MACHINE LEARNING ANSWERS

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 modal
7. D) Regularization
8. D) SMOTE
9. A) TPR and FPR
10. B) False
11. B) Apply PCA to project high dimensional data
12. A) We don't have to choose the learning rate AND B) It becomes slow when number of features is very large.
13. Regularization refers to techniques that are used to calibrate machine learning models in order to minimize the adjusted loss function and prevent overfitting or underfitting. There are two main types of regularization techniques: Ridge Regularization and Lasso Regularization. Ridge Regularization, it modifies the over-fitted or under fitted models by adding the penalty equivalent to the sum of the squares of the magnitude of coefficients. Lasso Regularization modifies the over-fitted or under-fitted models by adding the penalty equivalent to the sum of the absolute values of coefficients.
14. Ridge Regularization and Lasso Regularization
15. An error term is a residual variable produced by a statistical or mathematical model, which is created when the model does not fully represent the actual relationship between the independent variables and the dependent variables. As a result of this incomplete relationship, the error term is the amount at which the equation may differ during empirical analysis. A Linear Regression model's main aim is to find the best fit linear line and the optimal values of intercept and coefficients such that the error is minimized. Error is the difference between the actual value and Predicted value.

Statistics Answers

1. A) True
2. A) Central Limit Theorem
3. B) Modeling bounded count data
4. D) All of the mentioned
5. C) Poisson
6. B) False
7. B) Hypothesis
8. A) 0
9. C) Outliers cannot conform to the regression relationship
10. A normal distribution is a type of continuous probability distribution in which most data points cluster towards the middle of the range, while the rest taper off symmetrically towards either extreme. The middle of the range is also known as the mean of the distribution
11. Missing data can be dealt with in a variety of ways. I believe the most common reaction is to ignore it. Choosing to make no decision, on the other hand, indicates that your statistical programme will make the decision for you. Your application will remove things in a listwise sequence most of the time. Depending on why and how much data is gone, listwise deletion may or may not be a good idea.Another common strategy among those who pay attention is imputation. Imputation is the process of substituting an estimate for missing values and analysing the entire data set as if the imputed values were the true observed values.

Single or Multiple Imputation

- Single and multiple imputation are the two forms of imputation. When people say imputation, they usually mean single.
- The term "single" refers to the fact that you only use one of the seven methods to estimate the missing number outlined above.
- It's popular since it's simple to understand and generates a sample with the same number of observations as the complete data set.
- When listwise deletion eliminates a considerable amount of the data set, single imputation appears to be a tempting option. It does, however, have certain restrictions.
- Unless the data is Missing Completely at Random, certain imputation processes, such as means, correlations, and regression coefficients, result in skewed parameter estimations. The bias is frequently worse than with listwise deletion, which is most software's default.
- The level of the bias is determined by a number of factors, including the imputation technique, the missing data mechanism, the fraction of missing data, and the information in the data set

12. A/B testing is also known as split testing, refers to a randomized experimentation process where in two or more versions of variable (eg. Web page) are shown to

different segments of website visitors at the same time to determine which version leaves the maximum impact and drives business metrics.

13. Mean imputation is typically considered terrible practice since it ignores features correlation, consider the following scenario we have a table with age and fitness scores and an eight year old has a Missing fitness score. If we average the fitness scores of people between age 15 to 80, the eighty year will appear to have a significantly greater fitness level than he actually does.
14. Linear regression analysis is used to predict the value of a variable based on the value of another variable. The variable you want to predict is called the dependent variable. The variable you are using to predict the other variable's value is called the independent variable.
15. The two major areas of statistics are known as descriptive statistics, which describes the properties of sample and population data, and inferential statistics, which uses those properties to test hypotheses and draw conclusions.

Descriptive Statistics

Descriptive statistics mostly focus on the central tendency, variability, and distribution of sample data. Central tendency means the estimate of the characteristics, a typical element of a sample or population, and includes descriptive statistics such as mean, median, and mode. Variability refers to a set of statistics that show how much difference there is among the elements of a sample or population along the characteristics measured, and includes metrics such as range, variance, and standard deviation.

Inferential Statistics

Inferential statistics are tools that statisticians use to draw conclusions about the characteristics of a population, drawn from the characteristics of a sample, and to decide how certain they can be of the reliability of those conclusions. Based on the sample size and distribution statisticians can calculate the probability that statistics, which measure the central tendency, variability, distribution, and relationships between characteristics within a data sample, provide an accurate picture of the corresponding parameters of the whole population from which the sample is drawn.

Python worksheet

1. C) %
2. B) 0
3. C) 24
4. A) 2
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. A) in defining an iterator
9. A) _abc
10. D) all of the above

In [40]:
```
# 11. Write a python program to find the factorial of a number
# Factorial of a Number using Loop

num = 16
factorial = 1

if num < 0:
   print("Sorry, factorial does not exist for negative numbers")
elif num == 0:
   print("The factorial of 0 is 1")
else:
   for i in range(1,num + 1):
       factorial = factorial*i
   print("The factorial of",num,"is",factorial)
```

```
The factorial of 16 is 20922789888000
```

In [39]:
```
# 12. Write a python program to find whether a number is prime or composite


number = 9

if number > 1:
    for i in range(2,int(number/2)+1):
        if (number % i == 0):
            print(number, "is a composite Number")
            break
    else:
        print(number,"is a Prime number")
else:
    print(number,"is not a Prime number")
```

```
9 is a composite Number
```

In [42]:
```
# 12. Write a python program to find whether a number is prime or composite


number = 2

if number > 1:
    for i in range(2,int(number/2)+1):
        if (number % i == 0):
            print(number, "is a composite Number")
            break
    else:
        print(number,"is a Prime number")
else:
    print(number,"is a composite Number")
```

```
2 is a Prime number
```

In [43]:
```
# 13. Write a python program to check whether a given string is palindrome or not.

def isPalindrome(s):

    rev = ''.join(reversed(s))
    if (s == rev):
        return True
    return False
```

```
s = "noon"
ans = isPalindrome(s)

if (ans):
    print("Yes")
else:
    print("No")
```

```
Yes
```

In [46]:
```
# 14. Write a Python program to get the third side of right-angled triangle from two g
# Input: side1 = 4, side2 = 3



def pythagoras(opposite_side,adjacent_side,hypotenuse):
        if opposite_side == str("x"):
            return ("Opposite = " + str(((hypotenuse**2) - (adjacent_side**2))**0.5))
        elif adjacent_side == str("x"):
            return ("Adjacent = " + str(((hypotenuse**2) - (opposite_side**2))**0.5))
        elif hypotenuse == str("x"):
            return ("Hypotenuse = " + str(((opposite_side**2) + (adjacent_side**2))**0
        else:
            return "You know the answer!"

print(pythagoras(3,4,'x'))
print(pythagoras(3,'x',5))
print(pythagoras('x',4,5))
print(pythagoras(3,4,5))
```

```
Hypotenuse = 5.0
Adjacent = 4.0
Opposite = 3.0
You know the answer!
```

In [45]:
```
#15.Write a python program to print the frequency of each of the characters present ir

test_str = "helloworld"

all_freq = {}

for i in test_str:
    if i in all_freq:
        all_freq[i] += 1
    else:
        all_freq[i] = 1

print("Count of all characters in helloworld is :\n "
      + str(all_freq))
```

```
Count of all characters in helloworld is :
 {'h': 1, 'e': 1, 'l': 3, 'o': 2, 'w': 1, 'r': 1, 'd': 1}
```

In [ ]:
```

```