2.0. Unsupervised Learning

What Unsupervised Learning?

So far, all of the machine learning algorithms that you have seen are supervised learning. That is, the datasets have all been labeled, classified, or categorized. Datasets that have been labeled are known as labeled data, while datasets that have not been labeled are known as unlabeled data.

The table below (on left) shows an example of labeled data:

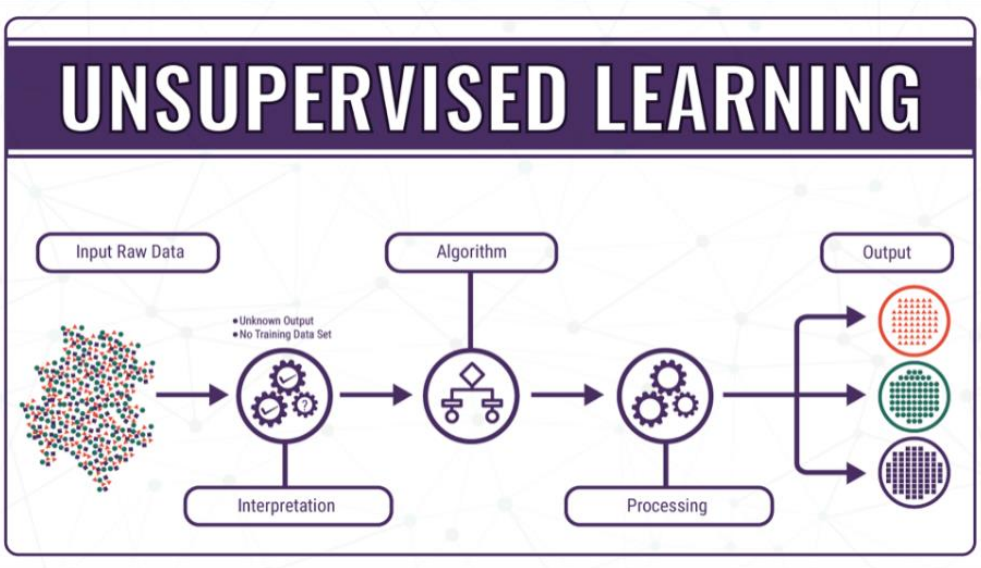
Features		*Label*
Size of House	**Year Built**	**Price Sold**

Features	
Waist Circumference	**Leg length**

Based on the size of the house and the year in which it was built, you have the price at which the house was sold. The selling price of the house is the label, and your machine learning model can be trained to give the estimated worth of the house based on its size and the year in which it was built.

Unlabeled data, on the other hand, is data without label(s). For example, Figure (on right) shows a dataset containing a group of people's waist circumference and corresponding leg length. Given this set of data, you can try to cluster them into groups based on the waist circumference and leg length, and from there you can figure out the average dimension in each group. This would be useful for clothing manufacturers to tailor different sizes of clothing to fit its customers.

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.



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:

1. 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.

2. Computer vision: Unsupervised learning algorithms are used for visual perception tasks, such as object recognition.

3. 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.

4. 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.

5. 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.

6. 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.

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:

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