Usupervised machine learning algorithms

Class 28

13/9/2025

Acknowledgement

The series of the IT & Japanese language course is Supported by AOTS and OEC.

Ministry of Economy, Trade and Industry

Overseas Employment Corporation

What you have Learnt Last Week

We were focused on following points.

- Usage of function, loop, and Numpy
- Software development Life cycle
- Importance of Security compliance, Bash Scripting, Ansible, docker and docker compose
- API testing with Postman and Introduction of Jira
- IAM Permission and S3 bucket
- Introduction to AWS, Azure and GCP
- Supervised Machine Learning Algorithms

What you will Learn Today

We will focus on following points.

1. Introduction to Un-Supervised Machine Learning
2. Overview of Popular Un-Supervised Learning Algorithms
3. Applications of Un-Supervised Machine learning algorithms
4. Types of hierarchical clustering
5. Q&A Session

Introduction to Unsupervised Machine Learning

Definition & Key Concepts

Definition: Machine learning on **unlabeled data** (no predefined output).

Goal: Find hidden patterns, groupings, or structure.

Key Concepts:

- **Clustering:** Group similar items together.
- **Dimensionality Reduction:** Reduce features while keeping information.

Examples: Customer segmentation, anomaly detection, topic modeling.

Overview of Popular Algorithms

Categories of Unsupervised Learning

Clustering Algorithms:

- K-Means
- DBSCAN
- Hierarchical Clustering
- Gaussian Mixture Models (GMM)

Dimensionality Reduction:

- PCA
- t-SNE
- UMAP

Applications of Unsupervised Learning

Real-World Use Cases

- **Healthcare:** Patient grouping, gene expression analysis.
- **Finance:** Fraud detection, risk segmentation.
- **E-Commerce & Marketing:** Customer segmentation, recommendation systems.
- **NLP:** Topic modeling, word embeddings.
- **Cybersecurity:** Anomaly detection in logs.

Clustering Techniques

Hierarchical vs. Partitional

Hierarchical Clustering: Builds a tree (dendrogram) of clusters.

- **Agglomerative:** Start with each point as a cluster, merge step by step.
- **Divisive:** Start with one big cluster, split into smaller clusters.

Partitional Clustering: Directly divides data into k groups (e.g., K-Means).

K-Means Clustering

Partitional Clustering Method

Steps:

1.Choose number of clusters (k).
2.Assign points to nearest cluster center.
3.Recalculate cluster centers → repeat until stable.

- **Strengths:** Simple, efficient.

- **Limitations:** Requires predefined k, sensitive to outliers.

- **Use Case:** Customer segmentation in marketing.

DBSCAN (Density-Based Clustering)

Density-Based Approach

Groups points that are closely packed together.

Can detect **arbitrary-shaped clusters** and noise.

Advantages: Finds outliers, no need to specify number of clusters.

Limitations: Struggles if density varies a lot.

Use Case: Anomaly detection in credit card fraud.

Hierarchical Clustering

Agglomerative & Divisive

Agglomerative (Bottom-Up):

- Each data point starts as its own cluster.
- Merge closest pairs until all points form one cluster.

- **Divisive (Top-Down):**
  - Start with one cluster → recursively split into smaller clusters.

- Visualized using **dendrograms** (tree-like diagrams).

Gaussian Mixture Models (GMM)

Probabilistic Clustering

Assumes data is generated from multiple Gaussian distributions.

Each cluster has a probability distribution.

- **Strengths:** Flexible, handles overlapping clusters.
- **Limitations:** Computationally expensive, may overfit.
- **Use Case:** Speech recognition, soft clustering in finance.

Dimensionality Reduction

Why Reduce Dimensions?

High-dimensional data = harder to visualize & process.

Goals:

- Reduce features while preserving structure.
- Improve efficiency & visualization.

Techniques: PCA, t-SNE, UMAP.

PCA, t-SNE & UMAP

Popular Dimensionality Reduction Methods

PCA (Principal Component Analysis):
- Projects data onto fewer dimensions.
- Best for linear patterns.

t-SNE:
- Non-linear, preserves local structure.
- Good for visualizing clusters.

UMAP:
- Similar to t-SNE but faster & preserves both local & global structure.

Applications: Data visualization, preprocessing for ML models.

Association Rule Learning

Discovering Relationships in Data

Definition: Finds hidden relationships between items in large datasets.

Rule Format: *IF (item A) THEN (item B)*.

Measures:

- **Support:** Frequency of itemset.
- **Confidence:** Likelihood of B given A.
- **Lift:** Strength of the rule compared to random chance.

