UNIT: 3

Dashrath Nandan

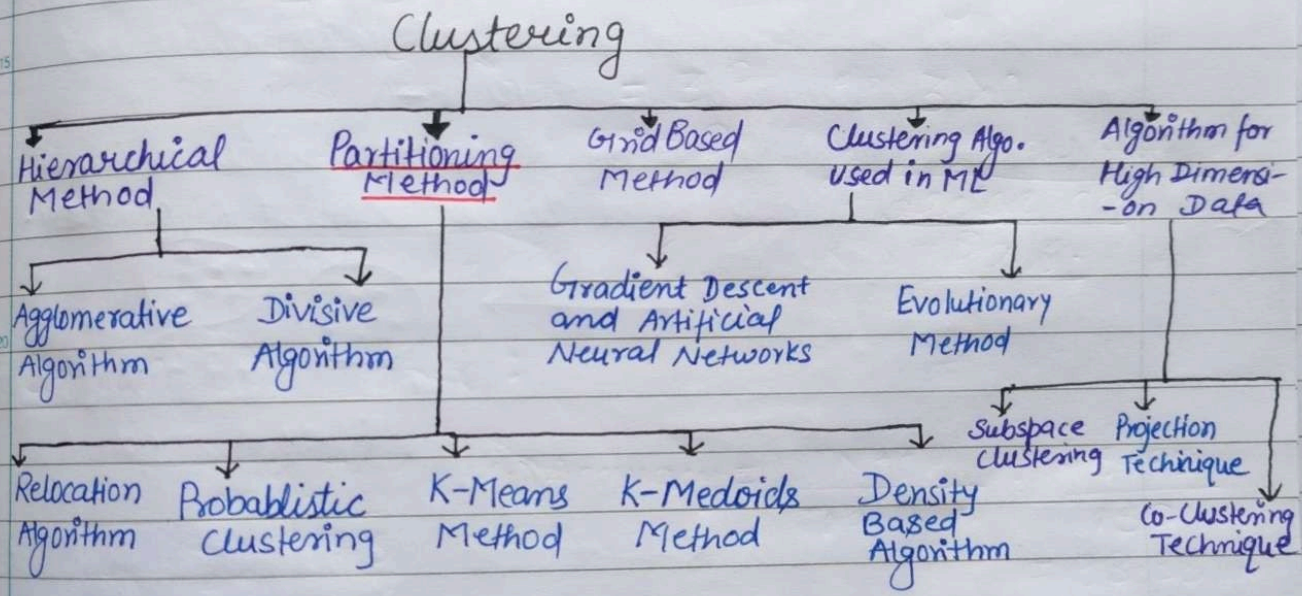
Date :

* Unsupervised Learning

Unsupervised learning is a machine learning technique in which a model learns patterns and relationships within a dataset without being provided with explicit labels or target values.

* Clustering

Clustering or cluster analysis is an unsupervised machine learning technique, which groups the unlabelled dataset. It is a way of grouping data points into different clusters, consisting of similar data points.



* Types of clusters: A cluster is a set of objects such that an object in cluster is closer to the "centre" of cluster, than to the centre of any other cluster.

1) Well-Seperated clusters:
2) Centre-based: The centre of a cluster is often centroid.
3) Contiguity-based
4) Density-based: A cluster is a dense region, which is seperated by low-density regions from other regions.
5) Conceptual cluster: Find cluster that share some common property or represent a particular concept.

Date :

* Aspects of Clustering

↳ A clustering algorithm : Partitional, Hierarchical clustering...
↳ A distance (similarity) function
↳ Clustering quality
  ↳ Inter-cluster distance ⇒ Maximized
  ↳ Intra-cluster distance ⇒ Minimized

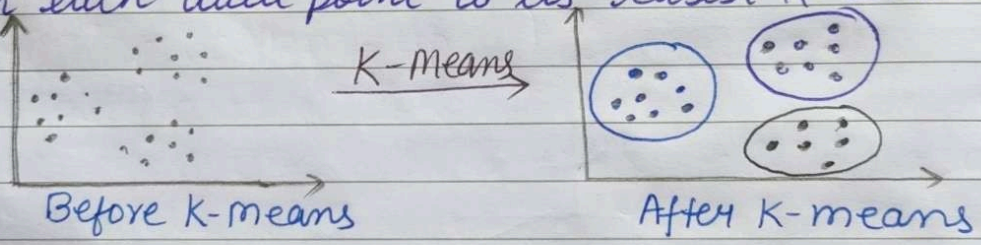
☆ Partitional Clustering

It is a type of clustering that divides the data into non-hierarchical group. It is also known as centroid-based method.

