MACHINE LEARNING

In Q1 to Q7, only one option is correct, Choose the correct option:

1. What is the advantage of hierarchical clustering over K-means clustering?
   A) Hierarchical clustering is computationally less expensive
   B) In hierarchical clustering you don't need to assign number of clusters in beginning
   C) Both are equally proficient D) None of these
   Ans :-B

2. Which of the following hyper parameter(s), when increased may cause random forest to over fit the data?
   A) max_depth B) n_estimators
   C) min_samples_leaf D) min_samples_splits
   Ans :-A

3. Which of the following is the least preferable resampling method in handling imbalance datasets?
   A) SMOTE B) RandomOverSampler
   C) RandomUnderSampler D) ADASYN
   Ans :-B

4. Which of the following statements is/are true about "Type-1" and "Type-2" errors?
   1. Type1 is known as false positive and Type2 is known as false negative.
   2. Type1 is known as false negative and Type2 is known as false positive.
   3. Type1 error occurs when we reject a null hypothesis when it is actually true.

   A) 1 and 2 B) 1 only
   C) 1 and 3 D) 2 and 3
   Ans :-C

5. Arrange the steps of k-means algorithm in the order in which they occur:
   1. Randomly selecting the cluster centroids
   2. Updating the cluster centroids iteratively
   3. Assigning the cluster points to their nearest center

   A) 3-1-2 B) 2-1-3
   C) 3-2-1 D) 1-3-2
   Ans :-A

6. Which of the following algorithms is not advisable to use when you have limited CPU resources and time, and when the data set is relatively large?
   A) Decision Trees B) Support Vector Machines
   C) K-Nearest Neighbors D) Logistic Regression
   Ans :-C

7. What is the main difference between CART (Classification and Regression Trees) and CHAID (Chi Square Automatic Interaction Detection) Trees?
   A) CART is used for classification, and CHAID is used for regression.
   B) CART can create multiway trees (more than two children for a node), and CHAID can only create binary trees (a maximum of two children for a node).
   C) CART can only create binary trees (a maximum of two children for a node), and CHAID can create multiway trees (more than two children for a node)
   D) None of the above
   Ans :-C

MACHINE LEARNING

In Q8 to Q10, more than one options are correct, Choose all the correct options:

8. In Ridge and Lasso regularization if you take a large value of regularization constant(lambda), which of the following things may occur?
   A) Ridge will lead to some of the coefficients to be very close to 0
   B) Lasso will lead to some of the coefficients to be very close to 0
   C) Ridge will cause some of the coefficients to become 0
   D) Lasso will cause some of the coefficients to become 0.
   Ans :-B, D

MACHINE LEARNING

9. Which of the following methods can be used to treat two multi-collinear features?
   A) remove both features from the dataset
   B) remove only one of the features
   C) Use ridge regularization
   D) use Lasso regularization
   Ans :-B, C, D

10. After using linear regression, we find that the bias is very low, while the variance is very high. What are the possible reasons for this?
   A) Overfitting
   B) Multicollinearity
   C) Underfitting
   D) Outliers
   Ans :-A, D

Q10 to Q15 are subjective answer type questions, Answer them briefly.

11. In which situation One-hot encoding must be avoided? Which encoding technique can be used in such a case?

Ans:- One-hot encoding is a popular technique used to encode categorical variables in machine learning. However, one-hot encoding may not always be the best choice and may need to be avoided in certain situations.

One situation where one-hot encoding may need to be avoided is when dealing with high cardinality categorical variables. High cardinality categorical variables have a large number of distinct categories, and encoding them using one-hot encoding can result in a large number of binary features, which can increase the dimensionality of the feature space and lead to the curse of dimensionality. This can make the model more complex and increase the risk of overfitting, and may also lead to computational issues and slower training times.

In such situations, we can use other encoding techniques like target encoding or frequency encoding. Target encoding replaces each category with the mean or median of the target variable for that category, while frequency encoding replaces each category with its frequency or count in the dataset. These techniques can be effective in reducing the dimensionality of the feature space and avoiding the curse of dimensionality.

Therefore, one-hot encoding must be avoided in situations where the categorical variable has high cardinality, and other encoding techniques like target encoding or frequency encoding can be used instead.

12. In case of data imbalance problem in classification, what techniques can be used to balance the dataset? Explain them briefly.

Ans:- Data imbalance occurs when the number of examples in one class is much higher or lower than the number of examples in another class in a classification problem. In such cases, the model may be biased towards the majority class and perform poorly on the minority class. To address this issue, we can use various techniques to balance the dataset. Some of these techniques are:

Undersampling: Undersampling involves removing examples from the majority class to balance the dataset. This can be done randomly or using techniques like Tomek links or edited nearest neighbors. The advantage of undersampling is that it can reduce the computation time and memory requirements of the model, but it can also lead to loss of information and may not work well with small datasets.

Oversampling: Oversampling involves adding examples to the minority class to balance the dataset. This can be done by replicating examples or using techniques like SMOTE (Synthetic Minority Over-sampling Technique) or ADASYN (Adaptive Synthetic Sampling). The advantage of oversampling is

MACHINE LEARNING

that it can improve the model's performance on the minority class, but it can also increase the risk of overfitting and lead to a larger dataset.

Class weight adjustment: Class weight adjustment involves assigning higher weights to the minority class and lower weights to the majority class during training. This can be done using techniques like inverse frequency weighting or balanced class weights. The advantage of class weight adjustment is that it can improve the model's performance on the minority class without changing the dataset, but it may not work well with highly imbalanced datasets.

Ensemble methods: Ensemble methods involve combining multiple models to improve the model's performance on the minority class. This can be done using techniques like bagging, boosting, or stacking. The advantage of ensemble methods is that they can improve the model's performance on both the majority and minority classes, but they can also increase the computation time and complexity of the model.

Therefore, to balance the dataset in case of data imbalance problem in classification, we can use techniques like undersampling, oversampling, class weight adjustment, and ensemble methods. The choice of technique depends on the dataset size, class imbalance ratio, computational resources, and the desired level of performance.

13. What is the difference between SMOTE and ADASYN sampling techniques?

Ans:- SMOTE (Synthetic Minority Over-sampling Technique) and ADASYN (Adaptive Synthetic Sampling) are both techniques used for handling imbalanced datasets in machine learning. They are both oversampling techniques that generate synthetic samples for the minority class to balance the class distribution.

The main difference between SMOTE and ADASYN is the way they generate synthetic samples. SMOTE creates new samples by interpolating between existing ones in the minority class, while ADASYN focuses more on areas where the density of minority samples is low by adding more noise to the new samples.

In SMOTE, synthetic samples are generated by randomly selecting a minority class sample and finding its k-nearest neighbors. New samples are then created by interpolating between the selected sample and its k-nearest neighbors. This process continues until the desired number of synthetic samples is generated.

On the other hand, ADASYN uses a density distribution-based approach to generate synthetic samples. It focuses more on the areas where the density of minority samples is low and generates more samples there. This is done by measuring the density distribution of the samples and generating synthetic samples proportionally to the local density of the minority class.

In summary, while SMOTE generates new samples by interpolating between existing ones in the minority class, ADASYN generates samples by adding more noise to the new samples in the areas of low density of the minority class**.**

14. What is the purpose of using GridSearchCV? Is it preferable to use in case of large datasets? Why or why not?

Ans:- GridSearchCV is a technique in machine learning used to find the best combination of hyperparameters for a given model. It is essentially a cross-validation technique that exhaustively searches through a specified parameter grid to find the best set of hyperparameters that optimize a specified evaluation metric.

The purpose of using GridSearchCV is to simplify the process of tuning hyperparameters by automating the search process. This saves time and reduces the risk of human error in selecting the best hyperparameters.

GridSearchCV is not always preferable to use in case of large datasets. The reason for this is that GridSearchCV performs an exhaustive search over all the possible combinations of

MACHINE LEARNING

hyperparameters in the specified grid. This can be computationally expensive and time-consuming, especially for large datasets. It may also be memory-intensive as it needs to store all the trained models in memory.

To address this issue, several techniques can be used to reduce the computational cost of GridSearchCV. One approach is to reduce the size of the grid search space by specifying a smaller range of hyperparameters to search over. Another approach is to use random search instead of GridSearchCV. In random search, the hyperparameters are randomly sampled from a specified distribution, which can be faster and more memory-efficient than GridSearchCV.

In summary, GridSearchCV is a useful technique for finding the best set of hyperparameters for a given model. However, it may not be preferable to use in case of large datasets due to its computational and memory requirements. Other techniques such as random search may be more appropriate for large datasets.

15. List down some of the evaluation metric used to evaluate a regression model. Explain each of them in brief.

Ans:- Here are some commonly used evaluation metrics for regression models:

a. Mean Squared Error (MSE): This metric calculates the average squared difference between the predicted and actual values. It is one of the most popular metrics used in regression analysis, and the lower the value, the better the model. However, since it involves squaring the errors, it may give more weight to large errors.

b. Root Mean Squared Error (RMSE): RMSE is the square root of MSE and is also a popular evaluation metric for regression models. It gives an idea of how far the predicted values are from the actual values. It is similar to MSE but provides a more interpretable scale.

c. Mean Absolute Error (MAE): This metric calculates the absolute difference between the predicted and actual values and takes the average of those differences. It provides an idea of the magnitude of the errors in the predictions.

d. R-squared (R^2): R-squared measures how well the model fits the data by comparing the variance of the predicted values to the variance of the actual values. It ranges from 0 to 1, where 1 indicates a perfect fit and 0 indicates a model that does not fit the data at all. However, it can be misleading when used with complex models or models with many predictors.

e. Adjusted R-squared: Adjusted R-squared is a modified version of R-squared that accounts for the number of predictors in the model. It penalizes the addition of predictors that do not improve the model fit.

f. Mean Absolute Percentage Error (MAPE): MAPE measures the percentage difference between the predicted and actual values. It is useful for evaluating models that predict percentage changes or ratios.

g. Coefficient of Determination (COD): COD is a measure of how well the model fits the data. It ranges from 0 to 1, where 1 indicates a perfect fit and 0 indicates a model that does not fit the data at all. COD is a preferred metric when the regression model is quadratic or cubic.

h. Mean Absolute Scaled Error (MASE): MASE compares the mean absolute error of the forecast with the mean absolute error of a naive forecast. It is useful when the time series data contains seasonal patterns.

It is important to choose the appropriate evaluation metric based on the problem and the data at hand.

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: A

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:

```
num = int(input("Enter 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. Write a python program to find whether a number is prime or composite.

Ans:

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

if num > 1:

   for i in range(2, num):
      if (num % i) == 0:
         print(num, "is composite")
         break
   else:
      print(num, "is prime")
else:
        print(num, "is neither prime nor composite")
```

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

Ans:

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

# convert to lowercase
string = string.lower()

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

# Check if the string and its reverse are equal
if string == reverse_string:
   print("The string is a palindrome")
else:
        print("The string is not a palindrome")
```

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

Ans: ddfdgcfc

 # In case hypotenuse is given

```
side1 = float(input("Enter the length of the first side: "))
hypotenuse = float(input("Enter the length of the hypotenuse: "))

# Use the Pythagorean theorem to find the length of the hypotenuse

side2 =(hypotenuse**2 - side1**2)**0.5

print("The length of the side2 is:", side2)
```

 # In case hypotenuse is unknown side

```
side1 = float(input("Enter the length of the first side: "))
side2 = float(input("Enter the length of the second side: "))

# Use the Pythagorean theorem to find the length of the hypotenuse
hypotenuse =(side1**2 + side2**2)**0.5

print("The length of the hypotenuse is:", hypotenuse)
```

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

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

# Create an empty dictionary to store the frequency of each character
frequency = {}

# Iterate over each character in the string
for char in string:
    # If the character is already in the dictionary, increment its count by 1
    if char in frequency:
        frequency[char] += 1
    # Otherwise, add the character to the dictionary with a count of 1
    else:
        frequency[char] = 1

print("frequency of all characters :\n "+ str(frequency))
```

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

In [9]:
```
num = int(input("Enter 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)
```

```
Enter a number: 5
The factorial of 5 is 120
```

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

In [10]:
```
num = int(input("Enter a number: "))

if num > 1:
    # Check for factors
    for i in range(2, num):
        if (num % i) == 0:
            print(num, "is composite")
            break
    else:
        print(num, "is prime")
else:
    print(num, "is neither prime nor composite")
```

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

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

In [12]:
```
string = input("Enter a string: ")

# convert to lowercase or uppercase
string = string.lower()
## string = string.upper()

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

# Check if the string and its reverse are equal
if string == reverse_string:
    print("The string is a palindrome")
else:
    print("The string is not a palindrome")
```

```
Enter a string: taT
The string is a palindrome
```

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

In [18]:
```
# In case hypotenuse is unknown side
side1 = float(input("Enter the length of the first side: "))
side2 = float(input("Enter the length of the second side: "))

# Use the Pythagorean theorem to find the length of the hypotenuse
hypotenuse =(side1**2 + side2**2)**0.5

# Print the result
print("The length of the hypotenuse is:", hypotenuse)
```

```
Enter the length of the first side: 3
Enter the length of the second side: 4
The length of the hypotenuse is: 5.0
```

In [20]:
```
# In case hypotenuse is given
side1 = float(input("Enter the length of the first side: "))
hypotenuse = float(input("Enter the length of the hypotenuse: "))

# Use the Pythagorean theorem to find the length of the hypotenuse
side2 =(hypotenuse**2 - side1**2)**0.5

# Print the result
print("The length of the side2 is:", side2)
```

```
Enter the length of the first side: 4
Enter the length of the hypotenuse: 5
The length of the side2 is: 3.0
```

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

In [22]:
```
string = input("Enter a string: ")

# Create an empty dictionary to store the frequency of each character
frequency = {}

# Iterate over each character in the string
for char in string:
    # If the character is already in the dictionary, increment its count by 1
    if char in frequency:
        frequency[char] += 1
    # Otherwise, add the character to the dictionary with a count of 1
    else:
        frequency[char] = 1
print("Count of all characters :\n "
      + str(frequency))
```

```
Enter a string: lfjkjojgcradf
Count of all characters :
 {'l': 1, 'f': 2, 'j': 3, 'k': 1, 'o': 1, 'g': 1, 'c': 1, 'r': 1, 'a': 1, 'd': 1}
```

In [ ]:
```
```

STATISTICS WORKSHEET-8

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

1. In hypothesis testing, type II error is represented by β and the power of the test is 1−β then β is:
   a. The probability of rejecting H_0 when H_1 is true
   b. The probability of failing to reject H_0 when H_1 is true
   c. The probability of failing to reject H_1 when H_0 is true
   d. The probability of rejecting H_0 when H_1 is true
   Ans: b

2. In hypothesis testing, the hypothesis which is tentatively assumed to be true is called the
   a. correct hypothesis
   b. null hypothesis
   c. alternative hypothesis
   d. level of significance
   Ans: b

3. When the null hypothesis has been true, but the sample information has resulted in the rejection of the null, a _____has been made
   a. level of significance
   b. Type II error
   c. critical value
   d. Type I error
   Ans: d

4.For finding the p-value when the population standard deviation is unknown, if it is reasonable to assume that the population is normal, we use
   a. the z distribution
   b. the t distribution with n - 1 degrees of freedom
   c. the t distribution with n + 1 degrees of freedom
   d. none of the above
   Ans: b

5. A Type II error is the error of
   a. accepting Ho when it is false
   b. accepting Ho when it is true
   c. rejecting Ho when it is false
   d. rejecting Ho when it is true
   Ans: a

6. A hypothesis test in which rejection of the null hypothesis occurs for values of the point estimator in either tail of the sampling distribution is called
   a. the null hypothesis
   b. the alternative hypothesis
   c. a one-tailed test

d. a two-tailed test

Ans: d

7. In hypothesis testing, the level of significance is
a. the probability of committing a Type II error
b. the probability of committing a Type I error
c. the probability of either a Type I or Type II, depending on the hypothesis to be tested
d. none of the above
Ans: b

8. In hypothesis testing, b is
a. the probability of committing a Type II error
b. the probability of committing a Type I error
c. the probability of either a Type I or Type II, depending on the hypothesis to be test
d. none of the above
Ans: a

9. When testing the following hypotheses at an α level of significance
H_0: p = 0.7
H_1: p > 0.7
The null hypothesis will be rejected if the test statistic Z is
a. $z > z_\alpha$
b. $z < z_\alpha$
c. z < -z
d. none of the above
Ans: a

10. Which of the following does not need to be known in order to compute the P-value?
a. knowledge of whether the test is one-tailed or two-tail
b. the value of the test statistic
c. the level of significance
d. All of the above are needed
Ans: c

11. The maximum probability of a Type I error that the decision maker will tolerate is called the
a. level of significance
b. critical value
c. decision value
d. probability value
Ans: a

12. For t distribution, increasing the sample size, the effect will be on
a. Degrees of Freedom
b. The t-ratio
c. Standard Error of the Means
d. All of the Above
Ans: d

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

13. What is Anova in SPSS?

Ans:

ANOVA (Analysis of Variance) is a statistical technique used to compare the means of three or more groups of data. It is used to test whether the means of the groups are significantly different from each other, and it can be used to identify which specific groups are different from each other.

In SPSS, ANOVA can be performed using the "General Linear Model" menu. This menu provides options for performing one-way ANOVA, factorial ANOVA, repeated measures ANOVA, and ANCOVA (Analysis of Covariance). The user can specify the variables to be used in the analysis, the grouping variables, and the dependent variables.

After performing the ANOVA analysis in SPSS, the output provides several tables of results, including the sum of squares, degrees of freedom, mean square, F-value, and p-value. These results can be used to determine whether there are significant differences between the groups, and to identify which specific groups are different from each other.

14. What are the assumptions of Anova?

Ans:

The assumptions of ANOVA (Analysis of Variance) are as follows:

1. Independence: The observations within each group are independent of each other. There should be no correlation between the observations within each group.
2. Normality: The distribution of the dependent variable should be approximately normal within each group. This assumption can be checked using a histogram or normal probability plot.
3. Homogeneity of variances: The variances of the dependent variable should be equal across all groups. This assumption can be checked using a statistical test such as Levene's test.
4. Random sampling: The observations should be sampled randomly from the population. This assumption ensures that the sample is representative of the population.

15. What is the difference between one way Anova and two way Anova?

Ans:

The main difference between one-way ANOVA and two-way ANOVA is the number of independent variables (factors) being considered in the analysis.

One-way ANOVA is used to test for differences in the means of a single dependent variable across multiple independent groups (or levels of a single independent variable). In other words, it examines the effect of one categorical variable on a continuous variable.

Two-way ANOVA, on the other hand, examines the effect of two categorical independent variables (factors) on a single continuous dependent variable. It tests for main effects of each factor, as well as any interaction between the two factors.

In simple terms, one-way ANOVA compares means across different groups while controlling for only one variable, whereas two-way ANOVA compares means across different groups while controlling for two variables.

Additionally, in one-way ANOVA, there is only one factor to be tested, while in two-way ANOVA, there are two factors that can be tested individually or together. The interaction effect between the two factors can be analyzed to understand if the effect of one factor varies across the levels of the other factor.