Unsupervised Machine Learning — An In-Depth Overview

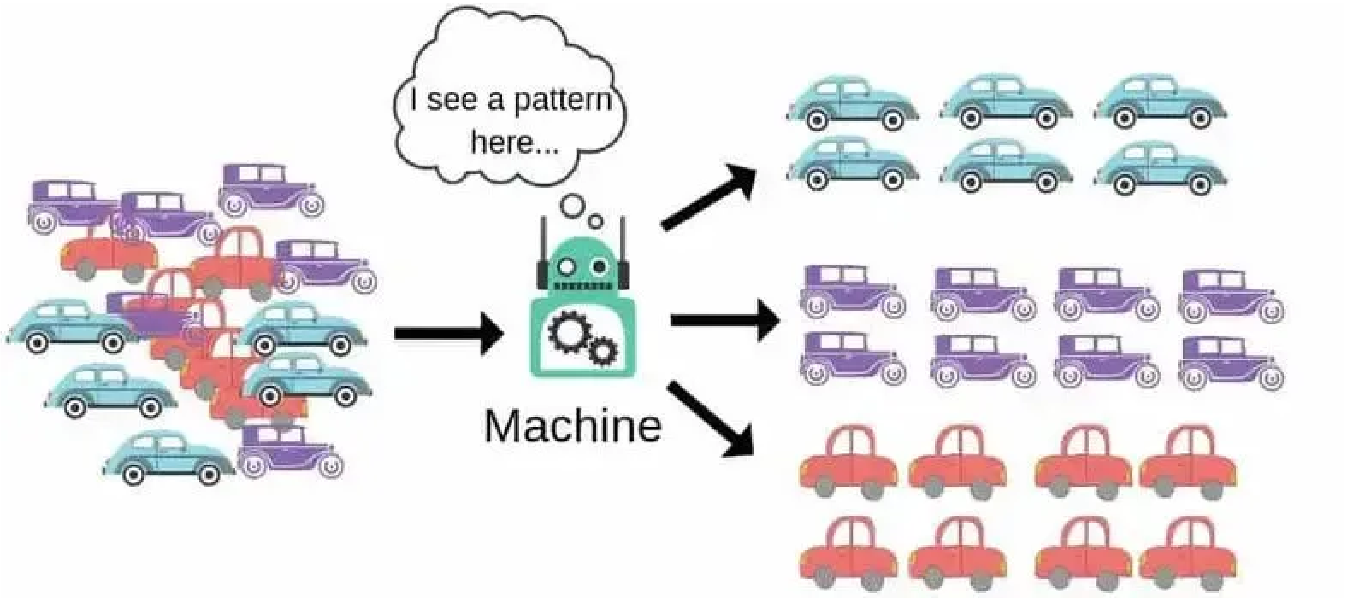
Unsupervised Machine Learning is one of the most commonly used Machine Learning modelling techniques as it has more relevance in real world applications. Read more about it in-depth in this article written by AI Club Research Member Sejal.Dubey Btech2022

AI Club - SIT · Follow

8 min read · Jul 31

Unsupervised machine learning is a type of machine learning where an algorithm learns patterns, relationships or structures in data without actually defining any instructions or stating labels . In simpler terms, we are asking it to find patterns or make sense of the data on its own by giving a bunch of data. Thus, the goal of unsupervised learning is to **find the principal structure of a given dataset, group that data according to similarities and represent that dataset in a squeezed format.**

HOW DOES IT WORK?



