UNIT 3

INTRODUCTION TO UNSUPERVISED LEARNING

Unsupervised learning is a **machine learning** problem type in which training data consists of a set of input vectors but no corresponding target values. The idea behind this type of learning is to group information based on similarities, patterns, and differences.

Unlike in **supervised learning** problems, unsupervised learning algorithms do not require input-to-output mappings to learn a mapping function—this is what is meant when we say, "no teacher is provided to the learning algorithm." Consequently, an unsupervised learning algorithm cannot perform classification or regression.

The role of an unsupervised learning algorithm is to discover the underlying structure of an unlabeled dataset by itself.

Types of Unsupervised Learning

In the introduction, we mentioned that unsupervised learning is a method we use to group data when no labels are present. Since no labels are present, unsupervised learning methods are typically applied to build a concise representation of the data so we can derive imaginative content from it.

For example, if we were releasing a new product, we can use unsupervised learning methods to identify who the target market for the new product will be: this is because there is no historical information about who the target customer is and their demographics.

But unsupervised learning can be broken down into three main tasks:

- Clustering
- Association rules
- Dimensionality reduction.

Let's delve deeper into each one:

Clustering

From a theoretical standpoint, instances within the same group tend to have similar properties. You can observe this phenomenon in the periodic table. Members of the same group, separated by eighteen columns, have the same number of electrons in the outermost shells of their atoms and form bonds of the same type.

This is the idea that's at play in clustering algorithms; **Clustering methods** involve grouping untagged data based on their similarities and differences. When two instances appear in different groups, we can infer they have dissimilar properties.

Clustering is a popular type of unsupervised learning approach. You can even break it down further into different types of clustering; for example:

- 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 using probability distribution.

Association Rule Mining

This type of unsupervised machine learning takes a rule-based approach to discovering interesting relationships between features in a given dataset. It works by using a measure of interest to identify strong rules found within a dataset.

We typically see association rule mining used for market basket analysis: this is a data mining technique retailers use to gain a better understanding of customer purchasing patterns based on the relationships between various products.

The most widely used algorithm for association rule learning is the Apriori algorithm. However, other algorithms are used for this type of unsupervised learning, such as the Eclat and FP-growth algorithms.

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 the data simplifies the modeling problem.

When more input features are being fed into the model, the model must learn a more complex approximation function. This phenomenon can be summed up by a saying called the "curse of dimensionality."

Unsupervised Learning Applications

Most executives would have no problem identifying use cases for supervised machine learning tasks; the same cannot be said for unsupervised learning.

One reason this may be is down to the simple nature of risk. Unsupervised learning introduces much more risk than unsupervised learning since there's no clear way to measure results against ground truth in an offline manner, and it may be too risky to conduct an online evaluation.

Nonetheless, there are several valuable unsupervised learning use cases at the enterprise level. Beyond 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.
- Recommendation Engines. Past purchase behavior coupled with unsupervised learning can be used to help businesses discover data trends that they could use to develop effective cross-selling strategies.

References:

https://www.datacamp.com/blog/introduction-to-unsupervised-learning