Applications: Market basket analysis, cross-selling.

Apriori & Eclat Algorithms

Techniques for Association Rules

Apriori Algorithm:

- Uses iterative approach.
- Generates frequent itemsets → derives association rules.

Eclat Algorithm:

- Uses vertical data format (item–transaction lists).
- More efficient for large datasets.

Both widely used in retail & recommendation systems.

Market Basket Analysis

Classic Application of Association Rules

Goal: Find products often bought together.

Example: “Customers who buy bread also buy butter.”

Business Use:

- Cross-selling strategies.
- Store layout & promotions.

Widely used in Amazon, Walmart, e-commerce platforms.

Anomaly Detection in Unsupervised Learning

Identifying Outliers

Definition: Detect unusual patterns that deviate from normal behavior.

Use Cases: Fraud detection, network intrusion, equipment failure.

Approaches:

- **Isolation Forest**
- **Autoencoders**

Isolation Forest

Tree-Based Anomaly Detection

Works by randomly partitioning data.

Outliers are easier to isolate (require fewer splits).

Advantages: Fast, scalable, effective on high-dimensional data.

Application: Credit card fraud detection, cybersecurity.

Autoencoders for Anomaly Detection

Neural Network Approach

Autoencoders: Neural networks that learn compressed representations.

Anomaly Detection:

- Train on normal data → low reconstruction error.
- High error = anomaly.

Applications: Fraud detection, medical imaging, sensor data.

Evaluation Metrics for Unsupervised Learning

Cluster Validation

Silhouette Score: Measures how similar a point is to its own cluster vs. others.

Davies-Bouldin Index: Lower = better clustering.

Dunn Index: Higher = better clustering separation.

Used to compare clustering quality without labels.

Challenges in Unsupervised Learning

Practical Issues

Choosing Number of Clusters: Hard to decide optimal k.

High-Dimensional Data: More features = harder clustering.

Interpretability: Results may be less intuitive for decision-making.

Scalability: Some algorithms struggle with big datasets.

Real-World Use Cases of Unsupervised Learning

Industry Applications

Customer Segmentation: Grouping users for targeted marketing.

Recommendation Systems: Suggesting products/content.

Fraud Detection: Identifying unusual transactions.

Document/Topic Clustering: Organizing news, research papers, or support tickets.

Future Trends in Unsupervised Learning

Emerging Directions

Self-Supervised Learning: Uses unlabeled data to generate pseudo-labels.

Hybrid Approaches: Combine unsupervised + reinforcement learning.

AI-Driven Automation: Smart assistants, autonomous systems.

Goal: More intelligent, adaptive, and explainable AI systems.

