– Unsupervised Algorithms –

Clustering Algorithms

ALGORITHM	DESCRIPTION & APPLICATION	ADVANTAGES	DISADVANTAGES
K-Means	Most common clustering approach which assumes that the closer data points are to each other, the more similar they are It determines K clusters based on Euclidean distances.	1. Scales to large datasets 2. Interpretable & explainable results 3. Can generate tight clusters	1. Requires defining the expected number of clusters in advance. 2. Not suitable to identify clusters with non-convex shapes.
DBSCAN	Density-Based Spatial Clustering of Applications with Noise can handle non-linear cluster structures, purely based on density. It can differentiate and separate regions with varying degrees of density, thereby creating clusters.	1. No assumption on the expected number of clusters. 2. Can handle noisy data and outliers 3. No assumptions on the shapes and sizes of the clusters 4. Can identify clusters with different densities	1. Requires optimization of two parameters 2. Can struggle in case of very high dimensional data
HDBSCAN	Family of the density-based algorithms and has roughly two steps: finding the core distance of each point, and expands clusters from them. It extends DBSCAN by converting it into a hierarchical clustering algorithm.	1. No assumption on the expected number of clusters 2. Can handle noisy data and outliers. 3. No assumptions on the shapes and sizes of the clusters. 4. Can identify clusters with different densities	1. Mapping of unseen objects in HDBSCAN is not straightforward. 2. Can be computationally expensive
Agglomerative Hierarchical Clustering	Uses hierarchical clustering to determine the distance between samples based on the metric, and pairs are merged into clusters using the linkage type.	1. There is no need to specify the number of clusters. 2. With the right linkage, it can be used for the detection of outliers. 3. Interpretable results using dendrograms.	1. Specifying metric and linkages types requires good understanding of the statistical properties of the data 2. Not straightforward to optimize 3. Can be computationally expensive for large datasets
OPTICS	Family of the density-based algorithms where it finds core sample of high density and expands clusters from them. It operates with a core distance (e) and reachability distance.	No assumption on the expected number of clusters. 2. Can handle noisy data and outliers. 3. No assumptions on the shapes and sizes of the clusters. 4. Can identify clusters with different densities. 5 Not required to define fixed radius as in DBSCAN.	1. It only produces a cluster ordering. 2. Does not work well in case of very high dimensional data. 3. Slower than DBSCAN.

Dimensionality Reduction Techniques

ALGORITHM	DESCRIPTION & APPLICATION	ADVANTAGES	DISADVANTAGES
PCA	Principal Component Analysis (PCA) is a feature extraction approach that uses a linear function to reduce dimensionality in datasets by minimizing information loss.	1. Explainable Interpretable results. 2 New unseen datapoints can be mapped into the existing PCA space 3. Con be used as dimensionality reduction technique as preliminary step to other machine learning tasks 4 Helps reduce overfitting 5. Helps remove correlated features	1. Sensitive to outliers 2. Requires data standardization
t-SNE	t-distributed Stochastic Neighbor Embedding is a non-linear dimensionality reduction method that converts similarities between data points to joint probabilities using the Student t-distribution in the low-dimensional space	1. Helps preserve the relationships seen in high dimensionality 2. Easy to visualise the structure of high dimensional data in or 3 dimensions 3. Very effective for visualizing clusters or groups of data points and their relative proximities	1. The cost function is not convex: different initializations can get different results. 2. Computationally intensive for large datasets. 3. Default parameters do not always achieve the best results
UMAP	Uniform Manifold Approximation and Projection (UMAP) constructs a high-dimensional graph representation of the data then optimizes a low-dimensional graph to be as structurally similar as possible.	1. It can be used as general-purpose dimension reduction technique as o preliminary step to other machine learning tasks. 2. Can be very effective for visualizing clusters or groups of data points and their relative proximities. 3 Able to handle high dimensional sparse datasets	1. Default parameters do not always achieve the best results
ICA	Independent Component Analysis (ICA) is a linear dimensionality reduction method that aims to separate a multivariate signal into additive subcomponents under the assumption that independent components are non-gaussian. Where PCA "compresses" the data, ICA "separates" the information.	1. Can separate multivariate signals into its subcomponents 2 Clear aim of the method: only applicable if there are multiple independent generators of information to uncover. 3. Can extract hidden factors in the data by transforming a set of variables to new set that maximally independent.	1. Without any prior knowledge, determination of the number of independent components or sources can be difficult. 2. PCA is often required as a pre-processing step.

Association Rules

ALGORITHM	DESCRIPTION & APPLICATION	ADVANTAGES	DISADVANTAGES
Apriori algorithm	The Apriori algorithm uses the join and prune step iteratively to identify the most frequent itemset in the given dataset. Prior knowledge (apriori) of frequent itemset properties is used in the process.	1. Explainable & interpretable results. 2. Exhaustive approach based on the confidence and support.	1. Requires defining the expected number of clusters or mixture components in advance 2. The covariance type needs to be defined for the mixture of component
FP-growth algorithm	Frequent Pattern growth (FP-growth) is an improvement on the Apriori algorithm for finding frequent itemsets. It generates a conditional FP-Tree for every item in the data.	1. Explainable & interpretable results. 2. Smaller memory footprint than the Apriori algorithm	1. More complex algorithm to build than Apriori 2. Can result in many (incremental) overlapping/trivial itemsets
FP-Max Algorithm	A variant of Frequent pattern growth that is focused on finding maximal itemsets.	1. Explainable & Interpretable results. 2. Smaller memory footprint than the Apriori and FP-growth algorithms	1. More complex algorithm to build than Apriori