CLUSTERING AND DIMENSIONALITY REDUCTION

ROLL NUMBER	NAME
13000121058	ARKAPRATIM GHOSH

CA 1 : MACHINE LEARNING (PEC-CS701E)

CSE : SEMESTER 7

CONTENT

INTRODUCTION

When using unsupervised machine learning, AI technology does all the work. Unlike supervised learning, you do not have to label each data before feeding it to the machine for grouping and analysis.

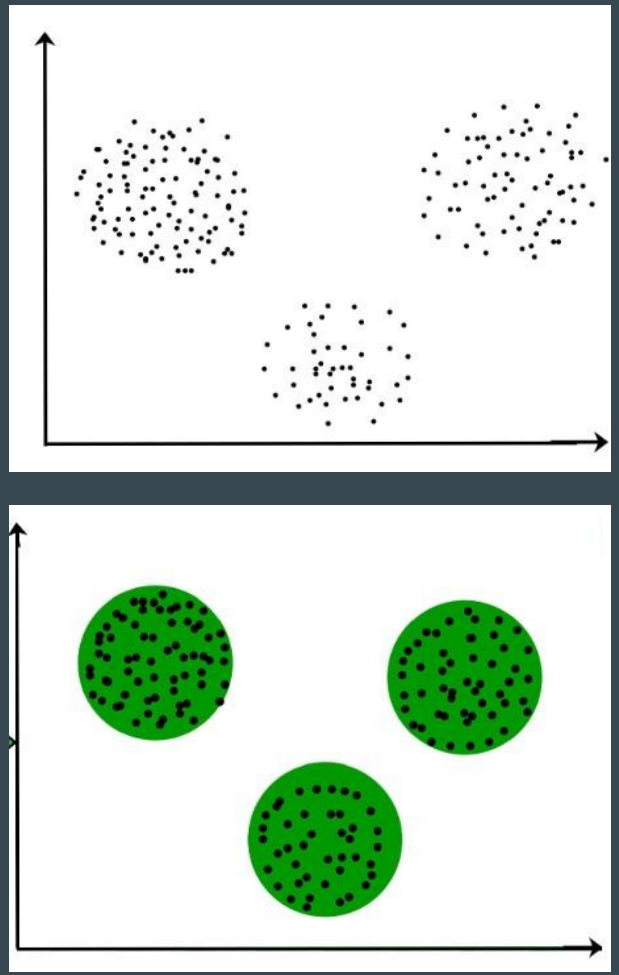
Instead, the machine uses algorithms to find differences, similarities, or patterns in the data and classifies them accordingly without supervision. While the output may not be very accurate, unsupervised learning is the best method to process complex, unlabelled data, which can be easily retrieved from a computer.

This ML type is primarily used in network analysis, anomaly detection, and singular value decomposition. Unsupervised machine learning has two major approaches: Clustering and Dimensionality Reduction, which are discussed below.

CLUSTERING

The task of grouping data points based on their similarity with each other is called Clustering or Cluster Analysis. This method is defined under the branch of Unsupervised Learning, which aims at gaining insights from unlabelled data points, that is, unlike supervised learning we don't have a target variable.

Clustering aims at forming groups of homogeneous data points from a heterogeneous dataset. It evaluates the similarity based on a metric like Euclidean distance, Cosine similarity, Manhattan distance, etc. and then group the points with highest similarity score together.