* K-means Clustering :

K-means clustering is an Unsupervised ML, which groups the unlabeled dataset into different cluster. It mainly perform two task-

i) Determine best value of K center points or centroids.
ii) Assign each data point to its closest K-center.



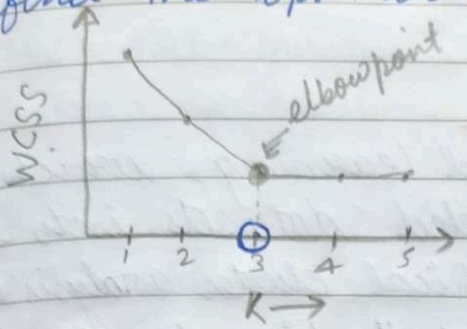
Algorithm :

Step 1: Select the number K to decide the number of clusters.
Step 2: Select Random K centroids.
Step 3: Assign each datapoints to closet Centroid (using Euclidean Distance)
Step 4: Calculate the varience and place a new centroid of each cluster.
Step 5: Repeate Step 3,
Step 6: If any reassignment occurs, then goto step 4 or else Finish.
Step 7: The model is ready.

Date :

Elbow Method:

The elbow method is one the most popular way to find the optimal number of clusters.



For $K=2$

$$WCSS = \sum (P_i \cdot C1)^2 + \sum (P_i \cdot C2)^2$$

WCSS: within Cluster Sum of Square

P_i in Cluster1 Distance

Advantages	Disadvantages
• Simple, understandable. Items automatically assigned to clusters.	• Must pick No. of cluster before hand. • All item forced into cluster. • Too sensitive to outliers.

* K-Medoids:

Similar to K-means. This method is less sensitive to outliers than K-means.

Medoids is a point in cluster from which the sum of distances to other points is minimal.

Algorithm

Step 1. Randomly choose 'K' points from the input data.

Step 2. Each data point gets assigned to the cluster to which its nearest medoids belong.

Step 3. For each data points of cluster, its distance from all other datapoints is computed and added. The point of i^{th} cluster for which computed dist. is Minimal is assigned as Medoids of cluster.

Step 4. Step 2 & 3 are repeated until medoids stop moving.

Total Cost of Swap: The total cost of swapping, S is the sum of all costs from non medoids objects. If total is negative, the swap is worthwhile doing.

Date :

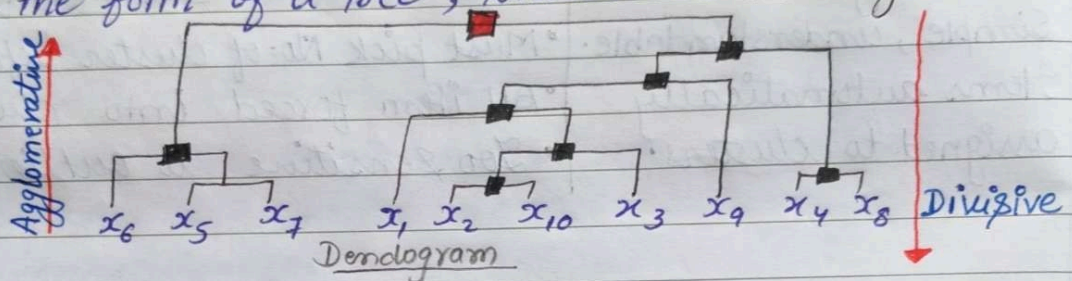
Advantages: Best for small datasets, Error Free, No overlapping.

Application:

★ Hierarchical Clustering

It is another Unsupervised learning, which is used to group the unlabeled datasets into a cluster, known as Hierachical cluster analysis or HCA.

