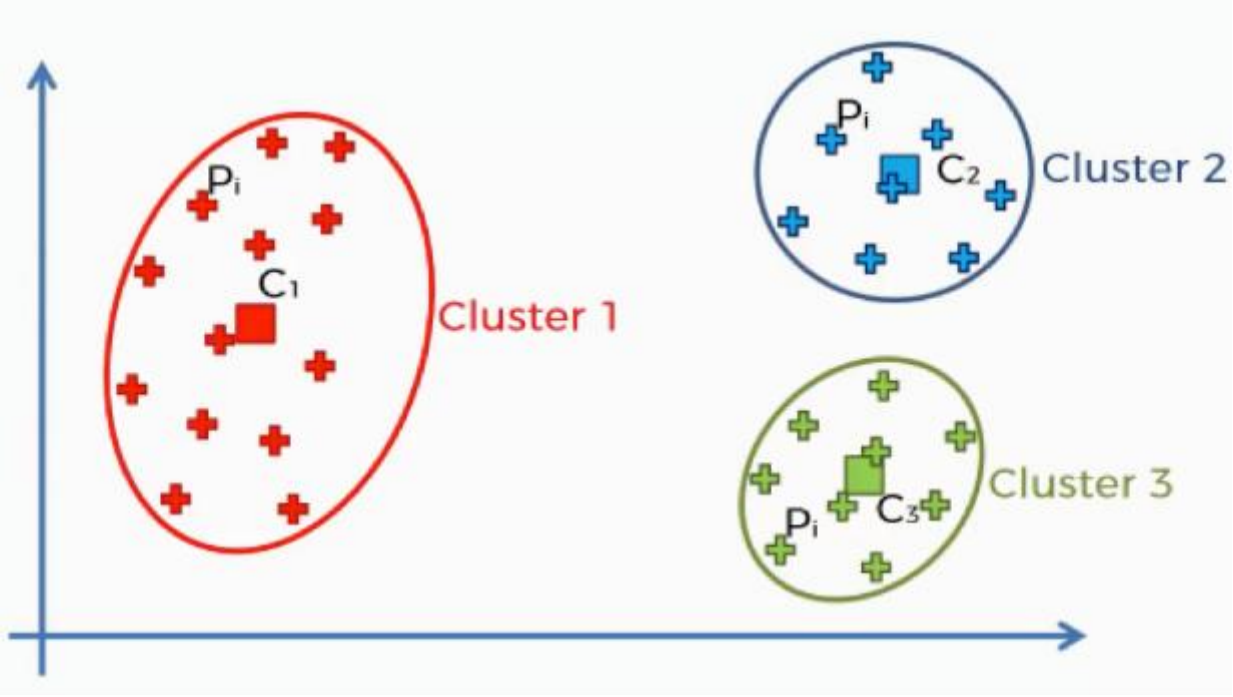
K-means (k = 3)



Cluster Assignment

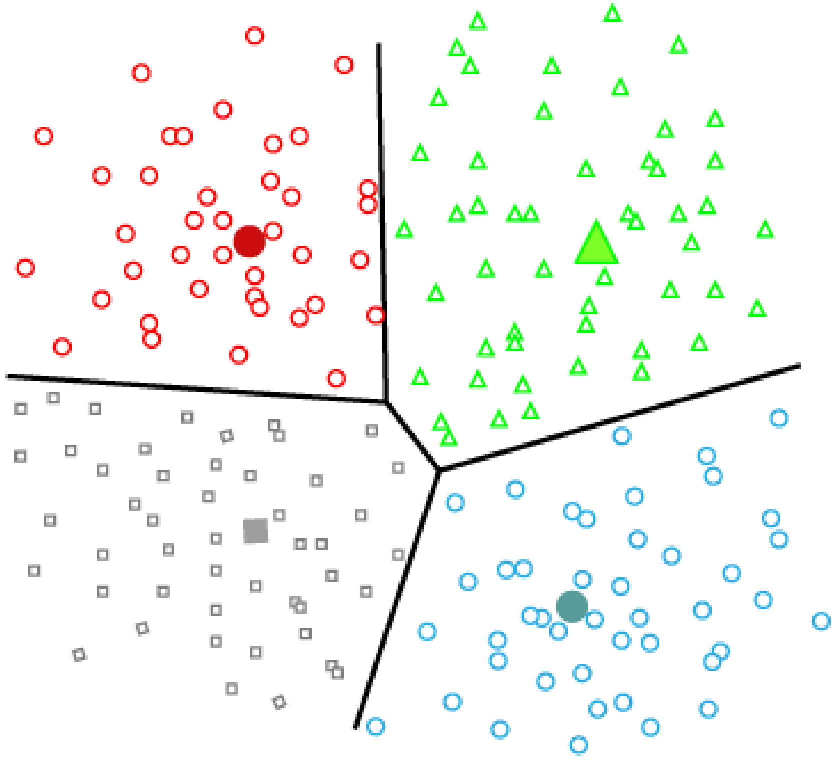
- Say $\alpha_i(j)$ denotes cluster assignment indicator for any given data vector
  - i – Cluster index - $\mathbf{1} \le \boldsymbol{i} \le \boldsymbol{k}$
  - j – Data vector index - $\mathbf{1} \le \boldsymbol{j} \le \boldsymbol{M}$
  - $\alpha_i(j) = \begin{cases} 1 & \bar{x}(j) \in \mathcal{C}_i \\ 0 & \bar{x}(j) \notin \mathcal{C}_i \end{cases}$