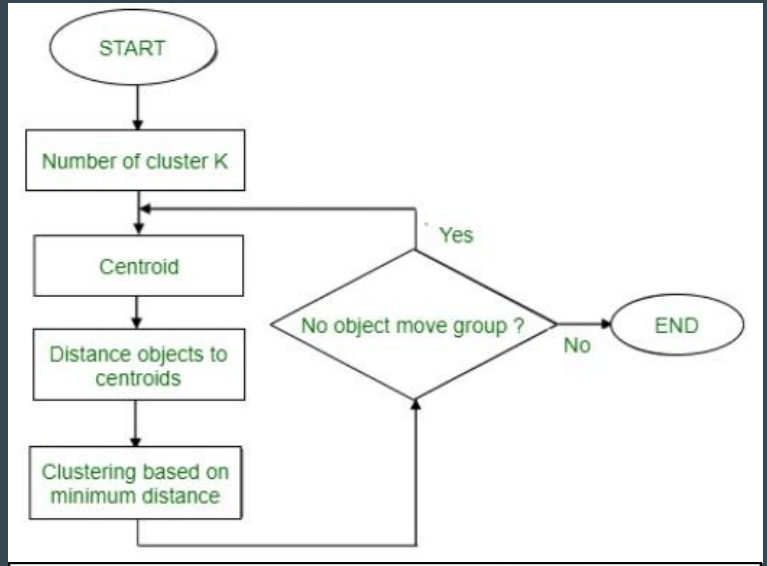
TYPES OF CLUSTERING

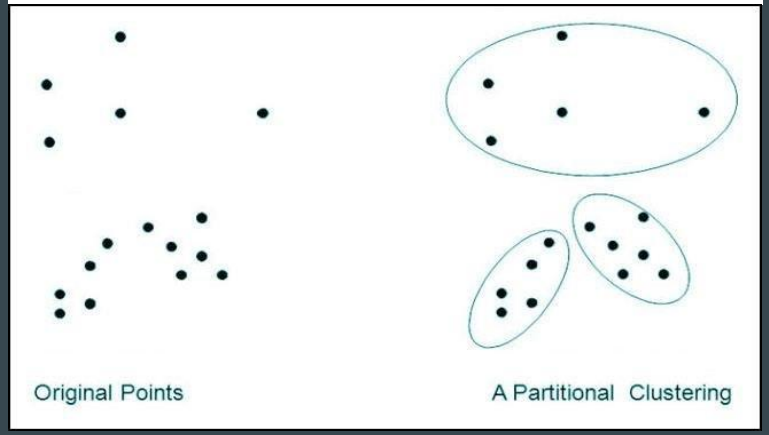
1. PARTITIONING CLUSTERING

Also known as hard clustering, where a data point can only belong to one group, partitioning clustering divides data into non-hierarchical groups. **The k-means clustering algorithm** is the most common form of this centroid-based method.

In K-means clustering, the dataset is divided into a pre-defined number of groups represented by K. The data are grouped according to their distance from the cluster center.

If the data has a large K value, the data has a smaller grouping. On the other hand, a smaller K value depicts a more extensive, less granular grouping.





