Parul® University
Vadodara, Gujarat

NAAC GRADE A++

UNIT- 4
Unsupervised Learning

Study Guide

Dr. Vinod Patidar
Associate Professor
CSE Department, PIT
Parul University

1. Unsupervised Learning

Unsupervised learning deals with discovering patterns, structures, and relationships in unlabeled data.

1.1 Types of Unsupervised Learning

Type	Description	Examples
Clustering	Grouping similar data points	K-means, DBSCAN
Dimensionality Reduction	Reducing features while preserving structure	PCA
Association	Finding rule-based relationships	Apriori

2. Clustering Basics

Clustering aims to group similar data points based on distance or density.

2.1 Partition-Based Clustering

Divides data into a fixed number of clusters.

K-Means Clustering.

Assigns points to the nearest centroid

Recomputes centroids iteratively

Algorithm Steps:

1. Choose k cluster centroids
2. Assign each point to the nearest centroid
3. Recalculate centroid of each cluster
4. Repeat until convergence

Figure: K-Means Workflow

Data Points → Choose k → Assign Points → Update Centroids → Repeat

Advantages: Simple, fast

Disadvantages: Needs k, sensitive to outliers

2.2 K-Modes Clustering

Used for categorical data.

Replaces mean with mode

Uses Hamming distance instead of Euclidean distance

Table: K-Means vs K-Modes

Feature	K-Means	K-Modes
Data Type	Numerical	Categorical
Distance	Euclidean	Hamming

Centroid	Mean	Mode

2.3 Hierarchical Clustering

Builds a cluster tree (dendrogram).

Types:

1) Agglomerative (bottom-up)
2) Divisive (top-down)

Figure: Dendrogram Example

```
     ┌──────── Cluster A
  ┌────────┤
  │   └──────── Cluster B
──┤
  │    ┌──────── Cluster C
  └─────────┤
     └──────── Cluster D
```

Distance Linkage Methods:

1. Single linkage
2. Complete linkage
3. Average linkage

2.4 Density-Based Clustering (DBSCAN)

Forms clusters using dense regions of points.

Key Concepts:

Core Point: Minimum points within radius

Border Point: Near a core point

Noise: Sparse points

Diagram: DBSCAN Concept

● = Core Point

○ = Border Point

x = Noise

●●●○ x ●●

●●○○○ ●●

Advantages: Detects arbitrarily shaped clusters

Disadvantages: Difficult to tune parameters

3. Self-Organizing Maps (SOM)

Neural-network-based clustering method mapping high-dimensional data to 2D grid.

Components:

- Input layer
- Output grid (usually 2D)

Workflow Diagram:

Input → Best Matching Unit → Update Neighboring Neurons → Repeat

Applications:

- Visualizing high-dimensional data
- Market segmentation

4. Expectation Maximization (EM)

Used in probabilistic clustering.

Often applied to Gaussian Mixture Models (GMMs).

Steps:

1. Expectation (E-step): Estimate probability of each point belonging to clusters
2. Maximization (M-step): Update parameters (means, variance)

Figure: EM Loop

E-Step →Update probs →M-Step →Update parameters →Repeat

5. Principal Component Analysis (PCA)

Used for dimensionality reduction.

Process:

1. Standardize data
2. Compute covariance matrix
3. Compute eigenvectors/eigenvalues
4. Select top components

Diagram: PCA Transformation

Original Axes →Rotate Axes →New PC1 & PC2

Table: PCA Uses

Use Case	Benefit
Visualization	Reduce to 2D/3D
Noise Removal	Drop low-variance components

6. Reinforcement Learning

Learning by interacting with environment.

Key Elements:

Element	Description
Agent	Learner/decision maker
Environment	World where agent interacts
Action	Moves taken by agent
Reward	Feedback signal
Policy	Strategy

Diagram: RL Loop

```
┌──────────────┐
│   Agent    │
└──────┬───────┘
       │ Action
       ▼
┌──────────────┐
│ Environment │
└──────┬───────┘
       │ Reward + State
       ▼
```

Types of RL:

1. Model-free: Q-learning, SARSA
2. Model-based: Uses learned environment model

- ## Additional Diagrams, Examples, and Solved Problems

1. K-Means – Solved Numerical Example

Dataset: Points: (1,1), (2,1), (4,3), (5,4)

Step 1: Choose k = 2 and initialize centroids

C1 = (1,1)

C2 = (5,4)

Step 2: Assign points to nearest centroid

Point	Dist to C1	Dist to C2	Cluster
(1,1)	0	5	C1
(2,1)	1	4.24	C1

(4,3)	3.61	1.41	C2
(5,4)	5	0	C2

Step 3: Recompute centroids

C1 = mean((1,1),(2,1)) = (1.5,1)

C2 = mean((4,3),(5,4)) = (4.5,3.5)

Step 4: Reassign (converges after one more iteration)

Final Clusters:

Cluster 1: (1,1),(2,1)

Cluster 2: (4,3),(5,4)

Diagram:

```
y
5|      ●(5,4)
4|    ●(4,3)
3|
2|
1| ●(1,1) ●(2,1)
0+---------------- x
```

2. Hierarchical Clustering – Example

Dataset: A(1), B(2), C(8), D(9)

Distances:

A-B = 1

C-D = 1

B-C = 6

Process:

Merge A & B

Merge C & D

Merge (AB) & (CD)

Dendrogram:

```
    ┌──── A
  ┌───┤
  │  └──── B
──┤
  │  ┌──── C
  └───┤
    └──── D
```

3. DBSCAN Example

Parameters: eps = 1.5, minPts = 3

Dataset: Points forming 2 dense blobs + 1 isolated point.