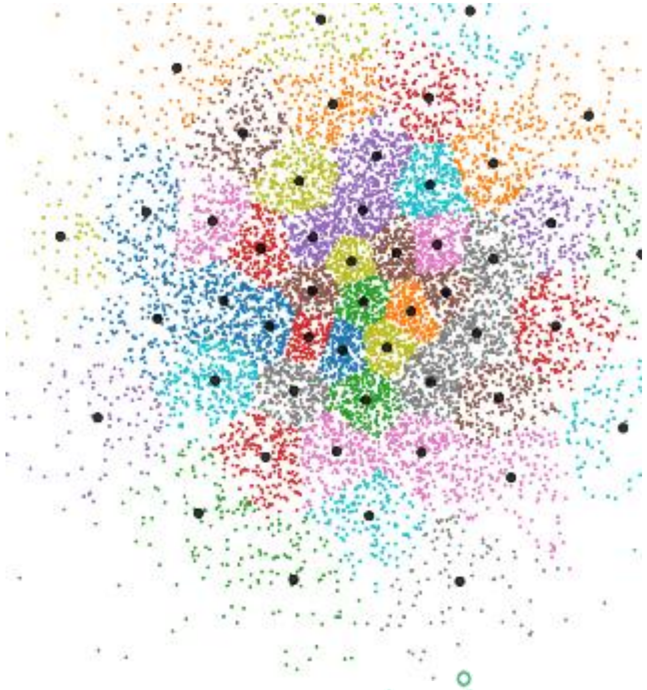
K-Means Clustering Algorithm

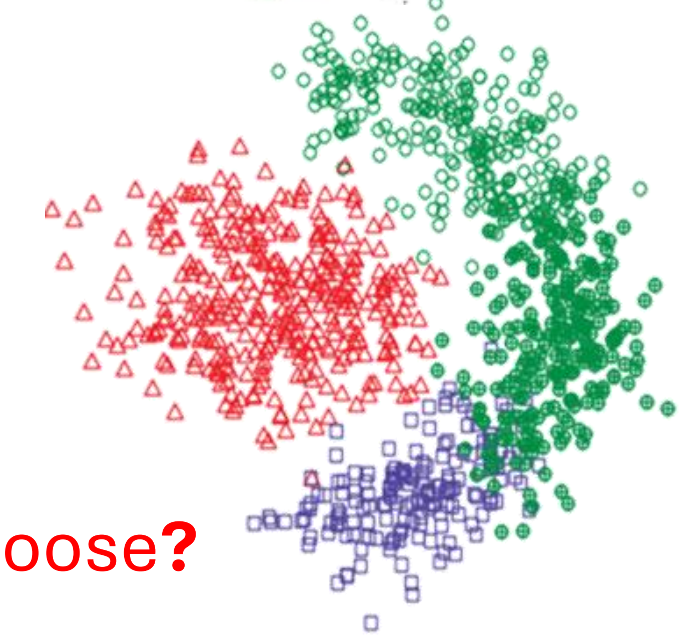
- Consider a dataset of $\boldsymbol{M}$ **data points/vectors** of each $\boldsymbol{n}$**-dimensions**
- Consider data can be organized into $\boldsymbol{k}$ clusters/groups - $\mathcal{C}_1, \mathcal{C}_2, \mathcal{C}_3, \dots \mathcal{C}_k$
- Suppose the center of these clusters or geometric objects - **centroids - are** $\boldsymbol{\bar{\mu}_1, \bar{\mu}_2, \bar{\mu}_3, \dots \bar{\mu}_k}$
- For the clusters, average or mean is the **central point or centroid** of a cluster – a point/vector of n-dimensional space
- Centroid as centre of mass
- For a cluster of points, the centroid is a point that minimizes the sum of squared distances between itself and all other points in the cluster



