Unsupervised Machine Learning

Data Science with Python

Cont..

- we learned supervised machine learning in which models are trained using labeled data under the supervision of training data.
- But there may be many cases in which we do not have labeled data and need to find the hidden patterns from the given dataset.
- So, to solve such types of cases in machine learning, we need unsupervised learning techniques.

What is Unsupervised Learning

- As the name suggests, unsupervised learning is a machine learning technique in which models are not supervised using training dataset.
- Instead, models itself find the hidden patterns and insights from the given data.
- *Unsupervised learning is a type of machine learning in which models are trained using unlabeled dataset and*

Cont…

- we have the input data but no corresponding output data.
- The goal of unsupervised learning is to **find the underlying structure of dataset, group that data according to similarities, and represent that dataset in a compressed format**.

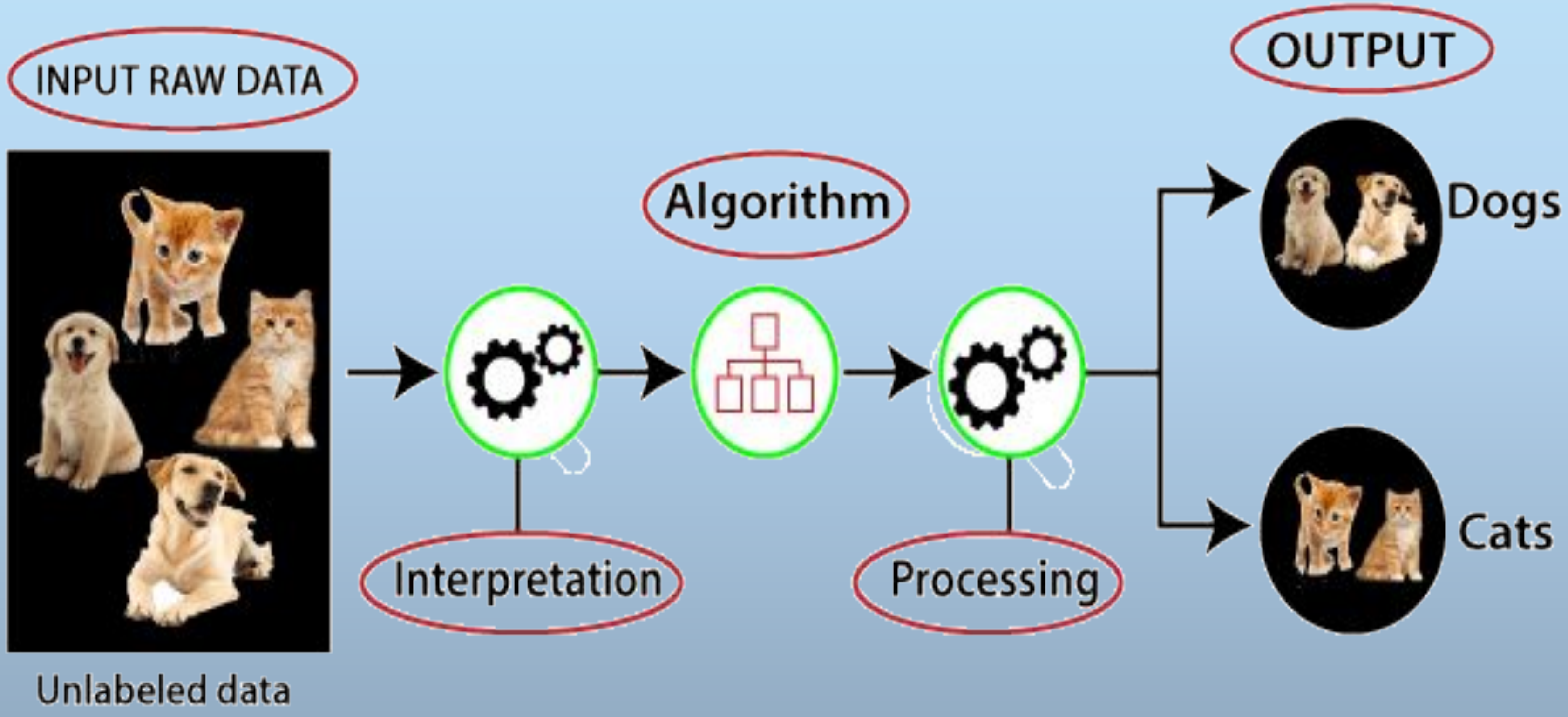
Example

- Suppose the unsupervised learning algorithm is given an input dataset containing images of different types of cats and dogs.
- 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. Unsupervised learning algorithm will perform this task by clustering the image dataset into the groups according to similarities between images

Why use Unsupervised Learning

- Unsupervised learning is helpful for finding useful insights from the data.
- Unsupervised learning is much similar as a human learns to think by their own experiences, which makes it closer to the real AI.
- Unsupervised learning works on unlabeled and uncategorized data which make unsupervised learning more important.
- In real-world, we do not always have input data with the corresponding output so to solve such cases, we need unsupervised learning.

Working of Unsupervised Learning

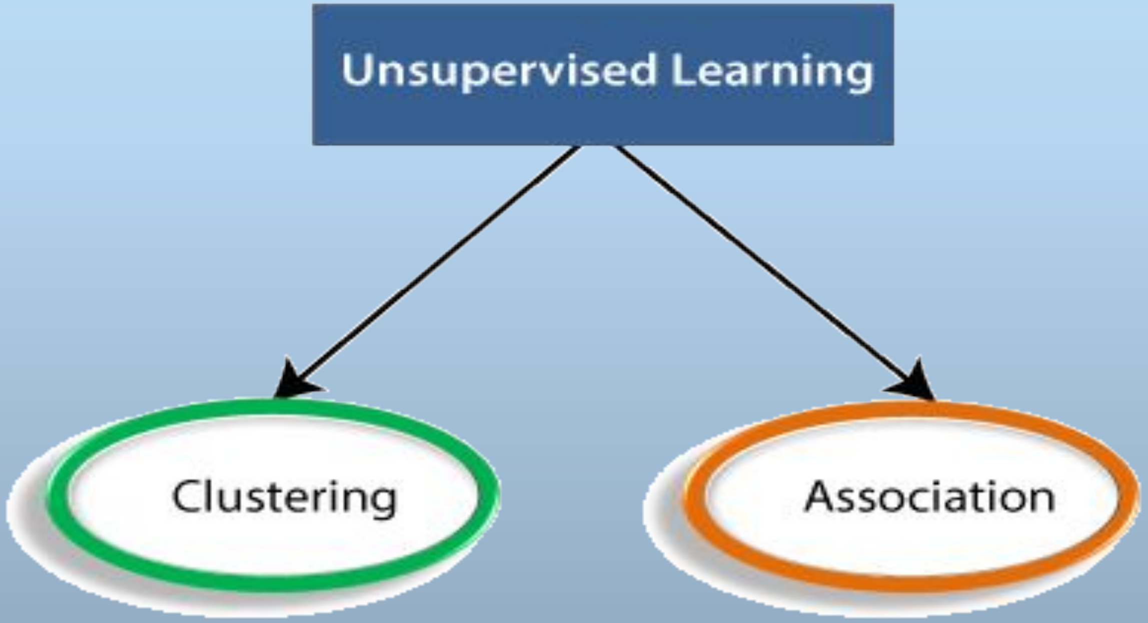


From the above example

- Now, this unlabeled input data is fed to the machine learning model in order to train it. Firstly, it will interpret the raw data to find the hidden patterns from the data and then will apply suitable algorithms such as k-means clustering, Decision tree, etc.
- Once it applies the suitable algorithm, the algorithm divides the data objects into groups according to the similarities and difference between the objects.

Types of Unsupervised Learning Algorithm:

- The unsupervised learning algorithm can be further categorized into two types of problems:



Clustering

- Clustering is a method of grouping the objects into clusters such that objects with most similarities remains into a group and has less or no similarities with the objects of another group.
- Cluster analysis finds the commonalities between the data objects and categorizes them as per the presence and absence of those commonalities.
- 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

different types of clustering

- **Exlcusive clustering**: Data is grouped such that a single data point exclusively belongs to one cluster.
- **Overlapping clustering**: A soft cluster in which a single data point may belong to multiple clusters with varying degrees of membership.
- **Hierarchical clustering**: A type of clustering in which groups are created such that similar instances are within the same group and different objects are in other groups.
- **Probalistic clustering**: Clusters are created

Hierarchical Clustering

- Hierarchical clustering is an algorithm which builds a hierarchy of clusters.
- It begins with all the data which is assigned to a cluster of their own.
- Here, two close cluster are going to be in the same cluster. This algorithm ends when there is only one cluster left.
- 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

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 centroid.
- The data points closest to a given centroid will be clustered under the same category.
- A larger K value will be indicative of smaller groupings with more granularity whereas a smaller K value will have larger groupings and less granularity.
- K-means clustering is commonly used in market segmentation, document clustering, image segmentation, and image compression.
- K-mean clustering further defines two subgroups:
- Agglomerative clustering

Association

- An association rule is an unsupervised learning method which is used for finding the relationships between variables in the large database.
- It determines the set of items that occurs together in the dataset. Association rule makes marketing strategy more effective.
- Such as people who buy X item (suppose a bread) are also tend to purchase Y (Butter/Jam) item.
- A typical example of Association rule is Market Basket Analysis.
- For example, people that buy a new home most likely to buy new furniture.

Other Examples for Association:

- A subgroup of cancer patients grouped by their gene expression measurements
- Groups of shopper based on their browsing and purchasing histories
- Movie group by the rating given by movies viewers

Dimensionality Reduction

- Popular algorithms used for dimensionality reduction include principal component analysis (PCA) and Singular Value Decomposition (SVD). These algorithms seek to transform data from high-dimensional spaces to low-dimensional spaces without compromising meaningful properties in the original data. These techniques are typically deployed during exploratory data analysis (EDA) or data processing to prepare the data for modeling.

- It's helpful to reduce the dimensionality of a dataset during EDA to help visualize data: this is because visualizing data in more than three dimensions is difficult. From a data processing perspective, reducing the dimensionality of

Unsupervised Learning algorithms

- K-means clustering
- KNN (k-nearest neighbors)
- Hierarchal clustering
- Anomaly detection
- Neural Networks
- Principle Component Analysis
- Independent Component Analysis
- Apriori algorithm
- Singular value decomposition

Advantages of Unsupervised Learning

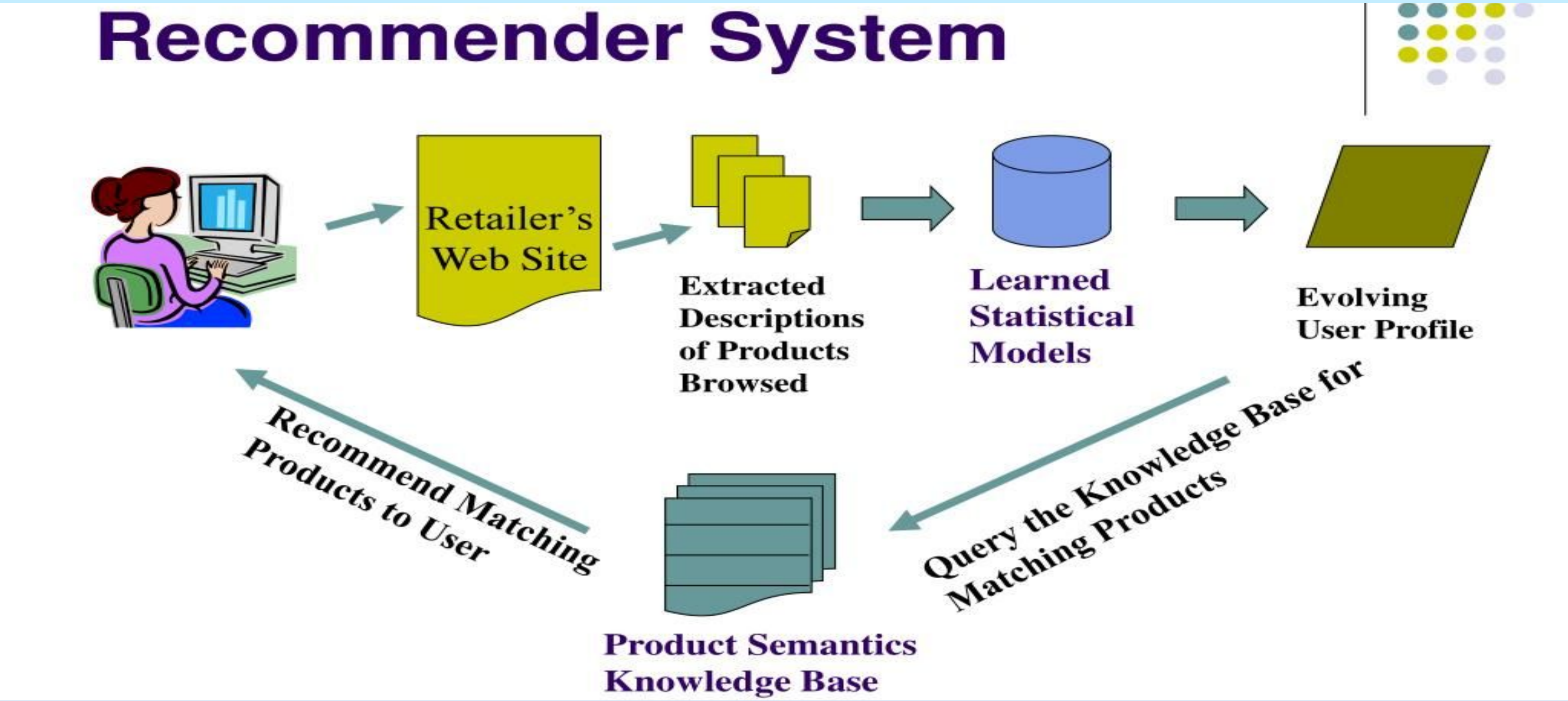
- Unsupervised learning is used for more complex tasks as compared to supervised learning because, in unsupervised learning, we don't have labeled input data.
- Unsupervised learning is preferable as it is easy to get unlabeled data in comparison to labeled data.

Unsupervised Learning Applications

- using unsupervised techniques to explore data, some common use cases in the real-world include:
- **Natural language processing (NLP)**. Google News is known to leverage unsupervised learning to categorize articles based on the same story from various news outlets. For instance, the results of the football transfer window can all be categorized under football.
- **Image and video analysis**. Visual Perception tasks such as object recognition leverage unsupervised learning.
- **Anomaly detection**. Unsupervised learning is used to identify data points, events, and/or observations that deviate from a dataset's normal behavior.
- **Customer segmentation**. Interesting buyer persona profiles can be created using unsupervised learning. This helps businesses to understand their customers' common traits and purchasing habits, thus, enabling them to align their products more accordingly.

Unsupervised Learning Example in Python

- Principal component analysis (PCA) is the process of computing the principal components then using them to perform a change of basis on the data. In other words, PCA is an unsupervised learning dimensionality reduction technique.
- It’s useful to reduce the dimensionality of a dataset for two main reasons:
- When there are too many dimensions in a dataset to visualize
- To identify the most predictive n dimensions for feature selection when building a predictive model.
- Example:



Recommender System

Retailer's Web Site

Extracted Descriptions of Products Browsed

Learned Statistical Models

Evolving User Profile

Query the Knowledge Base for Matching Products

Product Semantics Knowledge Base

Recommend Matching Products to User

A recommender system uses artificial intelligence (AI), machine learning, and big data to suggest items a user might be interested in. These systems are commonly used by e-commerce, media streaming, and social media sites to suggest or recommend other products and services to people.

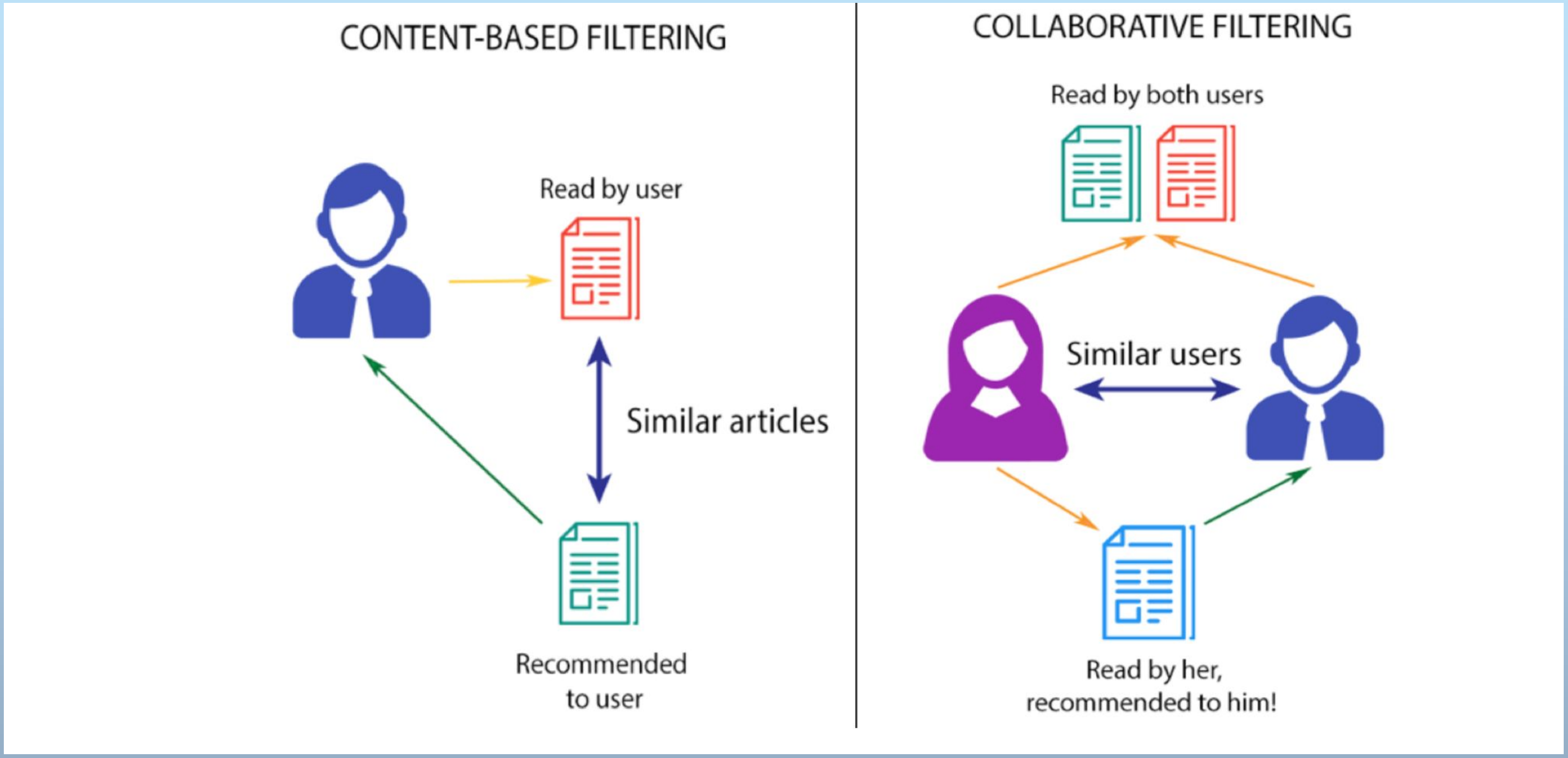
Cont..

- Recommender systems can use both supervised and unsupervised learning.
- Recommender systems operate using sophisticated machine learning algorithms that can harness data to understand – and predict – people's preferences based on their:
- search history
- demographic details, such as location, gender, or age.
- purchase history or other online interactions and metrics, such as clicks, likes, dislikes, and connections.

Types of recommender systems

- **Collaborative filtering**
- Collaborative filtering systems utilise past user behaviour to predict what the user might like in the future. These filtering systems are more commonly used because they are the easiest to implement.
- They work by finding similar users who have comparable interests and then recommend other items the other users have also liked, based on information such as user ratings, and so on

Types of recommender systems



Content-based filtering

- Content-based filtering systems are more complicated than collaborative filter systems, and are more difficult to implement because they require the system to understand the user's interests – but they are also typically more accurate than collaborative filtering.
- These interests are determined using additional information about

Hybrid filtering

- Hybrid recommender filtering methods are a combination of collaborative and content-based filtering, so users will receive recommendations based on their past preferences and historical data, their interests and the interests of users with similar tastes, and other

What are the benefits of recommender systems?

- Recommender systems are a helpful tool for people who want to find related content, whether it’s similar products and services, new music and films, or other accounts and articles.
- The level of personalisation offered by recommendation systems also supports content optimisation, creating a better user experience for people.
- Examples in python code:
- https://www.kaggle.com/code/basu369victor/recommender-system-using-un-supervised-learning

Understanding recommendation algorithms

- Recommendation engines are typically written in codes like Python and C++. They rely on various machine learning models, algorithms, data mining techniques, and other technologies in order to function.

- **Matrix factorization**
- **Deep learning**
- **K-nearest neighbours**
- **Neural networks**
- **Natural language processing (NLP)**
- **Matrix factorisation** algorithms are typically used for collaborative filtering. They break down the data that sits within what's known as the user-item interaction matrix to identify relationships between users and items.

- **Deep learning**
- Deep learning is a subset of machine learning, and its algorithms often support sophisticated recommendation models. It enables recommenders to delve into multiple layers of datasets to extract useful information, connections, and relationships in order to make more accurate, helpful recommendations for users.
- **K-nearest neighbours**
- According to IBM, the K-nearest neighbours algorithm uses proximity to make predictions about the grouping of an individual data point. It's typically used as a classification algorithm that works off the assumption that similar points can be found near one another – which is ideal for recommender systems – but it can also be used for regression problems

Examples of recommender systems

- **Amazon** :Amazon offers a compelling case study for recommender systems. Through its complex algorithms, it can recommend products to people based on their past purchases and search history, as well as the purchases of similar consumers, among many other variables.
- **Netflix**
- Streaming platform Netflix uses recommender systems to offer TV and movie recommendations based on collaborative and content-based filtering.
- **Spotify**
- Music platform Spotify makes user recommendations based on people's listening history, preferred artists and genres, and other user data.
- **MovieLens** :MovieLens recommends films to people based on their ratings of other movies.