Assignment Code: D-AG-008

Supervised Learning: Regression Models and Performance Metrics | **Solution**

Instructions: Carefully read each question. Use Google Docs, Microsoft Word, or a similar tool to create a document where you type out each question along with its answer. Save the document as a PDF, and then upload it to the LMS. Please do not zip or archive the files before uploading them. Each question carries 20 marks.

Total Marks: 200

Question 1 : What is Simple Linear Regression (SLR)? Explain its purpose.

Answer:

Simple Linear Regression (SLR) is a statistical technique used to examine the relationship between one independent variable (X) and one dependent variable (Y).

Purpose:

- To predict the value of Y based on X.
- To understand how X influences Y.
- To determine the best-fitting straight line (Y = a + bX) that explains the relationship between them.

Question 2: What are the key assumptions of Simple Linear Regression?

Answer:

Key Assumptions of Simple Linear Regression:

1. Linearity: The relationship between X and Y is linear.
2. Independence: The observations are independent of each other.
3. Homoscedasticity: The variance of errors is constant across all values of X.
4. Normality: The residuals (errors) are normally distributed.
5. No Multicollinearity: Since SLR has only one independent variable, this is automatically satisfied.

Question 3: Write the mathematical equation for a simple linear regression model and explain each term.

Answer:

Mathematical Equation:

$$Y=a+bX+e$$

Explanation of Terms:

- **Y**: Dependent variable (the variable to be predicted).
- X: Independent variable (the predictor).
- a: Intercept (the value of Y when X = 0).
- b: Slope (the change in Y for a one-unit change in X).
- e: Error term (the difference between the actual and predicted value of Y).

Question 4: Provide a real-world example where simple linear regression can be applied.

Answer:

Example:
Simple Linear Regression can be used to predict a student's exam score (Y) based on the number of study hours (X).

As study hours increase, exam scores are expected to increase linearly, making it a suitable case for applying SLR.

Question 5: What is the method of least squares in linear regression?

Answer:

Method of Least Squares:
It is a technique used to find the best-fitting line in linear regression by minimizing the sum of the squared differences between the actual values and the predicted values of Y.

In simple terms, it finds the line for which the total error (squared) is the smallest.

Question 6: What is Logistic Regression? How does it differ from Linear Regression?

Answer:

Logistic Regression:
It is a statistical method used to predict a categorical (usually binary) outcome, such as Yes/No or 0/1, based on one or more independent variables.

Difference from Linear Regression:

1. Output Type:
    - Linear Regression predicts continuous values.
    - Logistic Regression predicts categorical values (probabilities between 0 and 1).
2. Equation Used:
    - Linear uses a straight-line equation ($Y = a + bX$).
    - Logistic uses a sigmoid (S-shaped) curve.
3. Purpose:
    - Linear: For prediction of quantities.
    - Logistic: For classification tasks.

Question 7: Name and briefly describe three common evaluation metrics for regression models.

Answer:

Common Evaluation Metrics for Regression Models:

1. Mean Absolute Error (MAE):
   Measures the average of absolute differences between actual and predicted values.
2. Mean Squared Error (MSE):
   Calculates the average of squared differences between actual and predicted values (gives more weight to larger errors).
3. R-squared (R^2):
   Indicates how well the model explains the variation in the dependent variable; higher R^2 means a better fit.

Question 8: What is the purpose of the R-squared metric in regression analysis?

Answer:

Purpose of R-squared:
R-squared measures how well the independent variable(s) explain the variation in the dependent variable.

It shows the goodness of fit of the model — how closely the predicted values match the actual data.
Higher R-squared values indicate a better-fitting model.

Question 9: Write Python code to fit a simple linear regression model using scikit-learn and print the slope and intercept.
(Include your Python code and output in the code box below.)

Answer:

```
from sklearn.linear_model import LinearRegression
import numpy as np

X = np.array([1, 2, 3, 4, 5]).reshape(-1, 1)
y = np.array([2, 4, 5, 4, 5])

model = LinearRegression()
model.fit(X, y)

print("Slope (b):", model.coef_[0])
print("Intercept (a):", model.intercept_)

# OUTPUT

Slope (b): 0.6
Intercept (a): 2.2
```

Question 10: How do you interpret the coefficients in a simple linear regression model?

Answer:

Interpretation of Coefficients:

- Intercept (a): It represents the predicted value of Y when X = 0.
- Slope (b): It shows how much Y changes for a one-unit increase in X.

  If b is positive, Y increases as X increases.

  If b is negative, Y decreases as X increases.