K-Means Clustering Algorithm

- Initialize
  - k-cluster centroids $(\bar{\mu}_1, \bar{\mu}_2, \dots, \bar{\mu}_k)$ randomly
  - $\bar{\mu}_1^{(0)}, \bar{\mu}_2^{(0)}, \dots \bar{\mu}_k^{(0)}$
  - $\bar{\mu}_i^{(l)}$ - centroids in l^{th} iteration
- Cluster determination
  - Assign $\boldsymbol{\bar{x}}(\boldsymbol{j})$ to a cluster $\boldsymbol{\tilde{\imath}}$ whose centroid from previous iteration $\left(\bar{\mu}_{\tilde{\imath}}^{(l-1)}\right)$ is closest
  - Closest cluster - Distance between $\boldsymbol{\bar{x}}(\boldsymbol{j})$ and centroid of the cluster from previous iteration - $\boldsymbol{\bar{\mu}}_{\tilde{\imath}}^{(\boldsymbol{l}-\boldsymbol{1})}$ - is the lowest amongst all such distances with other cluster centroids
  - $\boldsymbol{\tilde{\imath} = arg \min_{i=1\ to\ k} \left\| \bar{x}(j) - \bar{\mu}_i^{(l-1)} \right\|^2}$
- New centroids determination
  - The mean/average of all points assigned to cluster $\boldsymbol{i}$ in this iteration $\boldsymbol{l}$
  - $\boldsymbol{\bar{\mu}_i^{(l)} = \dfrac{\sum_{j=1}^{M} \alpha_i^{(l)}(j)\bar{x}(j)}{\sum_{j=1}^{M} \alpha_i^{(l)}(j)} = \dfrac{\sum_{j:\bar{x}(j)\in c_i}^{M} \bar{x}(j)}{\sum_{j:\bar{x}(j)\in c_i}^{M} 1}}$
- Repeat
  - Above steps until cluster assignments DO NOT change - convergence

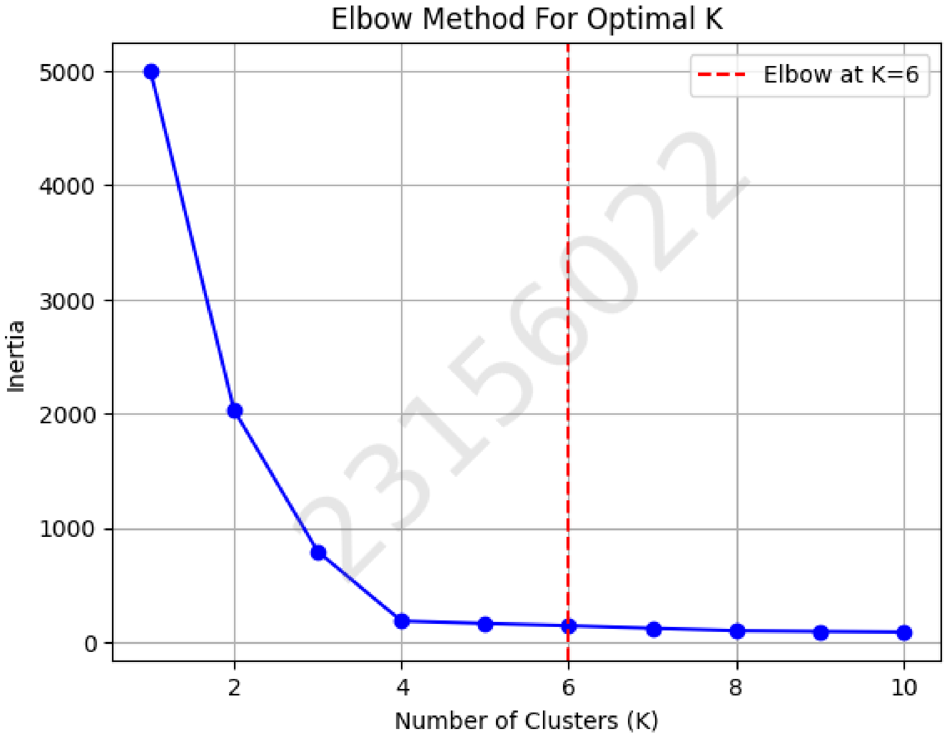




But what value for $\boldsymbol{k}$ and **distance measure** should be choose**?**

Elbow Method

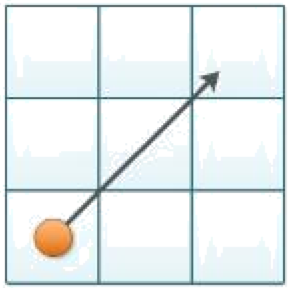
- Method to identify **optimal number of clusters** $(\boldsymbol{k})$
- By evaluating the model's performance across different values of $\boldsymbol{k}$
- Compute **Inertia** for various values of k
  - Inertia or WCSS : **Within-Cluster Sum of Squared Distances**
  - Sum of squared distances between each point and its centroid
  - $Inertia = \sum_{i=1}^{M} \|\bar{x}_i - \bar{\mu}_i\|^2$
- Fit data to each k, find $\bar{\mu}_k$ and find inertia for each k
- Elbow point is the point where the decrease in inertia starts to stabilize against number of clusters
  - This can be visually found in the plot - $\boldsymbol{k}$ Vs **Inertia** or
  - Solving first derivative equals to zero equation for



K-Nearest Neighbhour Distance Metric and Norms

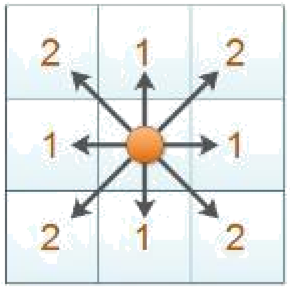
- The distance metric chosen plays a crucial role in determining the closeness of data points
- The distance between two vectors in vector spaces is measured using $norms$
- In mathematics - ***norm*** is a function from a real or complex vector space to the non-negative real numbers which behaves like the distance from the origin
- p-norm of $\bar{x} \in \mathbb{R}^n = \|\bar{x}\|_p = (|x_1|^p + |x_2|^p + \cdots + |x_n|^p)^{1/p}, p \geq 1$
- The most widely used norms are (level)1-norm, (level)2-norm,∞-norm
  - $\|\bar{x}\|_1 = (|x_1|^1 + |x_2|^1 + \cdots + |x_n|^1)^{1/1} = |x_1| + |x_2| + \cdots + |x_n|$ - Manhattan Distance
  - $\|\bar{x}\|_2 = (|x_1|^2 + |x_2|^2 + \cdots + |x_n|^2)^{1/2} = \sqrt{(|x_1|^2 + |x_2|^2 + \cdots + |x_n|^2)}$ – Euclidean Distance
  - $\|\bar{x}\|_\infty = \max_{1 \leq i \leq n} |x_i|$ - Chebyshev Distance

Euclidean Distance



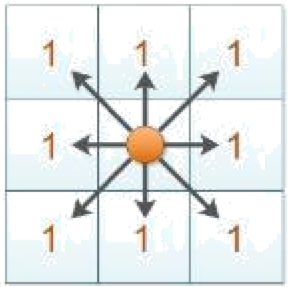
$\sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$

Manhattan Distance

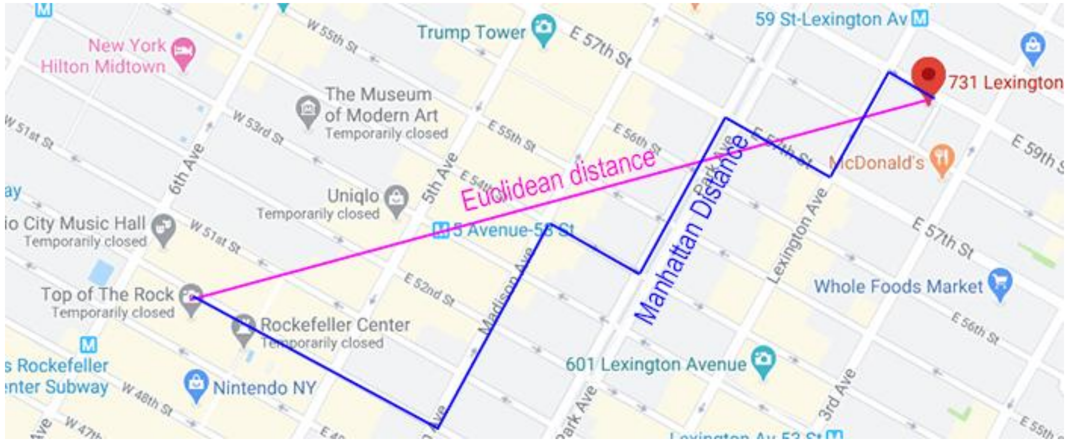


$|x_1 - x_2| + |y_1 - y_2|$

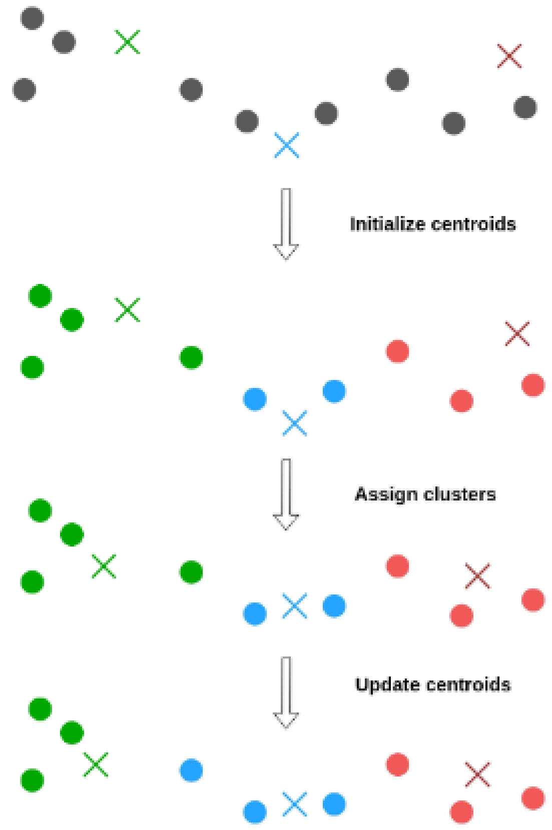
Chebyshev Distance

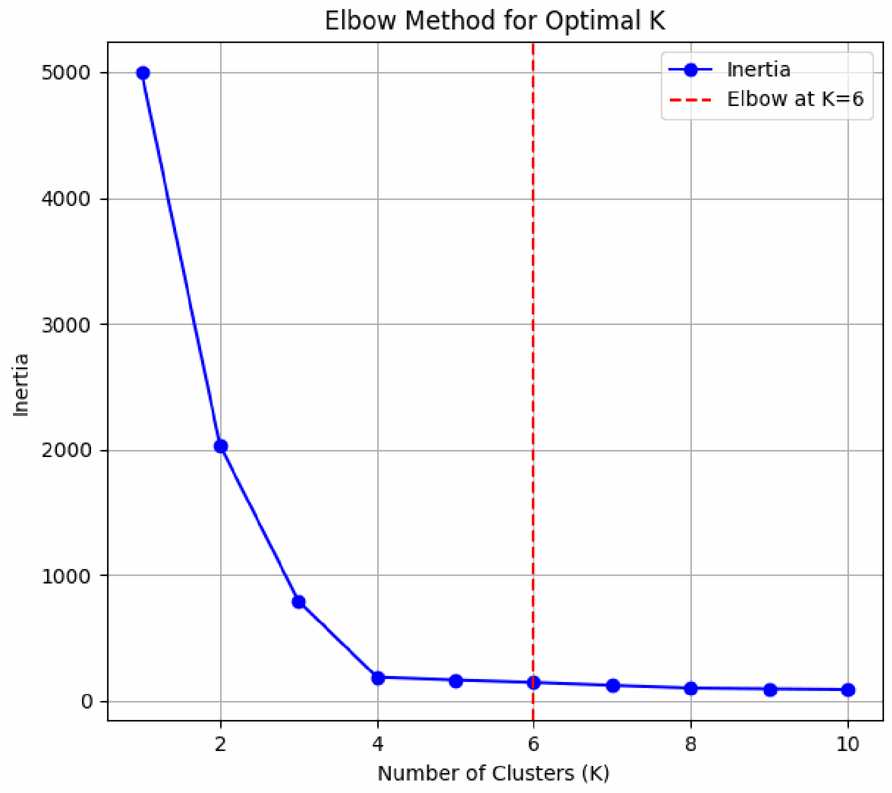


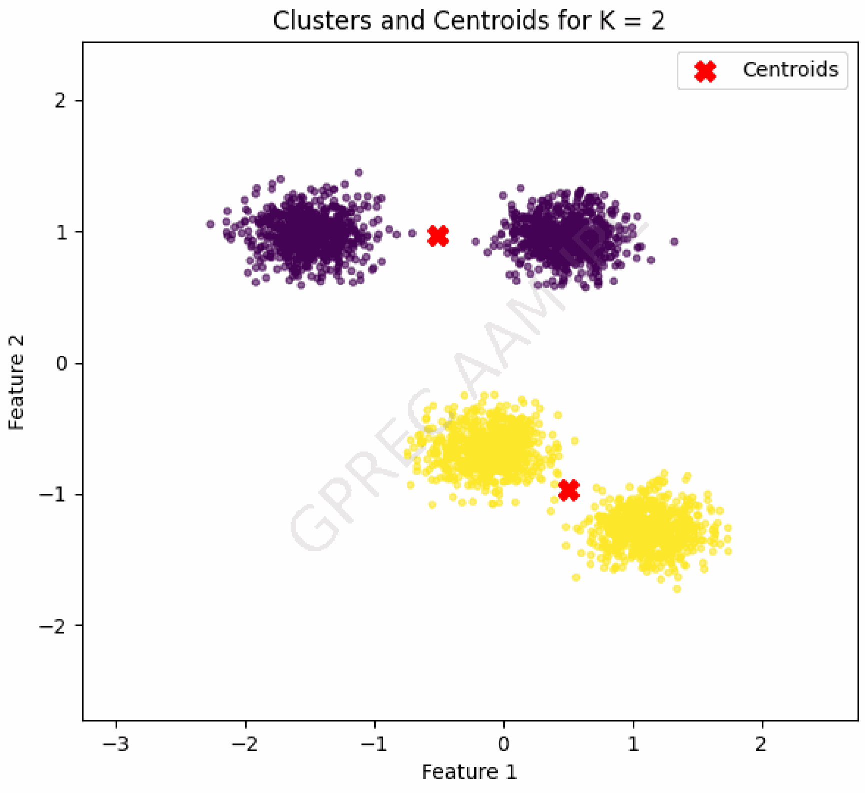
$\max(|x_1 - x_2|, |y_1 - y_2|)$



Centroid Convergence

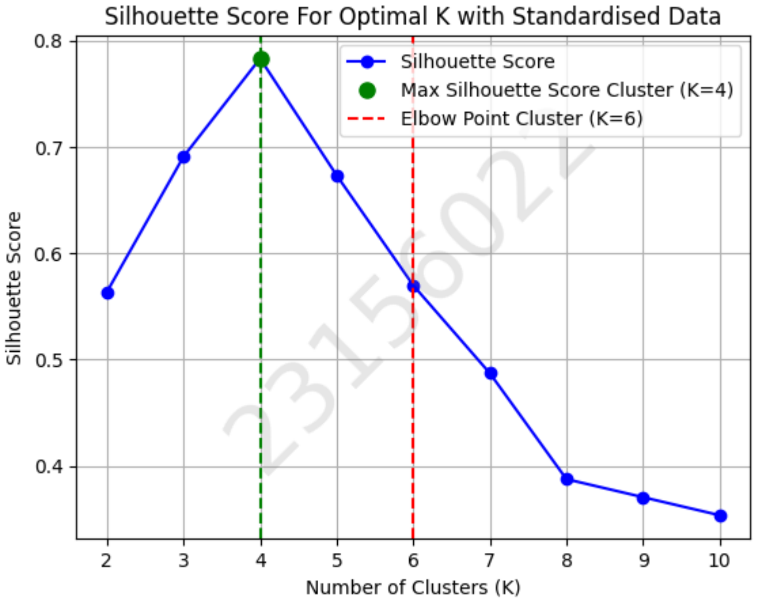






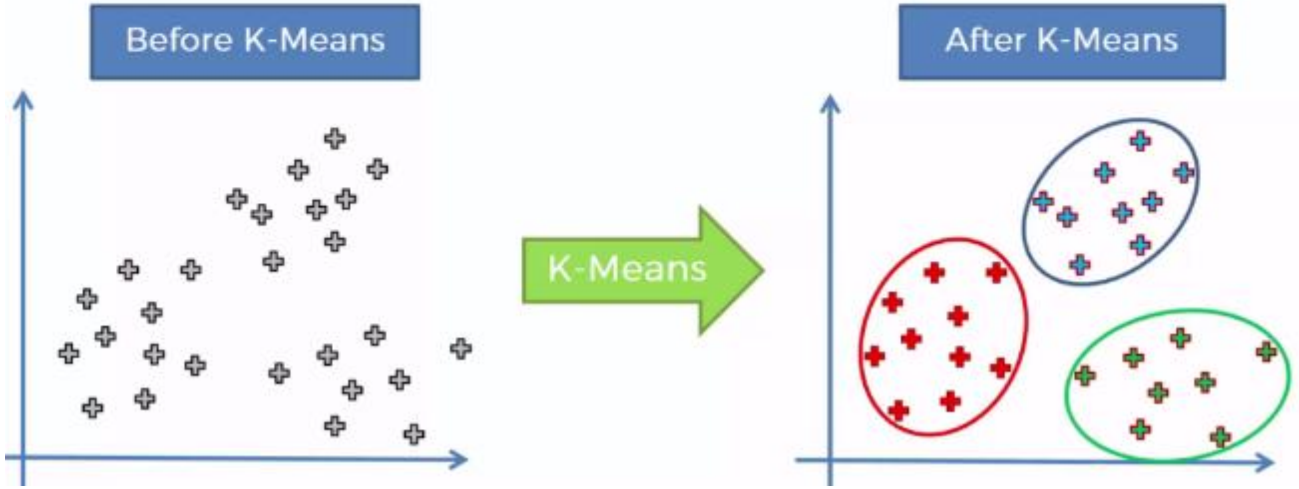
Quality of Clustering - Silhouette Score

