

Unsupervised Learning Models

HIGHLIGHTS

- Unsupervised Learning
 - Supervised learning vs. Unsupervised learning
 - Why Unsupervised Learning?
 - Some drawbacks of Unsupervised Learning
 - Parametric Unsupervised Learning
 - Non-parametric Unsupervised Learning
 - Unsupervised Learning Models
 - Types of Unsupervised Learning Algorithms
 - Classical unsupervised ML problem: clustering (explanation, motivation, etc)
 - Hierarchical clustering: Complete linkage, Single linkage, Average linkage, Centroid linkage
 - K-means Clustering
 - Principal Component Analysis
 - Association Rules
 - Summary
-

Unsupervised Learning

Unsupervised Learning is a machine learning technique in which there is no known target. Unlike supervised learning models, that works on a known target where users supervise the model, unsupervised learning techniques or models work with no known target or label. The users do not need to supervise these models. In unsupervised learning, the model works on its own to detect

unknown patterns and information in the data that was previously unknown. While supervised learning models work on labeled data or a known target, unsupervised learning models mainly deal with unlabeled data.

Unsupervised learning models try to find any similarities, differences, patterns, and structure in data by themselves with no human intervention.

As an example of unsupervised learning, consider a child in a family with a pet dog. The child is very familiar with the dog and knows what a dog looks like but does not know that there are lots of other dogs that are different. Now, when the child sees another dog, he will still be able to recognize a dog with its features (two ears, four legs, tail, appearance, etc.). This type of prediction is known as unsupervised learning, where the child is not told or trained as to what a dog looks like. If the child was trained to recognize a dog, it would be an example of a supervised learning problem.

Supervised learning vs unsupervised learning

The differences between unsupervised vs supervised learning are explained in Table 7.1.

Table 7.1 Differences between unsupervised vs supervised learning

	Unsupervised Learning	Supervised Learning
	The model tries to find patterns in the data without a label or target. This type of learning is done without human supervision.	The input data is labeled with answer keys telling the model the desired output.
	No label is provided (the data is unlabeled). Unsupervised learning is used to cluster information in datasets.	The data is labeled (the desired output is provided).
	The model is given input data (x) but no output or target. In other words, the data is not labelled.	The model is provided with input variables or features (x) and the output variable (desired output) or the target (y). The data is labelled in this case.
		Applied to Classification and Regression problems. Classification problems classify the

	Unsupervised learning methods are applied to clustering and association problems.	data into one group or another, or to which class an instance belongs. Regression problems are used to predict the future outcome (in many cases, the predict continuous variables, such as demand, sales, temperature, etc.
	The desired results may not be as accurate.	Provides better accuracy in the results.
Common algorithms	Common classification algorithms are: • Clustering (a technique that groups unlabeled data based on their similarities or differences). • k-means • Gaussian Mixture Models • Principal Component Analysis (Dimensionality reduction) • Association Rules	Common classification algorithms are Logistic Regression, Support Vector Machines, Decision Tree Classification, and K-Nearest Neighbours, and others outlined below: • Regression Models (different types) • Linear Discriminant Analysis • Support vector machines • Decision trees • Naïve Bayes • Random forest
	Challenges of Unsupervised Learning: Unsupervised learning is more difficult and involved than supervised learning as it does not have corresponding output. The unsupervised learning algorithm may be less accurate as input data is not labeled, and algorithms do not know the exact output.	Challenges of Supervised Learning: Requires a huge amount of data. Nowadays, machine learning algorithms are trained with millions of data. Collecting and labeling a huge quantity of data is involved and is a tedious process.
Applications	Data labeling: Using clustering, the data is clustered and subsequently labeled. Image compression: Unsupervised learning algorithms are used to reduce the number of dimensions in a dataset, preserving most of the information in the data. Anomaly detection: Unsupervised learning models are used for anomaly detection (a method to detect outliers and unusual data points). Some other applications include: • data exploration, • customer segmentation, • recommender systems, • target marketing campaigns, • data preparation and visualization, etc.	Supervised learning models have been used to assist doctors in diagnosing or predict disease in patients before they manifest. Other examples are spam detection (detect spam sent via email, sentiment detection, Self-driving cars, and many more

Why Unsupervised Learning?

Here, are prime reasons for using Unsupervised Learning in Machine Learning:

- Unsupervised learning is used to detect all kinds of unknown patterns in the data.
- Unsupervised learning is performed on unlabeled data that is easier to obtain and may be readily available from a computer.
- There are cases where it is difficult to determine the number of classes, or how many classes the data should be divided into.
- Unsupervised learning is very useful for raw data exploration needs.
- Unsupervised methods may help find features useful for categorization.
- Unsupervised methods are useful in searching for unknown similarities and differences in data and creating corresponding groups.
- The unsupervised method being used doesn't require training data to be labeled.

Some drawbacks of Unsupervised Learning:

- Unsupervised Learning is more complex compared to Supervised Learning tasks as no targets or labels in the data are available.
- It is difficult to be confident about the results as no labels are available to compare.
- Often external evaluation may be needed to justify the results.