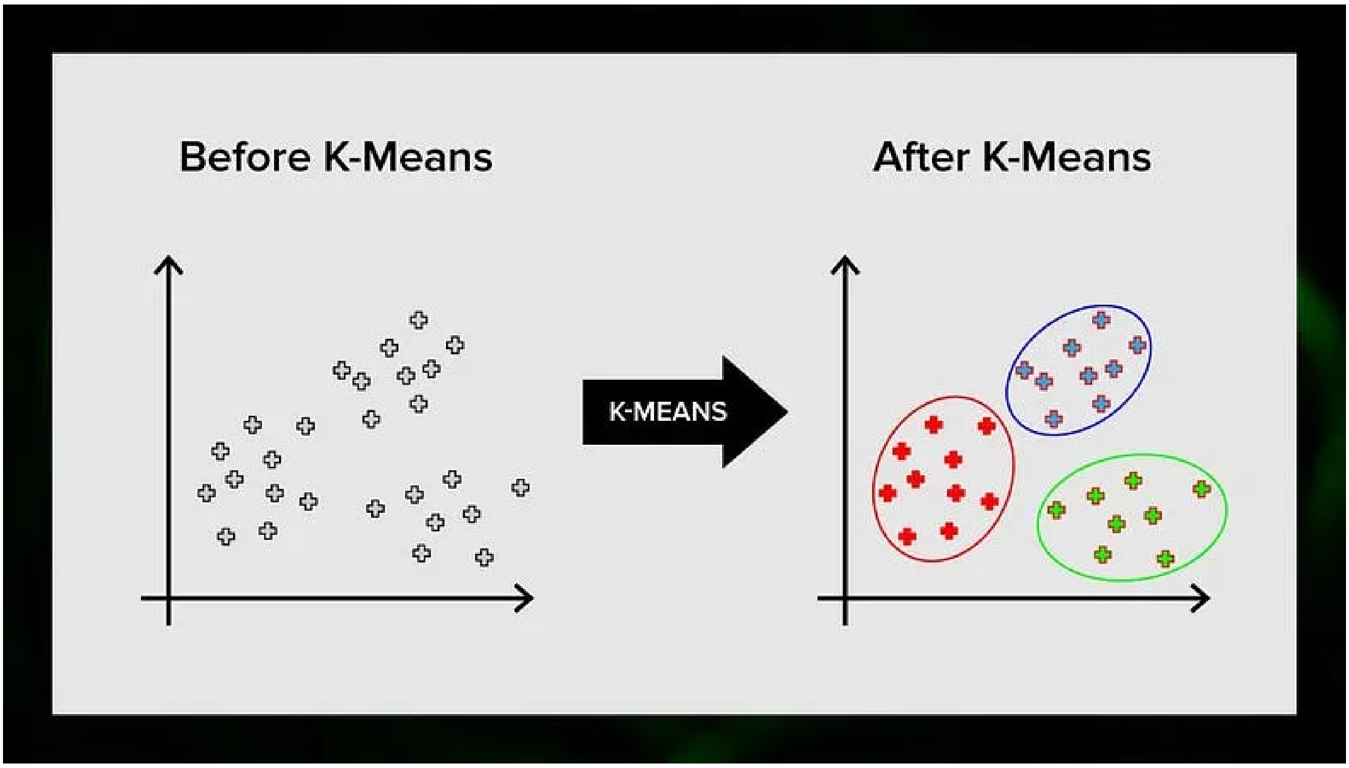
Suppose the unsupervised learning algorithm is given an input dataset containing images of different types of cars. The algorithm is never trained upon the given dataset, which means it does not have any idea about the features of the dataset. The task of the unsupervised learning algorithm is to identify the image features on their own. An unsupervised learning algorithm will perform this task by clustering the image dataset into groups according to similarities between images and then segregating it. There is no need for any manual supervision.

CLASSIFICATION OF UNSUPERVISED MACHINE LEARNING

1. **CLUSTERING**

Clustering is a technique where the algorithm groups similar data points together based on their built-in similarities or patterns. Without any prior knowledge of those clusters it aims to discover natural clusters within the data. There are subcategories of clustering in unsupervised machine learning including K-means clustering, hierarchical clustering, density-based clustering and Expectation-Maximization (EM) clustering.

a. K-means clustering:

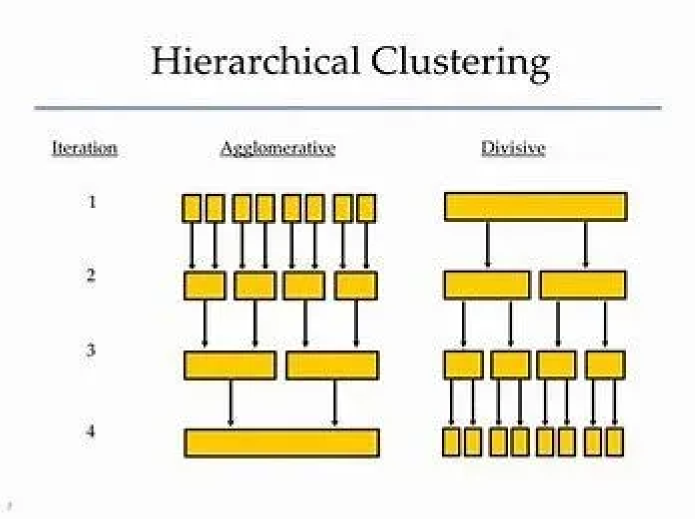


K-Means Clustering

It is a commonly used clustering algorithm that aims to partition the data into K clusters, where K is a predefined number chosen by the user. Here's how it works:

- **Initialization:** Randomly select K points from the dataset as initial cluster centroids.
- **Assignment:** Assign each data point to the nearest centroid, creating K clusters.
- **Update:** Recalculate the centroids by taking the mean of the data points within each cluster.
- **Iteration:** Repeat the assignment and update steps until convergence (when the centroids no longer change significantly) or after a fixed number of iterations.

b. Hierarchical clustering:

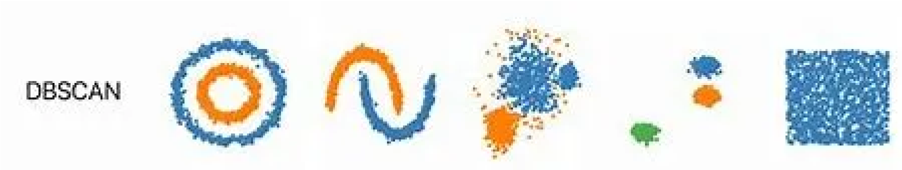


Hierarchical Clustering

Hierarchical clustering creates a hierarchy of clusters, also known as a dendrogram, by recursively merging or splitting clusters based on the similarity between data points. It doesn't require the number of clusters to be predefined. There are two main types of hierarchical clustering:

- **Agglomerative (Bottom-Up):** Starts with each data point as a separate cluster and iteratively merges the most similar clusters until reaching a desired number of clusters or a termination criterion.
- **Divisive (Top-Down):** Starts with all data points in one cluster and recursively splits it into smaller clusters based on dissimilarity measures until each cluster contains a single data point.

c. Density-based clustering:



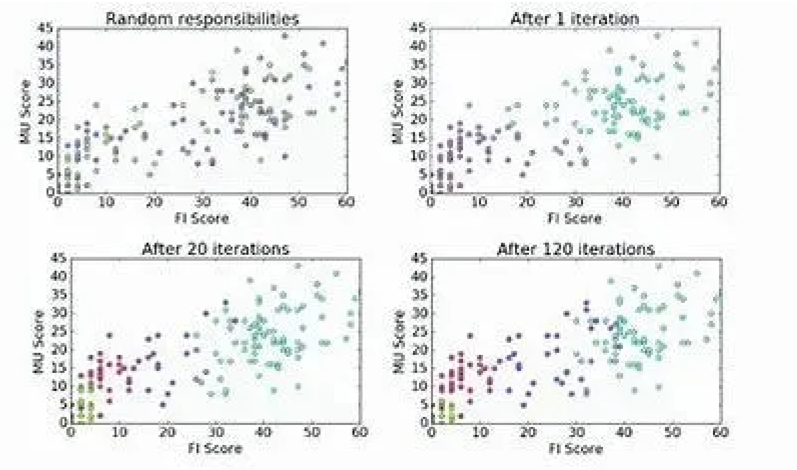
DBSCAN Clustering

Density-based clustering identifies clusters based on the density of data points in the feature space. It aims to discover regions of high-density separated by regions of low-density. One popular density-based clustering algorithm is DBSCAN (Density-Based Spatial Clustering of Applications with Noise). Here are the main steps of DBSCAN:

- **Core Points:** Identify core points that have a sufficient number of neighboring points within a specified radius.
- **Density-Reachable:** Expand the cluster by iteratively adding density-reachable points within the neighborhood.
- **Noise Points:** Assign data points that are not density-reachable from any cluster as noise points or outliers.

DBSCAN can find clusters of arbitrary shape, is robust to noise, and doesn't require specifying the number of clusters in advance.

d. EM(Expectation -Maximization) clustering:



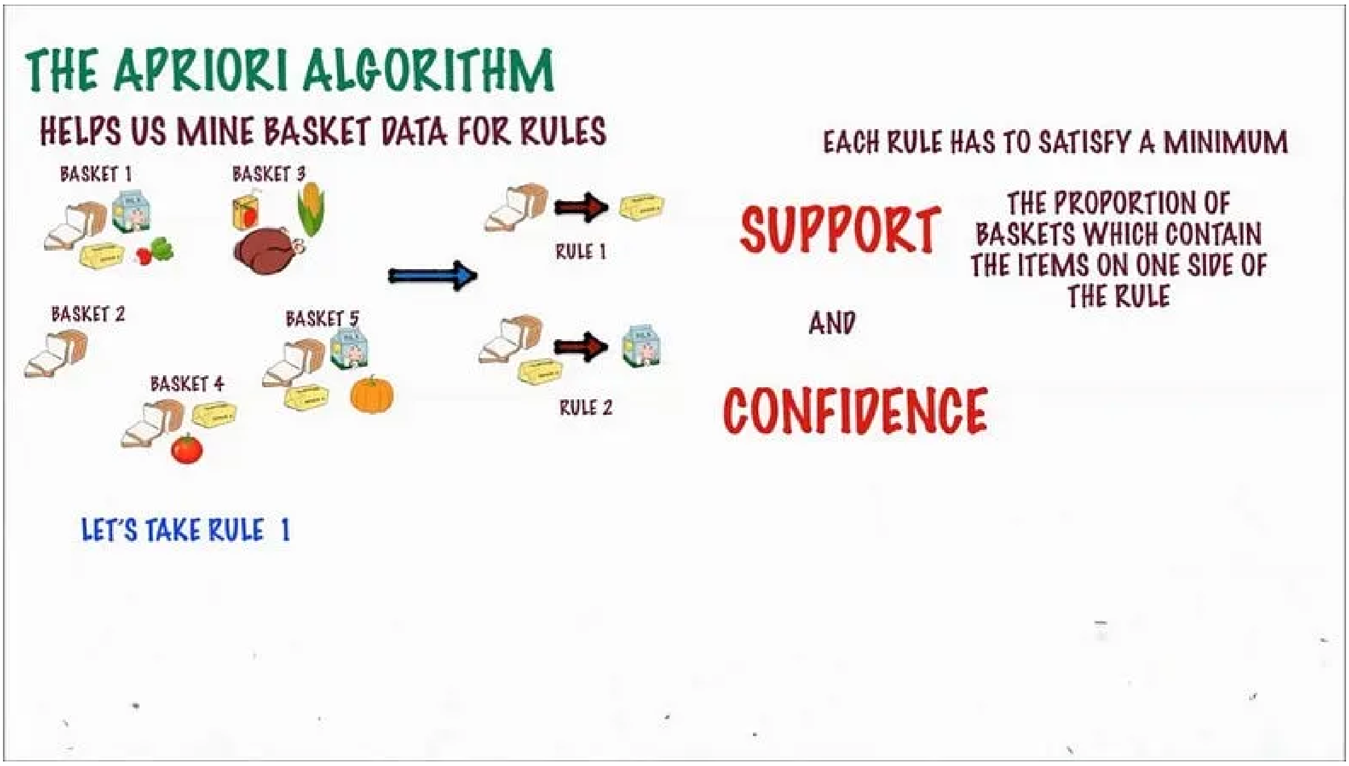
EM clustering is a probabilistic approach that assumes data points are generated from a mixture of probability distributions. It is commonly used for Gaussian Mixture Models (GMMs), where each cluster is modeled as a Gaussian distribution. Here are the working steps involved in EM clustering :

- **Initialization:** Randomly initialize the parameters of the Gaussian distributions (e.g., means and covariances).
- **E-step (Expectation):** Calculate the probability of each data point belonging to each cluster based on the current parameter estimates.
- **M-step (Maximization):** Update the parameters of the Gaussian distributions by maximizing the likelihood function.
- **Iteration:** Repeat the E-step and M-step until convergence or after a fixed number of iterations.

