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:

Ans-b. The probability of failing to reject H_0 when H_1 is true

2. In hypothesis testing, the hypothesis which is tentatively assumed to be true is called the

Ans-b. null hypothesis

3. When the null hypothesis has been true, but the sample information has resulted in the rejection of the null, a
_____ has been made

Ans-d. Type I error

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

c. the t distribution with n + 1 degrees of freedom

5. A Type II error is the error of
Ans-a. accepting Ho when it is false

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
Ans-d. a two-tailed test

7. In hypothesis testing, the level of significance is
Ans-d. none of the above

8. In hypothesis testing, b is
Ans-d. none of the above

9. When testing the following hypotheses at an α level of significance
Ans-a. $z > z_\alpha$

10. Which of the following does not need to be known in order to compute the P-value?
Ans-c. the level of significance

11. The maximum probability of a Type I error that the decision maker will tolerate is called the
a. level of significance

12. For t distribution, increasing the sample size, the effect will be on
d. All of the Above

13. What is Anova in SPSS?
Ans-Analysis of Variance, i.e. ANOVA in SPSS, is used for examining the differences in the mean values of the dependent variable associated with the effect of the controlled independent variables, after taking into account the influence of the uncontrolled independent variables.

14. What are the assumptions of Anova?

Ans-There are three primary assumptions in ANOVA: The responses for each factor level have a normal population distribution. These distributions have the same variance. The data are independent.

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

Ans-The only difference between one-way and two-way ANOVA is the number of independent variables. A one-way ANOVA has one independent variable, while a two-way ANOVA has two.

MACHINE LEARNING

ASSIGNMENT - 8

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?

Ans-D) None of these

2. Which of the following hyper parameter(s), when increased may cause random forest to over fit the data?
Ans-A) max_depth

3. Which of the following is the least preferable resampling method in handling imbalance datasets?
Ans-A) SMOTE

4. Which of the following statements is/are true about “Type-1” and “Type-2” errors?
Ans-C) 1 and 3

5. Arrange the steps of k-means algorithm in the order in which they occur:
Ans-D) 1-3-2

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?
Ans-B) Support Vector Machines

7. What is the main difference between CART (Classification and Regression Trees) and CHAID (Chi Square Automatic Interaction Detection) Trees?
Ans-D) None of the above

8. In Ridge and Lasso regularization if you take a large value of regularization constant(lambda), which
of the following things may occur?

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

9. Which of the following methods can be used to treat two multi-collinear features?
Ans-A) remove both features from the dataset

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?
Ans-A) Overfitting

11. In which situation One-hot encoding must be avoided? Which encoding technique can be used in such a case?
Ans-One-hot encoding creates d-dimensional vectors for each instance where d is the unique number of feature values in the dataset. **For a feature having a large number of unique feature values or categories**, one-hot encoding is not a great choice. Target Encoding is also known as the mean encoding is a popular technique used heavily by Kagglers, that represents a categorical variable into a numerical vector of a single

dimension. Each of the categories is the variable is replaced with the mean target value for that category.

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

Ans-. Resampling (Oversampling and Undersampling)

This technique is used to upsample or downsample the minority or majority class. When we are using an imbalanced dataset, we can oversample the minority class using replacement. This technique is called oversampling. Similarly, we can randomly delete rows from the majority class to match them with the minority class which is called undersampling. After sampling the data we can get a balanced dataset for both majority and minority classes. So, when both classes have a similar number of records present in the dataset, we can assume that the classifier will give equal importance to both classes.

13. What is the difference between SMOTE and ADASYN sampling techniques?
Ans-The key difference between ADASYN and SMOTE is that the former uses a density distribution, as a criterion to automatically decide the number of synthetic samples that must be generated for each minority sample by adaptively changing the weights of the different minority samples to compensate for the skewed distributions.

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 for **finding the optimal parameter values from a given set of parameters in a grid**. It's essentially a cross-validation technique. The model as well as the parameters must be entered. After extracting the best parameter values, predictions are made. For a large size dataset, Grid Search CV time complexity increases exponentially, and hence it's not practically feasible.

15. List down some of the evaluation metric used to evaluate a regression model. Explain each of them in brief.
Ans- There are three error metrics that are commonly used for evaluating and reporting the performance of a regression model; they are: Mean Squared Error (MSE). Root Mean Squared Error (RMSE). Mean Absolute Error (MAE).

Mean Squared Error, or MSE for short, is a popular error metric for regression problems. It is also an important loss function for algorithms fit or optimized using the least squares framing of a regression problem. Here "*least squares*" refers to minimizing the mean squared error between predictions and expected values.
The MSE is calculated as the mean or average of the squared differences between predicted and expected target values in a dataset.

The Root Mean Squared Error, or RMSE, is an extension of the mean squared error.

Importantly, the square root of the error is calculated, which means that the units of the RMSE are the same as the original units of the target value that is being predicted.

For example, if your target variable has the units "*dollars*," then the RMSE error score will also have the unit "*dollars*" and not "*squared dollars*" like the MSE.
As such, it may be common to use MSE loss to train a regression predictive model, and to use RMSE to evaluate and report its performance.

Mean Absolute Error, or MAE, is a popular metric because, like RMSE, the units of the error score match the units of the target value that is being predicted.
Unlike the RMSE, the changes in MAE are linear and therefore intuitive.

That is, MSE and RMSE punish larger errors more than smaller errors, inflating or magnifying the mean error score. This is due to the square of the error value. The MAE does not give more or less weight to different types of errors and instead the scores increase linearly with increases in error.

As its name suggests, the MAE score is calculated as the average of the absolute error values. Absolute or *abs()* is a mathematical function that simply makes a number positive. Therefore, the difference between an expected and predicted value may be positive or negative and is forced to be positive when calculating the MAE

STATISTICS WORKSHEET-7

1. A die is thrown 1402 times. The frequencies for the outcomes 1, 2, 3, 4, 5 and 6 are given in the following
table:

Outcome	1	2	3	4	5	6
Frequency	400	300	157	180	175	190

Find the probability of getting 6 as outcome:

Ans-b) 0.135

2. What will be the probability of getting a digit with unit place digit odd number that is 1, 3,5,7,9?

Ans-d) 0.53

3. If we buy a new tyre of this company, what is the probability that the tyre will last more than 9000

miles?

Ans-c) 0.745

4. Please refer to the case and table given in the question No. 3 and determine what is the probability that if we buy
a new tyre then it will last in the interval [4000-14000] miles?

Ans-b) 0.577

5. We have a box containing cards numbered from 0 to 9. We draw a card randomly from the box. If it is told to
you that the card drawn is greater than 4 what is the probability that the card is odd?

Ans-c) 0.6

6. We have a box containing cards numbered from 1 to 8. We draw a card randomly from the box. If it is told to
you that the card drawn is less than 4 what is the probability that the card is even?

Ans-a) 0.33

7. A die is thrown twice and the sum of the numbers appearing is observed to be 7. What is the conditional
probability that the number 6 has appeared at least on one of the die?

Ans-c) 0.33

8. Consider the experiment of tossing a coin. If the coin shows tail, toss it again but if it shows head, then throw a
die. Find the conditional probability of the event that ‘the die shows a number greater than 4’ given that ‘there is
at least one Head’.

Ans-b) 0.22

9. There are three persons Evan, Ross and Michelle. These people lined up randomly for a picture. What is the
probability of Ross being at one of the ends of the line?

Ans-a) 0.66

10. Let us make an assumption that each born child is equally likely to be a boy or a girl. Now suppose, if a family
has two children, what is the conditional probability that both are girls given that at least one of them is a girl?

Ans-a) 0.33

11. Consider the same case as in the question no. 10. It is given that elder child is a boy. What is the conditional
probability that both children are boys?

Ans-a) 0.33

12. We toss a coin. If we get head, we toss a coin again and if we get tail we throw a die. What is the probability of
getting a number greater than 4 on die?

Ans-a) 0.166

13. We toss a coin. If we get head, we toss a coin again and if we get tail we throw a die. What is the probability of
getting an odd number on die?

Ans-a) 0.345

14. Suppose we throw two dice together. What is the conditional probability of getting sum of two numbers found on
the two die after throwing is less than 4, provided that the two numbers found on the two die are different?

Ans-d) 0.06

15. A box contains three coins: two regular coins and one fake two-headed coin, you pick a coin at random and toss
it. What is the probability that it lands heads up?

Ans-a) 1/3

WORKSHEET 7 SQL

Q1 and Q2 have one or more correct answer. Choose all the correct option to answer your question.

1. The primary key is selected from the
Ans-B. Candidate keys

2. Which is/are correct statements about primary key of a table?

Ans-B. Primary keys cannot contain NULL values...

3. Which SQL command is used to insert a row in a table?

Ans-C. Insert

4. Which one of the following sorts rows in SQL?

Ans-C. ORDERBY

5. The SQL statement that queries or reads data from a table is

Ans-C. SELECT

6. Which normal form is considered adequate for relational database design?

Ans-C. 3NF

7. SQL can be used to

Ans-C. All of the above can be done by SQL

8. SQL query and modification commands make up

Ans-B. DML

9. The result of a SQL SELECT statement is a(n)

Ans-B. Table

10. Second normal form should meet all the rules for

Ans-B. 2 NF

11. What are joins in SQL?

JOINS in SQL are commands which are used to combine rows from two or more tables, based on a related column between those tables. There are predominantly used when a user is trying to extract data from tables which have one-to-many or many-to-many relationships between them.

12. What are the different types of joins in SQL?

Ans-Different Types of SQL JOINs

- (INNER) JOIN : Returns records that have matching values in both tables.
- LEFT (OUTER) JOIN : Returns all records from the left table, and the matched records from the right table.
- RIGHT (OUTER) JOIN : Returns all records from the right table, and the matched records from the left table.

13. What is SQL Server?

Ans-SQL stands for Structured Query Language, a language for manipulating and talking about data in databases. It first came into use in 1970 and became a standard in 1986 by IBM in conjunction with several projects for the US government and for many years it remained a government-only project. It's a programing language which is used to access data stored in a database.

14. What is primary key in SQL?

Ans-The primary key in SQL is a single, or a group of fields or columns that can uniquely identify a row in a table. Putting it simply, it is a column that accepts unique values for each row.

15. What is ETL in SQL?

Ans-ETL, which stands for "extract, transform, load," are the three processes that, in combination, move data from one database, multiple databases, or other sources to a unified repository—typically a data warehouse.

MACHINE LEARNING

1. Which of the following in sk-learn library is used for hyper parameter tuning?

Ans-A) GridSearchCV()

2. In which of the below ensemble techniques trees are trained in parallel?

Ans-A) Random forest

3. In machine learning, if in the below line of code:
*sklearn.svm.**SVC** (C=1.0, kernel='rbf', degree=3)*
we increasing the C hyper parameter, what will happen?

Ans-A) The regularization will increase

4. Check the below line of code and answer the following questions:
*sklearn.tree.**DecisionTreeClassifier**(*criterion='gini',splitter='best',max_depth=None, min_samples_split=2)*

Ans-C) both A & B

5. Which of the following is true regarding Random Forests?

Ans-D)None of the above

6. What can be the disadvantage if the learning rate is very high in gradient descent?

Ans-D) None of them

7. As the model complexity increases, what will happen?

Ans-B) Bias will decrease, Variance increase

8. Suppose I have a linear regression model which is performing as follows:
Train accuracy=0.95 and Test accuracy=0.75
Which of the following is true regarding the model?

Ans-C) model is performing good

9. Suppose we have a dataset which have two classes A and B. The percentage of class A is 40% and 60?
percentage of class B is 60%. Calcualate the Gini index and entropy of the dataset.

Ans-The Gini Index wil be A/A+B = 40/40+60= 2/5/2/5+3/5 = 2/5 and the entropy will be 2400.

10. What are the advantages of Random Forests over Decision Tree?

Ans-Random forest algorithm avoids and prevents overfitting by using multiple trees. The results are not accurate. This gives accurate and precise results. Decision trees require low computation, thus reducing time to implement and carrying low accuracy.

11. What is the need of scaling all numerical features in a dataset? Name any two techniques used for scaling.
Ans-The most common techniques of feature scaling are Normalization and Standardization. Normalization is used when we want to bound our values between two numbers, typically, between [0,1] or [-1,1]. While Standardization transforms the data to have zero mean and a variance of 1, they make our data unitless.

12. Write down some advantages which scaling provides in optimization using gradient descent algorithm.
Ans-Gradient descent is an optimization algorithm which is commonly-used to train machine learning models and neural networks. **Training data helps these models learn over time**, and the cost function within gradient descent specifically acts as a barometer, gauging its accuracy with each iteration of parameter updates.

13. In case of a highly imbalanced dataset for a classification problem, is accuracy a good metric to
measure the performance of the model. If not, why?
Ans-So, accuracy does not holds good for imbalanced data. In business scenarios, most data won't be balanced and so accuracy becomes poor measure of evaluation for our classification model.

14. What is “f-score" metric? Write its mathematical formula.
Ans-F-Measure or F-Score provides a way to combine both precision and recall into a single measure that captures both properties. This is the harmonic mean of the two fractions. The result is a value between 0.0 for the worst F-measure and 1.0 for a perfect F-measure. The traditional F measure is calculated as follows: **F-Measure = (2 * Precision * Recall) / (Precision + Recall)**

15. What is the difference between fit(), transform() and fit_transform()?
Ans-The fit() method helps in fitting the data into a model, transform() method helps in transforming the data into a form that is more suitable for the model. Fit_transform() method, on the other hand, combines the functionalities of both fit() and transform() methods in one step.