Worksheet Set 1 Solution

Name of Student: Mudassar Iqbal Shakil Ahmed Sayyed

Name of assignment: Machine Learning & Statistics

Internship Batch: DS2404

Machine Learning Solution

1. A) Least Square Error

2. A) Linear regression is sensitive to outliers

3. B) Negative

4. B) Correlation

5. C) Low bias and high variance

6. B) Predictive model

7. D) Regularization

8. D) SMOTE

9. A) TPR and FPR

10. B) False

11. B) Apply PCA to project high dimensional data

12. B) It becomes slow when number of features is very large.

    C) We need to iterate.

    D) It does not make use of dependent variable.

13. Regularization is a technique used in machine learning to prevent overfitting by adding a penalty term to the loss function, which penalizes large coefficients.

14. Some algorithms used for regularization include Ridge regression, Lasso regression, and Elastic Net.

15. The error present in a linear regression equation refers to the difference between the actual values and the predicted values of the label by the regression model.

Statistics Solution

1. a) True

2. a) Central Limit Theorem

3. b) Modeling bounded count data

4. c) The square of a standard normal random variable follows what is called chi-squared distribution

5. c) Poisson

6. b) False

7. b) Hypothesis

8. a) 0

9. c) Outliers cannot conform to the regression relationship

10. Normal Distribution is a probability distribution that is symmetric around the mean, forming a bell-shaped curve. It is characterized by its mean and standard deviation, and many natural phenomena follow this distribution.

11. Handling missing data involves various imputation techniques such as mean imputation, median imputation, mode imputation, or advanced methods like K-nearest neighbors imputation.

12. A/B testing is a statistical method used to compare two or more versions of a product or service to determine which one performs better. It is commonly used in marketing, web development, and other fields to optimize decision-making.

13. Mean imputation of missing data can introduce bias and underestimate variability, so it may not always be an acceptable practice, I only prefer mean imputation when the data is continuous, dataset is large and missing values are less.

14. Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables by fitting a linear equation to the observed data.

15. Various branches of statistics include descriptive statistics, inferential statistics, probability theory, Bayesian statistics, multivariate statistics, and nonparametric statistics. Each branch deals with different aspects of data analysis and inference.

Python Solution

1. C) %

2. B) 0

3. C) 24

4. D) 0

5. D) 6

6. C) the finally block will be executed no matter if the try block raises an error or not.

7. A) It is used to raise an exception.

8. C) in defining a generator

9. A) _abc

10. A) yield

    B) raise

I am writing the solution to python codes in a separate jupyter notebook titled “Worksheet Set 1 Solution - Mudassar Sayyed (DS2404).ipynb” uploaded in same folder named “Worksheet_set_1”.