>>>> Unsupervised learning algorithms by Suraj <<<<<<<<<<<<

Unsupervised learning

Unsupervised learning is a sort of machine learning in which an algorithm picks out relationships or patterns in data without the need of labelled samples or direct instruction. Unsupervised learning entails the algorithm discovering patterns or structures on its own within the data, as opposed to supervised learning, where the algorithm is given labelled data (i.e., input-output pairs). The programme uses methods like clustering, dimensionality reduction, and anomaly detection to find hidden patterns or structures in the data.

Table of contents

00 Clustering

Clustering is a common technique in unsupervised learning that involves grouping similar data points together based on their similarity or proximity in a given feature space. Let's dive into the math and code behind one of the most widely used clustering algorithms, K-means clustering.

> Mathematics for K-means clustering:

The goal of K-means clustering is to partition the data set into "k" separated clusters, where each data point belongs to the

cluster with the closest centroid. The algorithm minimizes the sum of squares of the distances between data points and their respective centroids, also known as cluster sum of squares (WCSS). The objective function for K-means clustering can be expressed mathematically as:

Objective function: $J = \sum_{i=1}^{k} \sum_{x \in c_i} \|x - \mu_i\|^2$

Where:

K: number of clusters
x: data point
μ_i: centroid of the ith cluster
c_i: set of data points belonging to the ith cluster
The goal is to find a center that minimizes this objective function.

K-means clustering is an iterative algorithm that can converge to different solutions depending on the initial location of the centroids. To improve the reliability of the results, it is common to run the algorithm several times with different initializations and choose the solution with the lowest WCSS. It is also important to pre-process the data and choose an appropriate value of 'K' to ensure meaningful results.

More types of clustering

Hierarchical Clustering: Hierarchical clustering is a clustering algorithm that creates a hierarchy of clusters by recursively splitting or merging clusters. There are two main types of hierarchical clustering: agglomerative and distributive. Agglomerative clustering starts with each data point as a single

cluster and then recursively joins them together based on a measure of similarity or distance, while divisive clustering starts with all data points in one cluster and then recursively splits the set into several small clusters. Hierarchical clustering produces a tree-like structure called a dendrogram that can be used to visualize cluster hierarchies.

DBSCAN (Density-Based Spatial Clustering for Noise Applied): DBSCAN is a density-based clustering algorithm that groups data points according to their density. It defines clusters as regions of dense points separated by regions of lower density of points and can identify noisy points that do not belong to any cluster. DBSCAN uses two main hyperparameters: epsilon (ε), which defines the radius around data points to form dense regions, and min_samples, which specifies the minimum number of points required to form dense regions.

Gaussian mixture model (GMM): GMM is a probabilistic model that models data points as a mixture of Gaussian distributions. It assumes that the data points are generated from a mixture of several Gaussian distributions, and the parameters of these distributions are estimated to determine the cluster structure. Compared to K-means clustering, GMM allows for more flexible cluster shapes because it models clusters as ellipsoidal shapes with different orientations and sizes. GMMs can also estimate the covariance between traits, making them suitable for capturing correlations among data variables.

Spectral Clustering: Spectral clustering is a graph-based clustering algorithm that uses eigenvalues and eigenvectors of a similarity matrix to perform clustering. It treats data points as nodes in a graph, and edges between nodes represent the

similarity or distance between data points. Spectral clustering first creates a similarity matrix and then uses spectral methods to obtain low-dimensional representations of the data, which are then clustered using traditional clustering algorithms such as k-means.

Fuzzy clustering: Fuzzy clustering is a clustering algorithm that assigns data points to multiple clusters with different degrees of membership, instead of assigning each data point to a single cluster, as in traditional clustering algorithms. Fuzzy clustering uses fuzzy logic to assign data points to clusters based on their similarity to multiple cluster centroids. This provides a more flexible allocation, suitable for cases where data points may belong to multiple clusters at the same time.

01 Dimensionality Reduction

Dimensionality reduction is a technique used to reduce the number of features or variables in a data set while preserving relevant information. Visualizing, analyzing, and modeling high-dimensional data can be challenging, and dimensionality reduction techniques help solve this problem by transforming the data into a lower-dimensional representation. There are two main types of size reduction techniques:

1. Selection of means. Feature selection techniques involve selecting a subset of raw features or variables from a data set based on specific criteria, such as their importance or relevance to a given problem. These methods preserve the original features but discard some of them, resulting in a reduced size set.

