MACHINE LEARNING

In Q1 to Q8, only one option is correct, Choose the correct option:

1. In the linear regression equation $y = \theta_0 + \theta_1 X$, θ_0 is the:
   A) Slope of the line　　B) Independent variable
   C) y intercept　　D) Coefficient of determination
   Ans: C

2. True or False: Linear Regression is a supervised learning algorithm.
   A) True　　B) False
   Ans: A

3. In regression analysis, the variable that is being predicted is:
   A) the independent variable　　B) the dependent variable
   C) usually denoted by x　　D) usually denoted by r
   Ans: B

4. Generally, which of the following method(s) is used for predicting continuous dependent variables?
   A) Logistic Regression　　B) Linear Regression
   C) Both　　D) None of the above
   Ans: B

5. The coefficient of determination is:
   A) the square root of the correlation coefficient　　B) usually less than zero
   C) the correlation coefficient squared　　D) equal to zero
   Ans: C

6. If the slope of the regression equation is positive, then:
   A) y decreases as x increases　　B) y increases as x increases
   C) y decreases as x decreases　　D) None of these
   Ans: B

7. Linear Regression works best for:
   A) linear data　　B) non-linear data
   C) both linear and non-linear data　　D) None of the above
   Ans: A

8. The coefficient of determination can be in the range of:
   A) 0 to 1　　B) -1 to 1
   C) -1 to 0　　D) 0 to infinity
   Ans: A

In Q9 to Q13, more than one options are correct, Choose all the correct options:

9. Which of the following evaluation metrics can be used for linear regression?
   A) Classification Report　　B) RMSE
   C) ROC curve　　D) MAE
   Ans: B, D

10. Which of the following is true for linear regression?
    A) Linear regression is a supervised learning algorithm.
    B) Linear regression supports multi-collinearity.
    C) Shape of linear regression's cost function is convex.
    D) Linear regression is used to predict discrete dependent variable.
    Ans: A, C

11. Which of the following regularizations can be applied to linear regression?
    A) Ridge　　B) Lasso

MACHINE LEARNING

C) Pruning D) Elastic Net

Ans: A, B, D

12. Linear regression performs better for:

A) Large amount of training samples with small number of features.

B) Same number of features and training samples

C) Large number of features

D) The variables which are drawn independently, identically distributed

Ans: A, D

13. Which of the following assumptions are true for linear regression?

A) Linearity B) Homoscedasticity

C) Non-Independent D) Normality

Ans: A, B

MACHINE LEARNING

Q14 and Q15 are subjective answer type questions, Answer them briefly.

14. Explain Linear Regression?

Ans:

Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables. It assumes a linear relationship between the dependent variable and each of the independent variables, and models this relationship using a linear equation of the form:

y = b0 + b1x1 + b2x2 + ... + bn*xn + e

where y is the dependent variable, x1, x2, ..., xn are the independent variables, b0, b1, b2, ..., bn are the coefficients (also known as the model parameters), and e is the error term (also known as the residual).

The goal of linear regression is to estimate the values of the coefficients that minimize the sum of the squared errors (SSE) between the predicted and actual values of the dependent variable. This is typically done using a method called ordinary least squares (OLS), which involves minimizing the sum of the squared residuals by finding the values of the coefficients that minimize the sum of the squared residuals.

Linear regression can be used for both simple and multiple regression problems, where simple regression involves only one independent variable, and multiple regression involves two or more independent variables. It can also be used for both continuous and categorical dependent variables, although in the latter case, logistic regression may be more appropriate.

Linear regression assumes several assumptions, such as linearity, homoscedasticity, independence of errors, normality of errors, and no multicollinearity between the independent variables. If these assumptions are not met, then the results of linear regression may not be reliable, and other methods such as generalized linear models, tree-based models, or neural networks may be more appropriate.

15. What is difference between simple linear and multiple linear regression?

Ans:

Simple linear regression is a statistical method used to model the relationship between a dependent variable and a single independent variable. The relationship between the dependent variable and independent variable is assumed to be linear.

Multiple linear regression, on the other hand, is a statistical method used to model the relationship between a dependent variable and two or more independent variables. The relationship between the dependent variable and each independent variable is assumed to be linear.

In simple linear regression, the regression equation has only one independent variable, while in multiple linear regression, the regression equation has two or more independent variables. The goal of both methods is to estimate the values of the coefficients that minimize the sum of the squared errors between the predicted and actual values of the dependent variable.

Simple linear regression is useful when we want to investigate the relationship between two variables and make predictions based on that relationship. Multiple linear regression is useful when we want to investigate the relationship between a dependent variable and several independent variables, and determine which independent variables have the strongest relationship with the dependent variable**.**

WORKSHEET 3 PYTHON

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

1. Which of the following will raise a value error in python?

A) int(32)
B) int(3.2)
C) int(-3.2)
D) int('32')

Ans: D

2. What will be the output of round(3.567)?

A) 3.5
B) 3.0
C) 4
D) 3

Ans: C

3. How is the function pow(a,b,c) evaluated in python?

A) a**b**c
B) (a**b)%c
C) (a**b)*c
D) (a**b)**c

Ans: B

4. What will be the output of **print(type(type(int)))** in python 3?

A) <class 'type'>
B) <type 'type'>
C) <class 'int'>
D) <type 'int'>

Ans: A

5. What will be the output of **ord(chr(65))**?

A) 'A'
B) 'a'
C) 65
D) TypeError

Ans: C

6. What is called when a function is defined inside a class?

A) Module
B) Function
C) _init_ function
D) Method

Ans: D

7. What will be the output of **all([1, 0, 5 ,7])**?

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

Ans: B

8. Is the output of the function abs() the same as that of the function math.fabs()?

A) Always
B) Sometimes
C) Never
D) None of these

Ans: A

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

9. Select all correct float numbers in python?

A) -68.7e100
B) 42e3
C) 4.2038
D) 3.0

Ans: A,B, C, D

10. Which of the following is(are) correct statement(s) in python?

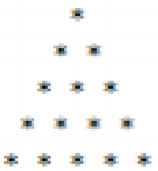
A) You can pass positional arguments in any order.

B) You can pass keyword arguments in any order.
C) You can call a function with positional and keyword arguments.
D) Positional arguments must be before keyword arguments in a function call
Ans: B, C

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

11. Write a python function print pyramid of stars. Level of the pyramid should be taken as an input from the user. E.g.
    Input = 5
    Output:

    Ans:

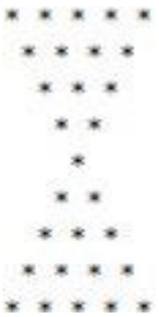


```
    *
   * *
  * * *
 * * * *
* * * * *
```

12. Write a python function print Hourglass pattern.
E.g.

Input = 5

Output:



```
* * * * *
 * * * *
  * * *
   * *
    *
   * *
  * * *
 * * * *
* * * * *
```

13. Write a python function to print Pascal’s Triangle. The number of levels in the triangle must be taken as input by the user. E.g.

Input = 5

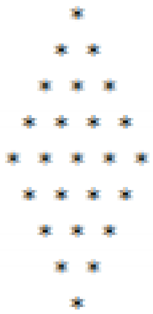
Output:

```
1
1 1
1 2 1
1 3 3 1
1 4 6 4 1
```

14. Write a python function to print Diamond Shaped Pattern shown below. Function must take integer input which represents the number of stars in the middle most line. E.g.:

Input = 5

Output:

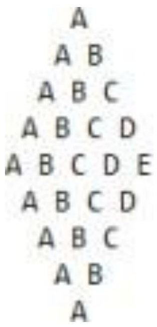


```
    *
   * *
  * * *
 * * * *
* * * * *
 * * * *
  * * *
   * *
    *
```

15. Write a python function to print Diamond Shaped Character Pattern shown below. Function must take integer input within range 1 to 26, which represents the rank of the alphabet. E.g.:

Input = 5

Output:



```
    A
   A B
  A B C
 A B C D
A B C D E
 A B C D
  A B C
   A B
    A
```

11. Write a python function print pyramid of stars. Level of the pyramid should be taken as an input from the user. E.g.

In [1]:

```python
def print_pyramid(levels):
    for i in range(1, levels + 1):
        print(' ' * (levels - i) + '*' * (2 * i - 1))

# Example usage:
print_pyramid(5)
```

```
    *
   ***
  *****
 *******
*********
```

12. Write a python function print Hourglass pattern

In [3]:

```python
def print_hourglass(levels):
    for i in range(levels, 0, -1):
        print(' ' * (levels - i) + '*' * (2 * i - 1))
    for i in range(2, levels + 1):
        print(' ' * (levels - i) + '*' * (2 * i - 1))

# Example usage:
print_hourglass(5)
```

```
*********
 *******
  *****
   ***
    *
   ***
  *****
 *******
*********
```

13. Write a python function to print Pascal’s Triangle. The number of levels in the triangle must be taken as input by the user

