Semester : VIII
Subject : DAV AIFB
Academic Year: 2024-25

K-MEANS CLUSTERING:

K-means clustering is an unsupervised machine learning algorithm used to group similar data points into K clusters.

Role of K-means clustering in financial Data Analysis:

(1) Customer Segmentation:

* Banks and financial institutions segment customers based on spending habits, credit scores and income levels.

* Helps in targeted marketing, personalized loan offers and credit risk assesment.

Example:-
A bank can group customers into high-income investors, middle class salaried individuals, and low-income customers for customized financial services.

(2) Credit Risk Analysis:

* Identifies groups of borrowers with risk profiles.
* Segments clients based on loan repayment history, credit scores, and transaction behaviours.
* Helps bank decide interest rates, loan eligibility and credit card limits.

Example:- Grouping customers into
Low-Risk → Regular payments, high credit score.
Medium-risk → Occasional late payments.
High-Risk → Frequent defaults, poor credit scores.

(3) Fraud Detection:

* K-means clusters normal vs. fraudulent transactions based on spending patterns.
* Anamolies (outliers) can indicate potential fraud, identity theft or suspicious activities.

Example:
If a customer usually spends $500 per month, but suddenly a $10,000 withdrawal happens from another country, K-means can flag it as fraud.

(4) Stock Market and Portfolio Clustering:

* Groups stocks based on volatility, returns and risk levels.
* Helps investors create diversified portfolio.

Example:
Stocks clustered into High Risk (Startups), Medium Risk (Tech stocks), and Low Risk (Blue-chip stocks).

(5) Anomaly Detection in Trading:

* Detect unusual trading patterns that may indicate insider trading or market manipulation.
* Helps regulatory bodies (SEC, RBI etc). ensure market integrity.

Semester: VIII Subject: ~~DAV~~ AIFB Academic Year: 2024-25

Steps for K-means Algorithm:

Step1: Initialization:
Randomly choose K centroids.

Step2: Assign Points
For each data point, calculate the distance to each centroid and assign it to the closest centroid.

Step3: Update Centroids
* For each cluster, sum the feature values of all points assigned to that cluster.
* Calculate the new centroid as the average of those points.

Step4: Repeat
Reassign points and update centroids. Continue until centroids no longer change significantly

Problem Statement:

A financial institute wants to segment 5 customers based on their Annual Income ($k) and Spending Score (out of 100) using K-means clustering (K=2)

Given dataset:

Customer	Annual Income ($K)	Spending Score
A	15	39
B	45	81
C	25	55
D	60	95
E	30	60

Semester: VIII   Subject: ~~DAV~~ AIFB   Academic Year: 2024-25

Solution:

Step 1: Choose Initial Centroids (Randomly)

We choose two initial centroids randomly from the dataset.

Centroid 1 (C_1) = (15, 39) → Based on Customer A

Centroid 2 (C_2) = (60, 95) → Based on Customer D

Step 2: Compute Euclidean distance:

The Euclidean distance between two points (x_1, y_1) and (x_2, y_2) is:

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

Compute the distance for each customer to both centroids:

Customer.	(15,39) C_1 - Distance	(60,95) C_2 - Distance	Assigned Cluster.
A(15,39)	0		Cluster 1
B(45, ~~16~~ 81)	$\sqrt{(45-15)^2+(81-39)^2} = 48.79$	$\sqrt{(45-60)^2+(81-95)^2} = 21.02$	Cluster 2
C(25,55)	$\sqrt{(25-15)^2+(55-39)^2} = 18.87$	$\sqrt{(25-60)^2+(55-95)^2} = 50$	Cluster 1
D(60,95)	$\sqrt{(60-15)^2+(95-39)^2} = 73.63$	0	Cluster 2
E(30,60)	$\sqrt{(30-15)^2+(60-39)^2} = 27.90$	$\sqrt{(30-60)^2+(60-95)^2} = 47.1$	Cluster 1

Step 3: Compute New Centroids

After the first iteration, the new clusters are:

Cluster 1 → A (15,39), C (25,55), E (30,60).

Cluster 2 → B (45,81), D (60,95)

Semester: VI   Subject: DAV   Academic Year: 2024-25

Compute New centroids:

New centroid (C_1'):

$$\left(\frac{15+25+30}{3}, \frac{39+55+60}{3}\right) = (23.33, 51.33)$$

New centroid (C_2'):

$$\left(\frac{45+60}{2}, \frac{81+95}{2}\right) = (52.5, 88)$$

Step 4: Recompute Distances and Reassign Clusters:

Recalculate distance using new centroids:

Customer	(23.33, 51.33) Distance	(52.5, 88) Distance	New cluster
A(15,39)	$\sqrt{(23.33-15)^2+(51.33-39)^2}$ = 15.02	$\sqrt{(15-52.5)^2+(39-88)^2}$ = 62.42	Cluster 1
B(45,81)	$\sqrt{(45-23.33)^2+(81-51.33)^2}$ = 36.17	$\sqrt{(45-52.5)^2+(81-88)^2}$ = 10.60	Cluster 2.
C(25,55)	$\sqrt{(25-23.33)^2+(55-51.33)^2}$ = 3.99	$\sqrt{(25-52.5)^2+(55-88)^2}$ = 42.60	Cluster 1
D(60,95)	$\sqrt{(60-23.33)^2+(95-51.33)^2}$ = 57.01	$\sqrt{(60-52.5)^2+(95-88)^2}$ 10.60	Cluster 2
E(30,60)	$\sqrt{(30-23.33)^2+(60-51.33)^2}$ = 11.13	$\sqrt{(30-52.5)^2+(60-88)^2}$ = 35.15	Cluster 1.

Since the clusters remain unchanged, the algorithm converges.

Semester : VIII   Subject : AIFB.   Academic Year: 2024-25

Final clusters:

Cluster 1: Low Income, Moderate Spending Customers

A(15,39)
C(25,55)
E(30,60)

Centroid: (23.33, 51.33)

Cluster 2: High Income, High - Spending Customers

B(45,81)
D(60,95)

Centroid: (52.5, 88)

SPARSITY AND CONNECTEDNESS OF UNDIRECTED GRAPH:

Graphs can be used to model various types of relationships, such as between assets, companies or individuals.

Understanding connectedness and sparsity of these graphs is crucial for analyzing network in finance, such as portfolio diversification, credit networks, market interactions and fraud detection.

Financial Asset Network (Portfolio Diversification)

In portfolio diversification, we can model the relationship between different financial assets (eg. stocks, bonds, commodities) using a graph where:

* Vertices (nodes) represent different financial assets.
* Edges represent relationships between assets, often based