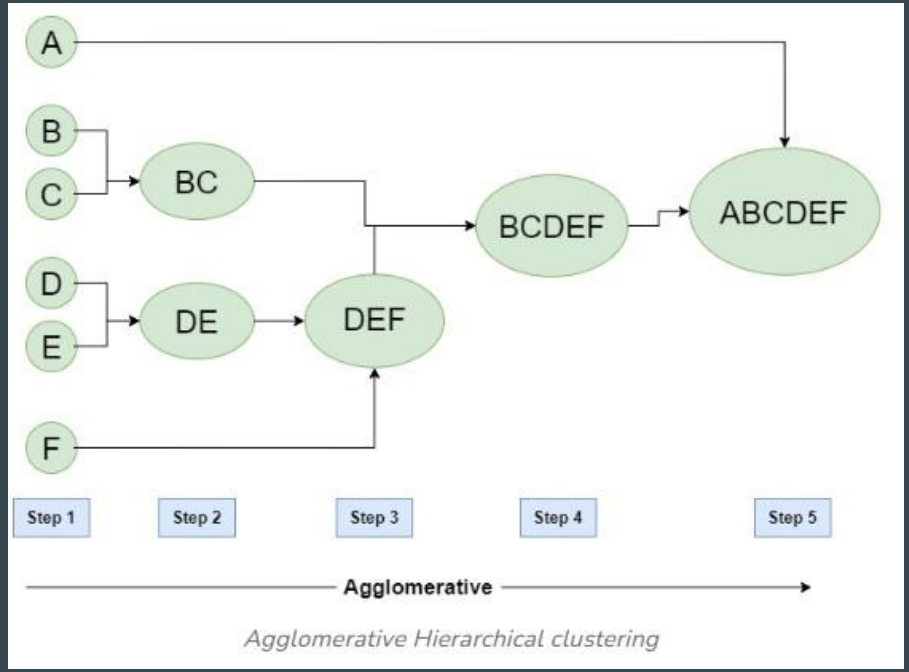
TYPES OF REGRESSION

2. HIERARCHICAL CLUSTERING

Hierarchical cluster analysis divides the data set into a dendrogram, having different levels. This soft clustering enables multiple grouping of data. Hierarchical clustering can either be agglomerative or divisive.

In agglomerative clustering, the data sets are first divided into small groups, then merged on higher levels based on the similarity as a cluster. This can be done through various methods, including Ward's, average, complete, or single linkage.

In divisive clustering, the data set is initially considered as one cluster, which is further divided into smaller groups in progressing levels based on the differences between the data points.



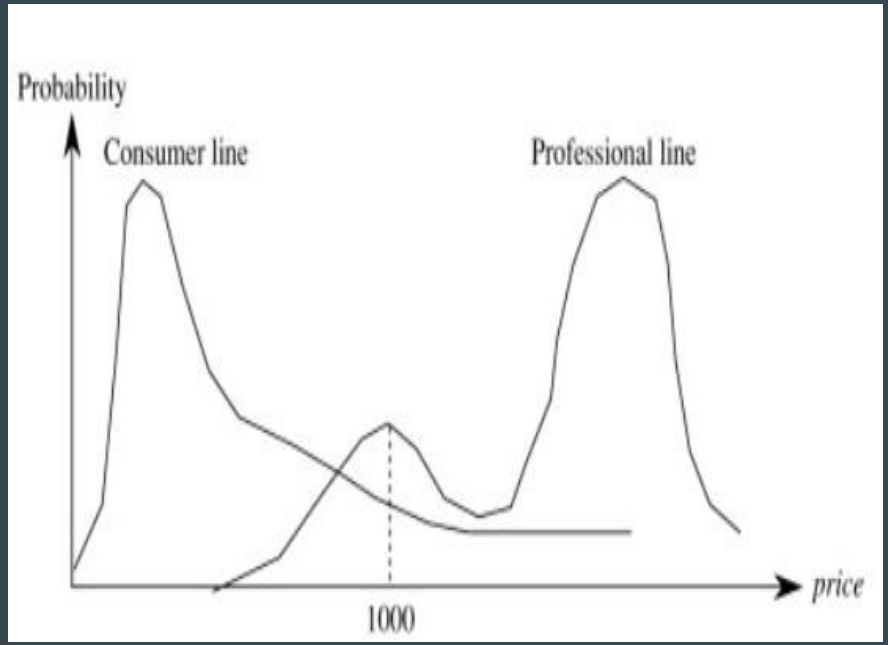
Agglomerative Hierarchical clustering

TYPES OF REGRESSION

3. PROBABILISTIC CLUSTERING

Probabilistic clustering or distribution model-based clustering divides data based on their probability of belonging to a particular distribution. The Gaussian Mixture Model (GMM) is the most common algorithm for probabilistic clustering.

GMM uses an unspecified number of distribution functions and divides data based on the probability of a data set. The variable for the clustering is assumed. Therefore, the expectation-maximization algorithm is commonly used with GMM.

