MACHINE LEARNING – Assignment 1

Q1- Which of the following methods do we use to find the best fit line for data in Linear Regression?

A) Least Square Error
B) Maximum Likelihood
C) Logarithmic Loss
D) Both A and B

Answer - Option A

Q2 – Which of the following statement is true about outliers in Linear Regression?

A) Linear Regression is sensitive to outliers
B) Linear Regression is not sensitive to outliers
C) Can't say
D) None of these

Answer – Option A

Q3 - A line falls from left to right if a slope is _____?

A) Positive
B) Negative
C) Zero
D) Undefined

Answer – Option B

Q4- Which of the following have symmetric relation between dependent variable and independent variable?

A) Regression
B) Correlation
C) Both of them
D) None of these

Answer – Option B

Q5 – Which of the following is the reason for overfitting condition?

A) High Bias and High Variance
B) Low Bias and Low Variance
C) Low Bias and High Variance
D) None of these

Answer – Option C

Q6 – If Output involves label then that model is called as:

A) Descriptive model
B) Predictive model
C) Reinforcement learning
D) All of the above

Answer- Option B

Q7 – Lasso and Ridge Regression techniques belong to _____?

A) Cross Validation
B) Removing outliers
C) SMOTE
D) Regularization

Answer – Option D

Q8 – To overcome with imbalance dataset which technique can be used?

A) Cross Validation
B) Regularization
C) Kernel
D) SMOTE

Answer - D

Q9 – The AUC Receiver Operator Characteristics (AUCROC) curve is an evaluation metric for binary classification problems. It uses _____ to make graph?

A) TPR and FPR
B) Sensitivity and Precision
C) Sensitivity and Specificity
D) Recall and Precision

Answer – A

Q10 – In AUC Receiver Operator Characteristics (AUCROC) curve for the better model area under the curve should be less

A) True
B) False

Answer- Option B

Q11 – Pick the feature extraction from below:

A) – Construction bag of words from a email
B) Apply PCA to project high dimensional data
C) Removing Stop words
D) Forward selection

Answer – option B

Q12 – Which of the following is true about Normal Equation used to compute the coefficients of the Linear Regression?

A) We don't have to choose the learning rate
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

Answer – A & B

Q13. Explain the term regularization?

Answer -

1- When we train our data by using linear regression then there is a good chance that overfitting will occur, and it will reduce the accuracy of our model, to overcome with this problem we use regularization and it will increase the accuracy of our model also.

2- Simply, by reducing the number of degrees of a polynomial function, regularization can be done, because in linear Equation, we don't want huge coefficients as a small change in coefficients can make a large difference for the dependent variables (y). So, Regularization constraint the huge coefficient within their limit to avoid overfitting.

3- To regularize the model, a shrinkage penalty is added to the cost function.

4- There are three types of regularization

a) – LASSO Regression (L1 Form)

b)- RIDGE Regression (L2 Form)

c)– ELASTICNET

Q14. Which particular algorithms are used for regularization?

Answer- Three algorithms are used for regularization

a) LASSO Regression (L1 Form)
LASSO Regression penalizes the model based on the sum of magnitude of the coefficients.
Regularization = shrinkage factor * Summation| Beta |

b) RIDGE Regression (L2 Form)
RIDGE Regression penalizes the model based on the sum of the square of the magnitude of the coefficients.
Regularization = shrinkage factor * Summation| Square of Beta |

c) ELASTICNET
It combines the advantage of both (Lasso & Ridge).

Q15. Explain the term error present in linear regression equation?