2.ASSOCIATION:

Association rule learning is a subcategory of unsupervised machine learning that focuses on discovering interesting relationships or associations among items in a dataset. It aims to identify frequent item sets and generate rules that describe the co-occurrence or dependency between items. Here are two commonly used algorithms within association rule learning:

a. **Apriori Algorithm:**



THE APRIORI ALGORITHM

HELPS US MINE BASKET DATA FOR RULES

EACH RULE HAS TO SATISFY A MINIMUM

SUPPORT

THE PROPORTION OF BASKETS WHICH CONTAIN THE ITEMS ON ONE SIDE OF THE RULE

AND

CONFIDENCE

LET'S TAKE RULE 1

The Apriori algorithm is a popular approach for mining association rules. It follows a “bottom-up” strategy to discover frequent item-sets in a dataset. The algorithm works as follows:

- Find frequent 1-itemsets: Scan the dataset to identify individual items that meet a minimum support threshold.
- Generate candidate item-sets: Use the frequent 1-itemsets to generate candidate item-sets of size 2 by combining pairs of items.
- Prune candidate item-sets: Remove candidate item-sets that contain subsets that are not frequent. This is known as the “downward closure property.”
- Calculate support for candidate item-sets: Scan the dataset again to count the occurrences of each candidate itemset.

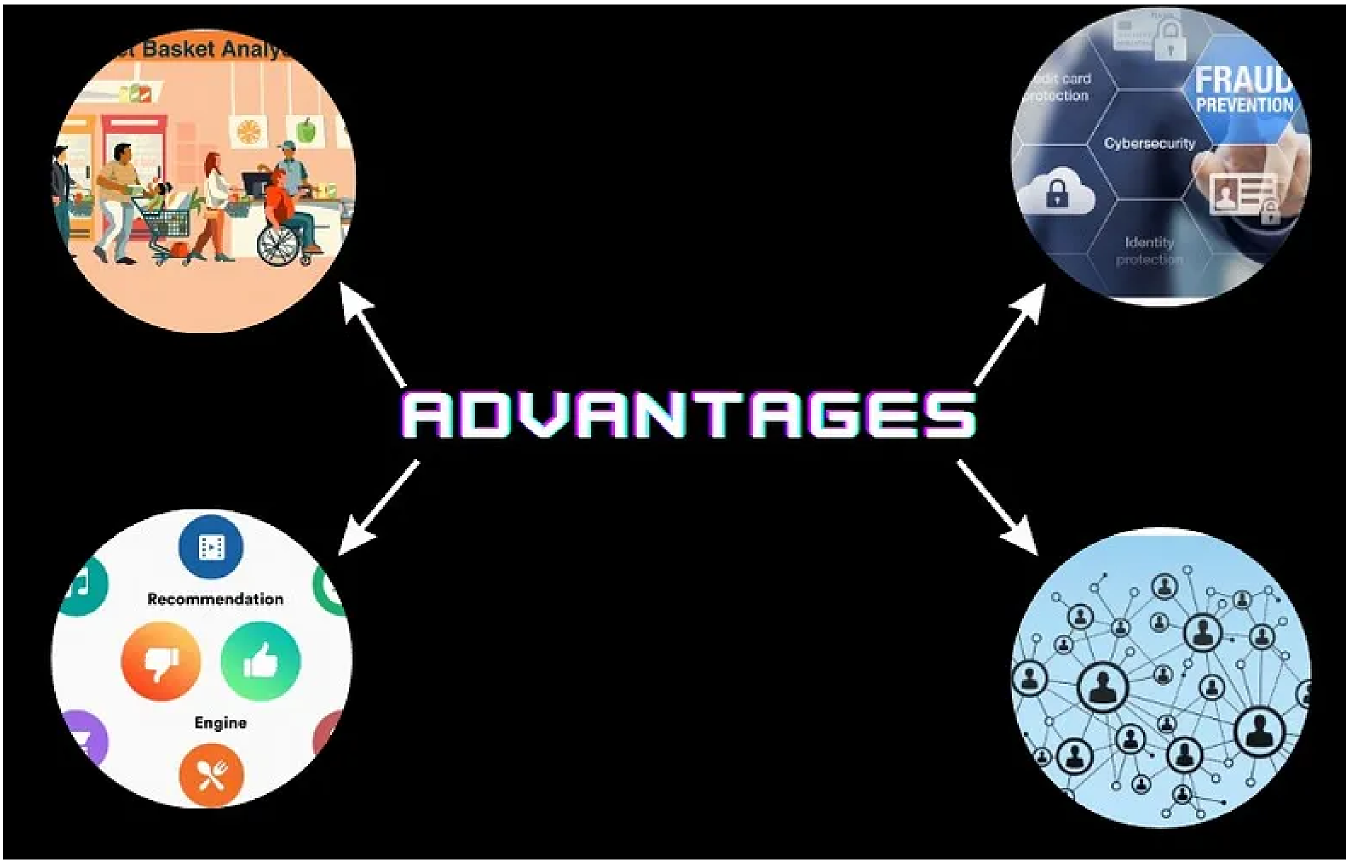
- Repeat Steps 2–4: Continue generating and pruning candidate item-sets of larger sizes until no more frequent item-sets can be found.