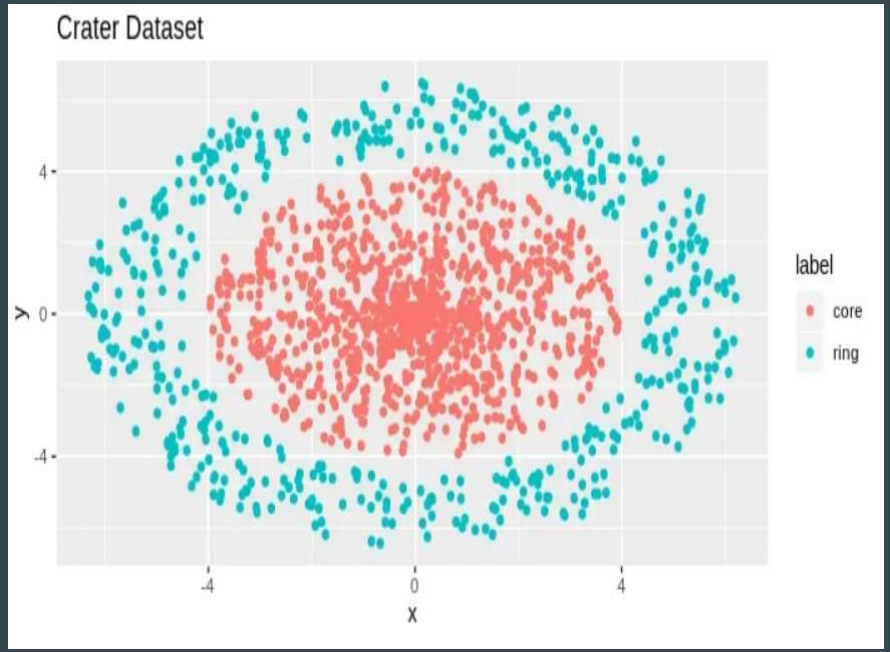


TYPES OF REGRESSION

4. DENSITY BASED CLUSTERING

The density-based clustering is a simple approach that divides data into groups based on their positioning. Highly dense areas are grouped as one cluster, forming arbitrary shapes.

Two clusters are divided by clear, sparser areas. However, this type of clustering is not helpful for data sets with varying densities.



TYPES OF REGRESSION

5. FUZZY CLUSTERING

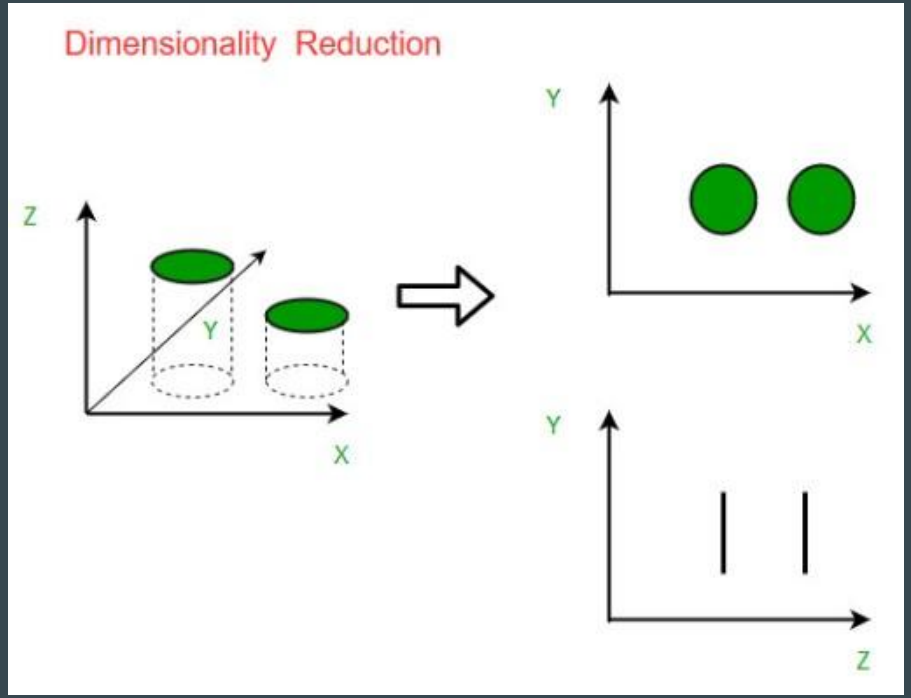
Another type of soft clustering, the data sets are assigned membership coefficients, determining their degree of belonging to different clusters. The most common method used for this is the Fuzzy C-means algorith

Clustering techniques have wide applications in the real world. These include identifying cancer cells, customer segmentation, search engine results, market segmentation, anomaly detection, and statistical data analysis.

