Name : Ravi kumar yadav
E-mail : kumaryadavravi016@gmail.com
Assignment name : Data structure
Drive : drive
Github : Github

1 Assignment : Feature Engineering

1.0.1 1) What is a parameter?

1.0.2 In the context of machine learning, a parameter is a variable that is learned from the data during training. These parameters determine how the model makes predictions. For example, in linear regression, the slope and intercept of the line are the model's parameters

1.0.3 2) What is correlation?

1.0.4 Correlation is a statistical measure that describes the strength and direction of a relationship between two variables. It indicates how closely related the two variables are. A positive correlation means that as one variable increases, the other tends to increase, and vice versa.

1.0.5 3) What does negative correlation mean?

1.0.6 Negative correlation means that as one variable increases, the other tends to decrease. In other words, there is an inverse relationship between the two variables. For example, as the temperature increases, the amount of clothing worn may decrease, which shows a negative correlation.

1.0.7 4) Define Machine Learning. What are the main components in Machine Learning?

1.0.8 Machine learning is a subset of artificial intelligence that enables systems to learn from data, identify patterns, and make decisions with minimal human intervention. The main components of machine learning are:

(i) Data: The input for training the model.

(ii)Algorithms: The techniques that learn patterns from data.

(iii)Model: The learned output of the algorithm, which makes predictions based on new data.

1.0.9 5) How does loss value help in determining whether the model is good or not?

1.0.10 The loss value is a measure of how well the model's predictions match the actual results. A high loss indicates that the model is making large errors, while a low loss suggests that the model is performing well. By minimizing the loss value during training, we can improve the model's accuracy.

1.0.11 6) What are continuous and categorical variables?

1.0.12 (i) Continuous variables are numerical values that can take any value within a range. Examples include height, weight, or temperature.

1.0.13 (ii)Categorical variables represent discrete categories or groups. Examples include gender, color, or type of car

1.0.14 7) How do we handle categorical variables in Machine Learning? What are the common techniques?

1.0.15 Categorical variables are typically transformed into a numerical form using methods like:

(i) One-Hot Encoding: Creates a new binary column for each category.

(ii) Label Encoding: Assigns a unique integer value to each category.

1.0.16 8) What do you mean by training and testing a dataset?

1.0.17 (i) Training Dataset: The portion of the dataset used to train the model. The model learns patterns from this data.

1.0.18 (ii)Testing Dataset: The portion of the dataset used to evaluate the model's performance after training.

1.0.19 9) What is sklearn.preprocessing?

1.0.20 sklearn.preprocessing is a module in scikit-learn that provides functions to preprocess and transform data before applying machine learning algorithms. Examples include scaling, normalizing, encoding, and imputation techniques.

1.0.21 10) What is a Test set?

1.0.22 A test set is a portion of the dataset that is used to evaluate the performance of the machine learning model after it has been trained. It is separate from the training set and helps determine how well the model generalizes to new, unseen data.

1.0.23 11) How do we split data for model fitting (training and testing) in Python?

1.0.24 In Python, the train_test_split() function from sklearn.model_selection is commonly used to split data into training and testing sets. The data can be splited in any ratio but the training data will not be less than 50% ,because it affects our accuracy.

1.0.25 12) How do you approach a Machine Learning problem?

(i) Understanding the Problem: Define the problem and determine the type of model needed.

(ii) Data Collection and Preparation: Gather and clean the data.

(iii)Exploratory Data Analysis (EDA): Analyze the data to uncover patterns and relationships.

(iv) Model Selection: Choose a suitable machine learning algorithm.

(v) Model Training: Train the model on the training data.

(vi) Evaluation: Assess the model using the test set and performance metrics.

(vii)Optimization and Tuning: Fine-tune the model for better performance.

(viii)Deployment: Deploy the model for use in real-world scenarios.

1.0.26 13) Why do we have to perform EDA before fitting a model to the data?

1.0.27 Exploratory Data Analysis (EDA) helps in understanding the data, finding patterns, identifying outliers, and determining the right preprocessing steps. It can also reveal relationships between variables that are important for choosing the right model.

1.0.28 14) What is correlation?

1.0.29 15) What does negative correlation mean?

1.0.30 A negative correlation between two variables means that as one variable increases, the other decreases, and vice versa. In other words, there is an inverse relationship between the two variables.

1.0.31 16) How can you find correlation between variables in Python?

(i) Use .corr() for continuous data to get Pearson correlation.

(ii) Use sns.heatmap() to visualize correlations.

(iii)For categorical variables, consider using the Chi-Square test.

(iv) Use Spearman or Kendall correlation methods for non-linear or ordinal data.

1.0.32 17) What is causation? Explain difference between correlation and causation with an example.

1.0.33 (i) Causation refers to a direct cause-and-effect relationship between two variables. If A causes B, then changing A will directly affect B.

1.0.34 (ii)Correlation means that two variables are related, but it does not imply a cause-and-effect relationship. For example, ice cream sales and drowning incidents are correlated in summer, but one does not cause the other; the common factor is the warm weather.

1.0.35 18) What is an Optimizer? What are different types of optimizers? Explain each with an example.

1.0.36 An optimizer is an algorithm used to minimize the loss function in machine learning by adjusting the model's parameters. Some common types of optimizers are:

(i) Gradient Descent: Iteratively adjusts parameters to minimize the loss by moving in the direction of the negative gradient.

1.0.37 (ii) Stochastic Gradient Descent (SGD): A variation of gradient descent where parameters are updated after processing each data point, making it faster but noisier.

1.0.38 (iii)Adam (Adaptive Moment Estimation): Combines the benefits of both gradient descent and momentum-based methods

1.0.39 19) What is sklearn.linear_model ?

1.0.40 sklearn.linear_model is a module in scikit-learn that provides linear models like LinearRegression, LogisticRegression, and Ridge. These models are used for regression and classification tasks.

1.0.41 20) What does model.fit() do? What arguments must be given?

1.0.42 The fit() method trains the machine learning model on the provided data. It takes two main arguments:

1.0.43 X: The features of the dataset.

1.0.44 y: The target variable (labels).

1.0.45 21) What does model.predict() do? What arguments must be given?

1.0.46 The predict() method uses the trained model to make predictions on new data. It typically takes X (features) as the argument and returns the predicted labels or values.

1.0.47 22) What are continuous and categorical variables?

1.0.48 (i) Continuous Variables : These are variables that can take an infinite number of values within a given range. They can be measured on a scale and have a meaningful order, where you can perform arithmetic operations like addition, subtraction, etc

1.0.49 (ii) Categorical Variables : These variables represent categories or groups and cannot be measured on a numerical scale. They describe qualities or characteristics and are typically used to classify items into groups.

1.0.50 23) What is feature scaling? How does it help in Machine Learning?

1.0.51 Feature scaling involves normalizing or standardizing the values of features so they are on the same scale. This helps algorithms that are sensitive to the magnitude of values, like distance-based algorithms (e.g., KNN, SVM), to perform better.

1.0.52 24) How do we perform scaling in Python?

1.0.53 Scaling can be done using scikit-learn's StandardScaler or MinMaxScaler from the sklearn.preprocessing module.

1.0.54 25) What is sklearn.preprocessing?

1.0.55 sklearn.preprocessing is a module that provides various utilities for preprocessing data, such as scaling, encoding, and normalizing.

1.0.56 26) How do we split data for model fitting (training and testing) in Python?

1.0.57 This can be done using train_test_split() from sklearn.model_selection, as shown earlier.

1.0.58 27) Explain data encoding?

1.0.59 Data encoding is the process of converting categorical data into a numerical

[ ]: