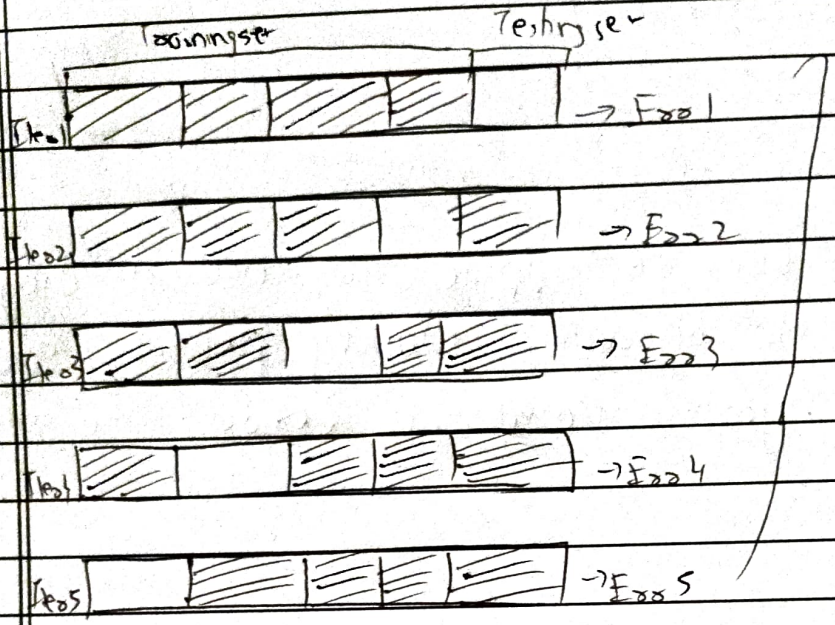
Name: Vinayak Pathak Roll no:- 1020226

Q1)




$$\text{Error} = \frac{1}{5}\sum_{i=1}^{5} \text{Error}_i$$

It is a technique commonly used in ML for assessing the performance of a predictive model and reducing the risk of overfitting. It helps in obtaining a more reliable estimate of model performance.

Steps:

1) Data Splitting
- Data is divided into k equal sized subset.

2) Model Training and Testing
- One of the K fold is used for testing
- Others are used for training

3) Performance metric
- Accuracy or MSE.

4) Averaged Performance
- All are averaged to get the model score.

5) Final Evaluation

Q2)

Ensemble learning is a machine learning techniques that combines predictions from multiple models to improve overall performance.

Scenario 1:

1) Medical Diagnosis:

Ensemble learning can be applied to medical diagnosis where the goal is to accurately classify patients into different disease categories based on various medical features.

Advantages:

Improved Accuracy:

Combining predictions from multiple models, ensemble reduce the risk of misdiagnosis.

Robustness

Medical data can be noisy and subject to variations. Ensemble are robust to outliers.

Generalization:

Ensemble models generalize well to new and unseen patient cases.

Scenario 2

2) Credit Scoring

In financial industry, banks and lending institutions use credit scoring to assess the creditworthiness of loan app

Advantages:
1) Higher Predictive Accuracy:
Ensembles combine the strength of multiple models reducing both false positives and false negatives.
2) Risk mitigation:-
The financial industry is highly regulated and making erroneous lending decisions can lead to significant financial losses
3) Adaptability:
Credit scoring models need to adapt to changing economic conditions and customer behaviors.

Q3)

→ Eigenvalues play a fundamental role in PCA. PCA can transform a high dimensional dataset into a lower dimensional space while retaining as much of the original variance as possible.

1) Covariance matrix:-
PCA begins with the calculation of the covariance matrix of the original data. The matrix represents the relationship between ~~different~~ diff. variable in dataset.
2) Eigenvectors
Eigenvectors are vectors that represent the direction of the maximum variance in the data.
3) Eigenvalues
Eigenvalues are associated with each eigenvector and represent the amount of variance explained by that particular eigenvector

4) Dimensionality Reduction:-
After calculating the eigenvalues and eigen vectors, PCA sorts the eigenvalues in descending order - The sorting provides a ranking of the principal components

5) Variance Retention
Eigenvalues also help in deciding how many principal components to keep.

~~In this~~ Eigenvalues in PCA are used to identify the principal components and quantify the amount of variance explained by each component.

Q4)
→

Dimensionality reduction techniques can be advantageous in real world applications for several reasons including improving computational efficiency, reducing noise, and enhancing the interpretability of data.

Scenario 1

Image and video Compression:

Example: Video streaming services like Netflix and Youtube.

Advantages:-
Dimensionality reduction techniques such as PCA or SVD can be applied to image and video data to reduce storage and bandwidth requirement for streaming

Biomedical Data Analysis
Example:- Analyzing Gene expression data for cancer diagnosis

Advantage:
In genomics, researchers often deal with high-dimensional data where each gene's expression level represents a dimension. Analyzing such data directly can be computationally expensive and prone to overfitting. Dimensionality reduction techniques like t-distributed Stochastic Neighbour embedding or UMAP to reduce the dimensionality of gene expression data.

Q6)

→ a) $J(A,B) = 1 - |A \cap B| / |A \cup B|$

$A = \{1, 2, 3, 4, 5, 8, 9\}$
$B = \{0, 1, 3, 4, 5, 6, 7\}$
$A \cap B = \{1, 3, 4, 5\}$
$A \cup B = \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$
$J(A,B) = 1 - \frac{|A \cap B|}{|A \cup B|}$
$= 1 - \frac{4}{10}$
$= 60\%$

b) Object 1 = {2, 3, 0, 5, 6}
Object 2 = {0, 3, 3, 8, 9}

Dist = $\sqrt{(2-0)^2 + (3-3)^2 + (10-3)^2 + (5-8)^2 + (6-9)^2}$

$= \sqrt{31} \approx 5.57$.

(iii) A = 111 000 111

B = 1111 0000

Hamming distance is 2.

iv) a = "Hello"

b = "Hellow"

No. of edit operation is 2.

v) obj 1 = {2, 3, 0, 5, 6}

obj 2 = {0, 3, 3, 8, 9}

Manhattan dist = $|2-0| + |3-3| + |10-3| + |5-8| + |6-9|$

$= 2 + 0 + 3 + 3 + 3 = 11$

Q7)

XGBoost:

XGBoost, short for extreme Gradient Boosting, is a powerful and popular machine learning algorithm that belongs to the ensemble learning family.

Key features:-

1) Gradient Boosting:
XGBoost is an extension of the gradient boosting algorithm which combine multiple weak learners to create a strong learner.

2) Regularization:
XGBoost incorporates L1 and L2 regularization techniques to prevent overfitting.

3) Tree Pruning:
It is a technique called tree pruning to control the depth of decision trees.

4) ~~Parallel~~ Parallel Processing
XGBoost is highly optimized for efficiency and can leverage parallel processing, making it faster.

Q5)

| Aspect | PCA | LDA | SVA |
|---|---|---|---|
| Purpose: | Dimensionality reduction, Data compression. | Dimensionality reduction, maximize class separability. | Identify and adjust for hidden src of variation. |
| Application | EDA, data preprocessing | Pattern recognition, classification. | Genomics |
| Supervised/ Unsupervised | Unsupervised | Supervised/ Unsupervised | Typically unsupervised |
| Use in dimensionality reduction. | Yes | Yes | No. |

Q5

→

Dimensionality reduction techniques are used in machine learning and data analysis to reduce the number of input variables in a dataset.

1) Feature Selection:

Purpose:

Feature selection aim to select a subset of the most relevant features from the original dataset while discarding less important ones. The primary purpose is to improve the model's performance.

Example:

Filter Methods:

Methods like correlation, mutual information and chi-squared test

Wrapper Methods

Wrapper methods evaluate feature subsets by training and testing a model on different combination of features.

2) Feature Extraction:

Purpose:

feature extraction techniques transform the original high-dimensional data into a lower-dimensional representation by creating new features.

that capture most of the variance or information in the data.

Example:

Principal Component Analysis (PCA):

PCA is a linear dimensionality reduction technique that identifies orthogonal axes that capture the maximum variance in the data.

Linear Discriminant Analysis (LDA):

LDA is a supervised dimensionality reduction method that seeks to find linear combinations of features that maximize class separability.

Q9)

→ Boosting and bagging are two ensemble learning approaches that differ in how they assign weights to misclassified sample and update their base model.

Boosting:

Weight Assignment to Misclassified samples

- In boosting more weight is assigned to misclassified samples during each iteration.
- Misclassified samples are given higher importance in subsequent rounds to focus on correcting the errors made by earlier base learners.

Updating models:

- Boosting sequentially trains weak learners, where each subsequent learner tries to correct the mistakes of the previous ones.
- Base learners are added iteratively, and their predictions are weighted when combined.
- Initially, all samples have equal weights, but weights are adjusted after each iteration to give higher importance to misclassified samples.

Bagging

Weight Assignment to Misclassified Samples:

- In bagging, each base learner is trained on a random subset of data with replacement.
- There is no specific emphasis on misclassified sample, all of them are treated equally within each base learner.

Updating Models:

- Bagging trains each base models independently and in parallel.
- The final prediction is typically made by aggregating the predictions of all base models, such as averaging for regression or majority voting for classification.

Q.10)

→ Graph based clustering is a data analysis technique used to group data points into clusters based on their relationships within a graph structure. It is particularly useful when dealing with data that can be represented as a network or graph, where nodes represent individual data points and edges represent relationships or connection between those data points.

Graph representation:

Graph with nodes represent data points, and edges represent some measure of similarity

Graph construction:

Depending on the nature of the data points, the data points can either be in high dimensional space or low dimensional space

Graph clustering
Several clustering algorithms are there:
- Modularity based clustering
- Spectral Clustering
- Agglomerative clustering
- Hierarchical clustering

Cluster Assignment
Once the cluster is identified, each data point is assigned to one of the clusters.

Visualization and analysis

To understand the clustering results, visualization techniques like t-SNE and PCA can be used.

Q11)

→ K-Fold cross validation is a technique used in machine learning for model evaluation and validation. Its purpose is to assess a model's performance, especially its ability to generalize unseen data.

Steps:
1) Splitting the Data
2) Training and testing
3) Performance metrics
4) Aggregating results

Purpose:
- Robustness:

K-Fold cross validation helps assess a model's performance

across different subsets of data, reducing the risk of obtaining overly optimistic or pessimistic performance estimates due to the randomness in data.

Generalization Assessment:

It provides a more reliable estimate of how well a model is likely to perform on unseen data because it evaluates the model on multiple independent test sets.

Hyperparameter Tuning:

It is often used in combination with hyperparameter tuning to select the best model config.