BUSINESS INTELLIGENCE AND ANALYTICS

ASSIGNMENT WEEK 5:

Total marks = (15 Qns * 1 mark = 15 marks)

1. In regression analysis, multicollinearity refers to: (1 Mark)
   a) The perfect linear relationship between the dependent and independent variables.
   b) The presence of outliers in the dataset that affect the regression coefficients.
   c) High intercorrelation among the independent variables, leading to unstable estimates of the regression coefficients.
   d) The variance in the residuals of the regression model.

   Answer: C) High intercorrelation among the independent variables, leading to unstable estimates of the regression coefficients.

2. What type of data transformation technique scales data to a specific range, such as 0 to 1? (1 Mark)

   a) Database normalization
   b) Aggregation
   c) Smoothing techniques
   d) Standardization/Normalization

   ANS: d) Standardization/Normalization

3. Which of the following statements about the coefficient of determination (R-squared) is true? (1 Mark)
   a) A higher R-squared value always indicates a lower model performance.
   b) A higher R-squared value always indicates better model performance, regardless of the number of predictor variables.
   c) R-squared ranges from 0 to 1 and represents the percentage of variation in the dependent variable explained by the independent variables.
   d) R-squared can only take positive values and is unaffected by the presence of multicollinearity in the regression model.

   Answer: C) R-squared ranges from 0 to 1 and represents the percentage of variation in the dependent variable explained by the independent variables.

4. What does Ordinary Least Squares (OLS) aim to minimize in the context of linear regression? (1 Mark)
   a) The sum of squared errors between the predicted and observed values of the dependent variable.
   b) The sum of squared residuals between the predicted and observed values of the independent variable.
   c) The total variance of the independent variables.
   d) The sum of squared errors between the predicted and observed values of the independent variable.

   Answer: A) The sum of squared errors between the predicted and observed values of the dependent variable.

5. The coefficient of determination (R-squared) value of 0.98 in a regression model implies: (1 Mark)

   A) The model has a high level of multicollinearity.

   B) 98% of the variability in the dependent variable is explained by the independent variable.

   C) The regression model is overfitting the data by 98 %.

   D) The residuals in the model are normally distributed with z value of 0.98

   Answer: B) 98% of the variability in the dependent variable is explained by the independent variable.

6. Prediction error in a model refers to: (1 Mark)
   a) The difference between actual and predicted values.
   b) The degree of overfitting in the model.
   c) The number of features used in the model.
   d) The variability of the target variable.

   Solution:A) The difference between actual and predicted values.

7. Which of the following statements is wrong with regards to Overfitting in a machine learning model? (1 Mark)
   a) The model is too simple to capture the underlying patterns in the data.
   b) The model performs well on training data but poorly on unseen data.
   c) The model fits the noise in the training data.
   d) None of the above

   Solution: A) The model is too simple to capture the underlying patterns in the data.

8. Underfitting in a machine learning model results in: (1 Mark)
   a) Low bias and high variance.
   b) High bias and low variance.
   c) High bias and high variance.
   d) Low bias and low variance.

   Solution: B) High bias and low variance.

9. When should one focus on reducing bias in a machine learning model? (1 Mark)
   a) When the model performs well on the training data but poorly on test data.
   b) When the model shows high variability in predictions.
   c) When the model consistently overfits the training data.
   d) When the model doesn’t fit the data well, and works poorly in explanatory/predictive performance

   Solution: D) When the model doesn’t fit the data well, and works poorly in explanatory/predictive performance

10. What is the bias-variance trade-off in machine learning? (1 Mark)
    a) Balancing the computational resources used in training with model accuracy.
    b) Aiming to minimize the difference between predicted and actual values in a model.

c) Finding the equilibrium between model complexity and its ability to generalize to unseen data.
d) Choosing the best algorithm that minimizes both bias and variance simultaneously.

Solution: C- Finding the equilibrium between model complexity and its ability to generalize to unseen data.

11. Training error refers to: (1 Mark)

a) Error calculated on the training dataset.
b) Error due to overfitting.
c) Error calculated on the testing dataset.
d) Error due to underfitting.

Solution: A) Error calculated on the training dataset.

12. What does Leave-One-Out Cross-Validation (LOOCV) do? (1 Mark)

a) It iteratively uses all but one sample as the test set and the remaining sample as the training set.
b) It divides the dataset into k subsets and uses each subset as the testing set in turn.
c) It creates a validation set from a small portion of the data.
d) It iteratively uses all but one sample as the training set and the remaining sample as the testing set.

Solution: A) It iteratively uses all but one sample as the test set and the remaining sample as the training set.

13. What is the primary purpose of cross-validation in machine learning? (1 Mark)

a) To fit the model to the training data efficiently.
b) To evaluate the model's performance on unseen data.
c) To increase model complexity for better predictions.
d) To reduce the number of features in the dataset.

Solution: B) To evaluate the model's performance on unseen data.

14 What are the three sources of error in predicted Y in machine learning? (1 Mark)

a) Measurement error, data preprocessing error, and feature selection error.
b) Model complexity error, parameter tuning error, and overfitting error.
c) Reducible error due to inaccurate estimation of f, irreducible error due to randomness, and test data variation.
d) Training error, validation error, and testing error.

Solution: C) Reducible error due to inaccurate estimation of f, irreducible error due to randomness, and test data variation.

15. Which of the following statements most accurately distinguishes supervised learning from unsupervised learning in machine learning? (1 Mark)

a) Supervised learning requires labelled data for training models to predict specific outcomes, while unsupervised learning uncovers patterns or structures in data without predefined outcomes.
b) Supervised learning primarily deals with clustering data points based on similarities, while unsupervised learning focuses on predicting future trends based on historical data.
c) Supervised learning utilizes human supervision to label data for analysis, while unsupervised learning relies on algorithms to classify data into distinct categories.
d) Supervised learning involves training models without any prior knowledge of the dataset, while unsupervised learning requires prior information about the characteristics of the data.

Solution: A) Supervised learning requires labelled data for training models to predict specific outcomes, while unsupervised learning uncovers patterns or structures in data without predefined outcomes.