↳ In this algorithm, we develop the hierachy of cluster in the form of a tree, known as dendrogram.



Dendogram

✱ Types of hierarchical clustering:

1.) Agglomerative (bottom-up.) clustering:

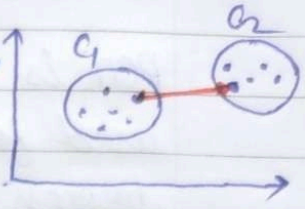
Agglomerative is a bottom-up approach, in which the algorithm starts with taking all data points as single clusters and merging them until one cluster is left.

2.) Divisive (Top-down) clustering:

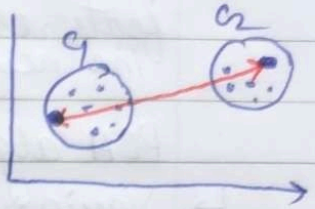
Divisive or Top-down clustering, where we start with observation in 1 cluster and then split clusters until we reach the leave.

* Measuring the distance of two clusters:

i) Single link: The distance b/w two clusters is the dist. b/w two Closet data points in the two cluster.

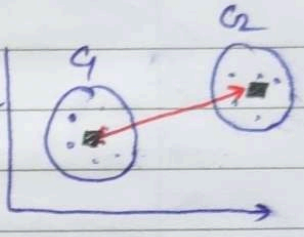


ii) Complete link: The dist. b/w two cluster is the dist. b/w two furthest data points.



iii) Average link: Each pair of datasets is added up then diuided by total no. of datasets.

iv) Centroid Method: The dist. b/w two cluster is the distance between their Centroids.



* Distance Function:

- Euclidean Distance = $dist(x_i, x_j) = \sqrt{(x_{i_1} - x_{j_1})^2 + (x_{i_2} - x_{j_2})^2 + \cdots + (x_{i_r} - x_{j_r})^2}$ For $H=2$,
- Manhattan Distance = $dist(x_i, x_j) = |x_{i_1} - x_{j_1}| + |x_{i_2} - x_{j_2}| + \cdots + |x_{i_r} - x_{j_r}|$ $h=1$

☆ Algorithm:

Agglomerative	Divisive
Step 1: Create each data points as a single cluster. No. of clusters will be N.	1. Split whole data into 2 clusters.
Step 2: Take two closest data points and merge them to form one cluster. (N-1)	2. Choose a current cluster, and split it as in Step 1.
Step 3: Again, take two closest cluster & merge them. There will be (N-2) cluster.	3. Repeate Step 2, until each clusters contains a point.
Step 4: Repeate step 3, until only one cluster left.	
Step 5: Develop the dendogram.	

Date :

★ Principal Component Analysis (PCA)

PCA is an Unsupervised learning algorithm that is used for dimensionality Reduction in ML. It is a statistical process that converts observations of correlated features into uncorrelated features with the help of orthogonal transformation. These new features are called principal component.

* Steps for PCA algorithm

1. Getting the dataset; 2. Representing data into a structure
3. Standardizing the data; 4. Calculating the covariance
5. Calculating the Eigen values and Eigen vectors.
6. Sorting Eigen vectors; 7. Calculating principal Component (New features)
8. Remove less or unimportant features.

* The PCA algorithm is based on some Mathematical concept:-
↳ Variance and Co-variance.
↳ Eigen values and Eigen vectors.

1. Normalize the data:

$$X_n = \frac{(X - X_{min})}{(X_{max} - X_{min})} \quad ; \quad Z(\text{Standardization}) = \frac{x - \mu}{\sigma} \quad \mu \hookrightarrow \text{mean}, \ \sigma \hookrightarrow \text{SD}$$

2. Build the Covariance Matrix:

	x	y	z	n
x	Var(x)	Cov(x,y)	Cov(x,z)	Cov(x,n)
y	Cov(x,y)			
z	Cov(x,z)			
n	Cov(x,n)			Var(n)

(Cov(x,x) = Var(x))

$$Q_{jk} = \frac{1}{N} \sum_{i=1}^{N} (X_{ij} - E(X_j))(X_{ik} - E(X_k))$$

($E(X_j)$ ↓ sample mean)

Variance: Measure of deviation from the mean for points in one dimension.

$$Cov(x, y) = Cov(y, x)$$

Date :

Covariance: Measure of how much of the dimensions vary from the mean with respect to each other.