2. Feature extraction: Feature extraction techniques generate new features or variants from raw features using mathematical transformations. These methods create a new feature set that captures the most important information from the original features, resulting in a low-dimensional representation of the data.

Commonly used Dimensionality reduction techniques

1. Principal Component Analysis (PCA): PCA is a widely used linear dimensionality reduction technique that transforms the original features into a new set of uncorrelated features called principal components, which are ordered by their explained variance. PCA finds the directions in the data with the highest variance and projects the data onto these directions to create a lower-dimensional representation.

2. t-SNE (t-distributed Stochastic Neighbor Embedding): t-SNE is a nonlinear dimensionality reduction technique that is particularly useful for visualizing high-dimensional data in a low-dimensional space. It preserves the local structure of the data by minimizing the divergence between pairwise similarities in the original data and in the reduced-dimensional representation.

3. LLE (Locally Linear Embedding): LLE is a nonlinear dimensionality reduction technique that assumes that the data lies on a locally linear manifold. It finds a lower-dimensional representation of the data by reconstructing each data point as a weighted linear combination of its neighbors in the original space.

4. Autoencoders: Autoencoders are a type of neural network-

dimensionality reduction technique that consists of an encoder and a decoder. The encoder maps the original features to a lower-dimensional representation, and the decoder maps the lower-dimensional representation back to the original features. Autoencoders are trained to minimize the reconstruction error, which encourages the model to learn a meaningful lower-dimensional representation of the data.

03. Anomaly detection

Anomaly detection

Anomaly detection, also known as outlier detection, is a form of unsupervised learning that aims to identify data points that are significantly different from the normal behavior of most data points. Anomalies are data points that are rare, unusual, or unusual compared to the rest of the data, and anomaly detection algorithms aim to identify these data points based on their unique characteristics. There are several types of anomaly detection algorithms, including:

1. Statistics-based methods: These methods assume that normal data points follow a certain statistical distribution, such as a Gaussian or Poisson distribution, and identify anomalies based on their deviation from this expected distribution. Common statistical anomaly detection methods include:

2. Z-score: Z-score, also known as standard score, measures the number of standard deviations by which a data point differs from the data mean. Data points with Z-scores above a certain threshold can be considered outliers.

3. Modified Z-score: Modified Z-score is a variant of Z-score that is more robust to bias. It uses the mean and mean absolute deviation (MAD) instead of the mean and standard deviation, which makes it less sensitive to extreme values.

4. Gaussian Mixture Model (GMM): As mentioned in the previous answer, GMM is a probabilistic model that can also be used for anomaly detection. It estimates the parameters of a mixture of Gaussian distributions from the data and identifies anomalies based on the probability of the data points in the learned distribution.

Distance-based methods: These methods measure the similarity or dissimilarity of data points and identify anomalies based on their distance from the majority of data points. Common distance-based anomaly detection methods include:

1. Euclidean distance: Euclidean distance is a widely used distance measure that measures the straight-line distance between two data points in object space. Data points with a large Euclidean distance from the centroid or the majority of data points can be considered anomalies.

2. Mahalanobis distance: The Mahalanobis distance is a measure of the distance between a data point and the center of the data, taking into account the covariance of the features. It is a more robust distance measure than the Euclidean distance because it takes into account the correlation between features.

3. Cluster-based methods: These methods group data points into clusters and identify anomalies as data points that do not belong

to any cluster or belong to very small clusters. Common methods for detecting cluster anomalies include:

4. DBSCAN: As mentioned in the previous answer, DBSCAN is a density-based clustering algorithm that can also be used for anomaly detection. Data points that do not belong to any cluster or clusters with very few points can be considered as anomalies.

5. K-means: K-means clustering, which is primarily used for clustering, can also be used to detect anomalies, and data points far from any cluster centroid are considered anomalies.

Ensemble methods: These methods combine multiple anomaly detection methods to improve the overall performance and robustness of the detection. Common ensemble methods for anomaly detection include:

1. Isolation Forest: Isolation Forest is an ensemble method that uses random forests to isolate anomalies. It recursively subdivides the data until anomalies are isolated to individual leaf nodes, enabling efficient anomaly detection.

2. Local Outlier Factor (LOF): LOF is a density-based ensemble method that measures the local density of a data point and compares it to the density of neighboring data points. Data points with a significantly lower density compared to their neighbors are considered anomalies.

*THE END