Answer –

1) The term error present in Linear Regression equation is the difference between the Predicted Value and the Actual Value.
2) It often uses MSE (Mean Squared Error) to calculate the error of present in the model.
3) The more the errors, the less of our models accuracy.
4) We use Gradient Dissent method, the main goal of Gradient Dissent is to minimize the cost function, or we can say that it minimise our loss function or error term.
5) To improve the accuracy and to reduce this error, we need to clean our data while cleaning our data, we need to note down proper documentation like
   1- missing values
   2- Outliers,
   3- check whether the Columns/ features are Linear or Non-Linear,
   4- make columns/features curve to proper normalized distribution,
   5- feature scaling is essential to perform,

PYTHON – WORKSHEET -1

Q1. Which of the following operators is used to calculate remainder in a division?
A) #
B) &
C) %
D) $
Answer - Option C

Q2. In python 2//3 is equal to?

A) 0.666
B) 0
C) 1
D) 0.67
Answer – Option B

Q3. In python, 6<<2 is equal to?

A) 36
B) 10
C) 24
D) 45
Answer- Option A

Q4. In python, 6&2 will give which of the following as output?
A) 2
B) True
C) False
D) 0
Answer- A

Q5. In python, 6|2 will give which of the following as output?
A) 2
B) 4
C) 0
D) 6
Answer- D

Q6. What does the finally keyword denotes in python?

A) It is used to mark the end of the code

B) It encloses the lines of code which will be executed if any error occurs while executing the lines of code in the try block.

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

D) None of the above
Answer – Option C

Q7. What does raise keyword is used for in python?

A) It is used to raise an exception.
B) It is used to define lambda function
C) it's not a keyword in python.
D) None of the above

Answer - Option A

Q8. Which of the following is a common use case of yield keyword in python?
A) in defining an iterator
B) while defining a lambda function
C) in defining a generator
D) in for loop.

Answer- Option C

Q9 and Q10 have multiple correct answers. Choose all the correct options to answer your question.

Q9. Which of the following are the valid variable names?
A) _abc
B) 1abc
C) abc2
D) None of the above

Answer- Option is A and C

Q10. Which of the following are the keywords in python?
A) yield
B) raise
C) look-in
D) all of the above

Answer- Option A and B

STATISTICS– WORKSHEET -1

1. Bernoulli random variables take (only) the values 1 and 0.
a) True
b) False
Answer - Option A

2. Which of the following theorem states that the distribution of averages of iid variables, properly normalized, becomes that of a standard normal as the sample size increases?
a) Central Limit Theorem
b) Central Mean Theorem
c) Centroid Limit Theorem
d) All of the mentioned
Answer - Option A

3. Which of the following is incorrect with respect to use of Poisson distribution?
a) Modeling event/time data
b) Modeling bounded count data
c) Modeling contingency tables
d) All of the mentioned
Answer - Option B

4. Point out the correct statement.
a) The exponent of a normally distributed random variables follows what is called the log- normal distribution
b) Sums of normally distributed random variables are again normally distributed even if the variables are dependent
c) The square of a standard normal random variable follows what is called chi-squared distribution
d) All of the mentioned
Answer - Option A

5. _____ random variables are used to model rates.
a) Empirical
b) Binomial
c) Poisson
d) All of the mentioned
Answer - Option C

6. Usually replacing the standard error by its estimated value does change the CLT?
a) True
b) False
Answer - Option B

7. Which of the following testing is concerned with making decisions using data?
a) Probability
b) Hypothesis
c) Causal
d) None of the mentioned
Answer - Option B

8. Normalized data are centered at_____and have units equal to standard deviations of the original data.
a) 0
b) 5
c) 1
d) 10
Answer - Option A

9. Which of the following statement is incorrect with respect to outliers?
a) Outliers can have varying degrees of influence
b) Outliers can be the result of spurious or real processes
c) Outliers cannot conform to the regression relationship
d) None of the mentioned
Answer - Option C

10. What do you understand by the term Normal Distribution?
Answer

- Normal Distribution is also known as 'Bell Curve" because it looks like a bell.
- Normal Distribution is also known as Gaussian distribution.
- And it is a most common distribution function for independent, randomly generated variables
- In this distribution, the value of Mean is equals to Median and they both are equals to Mode.
- It means, Mean = Median = Mode
- It has symmetry about the centre.
- 50% of the value is less than the mean and 50% of the value is greater than mean.
- The graph of the normal distribution is characterised by two parameters:
    - Mean or Average
    - Standard deviation
- **Mean** – it is the maximum of the graph and about which the graph is always symmetric.
- **Standard deviation** – it determines the amount of dispersion away from the mean.
- A small standard deviation as compared to mean produces a steep graph
- A large standard deviation as compared to mean produces a flat graph.

