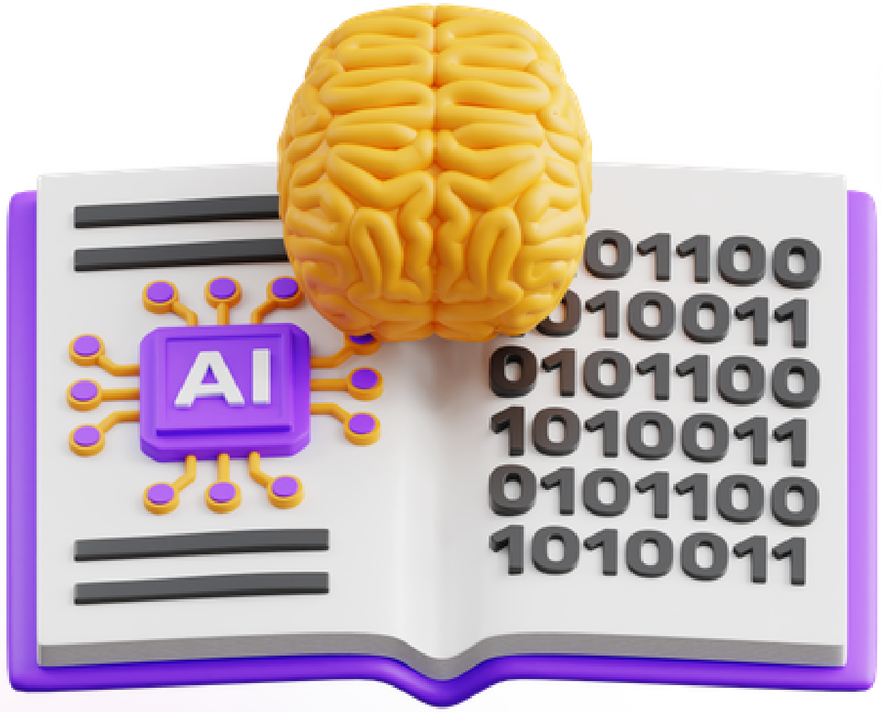


UNSUPERVISED LEARNING

About Unsupervised Learning

- Branch of Machine Learning
- Unsupervised learning is a branch of machine learning where models discover patterns, relationships, or structure in data without any labeled outputs (i.e., no target “answers” are provided during training). Instead, the algorithm explores the input data and tries to organize it in meaningful ways.
  - Models are not supervised using training dataset.
- Deals with unlabeled data (data that doesn’t have an output or category assigned to it.) and finds hidden patterns from the given dataset

“A type of machine learning in which models are trained using unlabeled dataset and are allowed to act on that data without any supervision. “

Key features

1. No supervision

2. Unlabeled Data

3. Pattern Discovery
   - The goal is to explore the data and identify meaningful insights, such as:
     - Groups or clusters
     - Anomalies (outliers)
     - Associations or correlations

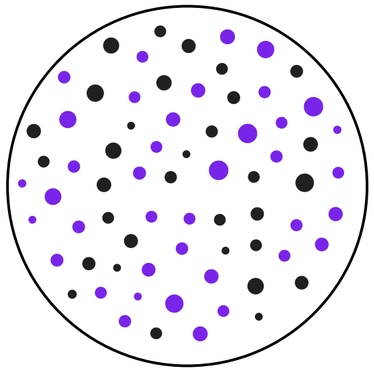
Goal

Find underlying structure of a dataset, group the data according to similarities, and represent the dataset in a compressed format.

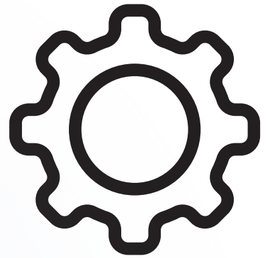
Why Unsupervised Learning

In real-world we do not always have input data with the corresponding output, so to solve such cases we need unsupervised learning

Raw input data



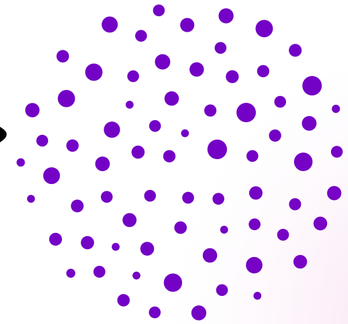
Learning Algiorthm

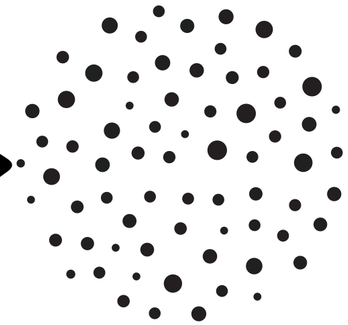


Interpretation

Unknown ouptut
No training data set

Interpretation





Output

Common approaches

Clustering

Clustering is a technique which groups unlabeled data based on their similarities or differences. Clustering algorithms are used to process raw, unclassified data objects into groups represented by structures or patterns in the information.

It finds some similar patterns in the unlabelled dataset such as shape,size, colour, etc., and divides them as per the absence or presence of those patterns.

After applying this technique, each cluster or group is provided with a cluster-ID which ML system use to simplifying large datasets

It can be categorized into a few types, specifically exclusive, overlapping, hierarchical, and probabilistic

Cluster Algorithms

- Based on how they assign points

Exclusive clustering stipulates a data point can exist only in one cluster. This can also be referred to as “hard” clustering. Example, K-Means

Overlapping Clustering differs from exclusive clustering in that it allows data points to belong to multiple clusters with separate degrees of membership (like probabilities or weights). Useful when boundaries between groups are fuzzy.

Hierarchical Clustering builds a hierarchy of clusters, either by merging or splitting them, an example is the agglomerative Hierarchical Clustering. The two main types are

- Agglomerative (bottom-up) , start by each data point as its own cluster then you merge the closest clusters until all points are in one cluster.
- Divisive (top-down), all data is one single cluster and then split to smaller clusters.
- Similarity Methods (used to decide how to merge/split): ward’s linkage, average linkage, complete linkage, single linkage

Probabilistic Clustering data points are clustered based on the likelihood that they belong to a particular distribution. The Gaussian Mixture Model (GMM) is the one of the most commonly used probabilistic clustering methods.

Association Rules

This technique finds out some very useful relations between parameters of a large data set.

They're particularly common in market basket analysis, where the goal is to discover how items are often purchased together or in recommendation system such as spotify to recommend content based on previous behaviour

How it works:

- First identify items that frequently occur together(Frequent item sets)
- Generate association rule that describes the relationship between the items
- Explore possible itemsets, making sure to only check combinations that meet a certain support threshold

Association Rules Algorithms

Apriori Algorithm, operates in a breadth-first manner. How it works:

- Scan the dataset for individual items and their frequencies, then set a minimum support threshold.
- Generate larger itemsets by combining frequent items iteratively.
- Remove itemsets that don’t meet the minimum support, using the Apriori Property (if an itemset is infrequent, so are its supersets).
- Create rules based on frequent itemsets,

Eclat Algorithm, operates in a depth-first search reducing the number of database scans. How it works:

- Create Tidset by generating a list of transaction IDs (Tidset) where each item appears.
- Calculate Support by finding intersecting Tidsets.
- Recursively combine itemsets, checking support by intersecting Tidsets, until no more frequent itemsets are found.
- Create rules based on frequent itemsets,

FP-Growth Algorithm, operates by building a frequent-pattern tree to compactly store itemsets and their counts.It then breaks the tree into smaller conditional trees to find frequent patterns efficiently. Finally, it extracts frequent itemsets and generates association rules based on a minimum support threshold.

Efficient Tree-based Algorithm, uses a tree-like structure (e.g., dendrogram) to group data into nested clusters.

Dimensionality Reduction

Dimensionality reduction is a technique used when the number of features, or dimensions, in a given dataset is too high. It reduces the number of data inputs to a manageable size while also preserving the integrity of the dataset as much as possible.

Commonly used in the preprocessing data stage.

Dimensionality Reduction Algorithms

- **Principal Component Analysis (PCA):** Reduces dimensions by transforming data into uncorrelated principal components.
- **Linear Discriminant Analysis (LDA):** Reduces dimensions while maximizing class separability for classification tasks.
- **Non-negative Matrix Factorization (NMF):** Breaks data into non-negative parts to simplify representation.
- **Locally Linear Embedding (LLE)**: Reduces dimensions while preserving the relationships between nearby points.
- **Isomap:** Captures global data structure by preserving distances along a manifold.

Challenges
Assumption Dependence: Algorithms often rely on assumptions (e.g., cluster shapes), which may not match the actual data structure.
Overfitting Risk: Overfitting can occur when models capture noise instead of meaningful patterns in the data.
Limited Guidance: The absence of labels restricts the ability to guide the algorithm toward specific outcomes.
Lack of Ground Truth: Unsupervised learning lacks labeled data, making it difficult to evaluate the accuracy of results.

Real World Applications
Natural Language Processing: Analyzing and understanding text data to identify topics, sentiments, and relationships. e.g, categorizing news articles by topic, identifying the sentiment of customers, or translating languages. Topic modeling, clustering and word embeddings
Image and Video Analysis: Analyzing images and videos to identify objects, scenes, and patterns using computer vision techniques. e.g., analyzing images in the medical field to automatically identify and analyze patterns in order to detect diseases early on. Clustering, dimensionality reduction, and autoencoders.
Customer segmentation: Grouping customers based on similar behaviors, demographics, or preferences to tailor marketing strategies. e.g., identifying distinct customer segments based on purchasing patterns, allowing for targeted advertising and product recommendations. K-means or hierarchical clustering.
Recommendation Systems: Recommending products, content, or services to users based on their preferences or past behavior. e.g., suggesting movies to watch , products to buy, or articles to read based on the user's history and preferences. Collaborative filtering and content-based recommendation
Anomaly Detection: Identifying unusual or abnormal patterns in data to detect fraud, network intrusions, or equipment malfunctions. e.g., detecting fraudulent credit card transactions by identifying unusual spending patterns. Isolation forests or one-class SVM

Comparison with Supervised Learning

Feature	Supervised Learning	Unsupervised Learning
Data	Labeled	Unlabeled
Goal	Predict output	Find patterns/structure
Examples	Regression, Classification	Clustering, PCA