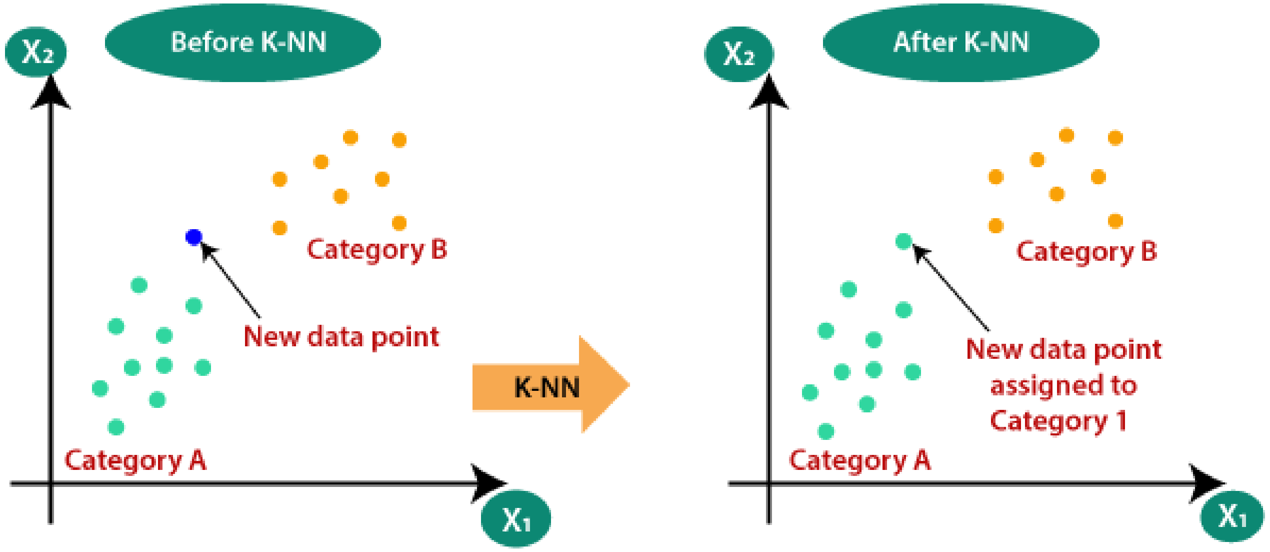
- Silhouette score evaluates the quality of clusters
- Measures how similar each data point is to its own cluster (cohesion) compared to other clusters (separation)
- For each data point i
  - Cohesion (a): Average distance between i and all other points in the same cluster
  - Separation (b): Average distance between i and points in the nearest/closest/neighbouring clusters
  - Silhouette Score = $\frac{(b-a)}{\max(a,b)}$
- Average silhouette score across all points provide an indicator of overall cluster quality
  - Close to 1
    - Well-separated, compact clusters with distinct boundaries.
  - Near 0
    - Poorly defined clusters; indicates overlap or ambiguity between clusters.
  - Negative Values
    - Misclassified points; possibly an indication to reconsider the clustering approach or the number of clusters.