Unsupervised learning may be parametric or non-parametric.

Parametric Unsupervised Learning

Parametric Unsupervised learning assumes that the population from which the data is extracted follows a certain probability distribution that has a fixed set of parameters. For example, if the normality assumption holds and is appropriate, then the distribution is defined by a normal curve with parameters mean (μ) and standard deviation (σ). This implies that if the underlying distribution can be safely assumed to be normal, then the probability of any future observation can be determined using the mean and standard deviation of the data.

Parametric unsupervised learning enables the construction of Gaussian Mixture Models in which the expectation-maximization algorithm can be used to predict the class of the sample in question. The parametric unsupervised model is involved and complex as there is no correct measure of accuracy in absence of labels.

Non-parametric Unsupervised Learning

In this version of unsupervised learning, the data is grouped into clusters, where each cluster provides some information about categories and classes present in the data. The method is commonly used to model data with small sample sizes. As the name suggests, there is no underlying assumption regarding the distribution of the population from which the data comes. Since there is no assumption about the distribution, the method is also referred to as a distribution-free method.

Unsupervised Learning Models

Types of Unsupervised Learning Algorithms

Clustering is the most widely used unsupervised learning problem. Unsupervised learning problems are generally grouped into clustering and association problems.

Clustering: a machine learning technique that groups *unlabeled* data based on their similarities or differences. Unsupervised learning is commonly used to cluster information in datasets. This method groups similar data points or objects into clusters that are not defined beforehand. In a cluster, the objects in the same group are more similar to each other than the objects in the other groups. The group of similar objects is called a **Cluster**.

In clustering, the technique or the model being used finds any patterns, similarities, and/or differences within uncategorized data. If any natural groups or classes exist in data, the model detects them. Cluster analysis seeks to find clusters in the data without having to prespecify a set of categories.

Types of Clustering Algorithms

The following are the commonly used clustering algorithms used in Machine Learning:

- Hierarchical clustering
- K-means clustering
- Singular Value Decomposition
- Independent Component Analysis

We will discuss some of these algorithms here.

A well-known clustering algorithm is K-means. K-means clustering is a common example of an exclusive clustering method where data points are assigned to K groups, where K is the number of clusters, based on the distance from each group's centroid. The data points closest to a given centroid will be clustered under the same category.

Hierarchical clustering

The hierarchical clustering algorithm puts similar objects into groups called clusters. The outcome is a set of clusters where each cluster is distinct from each other cluster and the objects within each

cluster are like each other. This method creates a hierarchy of clusters where two close clusters are in the same cluster. The algorithm ends when there is only one cluster left. Hierarchical clustering can be broadly divided into the following two groups:

- (1) Agglomerative: It is a bottom-up approach, in which the algorithm starts with each data point as a single cluster and merging them until one cluster is left.
- (2) Divisive: Divisive algorithm is the reverse of the agglomerative algorithm. it is a top-down approach.

Agglomerative Hierarchical Clustering is the most common type of hierarchical clustering for grouping objects into clusters based on their similarity. It's also known as AGNES (Agglomerative Nesting). It is a bottom-up approach where each observation is its own cluster, and pairs of clusters are merged as it moves up in the hierarchy. The steps for this process is as follows:

1. Take each data point as a single-point cluster thus forming N number of clusters.
2. Take the two closest data points and make them one cluster forming $N-1$ clusters.
3. Take the two closest clusters and make them one cluster, forming $N-2$ clusters.
4. Repeat step-3 until there is only one cluster.

Agglomerative Hierarchical Clustering is the most common type of hierarchical clustering for grouping objects into clusters based on their similarity. It's also known as AGNES (Agglomerative Nesting). It is a bottom-up approach where each observation is its own cluster, and pairs of clusters are merged as it moves up in the hierarchy. The steps for this process are as follows:

5. Take each data point as a single-point cluster, thus forming N number of clusters.
6. Take the two closest data points and make them one cluster, forming $N-1$ clusters.

7. Take the two closest clusters and make them one cluster, forming $N-2$ clusters.
8. Repeat step-3 until there is only one cluster.

Measuring the Distance between clusters

Different methods are used to decide the rules for forming clusters. These methods are known as Linkage Methods. The common linkage methods are:

- **Complete linkage:** in this method, the distance between two clusters is defined as the *largest* distance between two points in each cluster.
- **Single linkage:** in this method, the distance between two clusters is the *shortest* distance between two points in each cluster.
- **Average-linkage:** here the distance between two clusters is defined as the average distance between each point in one cluster and every point in the other cluster.
- **Centroid-linkage:** this method finds the centroid of each cluster, for example, cluster 1 and cluster 2, then calculates the distance between the two before merging.

The use of these methods is the choice of the analyst. There are no guiding rules for these methods, but different linkage methods may lead to different clusters.

These methods demonstrate how hierarchical clustering works. The entire process of this clustering method is usually demonstrated using a **dendrogram**, which is a tree diagram depicting the hierarchical relationships between sets of data. A Dendrogram memorizes or stores the steps of the hierarchical clustering algorithm. A dendrogram is a pictorial description that shows how the clusters are formed.