MACHINE LEARNING

In Q1 to Q11, only one option is correct, choose the correct option:

1. 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

Ans-> A

2. 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

Ans-> A

3. A line falls from left to right if a slope is ______?

A) Positive B) Negative

C) Zero D) Undefined

Ans-> B

4. Which of the following will have symmetric relation between dependent variable and independent variable?

A) Regression B) Correlation

C) Both of them D) None of these

Ans-> D

5. Which of the following is the reason for over fitting condition?

A) High bias and high variance B) Low bias and low variance

C) Low bias and high variance D) none of these

Ans-> C

6. If output involves label then that model is called as:

A) Descriptive model B) Predictive modal

C) Reinforcement learning D) All of the above

Ans-> B

7. Lasso and Ridge regression techniques belong to _________?

A) Cross validation B) Removing outliers

C) SMOTE D) Regularization

Ans-> D

8. To overcome with imbalance dataset which technique can be used?

A) Cross validation B) Regularization

C) Kernel D) SMOTE

Ans-> D

9. The AUC Receiver Operator Characteristic (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

Ans-> A

10. In AUC Receiver Operator Characteristic (AUCROC) curve for the better model area under the curve should be less.

A) True B) False

Ans-> B

11. 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

Ans-> A

In Q12, more than one options are correct, choose all the correct options:

12. Which of the following is true about Normal Equation used to compute the coefficient 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.

Ans-> B

Q13 and Q15 are subjective answer type questions, Answer them briefly.

13. Explain the term regularization?

Ans-> Overfitting occurs when a machine learning model learns the specific details of the training data too well, leading to poor performance on new, unseen data. Regularization techniques aim to address this issue by introducing additional constraints or penalties during the training process to prevent the model from overfitting.

There are several types of regularization techniques, but two commonly used methods are L1 regularization (Lasso) and L2 regularization (Ridge).

L1 regularization, adds a penalty proportional to the absolute value of the coefficients. This penalty encourages the coefficients to be sparse, meaning many of them will be exactly zero. As a result, L1 regularization can lead to feature selection, where only a few important features are selected for the model.

L2 regularization, on the other hand, adds a penalty proportional to the squared value of the coefficients. This penalty encourages the coefficients to be small but not necessarily zero. L2 regularization tends to distribute the weights more evenly across the features, preventing any single feature from dominating the model.

In summary, regularization techniques like L1 and L2 regularization help mitigate overfitting by penalizing large coefficient values, leading to more generalized and robust models.

14. Which particular algorithms are used for regularization?

Ans-> Regularization is a technique used to reduce overfitting and improve the generalization of machine learning models. There are three commonly used regularization algorithms:

L1 regularization (lasso regression):-

- Adds the absolute value of the model's coefficients as a penalty term to the loss function.

- Tends to shrink some coefficients to zero, resulting in a sparse model that selects only the most important features.

L2 regularization (ridge regression):-

- Adds the squared value of the model's coefficients as a penalty term to the loss function.

- Tends to shrink all coefficients by the same amount, resulting in a more stable model that avoids large coefficients.

Elastic Net:-

- Combines both L1 and L2 regularization terms, allowing for a balance between feature selection and coefficient shrinkage.

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

Ans-> In linear regression, the error term represents the unexplained variation in the dependent variable that is not accounted for by the independent variables included in the model. It is the difference between the predicted value of the dependent variable and the actual observed value.

The error term is crucial in linear regression because it allows us to assess the accuracy and reliability of our model. A smaller error term indicates a better fit of the model to the data, while a larger error term suggests that there is more unexplained variation in the dependent variable.

The error term is also essential for statistical inference in linear regression. It is used to calculate the standard errors of the regression coefficients, which are used to test the significance of the independent variables and construct confidence intervals for the predicted values.

In summary, the error term in linear regression represents the unexplained variation in the dependent variable and plays a critical role in assessing the accuracy, reliability, and statistical validity of the regression model.

PYTHON – WORKSHEET 1

Q1 to Q8 have only one correct answer. Choose the correct option to answer your question.

1. Which of the following operators is used to calculate remainder in a division?

A) # B) & C) % D) $

Ans-> C

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

A) 0.666 B) 0 C) 1 D) 0.67

Ans-> B

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

A) 36 B) 10 C) 24 D) 45

Ans-> C

4. In python, 6&2 will give which of the following as output?

A) 2 B) True C) False D) 0

Ans-> A

5. In python, 6|2 will give which of the following as output?

A) 2 B) 4 C) 0 D) 6

Ans-> D

6. 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

Ans-> C

7. 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

Ans-> A

8. 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.

Ans-> C

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

9. Which of the following are the valid variable names?

A) _abc B) 1abc C) abc2 D) None of the above

Ans-> A & C

10. Which of the following are the keywords in python?

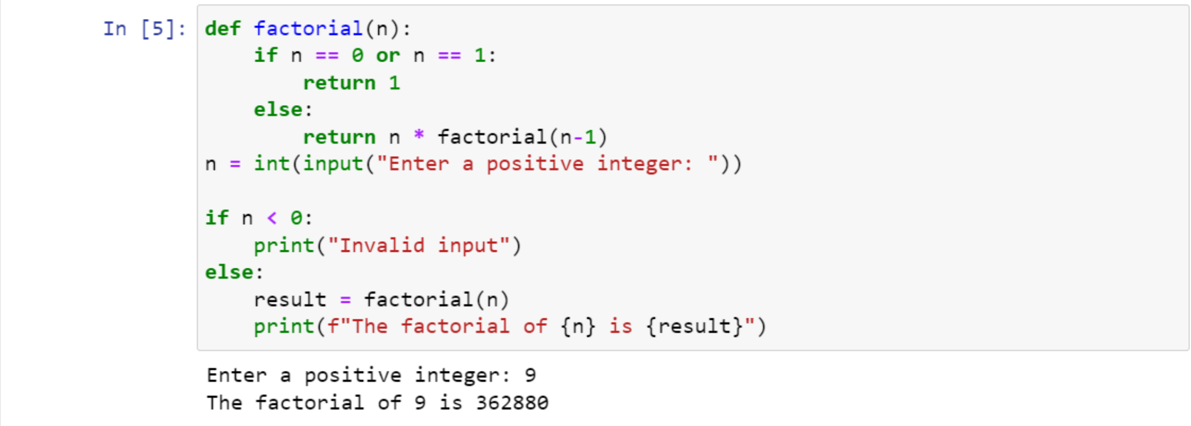
A) yield B) raise C) look-in D) all of the above

Ans-> A & B

Q11 to Q15 are programming questions. Answer them in Jupyter Notebook.

11. Write a python program to find the factorial of a number.

Ans->



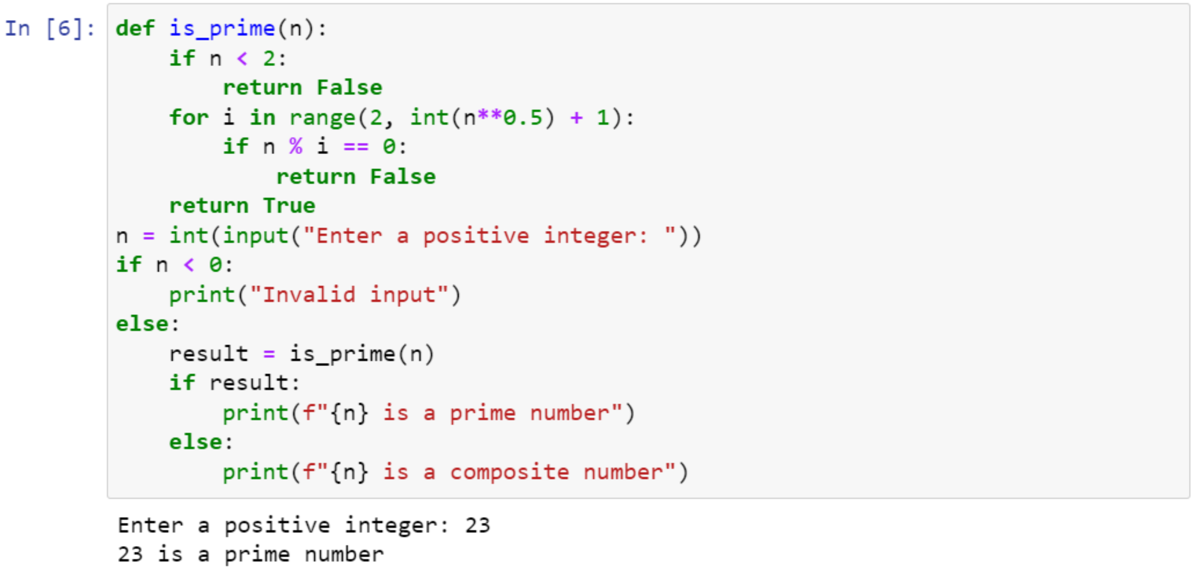
```
In [5]: def factorial(n):
            if n == 0 or n == 1:
                return 1
            else:
                return n * factorial(n-1)
        n = int(input("Enter a positive integer: "))

        if n < 0:
            print("Invalid input")
        else:
            result = factorial(n)
            print(f"The factorial of {n} is {result}")
```

```
Enter a positive integer: 9
The factorial of 9 is 362880
```

12. Write a python program to find whether a number is prime or composite.

Ans->

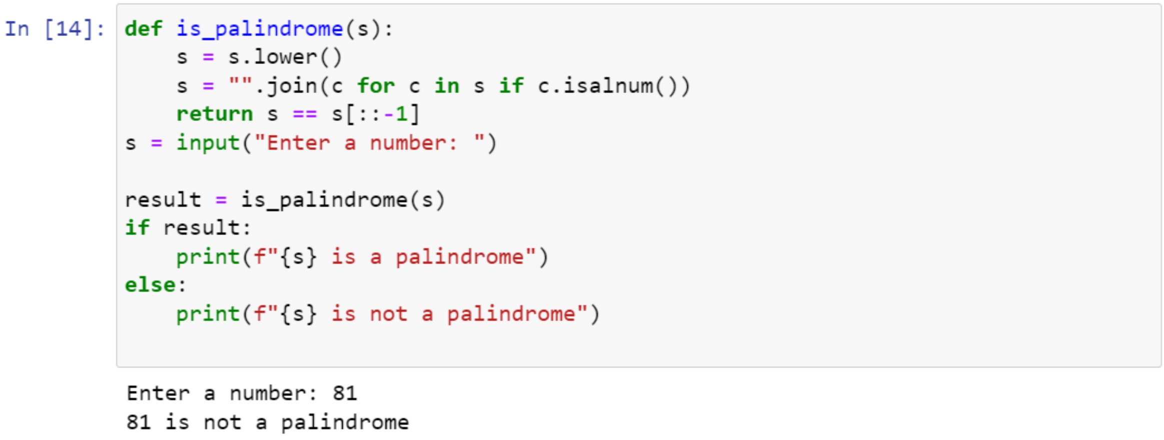


```
In [6]: def is_prime(n):
            if n < 2:
                return False
            for i in range(2, int(n**0.5) + 1):
                if n % i == 0:
                    return False
            return True
        n = int(input("Enter a positive integer: "))
        if n < 0:
            print("Invalid input")
        else:
            result = is_prime(n)
            if result:
                print(f"{n} is a prime number")
            else:
                print(f"{n} is a composite number")
```

```
Enter a positive integer: 23
23 is a prime number
```

13. Write a python program to check whether a given string is palindrome or not.

Ans->



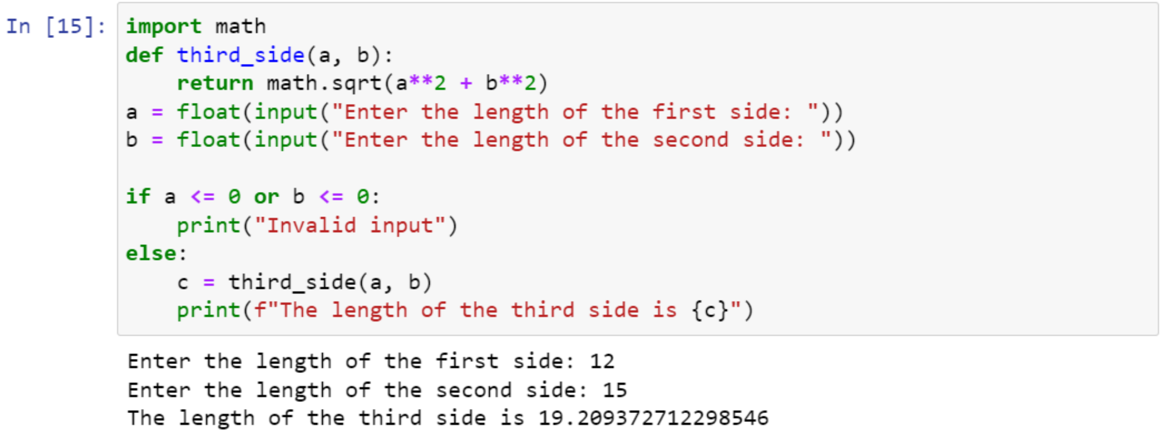
```
In [14]: def is_palindrome(s):
             s = s.lower()
             s = "".join(c for c in s if c.isalnum())
             return s == s[::-1]
         s = input("Enter a number: ")

         result = is_palindrome(s)
         if result:
             print(f"{s} is a palindrome")
         else:
             print(f"{s} is not a palindrome")
```

```
Enter a number: 81
81 is not a palindrome
```

14. Write a Python program to get the third side of right-angled triangle from two given sides.

Ans->



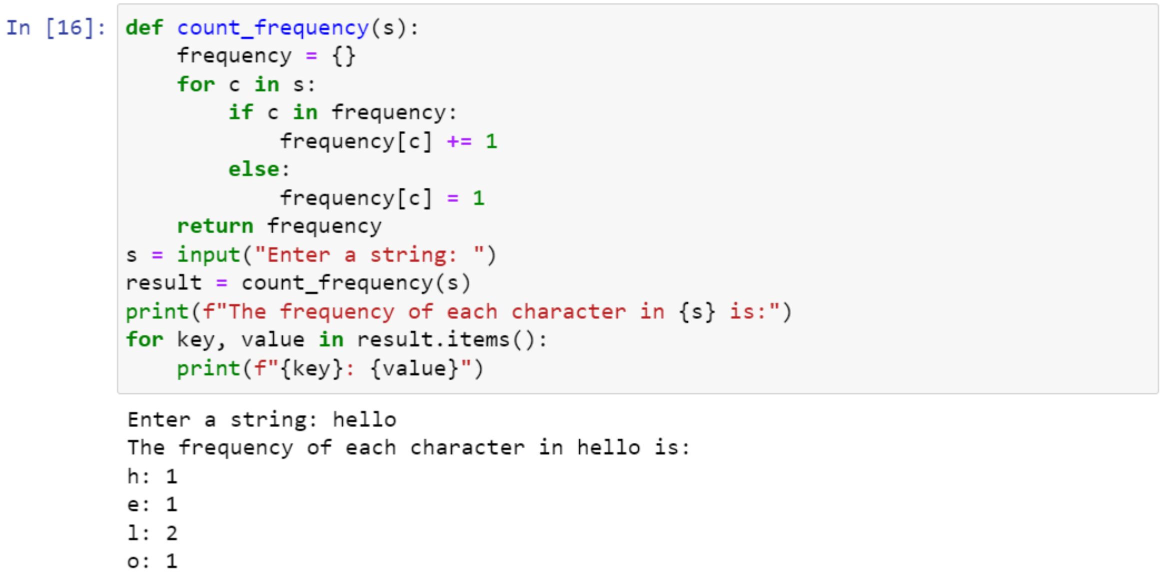
```
In [15]: import math
         def third_side(a, b):
             return math.sqrt(a**2 + b**2)
         a = float(input("Enter the length of the first side: "))
         b = float(input("Enter the length of the second side: "))

         if a <= 0 or b <= 0:
             print("Invalid input")
         else:
             c = third_side(a, b)
             print(f"The length of the third side is {c}")
```

```
Enter the length of the first side: 12
Enter the length of the second side: 15
The length of the third side is 19.209372712298546
```

15. Write a python program to print the frequency of each of the characters present in a given string.

Ans->



```
In [16]: def count_frequency(s):
             frequency = {}
             for c in s:
                 if c in frequency:
                     frequency[c] += 1
                 else:
                     frequency[c] = 1
             return frequency
         s = input("Enter a string: ")
         result = count_frequency(s)
         print(f"The frequency of each character in {s} is:")
         for key, value in result.items():
             print(f"{key}: {value}")

         Enter a string: hello
         The frequency of each character in hello is:
         h: 1
         e: 1
         l: 2
         o: 1
```

STATISTICS WORKSHEET-1

Q1 to Q9 have only one correct answer. Choose the correct option to answer your question.

1. Bernoulli random variables take (only) the values 1 and 0.

a) True b) False

Ans-> 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

Ans-> 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

Ans-> 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

Ans-> c

5. ______ random variables are used to model rates.

a) Empirical b) Binomial c) Poisson d) All of the mentioned

Ans-> c

6. Usually replacing the standard error by its estimated value does change the CLT.

a) True b) False

Ans-> 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

Ans-> 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

Ans-> 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

Ans-> c

Q10and Q15 are subjective answer type questions, Answer them in your own words briefly.

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

Ans-> A normal distribution is a kind of probability distribution that shows how a continuous random variable can have any value on the real number line, but most values are close to the average and fewer values are far from the average. A normal distribution has a smooth curve that looks like a bell and is balanced around the average. The average and the standard deviation are the two factors that affect the shape and position of the normal distribution. A normal distribution is helpful for studying many natural and social phenomena, such as heights, weights, IQ scores, errors, etc.

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

Ans-> Missing data can cause problems in data analysis, and different methods can be used to handle it, depending on the situation. Some of the methods are:

- Deleting the rows or columns with missing values. This is easy and fast, but it can lose information and make the sample smaller. This is only good when the missing data is random and small.
- Filling in the missing values with a constant or a statistic. This can keep the sample size and prevent bias, but it can also lower the variation and change the shape of the data. This is good when the missing data is random and moderate.
- Using a machine learning model that can work with missing values. This can use the information from the other features and avoid filling in errors, but it can also be costly and complicated. This is good when the missing data is not random and large.
- Using a multivariate imputation method that can create multiple filled-in datasets. This can account for the connections among the features and the uncertainty of the filling in, but it can also need more assumptions and resources. This is good when the missing data is not random and large.

There is no best method for handling missing data, and the method depends on the goal and the context of the analysis.

12. What is A/B testing?

Ans-> A/B testing is a method of comparing two or more versions of something, such as a web page, an email, or a product, to see which one performs better. A/B testing is used to test different aspects of the user experience, such as design, content, functionality, etc. The goal of A/B testing is to find the optimal version that maximizes a desired outcome, such as conversions, sales, engagement, etc. A/B testing is also known as split testing or bucket testing.

13. Is mean imputation of missing data acceptable practice?

Ans-> Mean imputation of missing data is a quick and easy method, but it has many problems and limitations. It can create bias, lower variation, change distributions, and influence multivariate analysis. It is not advised for most cases, especially when the missing data is not random or the amount is big. There are more sophisticated and trustworthy methods for dealing with missing data, such as machine learning algorithms or multivariate imputation techniques.

14. What is linear regression in statistics?

Ans-> Linear regression is a type of statistical analysis that explores how a dependent variable changes with one or more independent variables. It can help to test hypotheses, make predictions, and understand patterns in data. Linear regression can have different variations depending on the number and nature of the variables, such as simple, multiple, or logistic regression.

15. What are the various branches of statistics?

Ans-> Statistics has two main branches: descriptive statistics and inferential statistics. Descriptive statistics summarizes the features of a data set or a population, such as the average, variance, skewness, and kurtosis. Inferential statistics uses the data to make inferences or predictions about a larger group based on a sample. It can also check hypotheses and measure the relationship between variables.