DIMENSIONALITY REDUCTION

The data dimension refers to the number of variables, columns, or inputs. Any dataset with multiple features can make prediction modeling difficult due to high dimensionality that causes overfitting and make it harder to visualize the data sets.

Dimensionality reduction is an essential aspect of machine learning since it helps drive more accurate results for large data sets by helping reduce the number of features under scrutiny, making the data more manageable without removing any integral part. This helps avoid the curse of dimensionality and produce a better-fit predictive model.

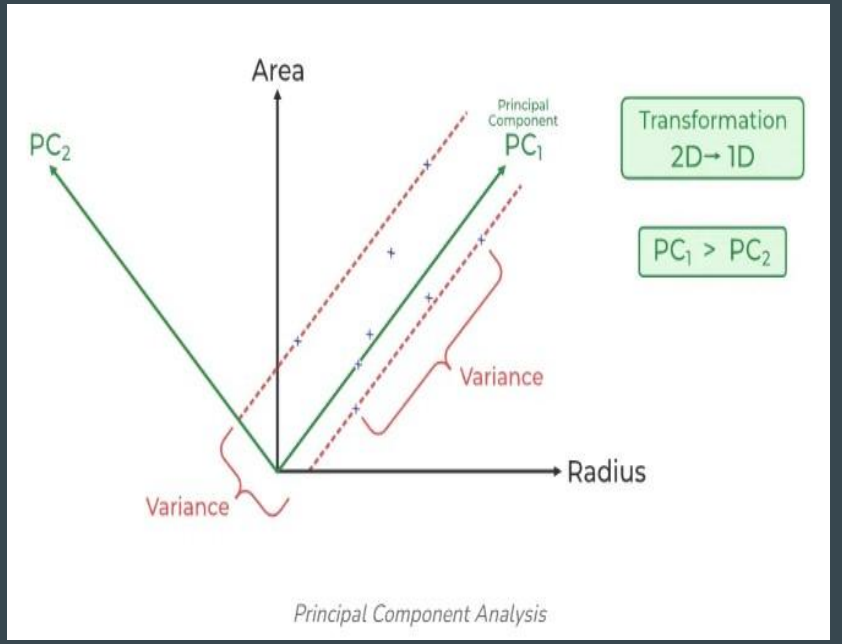


TECHNIQUES

1. PRINCIPAL COMPONENT ANALYSIS

For feature extraction, Principal Component Analysis (PCA) uses a linear transformation to produce a set of new principal components, reducing the number of dimensions to a minimum without information loss.

The process is repeated to find linear transformations which are entirely uncorrelated to each other in an orthogonal way. This helps maximize the variance of the data set.

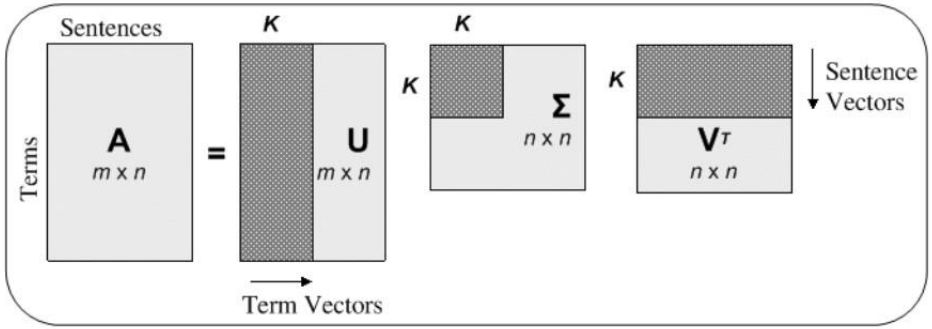


TYPES OF CLASSIFICATION

2. SINGULAR VALUE DECOMPOSITION

Singular Value Decomposition (SVD) divides a principal matrix into three lower matrices. It is generally based on the formula A = USVT, where U and V represent orthogonal matrices, and S represents a diagonal matrix.