Result:

Two clusters detected based on density

Outlier marked as noise

Diagram:

```
●●●○    x
●●○○●  ●●●
```

Where:

● = core points

○ = border points

x = noise

4. Self-Organizing Map (SOM) – Example

Dataset: Features of animals (size, speed)

SOM Grid Result:

```
+----------+----------+----------+
|  Cat    |  Dog    |  Wolf  |
+----------+----------+----------+
| Sparrow |  Parrot |   Hawk   |
+----------+----------+----------+
```

Animals with similar characteristics cluster together.

5. Expectation-Maximization – Numerical Example (Simplified)

Data: 1D points = {1, 2, 8, 9} Assume 2 Gaussians.

Initialization:

Means: $\mu1=2$, $\mu2=8$

E-Step: Assign soft probabilities based on distance.

M-Step: Update means:

$\mu1 = \text{mean}(1,2) = 1.5$

$\mu2 = \text{mean}(8,9) = 8.5$

Repeat until convergence.

6. PCA – Worked Example

Dataset:

X	Y
2	0
0	2

Step 1: Compute covariance Var(X)=2, Var(Y)=2, Cov(X,Y)=0

Step 2: Eigenvalues = 2,2 (axes equally important)

Step 3: Principal components = X and Y axes

Diagram:

Original Axes = PCA Axes (since uncorrelated)

7. Reinforcement Learning Example

Agent navigating a grid to reach a goal.

Grid:

S → ☐ → ☐ → G

S = Start, G = Goal

Rewards:

Move = −1

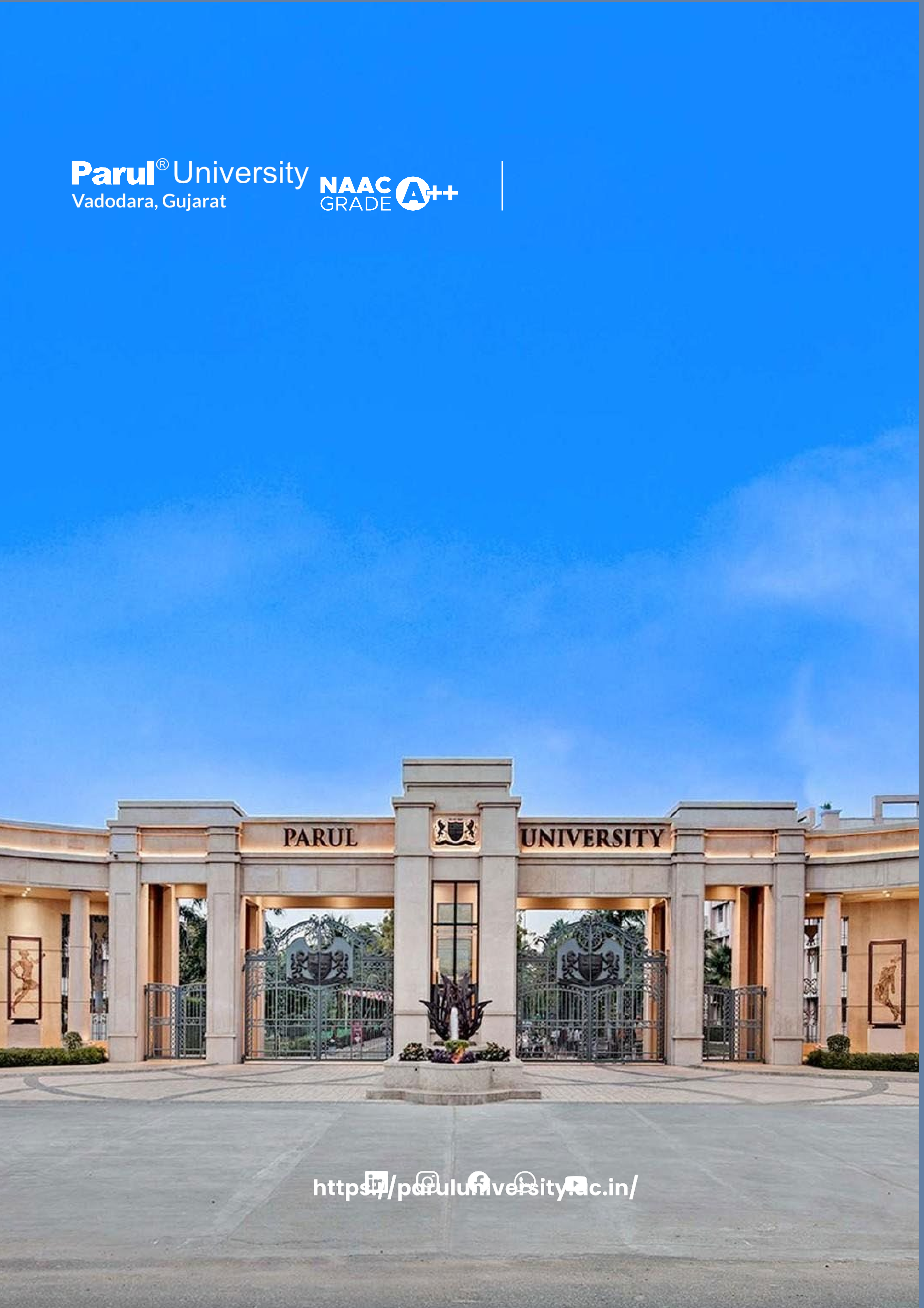
Reach goal = +10

Q-learning update: $Q(s,a) = Q + \alpha(r + \gamma \max Q' - Q)$

After multiple episodes, the agent learns shortest path.

Parul® University
Vadodara, Gujarat

NAAC GRADE A++



https://paruluniversity.ac.in/