In [4]:

```python
def pascals_triangle(levels):
    triangle = [[1]]
    for i in range(1, levels):
        row = [1]
        for j in range(1, i):
            row.append(triangle[i-1][j-1] + triangle[i-1][j])
        row.append(1)
        triangle.append(row)
    for row in triangle:
        print(' '.join(str(num) for num in row).center(levels*2))

pascals_triangle(5)
```

```
    1
   1 1
  1 2 1
```

```
 1 3 3 1
1 4 6 4 1
```

In [ ]:

```

```

14 Write a python function to print Diamond Shaped Pattern shown below. Function must take integer input which represents the number of stars in the middle most line

In [5]:

```python
def print_diamond(n):
    # Upper half of diamond
    for i in range(1, n+1):
        # Print spaces
        for j in range(n-i):
            print(" ", end="")
        # Print stars
        for j in range(2*i-1):
            print("*", end="")
        print()

    # Lower half of diamond
    for i in range(n-1, 0, -1):
        # Print spaces
        for j in range(n-i):
            print(" ", end="")
        # Print stars
        for j in range(2*i-1):
            print("*", end="")
        print()
print_diamond(5)
```

```
    *
   ***
  *****
 *******
*********
 *******
  *****
   ***
    *
```

15. Write a python function to print Diamond Shaped Character Pattern shown below. Function must take integer input within range 1 to 26, which represents the rank of the alphabet

In [9]:

```python
def print_diamond(rank):
    if rank < 1 or rank > 26:
        print("Rank should be within 1 to 26")
        return
    char = chr(ord('A') + rank - 1)  # get the character corresponding to the rank
    n = 2 * rank - 1  # number of rows in the diamond
    # upper half of diamond
    for i in range(1, rank+1):
        # print spaces
        print(' '*(rank-i), end='')
        # print characters
        print((chr(ord(char)-(rank-i)))*(2*i-1))
    # lower half of diamond
    for i in range(rank-1, 0, -1):
        # print spaces
        print(' '*(rank-i), end='')
        # print characters
        print((chr(ord(char)-(rank-i)))*(2*i-1))

print_diamond(10)
```

```
         A
        BBB
       CCCCC
      DDDDDDD
     EEEEEEEEE
    FFFFFFFFFFF
   GGGGGGGGGGGGG
  HHHHHHHHHHHHHHH
 IIIIIIIIIIIIIIIII
JJJJJJJJJJJJJJJJJJJ
 IIIIIIIIIIIIIIIII
  HHHHHHHHHHHHHHH
   GGGGGGGGGGGGG
    FFFFFFFFFFF
     EEEEEEEEE
      DDDDDDD
       CCCCC
        BBB
         A
```

In [ ]:

STATISTICS WORKSHEET-10

Q1 to Q12 have only one correct answer. Choose the correct option to answer your question.

1. Rejection of the null hypothesis is a conclusive proof that the alternative hypothesis is
a. True
b. False
c. Neither
Ans:- c

2. Parametric test, unlike the non-parametric tests, make certain assumptions about
a. The population size
b. The underlying distribution
c. The sample size
Ans:- b

3. The level of significance can be viewed as the amount of risk that an analyst will accept when making a decision
a. True
b. False
Ans:- a

4. By taking a level of significance of 5% it is the same as saying
a. We are 5% confident the results have not occurred by chance
b. We are 95% confident that the results have not occurred by chance
c. We are 95% confident that the results have occurred by chance
Ans:-c

5. One or two tail test will determine
a. If the two extreme values (min or max) of the sample need to be rejected
b. If the hypothesis has one or possible two conclusions
c. If the region of rejection is located in one or two tails of the distribution
Ans:-C

6. Two types of errors associated with hypothesis testing are Type I and Type II. Type II error is committed when
a. We reject the null hypothesis whilst the alternative hypothesis is true
b. We reject a null hypothesis when it is true
c. We accept a null hypothesis when it is not true
Ans:c

7. A randomly selected sample of 1,000 college students was asked whether they had ever used the drug Ecstasy. Sixteen percent (16% or 0.16) of the 1,000 students surveyed said they had. Which one of the following statements about the number 0.16 is correct?
a. It is a sample proportion.
b. It is a population proportion.
c. It is a margin of error.
d. It is a randomly chosen number.
Ans:- a

8. In a random sample of 1000 students, pˆ = 0.80 (or 80%) were in favour of longer hours at the school library. The standard error of pˆ (the sample proportion) is

a. .013

b. .160

c. .640

d. .800

Ans:- a

9. For a random sample of 9 women, the average resting pulse rate is $x = 76$ beats per minute, and the sample standard deviation is $s = 5$. The standard error of the sample mean is
a. 0.557
b. 0.745
c. 1.667
d. 2.778
Ans:- c

10. Assume the cholesterol levels in a certain population have mean $\mu= 200$ and standard deviation $\sigma = 24$. The cholesterol levels for a random sample of $n = 9$ individuals are measured and the sample mean x is determined. What is the z-score for a sample mean $x = 180$?
a. –3.75
b. –2.50
c. −0.83
d. 2.50
Ans:- b

11. In a past General Social Survey, a random sample of men and women answered the question “Are you a member of any sports clubs?” Based on the sample data, 95% confidence intervals for the population proportion who would answer “yes” are .13 to .19 for women and .247 to .33 for men. Based on these results, you can reasonably conclude that
a. At least 25% of American men and American women belong to sports clubs.
b. At least 16% of American women belong to sports clubs.
c. There is a difference between the proportions of American men and American women who belong to sports clubs.
d. There is no conclusive evidence of a gender difference in the proportion belonging to sports clubs.

Ans:- d

12. Suppose a 95% confidence interval for the proportion of Americans who exercise regularly is 0.29 to 0.37. Which one of the following statements is FALSE?
a. It is reasonable to say that more than 25% of Americans exercise regularly.
b. It is reasonable to say that more than 40% of Americans exercise regularly.
c. The hypothesis that 33% of Americans exercise regularly cannot be rejected.
d. It is reasonable to say that fewer than 40% of Americans exercise regularly.

Ans: b

Q13 to Q15 are subjective answers type questions. Answers them in their own words briefly.

13. How do you find the test statistic for two samples?

Ans:- To find the test statistic for two samples, you need to first determine the type of hypothesis test you will be conducting. The type of hypothesis test will depend on the specific research question you are investigating and the type of data you are working with.

Once you have identified the appropriate hypothesis test, you will need to calculate the test statistic based on the formula provided by that test. For example, if you are conducting a two-sample t-test to compare the means of two populations, the formula for the test statistic would be:

$$t = (\bar{x}_1 - \bar{x}_2) / (s\sqrt{(1/n_1 + 1/n_2)})$$

14. How do you find the sample mean difference?

Ans:- To find the sample mean difference, you need to take the difference between the means of two samples.

If you have two samples with n_1 and n_2 observations respectively, you would calculate the sample mean difference as:

sample mean difference = $\bar{x}_1 - \bar{x}_2$

where $\bar{x}_1$ is the sample mean of the first sample and $\bar{x}_2$ is the sample mean of the second sample.

For example, suppose you have collected data on the heights of two groups of people. The first group has a sample size of 50 and a sample mean height of 175 cm, while the second group has a sample size of 60 and a sample mean height of 170 cm. The sample mean difference in this case would be:

sample mean difference = $\bar{x}_1 - \bar{x}_2$
= 175 cm - 170 cm
= 5 cm

So, the sample mean difference between the two groups is 5 cm. This tells us that, on average, the first group is taller than the second group by 5 cm.

15. What is a two sample t test example?

Ans:- A two-sample t-test is a statistical test used to compare the means of two independent groups. Here is an example of a two-sample t-test:

Suppose a researcher wants to test whether a new weight loss program is more effective than the current weight loss program. The researcher randomly selects 50 individuals and randomly assigns them to either the new program or the current program. After 12 weeks on the program, the researcher measures the weight loss for each individual.

The researcher wants to compare the mean weight loss for the two groups using a two-sample t-test. The null hypothesis is that the mean weight loss for the new program is the same as the mean weight loss for the current program. The alternative hypothesis is that the mean weight loss for the new program is greater than the mean weight loss for the current program.

The researcher calculates the sample mean weight loss for the new program to be 8 pounds with a sample standard deviation of 2 pounds. The sample mean weight loss for the current program is 6 pounds with a sample standard deviation of 3 pounds. The researcher decides to use a significance level of 0.05.

To conduct the two-sample t-test, the researcher would calculate the test statistic using the following formula:

$t = (\bar{x}_1 - \bar{x}_2) / (s\sqrt{(1/n_1 + 1/n_2)})$

where $\bar{x}_1$ and $\bar{x}_2$ are the sample means, s is the pooled standard deviation, and n_1 and n_2 are the sample sizes.

Plugging in the sample values, the test statistic would be:

$t = (8 - 6) / (2.5\sqrt{(1/50 + 1/50)}) = 4.47$

The researcher would then compare this test statistic to a t-distribution with 98 degrees of freedom (50 + 50 - 2). At a significance level of 0.05, the critical t-value for a one-tailed test with 98 degrees of freedom is 1.66.

Since the calculated test statistic of 4.47 is greater than the critical t-value of 1.66, the researcher would reject the null hypothesis and conclude that the mean weight loss for the new program is greater than the mean weight loss for the current program.