What is unsupervised learning?

Learn how unsupervised learning works and how it can be used to explore and cluster data

Learn about watsonx.ai →

What is unsupervised learning?

Unsupervised learning, also known as unsupervised machine learning, uses machine learning algorithms to analyze and cluster unlabeled datasets. These algorithms discover hidden patterns or data groupings without the need for human intervention. Its ability to discover similarities and differences in information make it the ideal solution for exploratory data analysis, cross-selling strategies, customer segmentation, and image recognition.

Featured products

Watson Studio

IBM Cloud Pak for Data

Begin your journey to AI

Learn how to scale AI →

Explore the AI Academy →

Common unsupervised learning approaches

Unsupervised learning models are utilized for three main tasks—clustering, association, and dimensionality reduction. Below we'll define each learning method and highlight common algorithms and approaches to conduct them effectively.

Clustering

Clustering is a data mining 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. Clustering algorithms can be categorized into a few types, specifically exclusive, overlapping, hierarchical, and probabilistic.

Exclusive and Overlapping Clustering

Exclusive clustering is a form of grouping that stipulates a data point can exist only in one cluster. This can also be referred to as “hard” clustering. The K-means clustering algorithm is an example of exclusive clustering.

- **K-means clustering** is a common example of an exclusive clustering method where data points are assigned into K groups, where K represents the number of clusters based on the distance from each group's centr to a given centroid will be clustered under the same ca be indicative of smaller groupings with more granularit will have larger groupings and less granularity. K-mean used in market segmentation, document clustering, im image compression.

Overlapping clusters differs from exclusive clustering in that it allows data poin belong to multiple clusters with separate degrees of membership. “Soft” or fuzzy K-means clustering is an example of overlapping clustering.

Hierarchical clustering

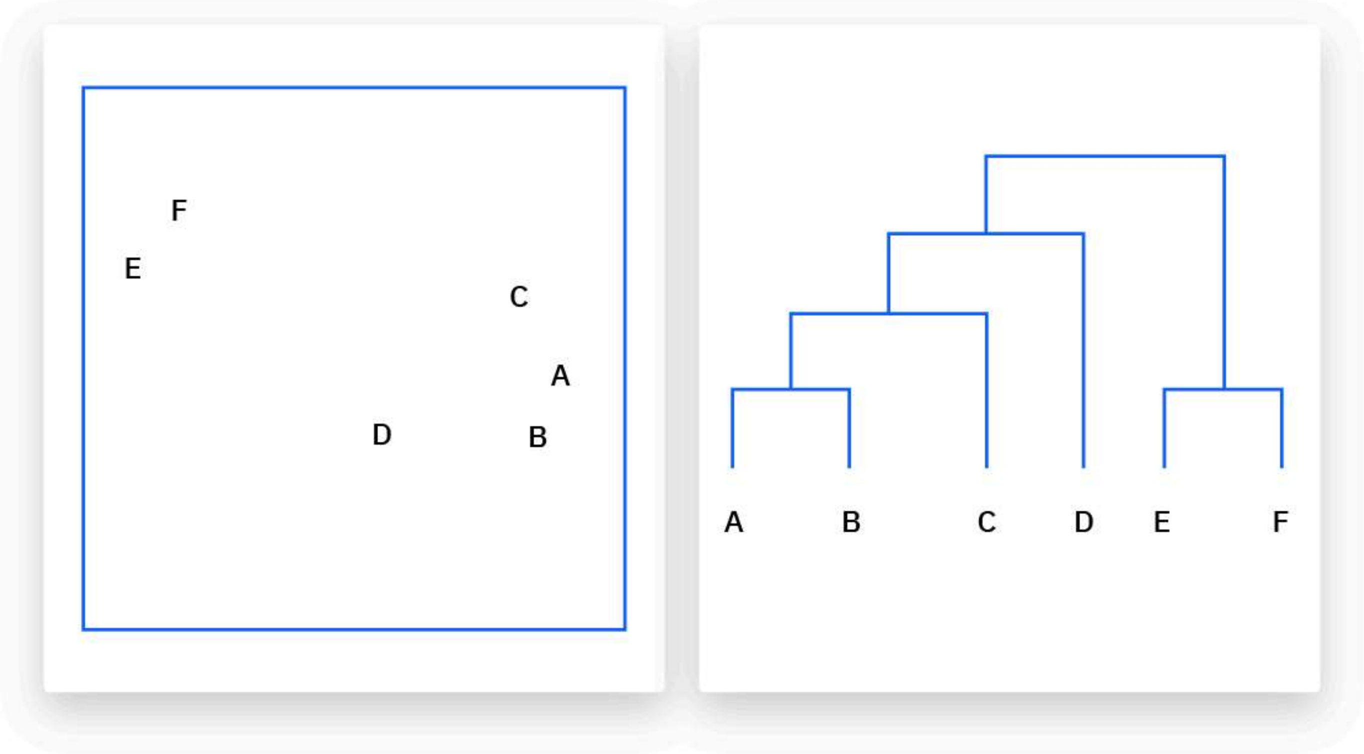
Hierarchical clustering, also known as hierarchical cluster analysis (HCA), is an unsupervised clustering algorithm that can be categorized in two ways: agglomerative or divisive.

