WORKSHEET-1

ANSWER

Python part.

ANSWER. 1

'%' is the operator used to calculate remainder in a division.

ANSWER.2

2//3 is equal to 0.

ANSWER.3

6<<2 is equal to 24.

ANSWER.4

6&2 is equal to 2.

ANSWER.5

6|2 will give 6 as an output.

ANSWER.6

Finally, keyword in python denotes that the finally block will be executed no matter if the try block raises an error or not. (Option c)

ANSWER.7

RAISE, keyword in python is used to raise an exception. (Option A)

ANSWER.8

A common use case of yield keyword in python is in defining a generator.

(Option C)

ANSWER.9

_abc and abc2 are the valid variable names. (Option A and C)

ANSWER.10

Yield and Raise are the keywords in python. (Option A and B)

ANSWER 11 TO 15 - are in the jupyter notebook.

MACHINE LEARNING PART

ANSWER 1.

Least Square Error is the method we use to find the best fit line for data in Linear Regression. (Option A)

ANSWER 2.

Linear regression is sensitive to outliers. (Option A)

ANSWER 3.

A line falls from left to right if a slope is negative. (Option B)

ANSWER 4.

Correlation Will have symmetric relation between dependent variable and independent variable. (Option B)

ANSWER5.

Low bias and high variance is the reason for over fitting condition. (Option C)

ANSWER 6.

If output involves label then that model is called Predictive model. (Option B)

ANSWER 7.

Lasso and Ridge regression techniques belong to Regularization. (Option D)

ANSWER 8.

To overcome with imbalance dataset SMOTE technique can be used. (Option D)

ANSWER 9.

The AUC Receiver Operator Characteristic (AUCROC) curve is an evaluation metric for binary classification problems. It uses TPR and FPR to make graph.

(Option A)

ANSWER 10.

FALSE.

ANSWER 11.

Construction bag of words from an email. (Option A)

ANSWER 12.

(Option B) -It becomes slow when number of features is very large.

(Option D)- It does not make use of dependent variable.

ANSWER 13.

It is one of the most important concepts of machine learning. 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. The main techniques of regularization are: -

1. RIDGE REGRESSION.
2. LASSO REGRESSION.

Ridge regression is **a model tuning method that is used to analyze any data that suffers from multicollinearity**. This method performs L2 regularization. When the issue of multicollinearity occurs, least-squares are unbiased, and variances are large, this results in predicted values being far away from the actual values.

Lasso regression is another variant of the regularization technique used to reduce the complexity of the model. It stands for **Least Absolute shrinkage and Selection Operator**.

ANSWER 14.

There are three particular algorithms that are used for regularization these three are: -

1. RIDGE REGRESSION.
2. LASSO REGRESSION.
3. Elastic-net regression

- Ridge regression

Ridge regression is **a model tuning method that is used to analyze any data that suffers from multicollinearity**. This method performs L2 regularization. When the issue of multicollinearity occurs, least-squares are unbiased, and

variances are large, this results in predicted values being far away from the actual values.

- Lasso regression

It is another variant of the regularization technique used to reduce the complexity of the model. It stands for **Least Absolute shrinkage** and **Selection Operator**. It is a regression analysis method that performs both variable selection and regularization in order to enhance the prediction accuracy and interpretability of the resulting statistical model.

- Elastic-net regression

Elastic net linear regression uses the penalties from both lasso and ridge techniques to regularize regression models. The technique combines both the lasso and ridge regression methods by learning from their shortcomings to improve the regularization of statistical models.

ANSWER 15.

Within a linear regression model tracking a stock's price over time, the error term is the difference between the expected price at a particular time and the price that was actually observed. When using regression, an error is also called as an intercept.

STATISTICS PART

ANSWER 1.

TRUE

ANSWER 2.

Central Limit Theorem - States that the distribution of averages of iid variables, properly normalized, becomes that of a standard normal as the sample size increases. (Option A)

ANSWER 3.

Modelling bounded count data is incorrect with respect to use of Poisson distribution. (Option B)

ANSWER 4.

All of the mentioned. (Option D)

ANSWER 5.

Poisson random variables are used to model rates. (Option C)

ANSWER 6.

False.

ANSWER 7.

Hypothesis testing is concerned with making decisions using data. (Option B)

ANSWER 8.

Normalized data are centred at 0 and have units equal to standard deviations of the original data. (Option A)

ANSWER 9.

Outliers cannot conform to the regression relationship- This statement is incorrect with respect to outliers. (Option C)

ANSWER 10.

Normal distribution, also known as the gaussian distribution is a probability distribution that is symmetric about the mean, showing that data near the mean are more frequent in occurrence than data far from the mean. In graphical from, the normal distribution appears as a “bell curve”. A normal distribution is perfectly symmetrical around its centre. That is, right side of the centre is a minor image of the left side. There is also only one mode, or peak, in a normal distribution.

ANSWER 11.

Best techniques to handle missing data

1. Use deletion methods to eliminate missing data. The deletion methods only work for certain database where participants have missing fields.
2. Use regression analysis to systematically eliminate data.
3. Data scientists can use data imputation techniques.

Imputation techniques like: -

- Deletions. Pairwise deletion. Listwise deletion/dropping rows. Dropping complete columns.
- Basic imputation techniques. Imputation with a constant value. Imputation using the statistics (mean, median, mode)
- K-Nearest neighbour imputation

This could be used.

ANSWER 12.

A/B testing in its simplest sense is an example on two variants to see which performs better based on a given metric. Typically, two consumer groups are exposed to two different versions of the same thing to see if there is a significant difference in metrics like sessions, click-through rate, and/or conversions.

ANSWER 13.

Mean imputation is typically considered terrible practice since it ignores feature Correlation. Second, mean imputation decreases the variance of our data while increasing bias. As a result of the reduced variance, the model is less accurate and the confidence interval is narrower.

ANSWER 14.

In statistics, linear regression is a linear approach for modelling the relationship between a scaler response and one or more explanatory variables (also known as dependent and independent variables).

ANSWER 15.

There are three real branches of statistics: Data Collection, Descriptive Statistics and Inferential Statistics.

Data collection –

In statistics, data collection is a process of gathering information from all the relevant sources to find a solution to the research problem. It helps to evaluate the outcome of the problem. The data collection methods allow a person to conclude an answer to the relevant question.

Descriptive statistics-

Descriptive Statistics is summarizing the data at hand through certain numbers like mean, median etc. so as to make the understanding of the data easier. It does not involve any generalization or inference beyond what is available. This

means that the descriptive statistics are just the representation of the data (sample) available and not based on any theory of probability.

Inferential statistics-

Inferential statistics are often used to compare the differences between the treatment groups. Inferential statistics use measurement from the sample of subjects in the experiments to compare the treatment groups and make generalization about the larger population of subjects.

MADE BY- KIRTARATH SINGH SAINI