Feature Engineering
Assignment Questions and Answers

1. What is a parameter?

A **parameter** is a measurable factor that defines a system or sets the conditions of its operation. In Machine Learning, parameters are the internal values learned from the training data, such as weights in linear regression.

2. What is correlation?

Correlation measures the strength and direction of a linear relationship between two variables.

What does negative correlation mean?

A **negative correlation** means that as one variable increases, the other decreases.

3. Define Machine Learning. What are the main components in Machine Learning?

Machine Learning (ML) is a field of AI that enables systems to learn patterns from data and make predictions. **Main components:** - Dataset - Features - Model - Loss function - Optimization algorithm

4. How does loss value help in determining whether the model is good or not?

A **loss value** measures how far the model's predictions are from actual values. Lower loss means a better-performing model.

5. What are continuous and categorical variables?

- **Continuous variables:** Numeric values with infinite possibilities (e.g., height, weight).
- **Categorical variables:** Non-numeric values representing categories (e.g., gender, city).

6. How do we handle categorical variables in Machine Learning? What are the common techniques?

Common techniques include: - **Label Encoding** - **One-Hot Encoding** - **Ordinal Encoding** - **Target Encoding**

7. What do you mean by training and testing a dataset?

- **Training dataset:** Used to teach the model.

- **Testing dataset:** Used to evaluate the model's performance.

8. What is sklearn.preprocessing?

`sklearn.preprocessin` is a module containing methods to scale, transform, and encode data before modeling.

9. What is a Test set?

A **test set** is a portion of the dataset used to evaluate the final model's accuracy after training.

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

Using scikit-learn:

```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

How do you approach a Machine Learning problem?

Steps: 1. Understand the problem 2. Collect and explore data 3. Preprocess data 4. Select a model 5. Train the model 6. Evaluate performance 7. Optimize 8. Deploy

11. Why do we have to perform EDA before fitting a model to the data?

EDA (Exploratory Data Analysis) helps understand patterns, detect missing values, identify outliers, and guide feature selection.

12. What is correlation?

Correlation is a statistical measure that shows how two variables are related (positive, negative, or zero).

13. What does negative correlation mean?

A negative correlation means as one variable rises, the other falls.

14. How can you find correlation between variables in Python?

```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt

corr = df.corr()
sns.heatmap(corr, annot=True)
plt.show()
```

15. What is causation? Explain difference between correlation and causation with an example.

Causation means one event directly affects another. - **Correlation:** Two variables move together. - **Causation:** One variable *causes* the change. Example: Ice cream sales and drowning deaths are correlated but ice cream does **not** cause drowning.

16. What is an Optimizer? What are different types of optimizers? Explain each with an example.

An **optimizer** adjusts model parameters to minimize loss. Common optimizers: - **SGD (Stochastic Gradient Descent):** Updates weights after each sample. - **Adam:** Combines momentum + RMSProp, adaptive learning rate. - **RMSProp:** Uses moving average of squared gradients to adjust learning rate.

Example:

```
from tensorflow.keras.optimizers import Adam
model.compile(optimizer=Adam(learning_rate=0.001), loss='mse')
```

17. What is sklearn.linear_model?

It is a module in Scikit-Learn that contains linear models such as LinearRegression, LogisticRegression, Ridge, Lasso, etc., used for regression and classification.

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

`model.fit()` trains the machine learning model using the training data. - Required arguments: **X (features)** and **y (target)**.

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

`model.predict()` uses the trained model to make predictions. - Required argument: **X_test (input features)**.

20. What are continuous and categorical variables?

- **Continuous variables:** Numeric values that can take any value (e.g., height, temperature).
- **Categorical variables:** Represent categories or labels (e.g., gender, color, city).

21. What is feature scaling? How does it help in Machine Learning?

Feature scaling transforms data so all features are on a similar scale. It helps by: - Speeding up training - Improving accuracy - Preventing dominance of large-valued features

22. How do we perform scaling in Python?

Using StandardScaler:

```
from sklearn.preprocessing import StandardScaler
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)
```

23. What is sklearn.preprocessing?

A module in Scikit-Learn that provides tools for: - Scaling - Normalization - Encoding - Imputation

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

Using train_test_split:

```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,
random_state=42)
```

25. Explain data encoding.

Data encoding converts categorical data into numeric form so that ML models can process it. Common types: - **Label Encoding**: Converts categories to numbers. - **One-Hot Encoding**: Creates binary columns for each category.