Agglomerative clustering is considered a “bottoms-up approach.” Its data points are isolated as separate groupings initially, and then they are merged together iteratively on the basis of similarity until one cluster has been achieved. Four different methods are commonly used to measure similarity:

1. **Ward’s linkage:** This method states that the distance between two clusters is defined by the increase in the sum of squared after the clusters are merged.
2. **Average linkage:** This method is defined by the mean distance between two points in each cluster.
3. **Complete (or maximum) linkage:** This method is defined by the maximum distance between two points in each cluster.
4. **Single (or minimum) linkage:** This method is defined by the minimum distance between two points in each cluster.

Euclidean distance is the most common metric used to calculate these distances; however, other metrics, such as Manhattan distance, are also cited in clustering literature.

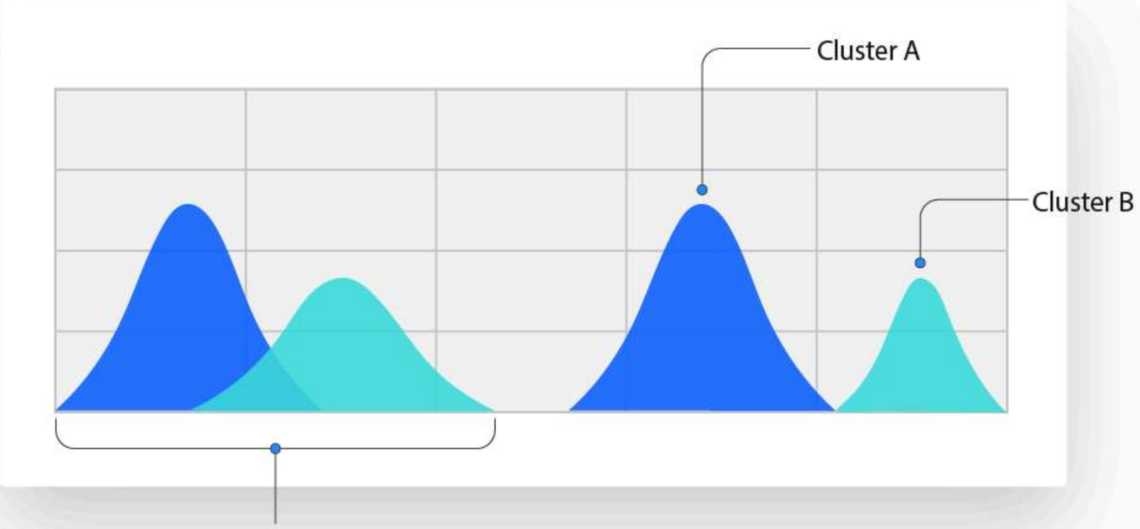
Divisive clustering can be defined as the opposite of agglomerative clustering; instead it takes a “top-down” approach. In this case, a single data cluster is divided based on the differences between data points. Divisive clustering is not commonly used, but it is still worth noting in the context of hierarchical clustering. These clustering processes are usually visualized using a dendrogram, a tree-like diagram that documents the merging or splitting of data points at each iteration.



Probabilistic clustering

A probabilistic model is an unsupervised technique that helps us solve density estimation or “soft” clustering problems. In 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.

- **Gaussian Mixture Models** are classified as mixture models, which means that they are made up of an unspecified number of probability distribution functions. GMMs are primarily leveraged to determine which Gaussian, or normal, probability distribution a given data point belongs to. If the mean or variance are known, then we can determine which distribution a given data point belongs to. However, in GMMs, these variables are not known, so we assume that a latent, or hidden, variable exists to cluster data points appropriately. While it is not required to use the Expectation-Maximization (EM) algorithm, it is a co the assignment probabilities for a given data point to a



Gaussian Mixture Models (GMMs) seek to group Cluster A and Cluster B accurately when distinct datasets are mixed together

Association Rules

An association rule is a rule-based method for finding relationships between variables in a given dataset. These methods are frequently used for market basket analysis, allowing companies to better understand relationships between different products. Understanding consumption habits of customers enables businesses to develop better cross-selling strategies and recommendation engines. Examples of this can be seen in Amazon's "Customers Who Bought This Item Also Bought" or Spotify's "Discover Weekly" playlist. While there are a few different algorithms used to generate association rules, such as Apriori, Eclat, and FP-Growth, the Apriori algorithm is most widely used.

Apriori algorithms

Apriori algorithms have been popularized through market basket analyses, leading to different recommendation engines for music platforms and online retailers. They are used within transactional datasets to identify frequent iter items, to identify the likelihood of consuming a product giv another product. For example, if I play Black Sabbath's rac their song "Orchid", one of the other songs on this channel song, such as "Over the Hills and Far Away." This is based as well as the ones of others. Apriori algorithms use a hash tree to count itemsets, navigating through the dataset in a breadth-first manner.

Dimensionality reduction

While more data generally yields more accurate results, it can also impact the performance of machine learning algorithms (e.g. overfitting) and it can also make it difficult to visualize datasets. 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. It is commonly used in the preprocessing data stage, and there are a few different dimensionality reduction methods that can be used, such as:

Principal component analysis

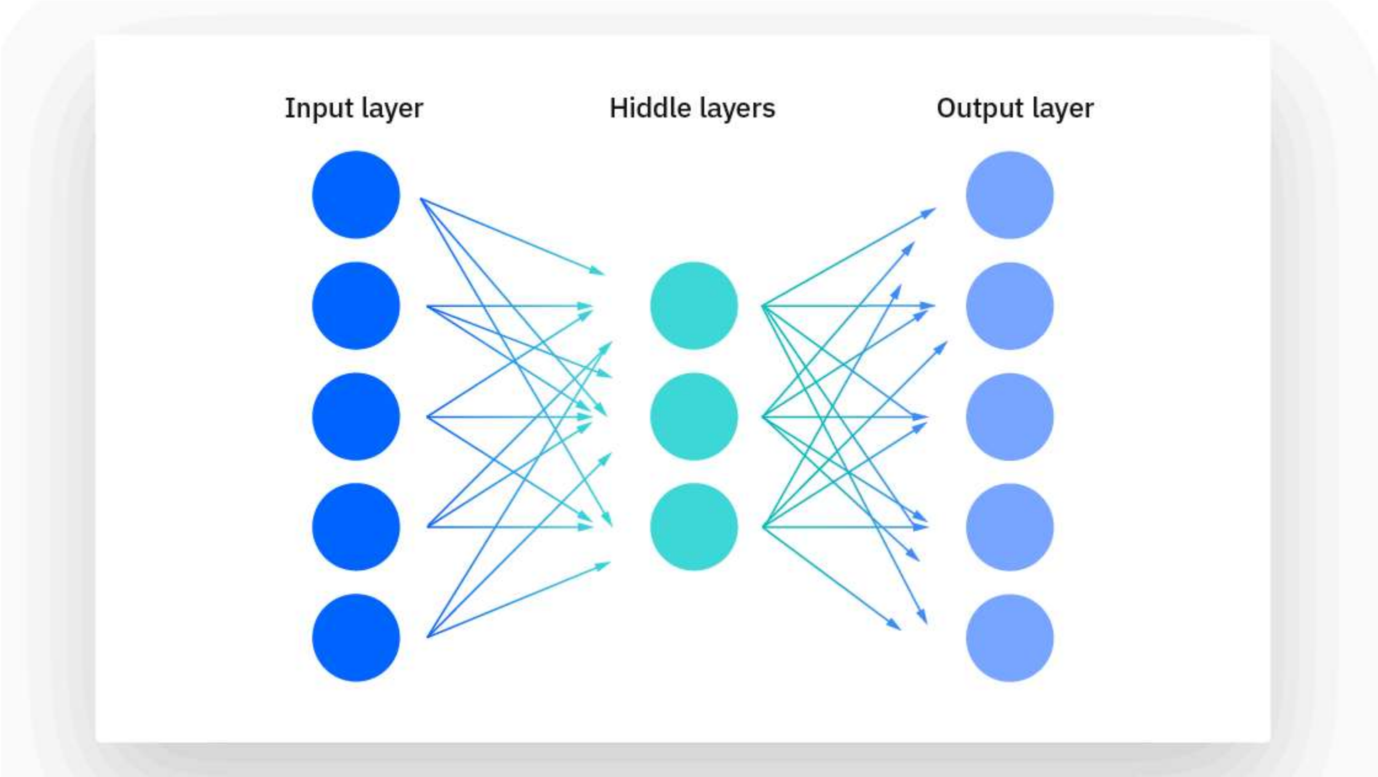
Principal component analysis (PCA) is a type of dimensionality reduction algorithm which is used to reduce redundancies and to compress datasets through feature extraction. This method uses a linear transformation to create a new data representation, yielding a set of "principal components." The first principal component is the direction which maximizes the variance of the dataset. While the second principal component also finds the maximum variance in the data, it is completely uncorrelated to the first principal component, yielding a direction that is perpendicular, or orthogonal, to the first component. This process repeats based on the number of dimensions, where a next principal component is the direction orthogonal to the prior components with the most variance.

Singular value decomposition

Singular value decomposition (SVD) is another dimensionality reduction approach which factorizes a matrix, A, into three, low-rank matrices. SVD is denoted by the formula, A = USVT, where U and V are orthogonal matrices. S is a diagonal matrix, and S values are considered singular values of matrix A. Similar to PCA, it is commonly used to reduce noise and compress data, such as image files.

Autoencoders

Autoencoders leverage neural networks to compress data and then recreate a new representation of the original data’s input. Looking at the image below, you can see that the hidden layer specifically acts as a bottleneck to compress the input layer prior to reconstructing within the output layer. The stage from the input layer to the hidden layer is referred to as “encoding” while the stage from the
layer is known as “decoding.”


Input layer
Hiddle layers
Output layer

Applications of unsupervised learning

Machine learning techniques have become a common method to improve a product user experience and to test systems for quality assurance. Unsupervised learning provides an exploratory path to view data, allowing businesses to identify patterns in large volumes of data more quickly when compared to manual observation. Some of the most common real-world applications of unsupervised learning are:

- **News Sections:** Google News uses unsupervised learning to categorize articles on the same story from various online news outlets. For example, the results of a presidential election could be categorized under their label for “US” news.
- **Computer vision:** Unsupervised learning algorithms are used for visual perception tasks, such as object recognition.
- **Medical imaging:** Unsupervised machine learning provides essential features to medical imaging devices, such as image detection, classification and segmentation, used in radiology and pathology to diagnose patients quickly and accurately.
- **Anomaly detection:** Unsupervised learning models can comb through large amounts of data and discover atypical data points within a dataset. These anomalies can raise awareness around faulty equipment, human error, or breaches in security.
- **Customer personas:** Defining customer personas makes it easier to understand common traits and business clients' purchasing habits. Unsupervised learning allows businesses to build better buyer persona profiles, enabling organizations to align their product messaging more appropriately.
- **Recommendation Engines:** Using past purchase behavior data, unsupervised learning can help to discover data trends that can be used to develop more effective cross-selling strategies. This is used to make relevant add-on recommendations to customers during the checkout process for online retailers.

Unsupervised vs. supervis
supervised learning

Unsupervised learning and supervised learning are frequently discussed togeth
Unlike unsupervised learning algorithms, supervised learning algorithms use labeled

data. From that data, it either predicts future outcomes or assigns data to specific categories based on the regression or classification problem that it is trying to solve. While supervised learning algorithms tend to be more accurate than unsupervised learning models, they require upfront human intervention to label the data appropriately. However, these labelled datasets allow supervised learning algorithms to avoid computational complexity as they don't need a large training set to produce intended outcomes. Common regression and classification techniques are linear and logistic regression, naïve bayes, KNN algorithm, and random forest.

Semi-supervised learning occurs when only part of the given input data has been labelled. Unsupervised and semi-supervised learning can be more appealing alternatives as it can be time-consuming and costly to rely on domain expertise to label data appropriately for supervised learning.

For a deep dive into the differences between these approaches, check out "Supervised vs. Unsupervised Learning: What's the Difference?"

Challenges of unsupervised learning

While unsupervised learning has many benefits, some challenges can occur when it allows machine learning models to execute without any human intervention. Some of these challenges can include:

- Computational complexity due to a high volume of training data
- Longer training times
- Higher risk of inaccurate results
- Human intervention to validate output variables
- Lack of transparency into the basis on which data was clustered

Related solutions

IBM Watson® Studio

Build and scale trusted AI on any cloud. Automate the AI lifecycle for ModelOps.

Explore IBM Watson® Studio →

IBM Cloud Pak® for Data

Connect the right data, at the right time, to the right people anywhere.

Explore Cloud Pak for Data →

IBM Cloud Solutions

Hybrid. Open. Resilient. Your platform and partner for digital transformation.

Explore Cloud Solutions →

Resources

How-to

Free, hands-on learning for generative AI technologies

Learn the fundamental concepts for

Blog

Supervised vs. Unsupervised Learning: What's the Difference?

Explore the basics of two data science

Article

The 3 mode mach learn

Learn about the three categories of

Article

Unsupervi on

Discover the th ideas behind

Schedule a call with an IBM sales representative who can assist you in finding the right products to meet your needs.

Let's talk

AI and generative AI, including prompt engineering, large language models and the best open source projects.

Learn more →

approaches: supervised and unsupervised. Find out which approach is right for your situation.

Read the post

algorithms: supervised, unsupervised, and reinforcement learning. See the ideas behind them and some key algorithms used for each

Read the article →

unsupervised learning and its applications. Read about the algorithms used in unsupervised learning for data classification

Read the article →

Take the next step

Train, validate, tune and deploy generative AI, foundation models and machine learning capabilities with IBM watsonx.ai, a next generation enterprise studio for AI builders. Build AI applications in a fraction of the time with a fraction of the data.

Explore watsonx.ai →

Book a live demo →

Schedule a call with an IBM sales representative who can assist you in finding the right products to meet your needs.

Let’s talk