3. Eigenvalues and Eigen vectors

for Square matrix, A ⇒

$$Ax = \lambda x \quad ; A = \text{Square matrix}$$

$$\Rightarrow (A - \lambda I)x = 0 \qquad \lambda = \text{Eigen Vector or Characteristic Vector}$$

$$x = \text{Eigen value.}$$

Note: The zero vector can't be eigen vector.
The value zero can be eigen value

The eigenvalues represent the scaling factor by which a vector is transformed when a linear transformation is applied, while eigen vector represent the direction in which transformation occurs.

for eigen value :→ $|A - \lambda I| = 0$, I = Identity matrix

for eigen vector :→ $(A - \lambda I)x = 0$

* Dimensionality Reduction:

1. PCA (Principal Component Analysis)
2. ICA (Idependent Component Analysis)
3. LDA Linear Discriminant Analysis)
4. Multidimensional Scaling.

* Characteristics of Principal Component:

↳ PC must be linear combination of original feature.
↳ These components are Orthogonal, i.e. correlation is zero.
↳ The importance of each component decreases when going 1 to n.

* Application : Image processing, Movie Recomm. System, etc.

Date :

★ Association Rule Mining (ARM)

ARM is a type of unsupervised learning technique that checks for the dependency of one data item and maps accordingly so that it can be more profitable.

↳ Also called as Market Basket analysis.

↳ if 'A' then 'B' { A ⇒ B }
antecedent ↵ ↳ consequent

* Association Measure Metrics :

- Support : Support is the frequency of A or how frequent an item appears in the dataset.

$$Supp(x) = \frac{Freq(x)}{T} \quad T \hookrightarrow \text{Transaction}$$

how often the items X and Y occur together in the dataset when the occurrence of X is already given

- Confidence = $\frac{Freq(X,Y)}{Freq(X)}$; Confidence indicates how often the rule has been found true, (strength of association)

- Lift = $\frac{Supp(X,Y)}{Supp(X) \times Supp(Y)}$; It is strength of any rule.

- Leverage = $Supp(X,Y) - Supp(X) * Supp(Y)$

* Types of ARM Algorithm :

1. Apriori Algorithm : This algorithm uses frequent dataset to generate association rules. It is designed to work on database that contains transaction.
   - This algorithm uses a BFS and Hash Tree.
   - It mainly used for basket analysis, and helps

Date :

to understand the products that can be bought together.

2. Eclat Algorithm: Equivalence Class Transformation, uses DFS technique to find frequent itemsets. It is faster than Apriori algorithm.

3. F-P Growth: Frequent-Pattern, is the improved version of the Apriori Algorithm. It represent database in the form of tree structure, known as frequent pattern or tree.

Examples → YouTube (on Apriori Algorithm)

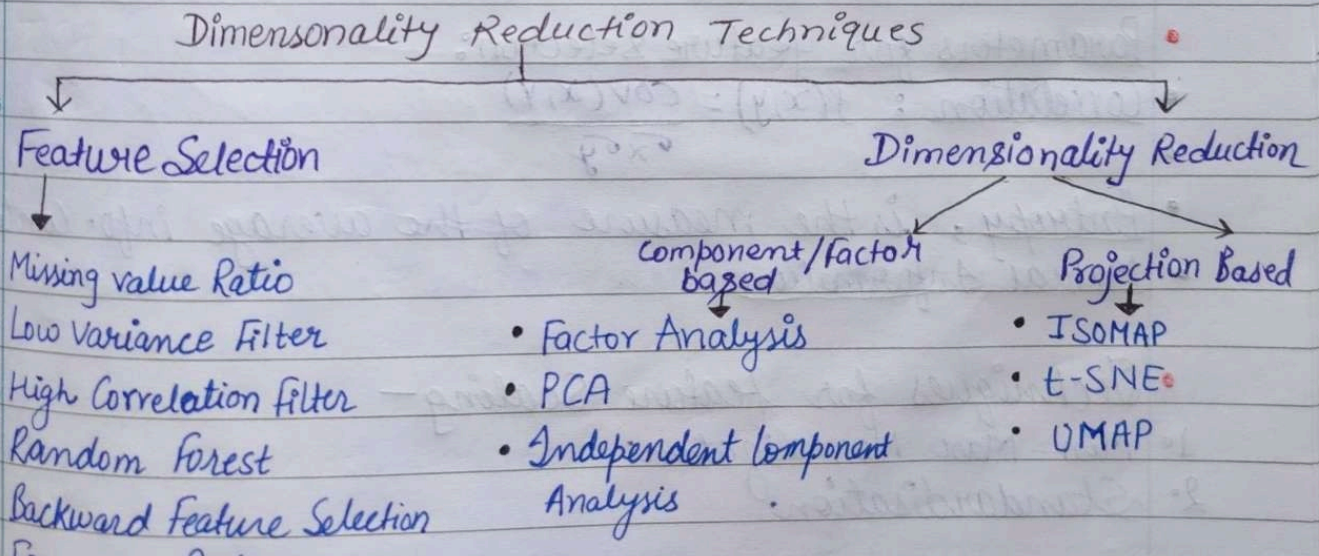
* Application:

i) Market Basket Analysis: Used by big retailers to determine the association b/w items.

ii) Medical Diagnosis: Helps in identifying prob. of illness.

iii) Protein Sequence: Helps in determing the Synthesis of artifical proteins.

☆ Dimensionality Reduction (DR)



Date :

Feature Reduction or dimensionality reduction is a way of converting the higher dimension dataset into lesser dimension dataset ensuring that it provides similar information

Application: Speech recognition, Signal processing, etc.

* Feature Selection:
Also called variable selection and attribute selection, it is used to identify and remove unneeded, irrelevant and redundant attribute from data.

- It has 4 approaches:

1) Wrapper Method: A predictive method used to evaluate a combination of features and assign a score based on Model accuracy. Eg. Best Fit Search.
2) Filter Method: Applies a statistical measure to assign a scoring to each feature. Eg. Chi Square Test.
3) Embedded Method: It selects the feature during the process of training the dataset.
4) Hybrid Method: Both Wrapper and filter based method.

- Parameters for feature selection:
- Correlation: $\rho(x,y) = \frac{cov(x,y)}{\sigma_x \sigma_y}$
- Entropy, is the measure of the average info. content.
- Mutual Information.

- Techniques for feature Scaling-
1. Min-Max Normalization.
2. Standardisation.

Date :

* Advantages of DR :
- It helps in data compression, hence reduced storage space.
- It reduces computation time.
- It removes redundant features.

* Disadvantage of DR :
- It may lead to some amount of data loss.
- PCA fails where mean and covariance are not enough to define dataset.

Semi-Supervised Learning

Semi-Supervised learning is a type of ML algorithm that represents the intermediate ground b/w Supervised and Unsupervised learning. It uses the combination of labeled and unlabeled dataset. It overcomes the drawbacks of S.ML and UnSupervised ML.

* Assumptions of Semi-SML :

1. Continuity Assumption: The object near each other tends to share same group or label.

2. Cluster Assumptions: Data are divided into different discrete clusters.

3. Manifold Assumptions: This assumptions helps to use distance and densities.

Date :

* Working:

↳ Firstly, it trains the model with less amount of training data.
↳ The algo. uses the unlabeled dataset with pseudo label in next step.
↳ Now, the labels from labeled training and pseudo label data are linked together.
↳ The input data are also linked from both dataset.
↳ In the end, again train model with the new combined input as did in the first step.

* Example or Real World Application of SSL:

i) Image and Speech Analysis:
ii) Web Content Classification: It is very critical and impossible to label each page on internet. Google uses SSL to rank a webpage.
iii) Text document classifier:

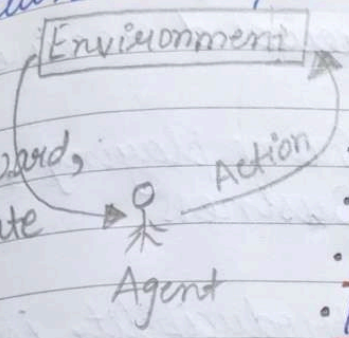
* Semi Supervised	Reinforcement
• Labels or output known for a subset of data.	• Focus on making decision based on previous experience.
• A blend of supervised and unsupervised learning.	• Policy making with feedback.
• Medical prediction.	• Game AI, Reward System.

Prashrak Nandan

Date :

Reinforcement Learning

It is a feedback-based learning method, in which a learning agents gets a reward for each right action and penalty for wrong action.



Common Terms or Elements of RL :

- Agent : Entity which perform action.
- Environment : A scenario thas agent faces.
- State (S) : Current situation returned by enviro.
- Reward (R) : St
- Policy (π) : Strategy which applies by agent.
- Value (V) : It is expected long-term return with discount.
- Q-Value ( ) : Similar to value, but take one parameter as a current action.
- Model

Types of RL :

adding something to increase the tendency that expected behavior would occur again. It impacts positively on the behavior of the agent and increases the strength of the behavior.

i) Positive RL : It increases the strength and the frequency of the behaviour and impact positively on agent's action.

ii) Negative RL : It is defined as strengthening of behaviour that occurs because of negative condition which should be avoided.

opposite to the positive reinforcement as it increases the tendency that the specific behavior will occur again by avoiding the negative condition.

Approaches to implement RL :

i) Value-Based : In this, RL method, you should try to maximize a value function V(s).

ii) Policy-based : In this RL method, you try to come up with such a policy that the action performed in every state helps to gain maximum reward in the future.

iii) Model-based : In this, a virtual model is created for the environment and agent explores that Enviro. to learn it.

Policy-based approach is to find the optimal policy for the maximum future rewards without using the value function

challenege:

Sample efficiency
Reproducibility issues
Performing in real-life scenarios
Sparse rewards
Offline reinforcement

Date :

* Characteristics or Key features of RL:

- In RL, the agent is not instructed about the environment and actions
- It is based on the hit and trial process.
- The agent may get a delayed reward receives.
- Applications : Robot navigation, Game playing, Finance Sector, Business.
- Challenges : Sample efficiency, Reproducibility issues. Sparse reward, Offline Reinforcement.

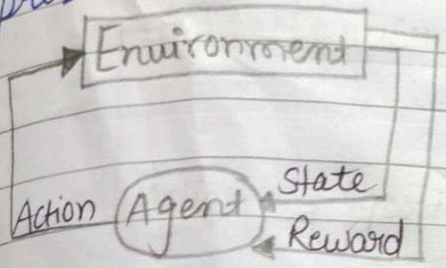
*

Parameter	Supervised ML	Reinforcement ML
Defination	ML by using labelled data.	Agent interacts with the environment by performing action.
Type of Data	labelled data.	No-predefined data.
Type of Problem	Regression and classification.	Reward and error based.
Aim	Calculate outcomes	learn a series of action.
Approach	Maps labelled input to known output	follow hit & trial method.
Application	Risk Evaluation, Forecast Sales.	Gaming, Self-driving cars.

Date :

Markov Decision Process (MDP):

MDP is used to formalize the reinforcement learning problems.



- MDP contains tuple of 4 elements (S, A, P_a, R_a)
- A set of finite States S.
- A set of finite Actions A.
- Reward received.
- Probability P_a.

- Markov Property: It says that "If the agent is present in the current state S_1, perform an action a_1 and move to state S_2, then the state transition from S_1 to S_2 only depends on current state and future action."

- The main used algorithm –
- Q-learning:
  - Q-learning is a value-based method of supplying info. to inform which action an agent should take.
  - It learns value function $Q(s,a)$.

- State Action Reward State Action (SARSA):
  - On-policy, control method select the action for each state while learning using a specific policy.

on-policy temporal difference learning method.

- Deep Q Neural Network (DQN):
  DQN, is a Q learning using neural network.