K-Means Clustering Vs K-Nearest Neighbhours

- Are they same?
- K-Nearest Neighbors - KNN
  - Supervised learning algorithm
  - Primarily used for classification and sometimes regression
  - A data point is classified based on the majority class of its 'K' nearest neighbors in the training set
  - A distance-based classification technique.
- K-Means Clustering - KMC
  - Unsupervised learning algorithm
  - Used for clustering – data points are grouped into 'K' clusters based on similarity
  - The algorithm iteratively assigns data points to the nearest cluster center then recalculates the cluster centers based on the points in each cluster, until the centroids stabilize





K-Means Clustering Vs K-Nearest Neighbhours

Feature	K-Means Clustering (KMC)	K-Nearest Neighbour (KNN)
Type of Algorithm	Unsupervised	Supervised learning
Purpose	Grouping similar data points into clusters	Classify or predict based on nearest neighbhours
Data Requirements	No labelled data required	Requires labelled training data
Computational Complexity	Iterative Process	Computationally intensive at prediction time
Output	Centroids and cluster assignments	Predicted labels or values

K-Means Clustering - Strengthens and Limitations

- Strengths
  - Simple and Efficient: Easy to understand and implement
  - Scalable: Handles large datasets
  - Versatile: Applicable to various data types
  - Interpretable: Results are relatively easy to interpret
- Limitations
  - Sensitive to Initialization: Initial centroid selection can affect results
  - Requires Predefined K: Number of clusters must be specified
  - Struggles with complex shapes
  - Sensitive to Outliers: Outliers can distort cluster assignments

Project Announcement

- Algorithm of Application
  - K-means Clustering
- Project Title
  - Clustering Analysis with K-Means
- Project Objective
  - Implement and analyze K-means clustering on a dataset of your choice to explore patterns and group data points based on similarity
- Dataset
  - Description
    - Students may choose any dataset relevant to clustering, such as customer segmentation, image compression, or geographical clustering
  - Dataset Details
    - Ensure the dataset has multiple features that can be visualized meaningfully in 2D space
    - Suggested examples include the Iris dataset, Mall Customer Segmentation, or any real-world dataset that allows for clustering
  - Features
    - List the features of your chosen dataset and their descriptions
  - Usage in Machine Learning
    - Describe the purpose of clustering within the context of your dataset
  - Data Source and Published By
    - Specify the source or URL of the dataset for reproducibility
  - Data Download Link
    - Provide the link to download the dataset

Project Implementation Steps

1. Data Loading and Preprocessing
   - Import the chosen dataset, clean, and preprocess it as necessary
2. Plotting Original Dataset
   - Create a scatter plot of the original data to visualize its structure
3. PCA Transformation and Plotting
   - Apply PCA to reduce dimensionality and visualize the data in 2D if the dataset has more than two features
4. Elbow Method for Optimal K
   - Implement the elbow method by plotting inertia against the number of clusters and identify the optimal number of clusters
5. Silhouette Score Calculation
   - Calculate and plot the silhouette score for each K value to evaluate clustering quality
6. K-Means Clustering and Visualization
   - Apply K-means with the optimal K value, visualize the resulting clusters, and annotate centroids
7. Final Analysis
   - Describe observations about cluster distributions and how they relate to your dataset

Project Implementation Demo

Q & A

Interested in building a Gen AI application?

Reach out to venkat@brillium.in

THANK YOU!

AAML-IPL Brought You in Partnership with:

Brillium Technologies
Transforming Lives Through Technology

Brillium Technologies

Sector 7, HSR Layout, Bengaluru 560102, Karnataka, India
Website: www.brillium.in | Email: connect@brillium.in