Machine Learning

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
D) Both A and B

2. Which of the following statement is true about outliers in linear regression?
A) Linear regression is sensitive to outliers

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

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

5. Which of the following is the reason for over fitting condition?
C) Low bias and high variance

6. If output involves label then that model is called as:
B) Predictive modal

7. Lasso and Ridge regression techniques belong to _________?
D) Regularization

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

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

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

11. Pick the feature extraction from below:
A) Construction bag of words from a email

12. Which of the following is true about Normal Equation used to compute the coefficient of the Linear
Regression?
B) It becomes slow when number of features is very large.
D) It does not make use of dependent variable.

13. Explain the term regularization?
Regularization is a technique used in machine learning and statistics to prevent overfitting by adding a penalty term to the model's loss function. This penalty discourages the model from fitting the training data too closely and encourages it to find simpler, more generalizable patterns.

14. Which particular algorithms are used for regularization?

1. Lasso Regression (L1 Regularization)
2. Ridge Regression (L2 Regularization)
3. Elastic Net
4. Dropout
5. Early Stopping
6. Data Augmentation
7. Batch Normalization
8. Weight Decay
9. Cross-Validation
10. Bootstrap Aggregating (Bagging)

15. Explain the term error present in linear regression equation?
The term "error" in the context of a linear regression equation refers to the difference between the actual observed values and the predicted values made by the linear regression model. It quantifies how well the model's predictions match the real data points and reflects the inherent variability or noise in the data that the model can't explain. The goal of linear regression is to minimize this error, often measured using metrics like Mean Squared Error (MSE) or Root Mean Squared Error (RMSE).

PYTHON – WORKSHEET 1

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

2. In python 2//3 is equal to?
B) 0

3. In python, 6<<2 is equal to?
C) 24

4. In python, 6&2 will give which of the following as output?
D) 0

5. In python, 6|2 will give which of the following as output?
B) 4

6. What does the finally keyword denotes in python?
C) the finally block will be executed no matter if the try block raises an error or not.

7. What does raise keyword is used for in python?
A) It is used to raise an exception.

8. Which of the following is a common use case of yield keyword in python?
A) in defining an iterator

9. Which of the following are the valid variable names?
A) _abc
C) abc2

10. Which of the following are the keywords in python?
A) yield
B) raise

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

```
def factorial(n):
    factorial = 1
    for i in range(1, n + 1):
        factorial *= i
    return factorial

print(factorial(5))
```

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

```
def is_prime(number):
    if number <= 1:
        return False

    for i in range(2, number):
        if number % i == 0:
            return False

    return True

print(is_prime(10))
print(is_prime(7))
```

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

```
def is_palindrome(string):
    reversed_string = string[::-1]
    return string == reversed_string

print(is_palindrome("racecar"))
print(is_palindrome("madam"))
print(is_palindrome("hello"))
```

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

```
def get_third_side(a, b):
       c = (a2 + b2)0.5
    return c

print(get_third_side(3, 4))
print(get_third_side(5, 12))
```

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

```
def get_character_frequency(string):

    character_frequency = {}
    for character in string:
        if character in character_frequency:
            character_frequency[character] += 1
        else:
            character_frequency[character] = 1
    return character_frequency

def print_character_frequency(character_frequency):

    for character, frequency in character_frequency.items():
        print(f"{character}: {frequency}")

string = "hello world"
character_frequency = get_character_frequency(string)
print_character_frequency(character_frequency)
```

STATISTICS WORKSHEET-1

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

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

3. Which of the following is incorrect with respect to use of Poisson distribution?
b) Modeling bounded count data

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

5. ______ random variables are used to model rates.
c) Poisson

6. 10. Usually replacing the standard error by its estimated value does change the CLT.
b) False

7. 1. Which of the following testing is concerned with making decisions using data?
b) Hypothesis

8. 4. Normalized data are centered at______and have units equal to standard deviations of the
original data.
a) 0

9. Which of the following statement is incorrect with respect to outliers?
c) Outliers cannot conform to the regression relationship

10. What do you understand by the term Normal Distribution?
Normal Distribution, also known as Gaussian Distribution, is a symmetric probability distribution that is characterized by its bell-shaped curve. In this distribution, the majority of data points cluster around the mean, with fewer points farther away towards the tails. It is widely used in statistics and modeling due to its mathematical properties and prevalence in various natural phenomena.

11. How do you handle missing data? What imputation techniques do you recommend?
Handling missing data involves various techniques:

1. Mean/Median Imputation: Simple, but may distort data.
2. Mode Imputation: For categorical data.
3. Forward Fill/Backward Fill: Suitable for time-series data.
4. KNN Imputation: Good for retaining relationships.

5. Multiple Imputation: Captures uncertainty.
6. Regression Imputation: Uses relationships between variables.
7. Domain-Specific Imputation: Knowledge-driven imputation.

12. What is A/B testing?

A/B testing is an experimental method used to compare two versions of a webpage, app, or other elements to determine which one performs better in terms of user engagement, conversions, or other predefined metrics. It involves randomly assigning users to either version (A or B), measuring their responses, and analyzing the results to make informed decisions about changes or optimizations.

13. Is mean imputation of missing data acceptable practice?

Mean imputation is a simple technique, but it has limitations. It can introduce bias and distort data distributions. It's generally not recommended for datasets with substantial missing values or when missingness is not random. More advanced imputation methods might be preferable.

14. What is linear regression in statistics?

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 and aims to find the best-fitting line (or hyperplane) that minimizes the difference between observed and predicted values. It's commonly used for prediction, understanding relationships, and making inferences in data analysis.

15. What are the various branches of statistics?

Various branches of statistics include:

1. Descriptive Statistics: Summarizing and describing data.
2. Inferential Statistics: Making predictions and inferences about populations from samples.
3. Probability Theory: Studying uncertainty and randomness.
4. Biostatistics: Applying statistical methods to biology and medicine.
5. Econometrics: Applying statistics to economic data.
6. Social Statistics: Analyzing social phenomena and trends.
7. Statistical Modeling: Creating mathematical models to represent data relationships.
8. Time Series Analysis: Analyzing data points collected at regular intervals.
9. Multivariate Statistics: Analyzing data with multiple variables.
10. Experimental Design: Planning experiments to gather meaningful data.