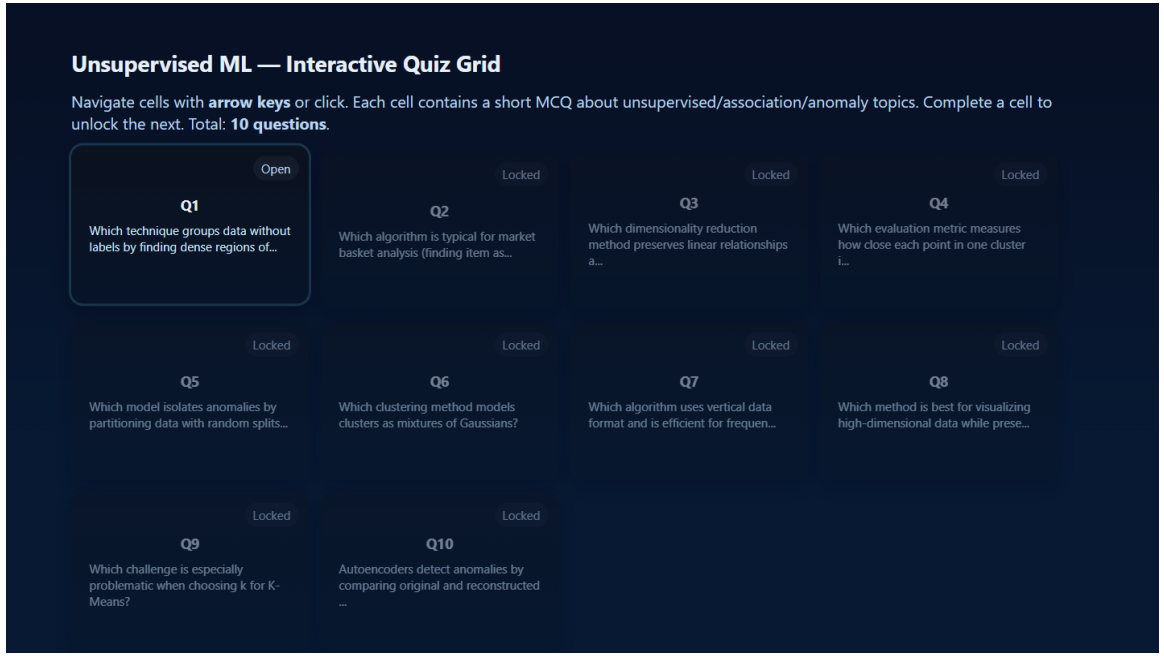
Game 1

Cloud Fundamentals Maze

Step1: Start the Game by Clicking the Link

Step2: Click on the Game It will Start

https://codepen.io/HT-Design/full/EaVzBOg



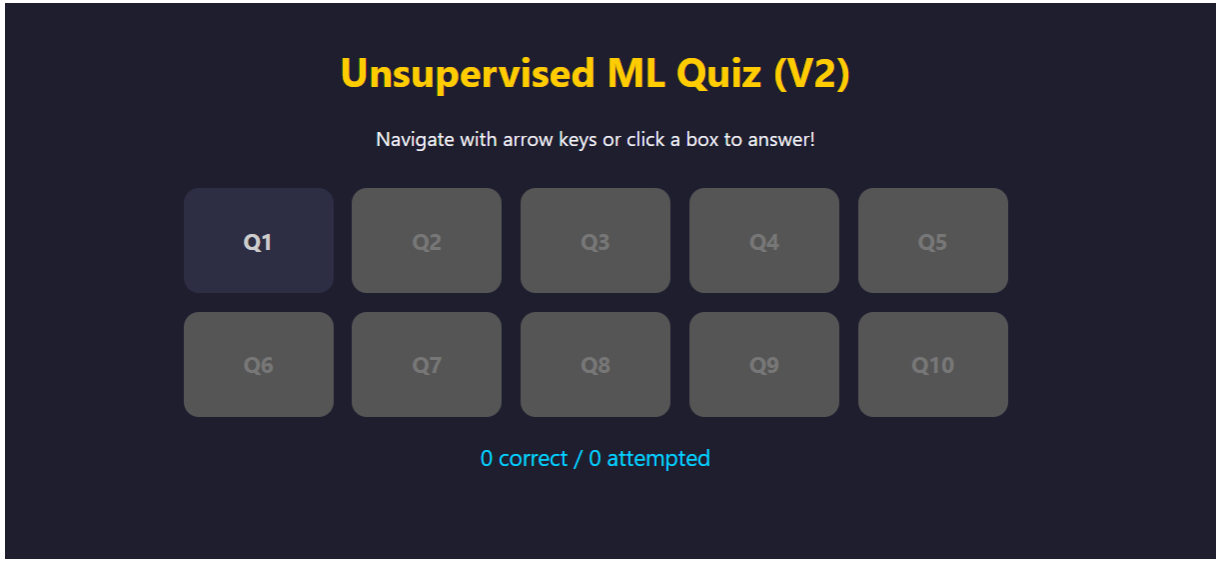
Game 2

Cloud Services Maze

Step1: Start the Game by Clicking the Link

Step2: Click on the Game It will Start

https://codepen.io/HT-Design/full/wBKbVvZ



Assignment

Quiz Section

Quiz

Everyone student should click on submit button before time ends otherwise MCQs will not be submitted

[Guidelines of MCQs]

1. There are 20 MCQs
2. Time duration will be 10 minutes
3. This link will be share on 12:25pm (Pakistan time)
4. MCQs will start from 12:30pm (Pakistan time)
5. This is exact time and this will not change
6. Everyone student should click on submit button otherwise MCQs will not be submitted after time will finish
7. Every student should submit Github profile and LinkedIn post link for every class. It include in your performance

Assignment

Assignment should be submit before the next class

[Assignments Requirements]

1. Create a post of today's lecture and post on LinkedIn.
2. Make sure to tag @Plus W @Pak-Japan Centre and instructors LinkedIn profile
3. Upload your code of assignment and lecture on GitHub and share your GitHub profile in respective your region group WhatsApp group
4. If you have any query regarding assignment, please share on your region WhatsApp group.
5. Students who already done assignment, please support other students

Q&A Session

ありがとうございます。
Thank you.
شکریا

For the World with Diverse Individualities