Like PCA, it is generally used to reduce noise and compress data, such as in image files.



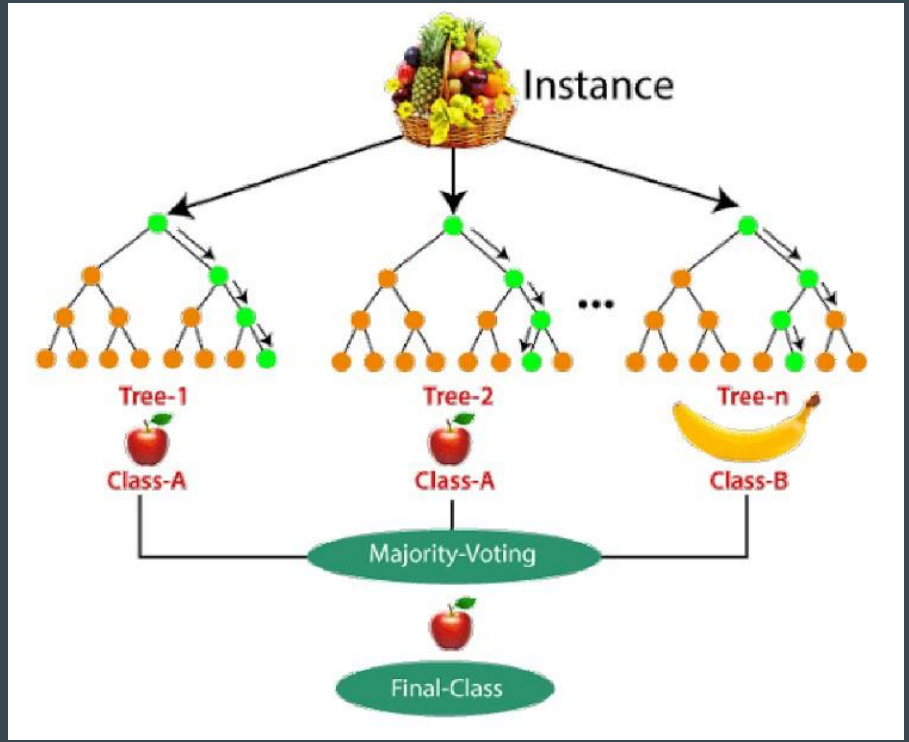
The Singular Value Decomposition (SVD) of an m × n rectangular matrix A

TYPES OF CLASSIFICATION

3. RANDOM FOREST

Another popular dimensionality reduction method in machine learning, the random forest technique, has an in-built algorithm for generating feature importance. It uses statistics of each attribute to find the subset of features.

However, this algorithm only accepts numerical variables. Therefore, the data has to be first processed using hot encoding.



CONCLUSION

Clustering and dimensionality reduction are essential techniques in data analysis. Clustering groups similar data points together, revealing patterns and structures within the data. Dimensionality reduction simplifies complex datasets by reducing the number of features, enhancing visualization, and improving the performance of machine learning models. Together, they enable more efficient data processing and insightful analysis, driving better decision-making.

REFERENCES

- https://ifacet.iitk.ac.in/knowledge-hub/machine-learning/unsupervised-learning-clustering-and-dimensionality-reduction-techniques/
- https://www.geeksforgeeks.org/clustering-in-machine-learning/
- https://www.geeksforgeeks.org/dimensionality-reduction/