2. FP-Growth Algorithm:

The FP-Growth algorithm is an alternative approach to mining association rules. It employs a “divide-and-conquer” strategy by constructing a compact data structure called an FP-tree. The algorithm proceeds as follows:

- Build the FP-tree: Scan the dataset to construct the FP-tree, which represents the frequency of item-sets in a compressed manner.
- Mine frequent item-sets: Traverse the FP-tree to extract frequent item-sets by recursively combining item-sets and their conditional patterns.
- Generate association rules: Similar to the Apriori algorithm, association rules can be derived from the frequent item-sets by applying support, confidence, and lift thresholds.

PROS OF UNSUPERVISED MACHINE LEARNING:



Unsupervised learning offers several advantages that make it a powerful tool in data analysis and problem-solving. This allows businesses to tap into vast amounts of untapped information and uncover hidden relationships that may not be apparent through manual inspection. Some of the advantages are as follows:

a. Market basket analysis: Unsupervised learning algorithms, such as association rules mining, enable businesses to understand the relationships between products frequently purchased together. It allows businesses to optimize inventory management, create personalized bundles, and improve overall sales performance.

b. Recommendation systems: Unsupervised learning techniques analyzes user behavior and item characteristics and suggest relevant products, services, or content to users. This helps businesses enhance customer experience, drive sales, and increase user engagement.

c. Fraud detection: By learning from normal behavior and identifying deviations, unsupervised learning algorithms can detect fraudulent activities or unusual patterns or outliers in transactional data. Thus it flags suspicious transactions, potentially saving businesses from financial losses and reputational damage.

d. Network analysis: Unsupervised learning algorithms can analyze complex networks, such as social networks or transportation networks, to uncover hidden structures and communities. This information can be utilized for targeted marketing, influencer identification, or optimizing supply chain networks.

In conclusion, the vast advantages of unsupervised learning across diverse domains make it an indispensable tool for businesses seeking to extract meaningful insights and unlock the hidden potential within their data. Also its versatility and applicability make it a valuable tool across other fields such as healthcare, finance, academia, social sciences, and more.

LIMITATIONS OF UNSUPERVISED LEARNING:

Unsupervised learning has several limitations that are important to consider:

a. Difficulty in Generalization: Unsupervised learning algorithms focus on finding patterns within the observed data. However, they may struggle to generalize these patterns to new, unseen data points. This limits the applicability of unsupervised models when it comes to making predictions or classifying new instances.

b. Sensitivity to Data Quality and Preprocessing: Unsupervised learning algorithms can be sensitive to noise, outliers, and irrelevant features in the data. If the data contains inconsistencies or outliers, they can have a significant impact on the learned patterns or clusters. It becomes crucial to preprocess and clean the data appropriately.

c. Difficulty in Assessing Performance: Evaluating the effectiveness of unsupervised models can be subjective and dependent on domain knowledge, making it challenging to compare different approaches or determine the optimal solution.

d. Biased Results: Unsupervised learning algorithms can unintentionally reinforce or amplify existing biases present in the data. If the input data contains biased or unrepresentative samples, the learned patterns or clusters may also reflect those biases, leading to biased results.

Despite these limitations, unsupervised learning remains a valuable tool for exploratory data analysis, pattern discovery, and data preprocessing. It complements supervised learning approaches and can provide insights into the structure and relationships within data when labeled examples are not available.

Some useful links: