MIDS & ANN Viva Questions with Answers

Q: What is data preprocessing? Why is it important?

A: Data preprocessing is the process of converting raw data into a clean and usable format for machine learning models. It involves handling missing values, noise, and inconsistencies. It is crucial because most real-world datasets are incomplete or inconsistent, which can negatively affect model performance.

Q: How do you handle missing data?

A: Missing data can be handled by:

- Imputation (mean, median, or mode)
- Deletion of rows or columns
- Prediction using algorithms
- Using models that handle missing data like XGBoost.

Q: What is feature scaling and why is it needed?

A: Feature scaling normalizes the range of independent variables. It's needed because many ML algorithms (like SVM, KNN, and K-means) are sensitive to the scale of data. Techniques include Min-Max normalization and Standardization (Z-score).

Q: What is document polarity? How is it different from sentiment?

A: Document polarity analysis detects if a document is positive, negative, or neutral. Sentiment analysis is a broader term that includes emotion analysis, tone, and context. Polarity focuses only on the direction of sentiment.

Q: Explain how Naive Bayes works for sentiment analysis.

A: Naive Bayes applies Bayes' theorem assuming independence among features. It calculates the probability of each class (positive/negative) given the words in the text and assigns the class with the highest probability.

Q: What is TF-IDF and how is it used in text classification?

A: TF-IDF (Term Frequency-Inverse Document Frequency) is a method to convert text into

numerical features. TF measures how often a word appears in a document. IDF measures how unique or rare a word is across all documents.

Q: What is the difference between regression and classification?

A: - Regression predicts continuous numeric values (e.g., house price).

- Classification assigns input data to a class (e.g., spam or not spam).

Q: Why do we use label encoding on the 'Sex' column?

A: Label encoding converts categorical variables like 'M', 'F', and 'I' into numerical form so that machine learning algorithms can process them.

Q: What are the evaluation metrics you used for regression?

A: Metrics include:

- Mean Absolute Error (MAE)

- Mean Squared Error (MSE)

- R^2 Score

- Explained Variance Score.

Q: Explain the working of K-means algorithm.

A: 1. Select k initial centroids

2. Assign each point to the nearest centroid

3. Recalculate centroids

4. Repeat until convergence.

Q: What is the role of activation functions in neural networks?

A: Activation functions introduce non-linearity into the network, allowing it to learn complex patterns. Without them, neural networks would behave like linear models.

Q: Difference between Sigmoid, ReLU, and Tanh?

A: - Sigmoid: Output range 0 to 1. Used in binary classification.

- Tanh: Output range -1 to 1. Zero-centered.

- ReLU: 0 for negative inputs, x for positive inputs.

Q: Why can't a single-layer perceptron solve XOR?

A: XOR is not linearly separable. A single-layer perceptron cannot create a nonlinear decision boundary. A multi-layer perceptron is required.

Q: Explain how backpropagation works.

A: Backpropagation is a training algorithm where:

1. Forward pass calculates outputs

2. Compute error

3. Gradients are calculated

4. Weights are updated.

Q: What is ART (Adaptive Resonance Theory) and how does it differ from other clustering methods?

A: ART allows incremental learning without forgetting old data using a vigilance parameter. Unlike K-means, ART adapts without retraining.

Q: What is the perceptron algorithm?

A: A simple binary classification algorithm that updates weights based on prediction error and adjusts the decision boundary iteratively.

Q: What does the decision boundary represent?

A: It is the line or surface that separates different classes. The model uses it to decide the class of a new data point.