MAJOR ASSIGNMENT - 2

Machine Learning Concepts

Machine Learning Fundamentals :-

(1.) Describe the main differences between supervised and unsupervised learning.

Ans. The main difference between supervised and unsupervised learning is stated as :-

Supervised Learning	Unsupervised Learning
* In supervised learning, the training data, feed to the algorithm includes desired solution called label data.	* In unsupervised learning, the training data is unlabelled.
* High accuracy	* Less accurate
* Prone to overfitting	* Subjective interpretation

(2.) Explain how semi-supervised learning works and its applications.

Ans. In semi-supervised learning, algorithms can deal with partially labelled training data, usually a lot of unlabelled data and a bit of labelled data. It learns from labelled data and leverages the unlabelled data.

Applications :-

- Text classification
- Speech recognition
- Fraud detection
- Medical image classification

(3.) Compare the advantages and disadvantages of supervised and unsupervised learning.

Ans. Advantages of supervised learning :-

* Supervised learning allows collecting data and produces data output from previous.
* Helps to optimize performance criteria with the help of experience.
* It performs classification and regression tasks.

Disadvantages of supervised learning :-

* Classifying big data can be challenging.
* Computation time is vast for supervised learning.
* Supervised learning cannot handle all complex tasks in machine learning.

Advantages of unsupervised learning :-

* It doesnot require training data to be labelled.
* Dimensionality reduction can be easily accomplished.
* Capable of finding previously unknown patterns in data.

Disadvantages of unsupervised learning :-

* Less accurate results.
* Unsupervise data can be noisy and contains outliers.

(4.) What is Transfer learning, and how does it relate to supervised learning?

Ans. Transfer learning is a technique where knowledge gained through one task or dataset is used to improve model performance on a similar related task to it.

When we use supervised learning, the learning from pre-trained model helps to leverage the knowledge and improve the performance in target task.

(5) Explain the concept of clustering in unsupervised learning.

Ans Clustering is an unsupervised machine learning technique that involves grouping similar data points together. Unlike supervised learning, where we have labelled data to train a model in unsupervised learning, we work with unlabelled data. The goal of clustering is to discover hidden patterns and structure within the data without any prior knowledge.

(6) Define reinforcement learning with an example.

Ans Reinforcement learning is technique where an agent learns to make decision by interacting with environment by getting rewards in returns or penalties in form of negative rewards.

For example – A Robot learning to walk, for every correct movement, it get positive rewards and every negative rewards for wrong movements.

(7) Define imbalanced data and give examples.

Ans Imbalanced data refers to datasets where the distribution of classes is not uniform. In classification problems, this means that one or more classes are significantly underrepresented compared to others.

Examples of imbalanced data :-

* Fraud Detection
* Medical Diagnosis
* Spam Detection
* Churn Prediction
* Earthquake Prediction

(8.) Explain the impact of imbalanced data on model performance.

Ans. Imbalanced data can significantly impact the performance of machine learning models, often leading to biased predictions and poor generalization, especially for the minority class. The key impacts are :-

* Bias toward the majority class
* Poor minority class performance
* Misleading Evaluation Metrics
* Delayed convergence during training.
* Reduced Generalization ability.

(9.) Provide examples of how to handle imbalanced data in real-world scenarios.

Ans i) Fraud Detection :- Fraudulent transactions are often a small fraction of all transactions.

Handle → oversampling minority data
under sampling majority data.

ii) Medical Diagnosis :- Rare diseases have limited data compared to common disease.

Handle → data augmentation
transfer learning

iii) Spam Detection :- Spam emails may constitute only 20 % of all emails.

Handle → undersampling
hybrid techniques
ensemble methods.

(10.) Explain hyperparameter tuning using grid search method.

Ans. Hyperparameter using grid search method :-

* System approaches to find optimal hyperparameter for models.
* Creates a grid of hyper values.
* Trains the models for each combination value.
* Selects the best performance value.