11. How do you handle missing data? What imputation techniques do you recommend?
Answer

To handle missing value is very crucial thing. We can handle missing values by many techniques:

- **Deleting rows with missing values** but it is not feasible because we may lose some of our important data which may reduce our model accuracy.
- **Deleting column with missing values** but it is also not feasible because some columns are related to other columns or features or it may relate to our target value also.
- **Mean / Median / Mode imputation**, this is very simple and easiest to implement.
- **Deductive imputation** – it is accurate but It cannot be applied to all datasets and it consumes more time and might require specific coding.
- **Regression imputation** – this approach replaces missing values with a predicted value based on a regression line.
- **Using algorithm which supports missing values,** for example KNN, it is a machine learning algorithm which works on the principle of distance measure. This algorithm is used when there are nulls present in the datasets.
- **Predicting the missing value,** using that feature do not contain missing values In that way we can predict nulls by using machine learning algorithm.

12. What is A/B testing?

Answer

- It is used to compare two versions of a variable to find out which performs better in a controlled environment.
- It is nothing but a most widely used statistical tools.
- It is used to compare the two versions of a variable to find out which performs better.
- For making decisions, it estimates population parameters based on sample statistics.
- Here population is Total population and the sample is number of people that participated in the test.
- It also includes statistical hypothesis testing.

13. Is mean imputation of missing data acceptable practice?

Answer

No, it is not generally acceptable, Although, it is a easy way to handle missing values.
It is not acceptable because it has some drawbacks.

- It does not preserve the relationships among variables.
- It decreases the variance of our data while increasing the bias. As a result, it will decrease the accuracy of our model which is not acceptable.
- It shrinks standard errors, which invalidates most hypothesis tests and the calculation of confidence intervals.

So, Mean imputation should be avoided because it has some serious drawbacks, there are many methods which we can adopt to remove missing values

14. What is linear regression in statistics?

Answer

- It is basic and most commonly used type of predictive analysis.
- These regression estimates are used to explain the relationship between one dependent variable and one or more than independent variables.
- The simplest form of one dependent variable and one or more independent variable is **Y = b*x + c**
- Where y = estimated dependent variable, x = estimated independent variable, b = regression coefficient, c = constant
- Dependent variable is also called Outcome variable, Endogenous variable, Criterion variable or Regressand.
- Independent variables are also called exogenous variables, predictors variables or Regressors.
- There are three major uses for regression analysis are:
    1) Determining the strength of the predictors
    2) Forecasting an effect
    3) Trend forecasting

There are many types of linear regression

- Simple linear regression
- Multiple linear regression
- Logistic regression
- Ordinal regression
- Multinomial regression
- Discriminant analysis

15. What are the various branches of statistics?

Answer

There are two main branches of statistics

1) Descriptive statistics
2) Inferential statistics

Descriptive statistics

- The data is described in a summarized way.
- The summarization is done from the sample of the population using different parameters like mean or standard deviation.
- It is used when we don't have hypothesis in our research.
- It can be categorized into
    1) Measures of central tendency
    2) Measures of variability

Inferential statistics

- It is used to gathered information from a sample to make inferences, decisions, or predictions about a given population.
- It often talks about probability terms by using descriptive statistics.
- These techniques are majorly used by statisticians to analyse data, make estimates and draw conclusions from the limited information which is obtained by sampling and testing data.
- There are different types of calculations in inferential statistics are:
    1) Regression analysis
    2) Analysis of variance
    3) Analysis of covariance
    4) Statistical significance
    5) Correlation analysis