MACHINE LEARNING

QUESTION NO.	ANSWER
1.	D) Both A and B.
2.	A) Linear regression is sensitive to outliers.
3.	A) Negative.
4.	B) Correlation.
5.	C) Low bias and high variance.
6.	B) Predictive model.
7.	D) Regularization.
8.	D) SMOTE.
9.	A) TPR and FPR.
10.	B) False.
11.	A) Construction bag of words from an email.
12.	A) We don't have to choose the learning rate.

13. Regularization is a technique used in machine learning to prevent overfitting and improve the generalization ability of a model. It involves adding a penalty term to the cost function, which discourages the model from fitting the data too closely and encourages it to learn simpler patterns.
In linear regression, regularization is typically implemented by adding a term to the cost function that penalizes the magnitude of the coefficients. There are two commonly used types of regularization: L1 regularization (also known as Lasso) and L2 regularization (also known as Ridge).
L1 regularization adds the sum of the absolute values of the coefficients as a penalty term to the cost function, while L2 regularization adds the sum of the squared values of the coefficients. By adding these penalty terms, the model is encouraged to learn coefficients that are smaller in magnitude and thus more generalizable to new data.
Regularization is particularly useful when the number of features in the dataset is large compared to the number of training examples. In such cases, the model may overfit the training data by learning complex relationships between features that are specific to the training set but do not generalize well to new data. Regularization can help mitigate this problem by encouraging the model to learn simpler relationships that are more likely to generalize to new data.

14. Regularization can be applied to a variety of machine learning algorithms, including linear regression, logistic regression, support vector machines (SVM), neural networks, and more.
In linear regression, L1 regularization (also known as Lasso) and L2 regularization (also known as Ridge) are commonly used. In Lasso regression, the L1 penalty term is added to the cost function, while in Ridge regression, the L2 penalty term is added.
In logistic regression, L1 regularization and L2 regularization can also be applied to the cost function.
In SVMs, regularization is achieved by adding a penalty term to the objective function that encourages the model to learn a decision boundary with a larger margin.

In neural networks, regularization can be implemented in several ways, such as weight decay, dropout, and early stopping.
Overall, regularization is a powerful technique that can improve the generalization ability of a wide range of machine learning models.

15. In linear regression, the error (also called the residual) represents the difference between the actual value of the dependent variable and the predicted value of the dependent variable based on the regression equation.
The regression equation is typically of the form $Y = b0 + b1X1 + b2X2 + ... + bnxn$, where Y is the dependent variable, X1, X2, ..., Xn are the independent variables, and b0, b1, b2, ..., bn are the coefficients that are estimated during the training phase. The predicted value of Y based on this equation is called the fitted value.
The error term for each data point is then calculated as the difference between the actual value of Y and the fitted value of Y for that data point. Mathematically, the error can be written as $e = y - \hat{y}$, where y is the actual value of Y and $\hat{y}$ is the predicted value of Y based on the regression equation.
The goal of linear regression is to minimize the sum of squared errors (SSE), which is the sum of the squared errors for each data point. Minimizing SSE means finding the values of the coefficients that provide the best fit to the data, in the sense that they minimize the overall difference between the actual values and the predicted values.
In summary, the error term in linear regression represents the difference between the actual value of the dependent variable and the predicted value of the dependent variable based on the regression equation. The goal of linear regression is to minimize the sum of squared errors, which is achieved by finding the values of the coefficients that provide the best fit to the data.

PYTHON WORKSHEET – 1

QUESTION NO	ANSWER
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.	C) in defining a generator.
9.	A) _abc and C) abc2.
10.	A) yield and B) raise.

11.

```
num = int(input("Enter a number: "))
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)
```

12.

```
def is_prime(n):
    # 0 and 1 are not prime nor composite
    if n < 2:
        return False
    # check if n is divisible by any number between 2 and n-1
    for i in range(2, n):
        if n % i == 0:
            return False
    # if n is not divisible by any number between 2 and n-1, it is prime
    return True

# take input from user
num = int(input("Enter a number: "))

# check if the number is prime or composite
if is_prime(num):
    print(num, "is a prime number")
else:
    print(num, "is a composite number")
```

13.

```python
string = input("Enter a string: ")

# Reverse the string
reverse_string = string[::-1]

# Compare the original string with the reversed string
if string == reverse_string:
    print("The given string is a palindrome.")
else:
    print("The given string is not a palindrome.")
```

14.

```python
import math

def get_third_side(side1, side2):
    """Returns the length of the third side of a right-angled triangle"""
    return math.sqrt(side1**2 + side2**2)

# Example usage:
side1 = 3
side2 = 4
third_side = get_third_side(side1, side2)
print("The length of the third side is:", third_side)
```

15.

```python
string = input("Enter a string: ")
freq = {}

for char in string:
    if char in freq:
        freq[char] += 1
    else:
        freq[char] = 1

print("Frequency of each character in the given string:")
for char in freq:
    print(char, ":", freq[char])
```

STATISTICS WORKSHEET

QUESTION NO.	ANSWER
1.	True
2.	Central Limit Theorem
3.	Modeling bounded count data
4.	The square of a standard normal random variable follows what chi-squared distribution
5.	Poisson
6.	False
7.	Hypothesis
8.	0
9.	Outliers cannot conform to the regression relationship

10. Normal distribution, also known as Gaussian distribution, is a probability distribution that is symmetric and bell-shaped. It is characterized by its mean (average) and standard deviation, and many natural phenomena follow a normal distribution, such as the heights of individuals in a population, IQ scores, and errors in measurements. The normal distribution is important in statistics because of the central limit theorem, which states that the distribution of averages of independent and identically distributed random variables approaches a normal distribution as the sample size increases. The normal distribution is widely used in statistical inference, hypothesis testing, and modeling in various fields such as finance, engineering, and social sciences.

11. Handling missing data is an important part of data preprocessing in machine learning. Here are some common imputation techniques that can be used to handle missing data:

1. Mean/median imputation: In this technique, missing values are replaced with the mean or median of the available data for that variable. This is a simple and commonly used technique, but it can lead to biased results if there are too many missing values or if the data is not normally distributed.
2. Mode imputation: For categorical data, missing values can be replaced with the mode (most common value) of the available data for that variable.
3. Regression imputation: Regression models can be used to predict missing values based on other variables in the data. This can be a more accurate method of imputation, but it requires a significant amount of data and can be computationally expensive.
4. K-nearest neighbor imputation: In this technique, missing values are replaced with the values of the k-nearest neighbors based on the other variables in the data. This can be a useful method for small datasets, but it can be sensitive to outliers and can lead to biased results if the data is not representative.

5. Multiple imputation: This technique involves creating multiple imputed datasets, each with different imputed values for the missing data. This can help to account for uncertainty in the imputation process and produce more accurate results.

12. A/B testing is a statistical technique used in marketing, product development, and other fields to compare two versions of a product, service, or marketing campaign. In A/B testing, two different versions of a product or service are presented to different groups of users, and statistical analysis is used to determine which version is more effective in achieving a desired outcome, such as increasing sales or improving user engagement.

For example, in a website A/B test, one group of users may see a website with a red call-to-action button, while another group sees a website with a blue call-to-action button. By tracking user behavior, such as click-through rates or conversion rates, statisticians can determine which version of the website is more effective in achieving the desired outcome.
There are many different variations of A/B testing, including multivariate testing, where multiple versions of a product or service are tested simultaneously, and sequential testing, where the test is designed to end early if one version clearly outperforms the other. A/B testing is a powerful tool for improving the effectiveness of marketing campaigns and other business strategies, and is widely used across industries to optimize decision-making.

13. Mean imputation is a commonly used method for handling missing data. It involves replacing the missing values with the mean of the available values for that variable. While it is a simple method, it has several drawbacks.

One of the main concerns with mean imputation is that it assumes that the missing values are missing at random (MAR) and that the missingness is unrelated to the value of the missing variable. However, this assumption is often not valid in practice, and missingness may be related to unobserved variables or variables that are missing.

Mean imputation can also lead to biased estimates of the mean, variance, and covariance of the variables, which can affect downstream analyses. Additionally, mean imputation can lead to an underestimation of the standard error and overestimation of the statistical significance of the results.

Therefore, while mean imputation may be a convenient method for handling missing data, it is generally not recommended. Alternative imputation techniques, such as multiple imputation or maximum likelihood estimation, are preferred as they account for the uncertainty in the missing data and produce more accurate estimates.

14. In statistics, linear regression is a commonly used approach for modeling the relationship between a dependent variable (often denoted as "y") and one or more independent variables (often denoted as "x"). The goal of linear regression is to find the best linear relationship between the dependent and independent variables. This is typically done by fitting a straight line (or hyperplane, in the case of multiple independent variables) to the data that minimizes the sum of the squared differences between the predicted values and the actual values of the dependent variable. The resulting linear regression equation can then be used to predict values of the dependent variable for new values of the independent variable(s). Linear regression is a widely used technique in fields such as economics, finance, psychology, and many others.

15. Statistics is a broad field that has several subfields or branches. Some of the major branches of statistics include:

1. Descriptive Statistics: Descriptive statistics is concerned with describing and summarizing the characteristics of a dataset.

2. Inferential Statistics: Inferential statistics is concerned with making generalizations about a population based on a sample of data.

3. Biostatistics: Biostatistics is the application of statistical methods to biological and medical data.

4. Econometrics: Econometrics is the application of statistical methods to economic data.

5. Psychometrics: Psychometrics is the application of statistical methods to psychological data.

6. Statistical Computing: Statistical computing is concerned with developing and implementing algorithms for statistical analysis.

7. Bayesian Statistics: Bayesian statistics is a framework for statistical inference that involves updating probabilities based on new data.

8. Time Series Analysis: Time series analysis is concerned with modeling and forecasting data that varies over time.

9. Spatial Statistics: Spatial statistics is concerned with modeling and analyzing data that are geographically referenced.

10. Multivariate Statistics: Multivariate statistics is concerned with the analysis of datasets that have more than one variable.