(11.) Define overfitting and underfitting.

Ans. Overfitting occurs when a machine learning model becomes too complex and fits the noise in the training data rather than the underlying patterns. This results in poore generalization to new data.

Underfitting occurs when a model is too simple to capture the underlying patterns in the data.

(12.) Explain bias - variance tradeoff.

Ans. Getting the right balance between the bias and variance tradeoff is fundamental of effective machine learning algorithms.

Bias : error due to incorrect assumption

Variance : error due to sensitivity to training data.

- High bias, low variance (Underfitting)
- Low bias, high variance (Overfitting)

(13.) Explain the following exploratory data analysis methods.

a) Regression plot :-

A regression plot visualizes the relationship between two variables along with a fitted regression line. It shows how well one variable predicts another.

It is used to analyze the trend or correlation between variables and detect deviations from the trend.

b) **Histograms :-**

A histogram represents the distribution of a single variable by dividing the data into intervals (bins) and displaying the frequency of data points in each bin.
It is used to understand the data distribution (e.g., normal, uniform)

c) **Scatter plot :-**

A scatter plot is a graph of points where each point represents the values of two numerical variables.
It is used to visualize relationships, trends, or clusters between two variables.

d) **Joint Plot :-**

A joint plot combines a scatter plot with histograms or density plots for the x and y axes to visualize their relationship and individual distributions.
It is used to analyze both the relationship between two variables and their marginal distributions.

e) **Pair Plot :-**

A pair plot creates scatter plots for all possible pairings of numerical variables in a dataset, often including histograms or density plots along the diagonal.
It is used to explore pairwise relationships and distributions among multiple variables.

f) **Correlation :-**

Correlation measures the strength and direction of a linear relationship between two variables, often visualized using a heatmap.
It is used to identify which variables are strongly related and guide feature selection or engineering.

g) **Box plot** :-

A box plot summarizes the distribution of a dataset using five statistics : minimum, first quartile (Q1), median (Q2), third quartile (Q3) and maximum. It also highlight outliers. It is used to detect outliers and understand the spread and skewness of data.

(14.) Explain how outliers can be determined in ML tasks.

Ans Outliers can be determined in ML Tasks :-

* **Statistical Method** :-

→ **Z-score** : Measures how many standard deviation datapoints is from the mean. Datapoints with z-scores exceeding a certain threshold are considered outliers.

→ **Interquartile range** : Identifies outliers based on their distance from the first and third quartiles. Data points falling outside the range of (Q1 - 1.5) x IQR and (Q3 - 1.5) x IQR are considered outliers.

* **Visualization Methods** :-

→ Box plots
→ Scatter plots
→ Histograms

* **Model-Based Methods** :-

→ Isolation Forest
→ One-class SVM
→ K-nearest Neighbours (KNN)

* Domain Knowledge.

(15) Explain following methods of handling outliers in ML.

a) Boxplot Method

b) Z-score method

c) Inter Quantile Range (IQR) Method.

Ans a) **Boxplot Method** :- Handling outliers by :-

* **Remove** : Drop outliers from the dataset.
* **Cap/Clamp** : Replace outliers with nearest whisker value.
* **Transform** : Use log or square root transformations to reduce the impact of outliers.

b) **Z-score method** :- Handling outliers by :-

* **Remove** : Exclude data points with extreme Z-scores.
* **Impute** : Replace outliers with mean, median or mode.
* **Cap** : Limit extreme values to the threshold.

c) **Interquartile Range (IQR) Method** :- Handling outliers by :-

* **Remove** : Drop points beyond the bounds.
* **Cap** : Replace outliers with nearest bound value.

Linear Regression :-

(1.) What assumptions must hold for a Linear Regression model to be considered valid ? Explain their significance .

Ans. For a linear regression model following assumptions must hold true :

* Linearity :- If this assumption does not hold, the model will not accurately capture the true relationship, leading to biased prediction .
* Independence of Error :- Violation of this assumption, such as in time-series data where errors are autocorrected .
* Normality of Residuals :- While this assumption is not critical for estimating coefficients .

(2.) How do you handle multicollinearity in Linear Regression ?

Ans Handling multicollinearity in linear regression is essential because high multicollinearity can lead to unstable coefficient estimates .

Strategies to handle are :-

* Remove highly correlated predictors .
* Combine predictors .
* Regularization Technique .
* Centre and Scale predictors .

(3.) How can you detect and address heteroscedasticity in a Linear Regression model ?

Ans ⇒ Heteroscedasticity can be detected by :-

* Residual Plots
* Statistical Tests
* Scale-location plot
* Leverage influence metrics .

⇒ Addressing Heteroscedasticity :-

* Transform the dependent variable
* Weighted least squares (WLS)
* Robust standard errors
* Model the variance.

(4.) Explain the concept of residuals in Linear Regression and their importance in model validation.

Ans. Residuals are the difference between the observed values (y_i) and the predicted values ($\hat{y}_i$) from a regression model.

$$e_i = y_i - \hat{y}_i$$

Importance of Residuals in model validation :-

* Checking model Assumption
* Identifying Model Fit
* Identifying outliers.
* Deflecting influential points.
* Improving model performance

(5.) How would you handle categorical ~~values~~ variables in a Linear Regression model.

Ans. To use categorical variables in a linear Regression model, we need to convert them into numerical variables that can be used in the model. There are several techniques for doing this, including one-hot encoding, label encoding, etc.

(6.) Explain the concept of gradient descent and how it achieves minimum of a given cost function for linear regression.

Ans. Gradient descent is an optimization algorithm used to minimize a cost function by iteratively adjusting the model parameters in the direction of steepest descent of cost function. In the context of linear regression, the goal of gradient descent is to minimize the mean squared error (MSE) which is the cost function commonly used.

The cost function

$$J(\theta) = \frac{1}{2m} \sum_{i=1}^{m} (y_i - \hat{y}_i)^2$$

$J(\theta)$: cost function value
m : number of training examples.

(7.) How do you decide which features to include in a Linear Regression model ?.

Ans Selecting the right feature is crucial for building an effective Linear Regression model. To decide which feature to include in Linear Regression these strategies are choosen.

* Domain knowledge
* Statistical Methods for Feature selection
* Automated Feature selection methods.
* Regularization techniques.

(8.) Explain batch learning; stochastic gradient learning, and minibatch learning used in linear regression.

Ans Batch Learning :-

In batch learning, the entire dataset is used to compute the gradients and update the model parameters during each iteration. All training examples are processed in one go, which ensures stable convergence of the algorithm.

Stochastic Gradient Learning :-

In stochastic gradient descent, only one training example is used to compute the gradient and update the model parameters in each iteration.
This leads to more frequent updates, which can make the training process faster.

Minibatch Learning :-

Minibatch learning is a compromise between batch learning and stochastic ~~learning~~ gradient descent. The dataset is divided into small batches, and each batch is used to compute the gradient and update the model parameters. This approach balances the stability of batch learning and the efficiency of stochastic gradient descent.

(9) Discuss the different types of regularization techniques used in linear regression.

Ans Regularization is a technique used to reduce errors by fitting the function appropriately on the given training set and avoiding overfitting.

The commonly used regularization technique are :-

* Lasso Regularization - L1 Regularization
* Ridge Regularization - L2 Regularization

(10) Explain the following performance measures in linear regression.

a) Absolute Error
b) Mean Absolute Error
c) Squared Error
d) Mean Squared Error
e) R2 score

Ans. a) **Absolute Error** :- Quantifies the absolute difference between the observed values (y_i) and the predicted values ($\hat{y}_i$) in a regression model.

b) **Mean Absolute Error** :- MAE is a very simple metric which calculates the absolute difference b/w actual and predicted values

$$MAE = \frac{1}{N}\sum_{i=1}^{N} |y_i - \hat{y}_i|$$

c) **Squared Error** :- SE is a measure of the difference between the actual observed values (y_i) and the predicted values ($\hat{y}_i$) in a regression model.

$$SE = (y_i - \hat{y}_i)^2$$

d) **Mean Squared Error** :- MSE is the most used and very simple metric with a little bit of change in mean absolute error. Mean squared error states that finding the squared difference.

$$MSE = \frac{1}{N}\sum_{i=1}^{N} (y_i - \hat{y}_i)^2$$

e) **R2 score** :- R2 score is a metric that tells the performance of your model, not the loss in a absolute sense that how many wells did your model perform.

$$R2 = 1 - \frac{SS1}{SSR}$$

Logistic Regression :-

(1.) What is the logit function in logistic regression ?.

Ans. The logit function in logistic regression is the link function that connects the linear model to the probability of an outcome. It is defined as the natural logarithm of the odds (log-odds) of the binary dependent variable Y being 1 (success) Vs 0 (failure).

$$\text{logit}(p) = \ln\left(\frac{p}{1-p}\right), \quad 0 < p < 1$$

(2.) Explain the concept of maximum likelihood function.

Ans. The maximum likelihood function is a fundamental concept in statistics and machine learning, used to estimate the parameters of a model.
The idea is to find the set of parameters that make the observed data most likely under the assumed model.

Mathematically, $L(\theta) = P(x|\theta)$

Properties :-

* Consistency
* Efficiency
* Invariance

(3) Explain the significance of log likelihood function used in logistic function.

Ans. The log-likelihood function in logistic regression is a key concept that quantifies how well the logistic model explains the observed data. It serves as the foundation for parameter estimation through maximum likelihood estimation (MLE).

Properties :-

* It measures model fit.
* Enables Maximum Likelihood function
* Simplifies Optimization
* Basis for model comparison.

(4.) Explain the concept of the sigmoid function in Logistic Regression and how it helps in making Predictions.

Ans. The sigmoid function is a crucial mathematic function in logistic Regression that transforms a linear combination of inputs into a probability value bounded b/w 0 and 1. This transformation enables logistic regression to model binary outcomes effectively.

Mathematically, $\sigma(z) = \frac{1}{1+e^{-z}}$

Properties :-

* Transform Raw scores to probabilities.
* Interpretable output.
* Facilitates Gradient-based optimization.

(5.) Explain how the F1 score is useful for imbalanced data.

Ans. The F1-score is highly useful for imbalanced data because it ensures that a model performs well on the minority class by balancing precision and recall. This makes it a better metric than accuracy for scenarious where the cost of misclassifying the minority class is high.
It focuses on positive class and is insensitive to class imbalance.

(6.) Discuss the use of ROC curve in imbalanced datasets.

Ans. The ROC curve is a valuable tool for evaluating classifiers on imbalanced datasets as it provides insights into the trade-offs b/w TPR and FPR across thresholds.
It is a graphical representation used to evaluate the performance of a binary classification model by illustrating its ability.

For e.g. :- In a medical diagnosis setting,
Positive class : Disease presence
Negative class : Healthy

(7) What is the role of a validation set ?.

Ans. A validation set is a subset of data used during model development to evaluate a model's performance on unseen data. Its primary role to help tune model parameters, select the best model, and avoiding overfitting, ensuring that the model generalizes well to new data.

Properties :-
* Hyper parameter Tuning
* Performance estimation.

(8.) What is the purpose of a confusion matrix ?.

Ans. The confusion matrix is a tabular representation used to evaluate the performance of a classification model. It provides a detailed breakdown of the model's predictions by comparing them to the actual values, offering insights into its strengths and weakness.

Example :-
* Medical Diagnosis
* Fraud Detection
* Spam Filtering

(9.) How is the accuracy metrix calculated ?.

Ans. The accuracy metric in machine learning is calculated as the ratio of correct predictions to the total number of predictions made by the model. It measures how often the model correctly classifies instances.

$$\text{Accuracy} = \frac{\text{True Positive (TP)} + \text{True Negative (TN)}}{\text{TP} + \text{TN} + \text{False Positive (FP)} + \text{False Negative (FN)}}$$

(10.) Explain the terms sensitivity and specificity concerned with classification.

Ans. Sensitivity measures the proportion of actual positive cases that the model correctly identifies. It focus on detecting positive instances.

$$\text{Sensitivity} = \frac{TP}{TP + FN}$$

Specificity measures the proportion of actual negative cases that the model correctly identifies. It focus on detecting negative instances.

$$\text{Specificity} = \frac{TN}{TN + FP}$$

Support Vector Machines :-

(1.) Explain the concept of the margin in SVM and why maximizing it is important.

Ans. The margin in Support Vector Machines (SVM) refers to the distance between the decision boundary (or hyperplane) and the nearest data points from each class. These nearest data points are called support vectors because they are critical in defining the position and orientation of decision boundary.

Maximizing the margin is important for :-

* Better Generalization
* Robustness to Noise
* Minimizing Overfitting
* Optimal Separation.

(2.) What are the differences between hard margin and soft margin SVMs, and in what scenarios would you use each?

Ans. Hard margin SVM strictly enforces that all training data points must be correctly classified and lie outside or on the correct side of the margin. No misclassification or margin violations are allowed.

It is used when :-

* The data is linearly separable.
* There is no noise or outliers in the dataset.

Soft margin SVM allows for some violations of the margin and even misclassification of some data points. These violations are controlled by introducing a slack variable $\xi_i \geq 0$ for each data point.

It is used when :-

* The data is not linearly separable.
* There is noise, overlap between classes, or outliers in the dataset.

(3.) How does the kernel trick in Support Vector Machines (SVM) help in handling non-linear data ?.

Ans. The kernel trick in Support Vector Machines (SVM) is a mathematical technique that enables the model to handle non-linear data by implicitly mapping it into a higher-dimensional feature space where it becomes linearly separable. This is achieved without explicitly computing the transformation saving computational resources and enabling efficient training.

(4.) What are the key parameters in SVM that need to be tuned and how do they impact model performance ?.

Ans. Tuning the key parameters of a SVM is critical to achieving optimal performance. The key parameters and their impact on the model are :-

* Regularization Parameter (C) :- It controls the trade-off between achieving a low error on the training data and maximizing the margin.

* Gamma (RBF Kernel) :- It determines the influence of a single training example and controls the smoothness of the decision boundary.

* Kernel type & parameter :- It defines data transformation and separation capability.

* Class Weight :- It helps prevent the SVM from being biased toward the majority class and improves performance on imbalanced datasets.

* Decision Boundary Tolerance (tol) :- It helps in precision of optimization and computational efficiency.

(5.) What are the advantages and disadvantages of using SVM compared to other classification algorithms ?.

Ans Advantages of SVM are :-

* Effective in High-Dimensional Spaces.
* Works well with Non-Linear Boundaries.
* Handles Small Datasets effectively.
* Versatile Kernel Options.
* Optimal Decision Boundary.

Disadvantages of SVM are :-

* Computationally Intensive.
* Inefficient with large datasets.
* Sensitive to feature scaling.
* Limited Interpretability.
* Not ideal for Noisy Data.

Ensemble Models :-

(1.) Compare and contrast Bagging and Boosting as ensemble techniques. What are their respective advantages and disadvantages ?.

Ans.

Aspect	Bagging	Boosting
Objective	Reduce variance	Reduce bias and variance.
Methoddogy	Builds models in parallel.	Builds model sequentially.
Focus on data	Uses random subsets of data for training	Focuses on difficult-to-predict instances.
Weight Adjustment	Equal weights for all models.	Adjusts weights based on model performance.
Model Independence	Models are independent of each other.	Models are dependent on previous iterations.
Examples	Random Forest	Ada Boost

Advantages of Bagging :-

* Reduces Variance
* Parallelizable
* Simple and Effective

Disadvantages of Bagging :-

* Doesnot reduce Bias
* Requires larger datasets.

Advantages of Boosting :-

* Reduces both Bias and Variance
* Strong Performance
* Flexibility.

Disadvantages of Boosting :-

* Prone to Overfitting
* Sequential Process
* Sensitive to Outliers.

(2.) How do you decide on the number of trees to use in a Random Forest model ?.

Ans. Deciding on the number of trees to use in a Random Forest model is an important aspect of model tuning. Here are some key considerations :-

* Bias - Variance Tradeoff :- Adding more trees reduces variance but doesnot significantly impact bias.
* Performance Plateau :- Beyond a certain point, adding more trees yields diminishing returns in performance improvement.
* Cross - validation :- Use cross - validation to evaluate the model's performance with different no. of trees.
* Computational Resources :- More trees requires more computational power and memory.

(3.) What are the key advantages of using ensemble methods in machine learning ?.

Ans The advantages of using ensemble methods in machine learning are :-

* Improved Accuracy
* Reduced Overfitting
* Reduced Variance
* Improved Generalization
* Versatility
* Robustness to Noise
* Better Handling of Complex Data.
* Stability and Reliability.

(4.) Explain the concept of feature importance in the context of Random Forests. How can this information be used in feature selection ?.

Ans. Feature Importance in Random Forests refers to the measure of how much each feature contributes to the predictive performance of the model. This is calculated by evaluating the decrease in impurity (e.g., Gini impurity or entropy for classification, variance for regression) or by assessing the impact on model accuracy when a feature is excluded or randomized.

This information can be used in feature selection by :-

* Mean Decrease in Impurity (MDI)
* Mean Decrease in Accuracy (MDA)

(5.) Explain the concept of stacking in ensemble learning and how it differs from Bagging and Boosting.

Ans. Stacking is an ensemble learning technique that combines the predictions of multiple base models to improve overall performance. Instead of aggregating predictions through averaging or voting (as in Bagging) or iteratively improving models (as in Boosting), stacking employs a meta-model to learn how to best combine the prediction of base models.

Aspect	Stacking	Bagging	Boosting
Objective	Combine diverse models using a meta-model	Reduce variance through averaging	Reduce bias & variance by iterative learning
Model Diversity	Encourages diversity	Often use identical base models.	Uses weak learners iteratively.
Combination Method	Trained meta-model	Voting or averaging	Sequential weighted updates.
Error Handling	Learns from strengths of base models.	Reduces variance by aggregation	Reduces bias by focusing on errors.

Dimensionality Reduction :-

(1.) How does Principal Component Analysis (PCA) reduce dimensionality of a dataset while preserving as much variance as possible ?.

Ans. PCA reduces dimensionality by identifying principal components, which are orthogonal (functions) directions capturing the maximum variance in the data. It ranks these components by their explained variance and projects the dataset onto the top components, preserving as much ~~as~~ variance as possible while discarding less information dimensions.

(2.) What are the limitations of PCA and how can you address them in practical applications ?.

Ans. Limitations :-

* Sensitive to Scale : PCA assumes data is standardized ; otherwise features with large magnitude dominate.
* Linear technique : It cannot capture non-linear relationships.
* Interpretability : Principal components are linear combinations of original features, making them harder to interpret.

Solutions :-

* Standardize the data before applying PCA.
* Use non-linear dimensionality reduction techniques like t-SNE or UMAP for non-linear patterns.
* Combine PCA with domain knowledge to interpret components better.

(3.) How do you decide the number of dimensions to reduce to in a dimensionality reduction technique ?.

Ans. We decide the no. of dimensions on the basis of :-

* Variance Retention :- Choose the number of components that retain a desired percentage of variance (eg 95%).

* Elbow Method :- Plot cumulative explained variance v/s. number of components and select the elbow point.

* Model Performance :- Experiment with different no. of components and evaluate model accuracy or performance metrics.

(4.) What are the trade-offs between preserving variance and computational efficiency in dimensionality reduction ?.

Ans. * Preserving Variance :- Retaining more components preserves more information but increases computational cost and storage requirement.

* Computational Efficiency :- Reducing dimensions significantly speeds up computation and reduces storage but risks losing important data patterns.

Optimal trade-offs depend on the application's tolerance for information loss v/s. performance constraints.

(5.) How can you use dimensionality reduction techniques to improve the performance of machine learning models ?.

Ans. We can use dimensionality reduction techniques by :-

* Reducing Noise :- Eliminates less informative features, reduces overfitting.

* Improved Training Speed :- Reduces computational cost and memory usage for large datasets.

* Better Visualization :- Helps in understanding and interpreting high dimensional data.

* Handling Multicollinearity :- Removes correlated features, improving model-robustness and interpretability.

Clustering :-

(1) Explain the K-means algorithm. How do you determine the optimal number of clusters (K) in a dataset ?.

Ans The K-means algorithm is a machine learning algorithm that partitions the dataset into a set of cluster. The goal of the K-means algorithm is to minimize the distance between data points and their assigned cluster centroid. The optimal no. of clusters, K, can be determined using the elbow method. The algorithm works by :-

* Selecting K : Choosing the number of cluster K
* Randomly selecting centroids :- Picking K points at random to represent the cluster.
* Assigning data points :- Assigning each data point to the closest centroid.
* Calculating variance :- computing the variance for each cluster.
* Updating centroids :- placing the new centroid for each cluster.
* Repeating :- Reassigning data points to closest centroid and repeating until there are no more reassignments.

(2) What are the strengths and weaknesses of hierarchical clustering compared to K-means clustering.

Ans	Hierarchical Clustering	K-means clustering
STRENGTHS:-	* No need to specify the number of cluster. * Works well with small dataset. * Distance-based & versatile.	* Efficient & scaleable * Reassignment of points * Memory Efficient.
WEAKNESS :-	* High computational complexity * Sensitive to Noise and outliers. * Scalability issue.	* Hard partitioning * Assume spherical cluster * Requires pre-specification of K.

(3.) How does the DBSCAN algorithm differ from K-means and hierarchical clustering ?.

Ans. DBSCAN (Density Based Spatial Clustering of Application with Noise) is a density-based algorithm, meaning it identifies cluster based on the local density of datapoints, allowing it to find clusters of arbitrary shapes without requiring the user to specify.

Key difference :-

* Cluster shape :- DBSCAN can find clusters of any shape whereas K-means and hierarchical clustering often struggle with non-spherical cluster.
* Number of clusters :- DBSCAN automatically determines the number of clusters based on density.
* Outlier Handling :- DBSCAN explicitly identifies and labels outliers points as "noise", while K-means and hierarchical clustering may misclassify outliers as part of a cluster.

(4.) Explain the concept of cluster validity indices and how they are used to assess clustering performance.

Ans. Cluster validity indices are qualitative measures that evaluate the performance of clustering algorithm by assessing the compactness of cluster within and the separation of clusters. They can help determine the optimal number of clusters of a given dataset cluster validity technique help find a set of cluster that best fits natural partition of a dataset without any prior class information. Index of validity is a number or magnitude which shows the levels of a tests ability to measure what it intends to measure.

(5.) What is the role of the distance metric in clustering, and how do you choose an appropriate one ?.

Ans In clustering, a distance metric is a mathematical function that determine how similar or different two data points are. The right distance metric is important because it directly affects the quality of the cluster and the insights derived from them. A poorly choosen distance metric can result in misleading or irrelevant cluster.

A appropriate distance metric can be choosen by considering the nature of your data and goals of your analysis.

* Euclidien distance
* Manhattan distance
* Cosine similarity
* Jaccard Index
* Minkowski distance