Introduction

Imagine your email automatically sorting out spam, your music app predicting your next favorite song, or your car driving itself through rush hour traffic. Behind these wonders lies the art of teaching machines to learn, but how does it work? I will try to briefly describe these concepts. Machine learning (ML) is a type of artificial intelligence that allows machines to learn and improve from experience without being explicitly programmed. There are three kinds of Machine Learning:

Supervised machine learning: is defined by its use of *labeled datasets* to train algorithms to classify data or predict outcomes accurately. As input data is fed into the model, the model adjusts its weights until it has been fitted appropriately. Some methods used in supervised learning include neural networks, naïve bayes, linear regression, logistic regression, and others.

Unsupervised machine learning: uses machine learning algorithms to analyze and cluster *unlabeled datasets* (subsets called clusters). These algorithms discover hidden patterns or data groupings without the need for human intervention. This method's ability to discover similarities and differences in information make it ideal for exploratory data analysis, customer segmentation, and image and pattern recognition.

Semi-supervised learning offers a happy medium between supervised and unsupervised learning. During training, it uses a smaller labeled data set to guide classification and feature extraction from a larger, unlabeled data set. *Semi-supervised learning can solve the problem of not having enough labeled data for a supervised learning algorithm. It also helps if it's too costly to label enough data.*

literature overview

Here we mainly focus on DBSCAN clustering algorithms. It, like any clustering algorithm, is basically an Unsupervised learning method that divides the data points into a number of specific batches or groups, such that the data points in the same groups have similar properties and data points in different groups have different properties in some sense.
What makes it different from other algorithms is the method it uses, for instance K-Means uses distance between points, Affinity propagation uses graph distance, Mean-shift uses distance between points, and when it comes to DBSCAN, it uses distance between nearest points to classify data points into different clusters. It divides your entire dataset into dense regions separated by sparse regions. DBSCAN goes beyond traditional clustering methods, offering a unique approach to identifying clusters based on the density of data points, but What makes it more valuable than other clustering algorithms. Before starting with its advantages let's understand some basic terms and its implementation.

1. We mentioned DBSCAN clusters based on density, but how do you measure density around a point? One answer is by defining two parameters:
    - ➢ *Epsilon (Eps):* a radius distance around a point.
    - ➢ *Minimum points(MinPts)*: the number of data points in a given Epsilon around a point.

   Based on these parameters we define a region around the point and assess the number of points within that designated area.
2. Now we know how to represent density lets understand the types of data points:
    - ➢ *Core points*: is a data point that has surrounding data points more than Minpts in given Eps radius.
    - ➢ *Border points*: is a data point that has a surrounding data point less than Minpts, but it is a neighbor for a core point.
    - ➢ *Noise* (*outliers*): A noise point is a data point which can neither be a core point nor a border point.
3. Finally let's look at the relationship between these data points:
    - ➢ *Directly Density Reachable*: A point P is directly density-reachable from a point Q given Eps, MinPts if first P is in the Eps-neighborhood of Q, second Both P and Q are core points.
    - ➢ *Densely Connected points*: A point P is density connected to Q given Eps, MinPts if there is a chain of points P1, P2, P3 …… Pn, P1 = P and Pn = Q such that Pi+1 is directly density reachable from Pi.

Now we can design an algorithm for DBSCAN using the above terms:

- ➔ Identify all points as either core point, border point or noise point.
- ➔ For all of the unclustered core points. Create a new cluster.
- ➔ Add all the points that are unclustered and density connected to the current point into this cluster.
- ➔ For each unclustered border point assign it to the cluster of nearest core point.
- ➔ Leave all the noise points as it is.

Advantages of DBSCAN clustering Algorithm

- It is robust to outliers as it defines clusters based on dense regions of data, and isolated points are treated as noise.
- Unlike some clustering algorithms, DBSCAN does not require the user to specify the number of clusters beforehand, making it more flexible and applicable to a variety of datasets.
- DBSCAN can identify clusters with complex shapes and is not constrained by assumptions of cluster shapes, making it suitable for data with irregular structures.

Disadvantages of DBSCAN clustering Algorithm

- The performance of DBSCAN can be sensitive to the choice of its hyperparameters, especially the distance threshold (eps) and the minimum number of points (min_samples). Suboptimal parameter selection may lead to under-segmentation or over-segmentation.
- DBSCAN struggles with clusters of varying densities. It may fail to connect regions with lower point density to the rest of the cluster, leading to suboptimal cluster assignments in datasets with regions of varying densities.

Application Areas of DBSCAN

★ DBSCAN is particularly well-suited for spatial data clustering due to its ability to find clusters of arbitrary shapes, which is common in geographic data. It's used in applications like identifying regions of similar land use in satellite images or grouping locations with similar activities in GIS (Geographic Information Systems).

★ The algorithm's effectiveness in distinguishing noise or outliers from core clusters makes it useful in anomaly detection tasks, such as detecting fraudulent activities in banking transactions or identifying unusual patterns in network traffic.

★ More broadly, in the fields of machine learning and data mining, DBSCAN is employed for exploratory data analysis, helping to uncover natural structures or patterns in data that might not be apparent otherwise.

Conclusion

DBSCAN is a powerful clustering algorithm that identifies clusters based on data density, making it robust to noise and effective for datasets with irregular cluster shapes. Its ability to work without predefined cluster counts and handle outliers makes it ideal for spatial data analysis, anomaly detection, and exploratory tasks. However, its performance depends on careful selection of parameters like Eps and MinPts and can struggle with clusters of varying densities. Despite these limitations, DBSCAN remains a versatile tool in unsupervised learning for uncovering meaningful patterns in complex datasets.

References

- https://www.geeksforgeeks.org/dbscan-clustering-in-ml-density-based-clustering/
- https://scikit-learn.org/1.5/modules/generated/sklearn.cluster.DBSCAN.html
- https://www.youtube.com/watch?v=RDZUdRSDOok