ASSIGNMENT ON PYTHON – WORKSHEET 1

The answers are for the questions given are as follows:

|  |  |
| --- | --- |
| Question | Answer |
| 1 | C |
| 2 | B |
| 3 | C |
| 4 | A |
| 5 | D |
| 6 | C |
| 7 | A |
| 8 | C |
| 9 | A, C |
| 10 | A, B |

Please See answers to **“Assignment on Machine Learning 1”**

ASSIGNMENT ON MACHINE LEARNING

The answers are to the questions given are as follows:

|  |  |
| --- | --- |
| Question | Answer |
| 1 | A |
| 2 | A |
| 3 | A |
| 4 | C |
| 5 | C |
| 6 | A |
| 7 | B |
| 8 | D |
| 9 | A |
| 10 | B |
| 11 | B |
| 12 | A and B |

13. Regularization is a machine technique used to reduce the number of errors (outliers or noise) that maybe present in a dataset, and this is done by fitting the functions accurately. By doing this, the machine checks on overfitting, underfitting, bias and variance. The checking of the above is to give a good balance as an output where both bias and variance turns out to be low.

14. The algorithms that used in regularization include are

1. Least Absolute Shrinkage and Selection Operator (Lasso)

2. Ridge Regression L1

3. Elastic Net Regression which combines both L1 and L2 to find the ratio of both.

15. The explanation of error in linear regression equation refers to the sum of deviations which in the regression line, which provides on explanation for the difference between the theoretical value of m0del and the actual observed result. To make it simpler, this error term explains why all the y values do not lie perfectly on the regression line.

Please See answers to **“Assignment on Statistics Worksheet 1”**

Assignment on Statistics Worksheet 1

The answers are to the questions given are as follows:

|  |  |
| --- | --- |
| Question | Answer |
| 1 | A |
| 2 | A |
| 3 | C |
| 4 | A |
| 5 | C |
| 6 | B |
| 7 | B |
| 8 | A |
| 9 | C |

# 10

Normal distribution refers the distribution of statistics that display data in a continuous manner, symmetrical around its mean, with most observations clustering around the central apex, while other values of the same sample are seen distributed further away from the normal on both sides of the mean left and right. Because of it evenness and normal distribution nature, graphical representation often has an admirable bell-shape. Its importance is mostly seen when engaged in statistical analysis of both independent and random variables.

11.

Handling of missing data is a serious issue that needs absolute carefulness. Therefore, it’s important to know that there’s no good way to handle missing data. The essence of Data Cleaning/Exploratory Analysis is to take care of the missing values. Hence, removing or replacing data in any given dataset calls for carefulness to avoid deleting necessary information especially where the dataset is small.

Principally, I’ll recommend two imputation techniques to include:

1.Deletion which can be done in three ways via deleting the rows (listwise deletion), column and pairwise deletion

2. Imputation that deals with both general and time series problems.

12.

In the field of statistics, A/B tests is a random experiment that mostly involves two possible variants say (A and B), although not limited, as this can also be handled where multiple variants of the same variable applied. To conduct a statistical analysis in a given data as in (random experiment), there’s need for the application of statistical hypothesis testing of the two-samples (A and B) where both hypotheses are subjected ***Ho* null and *H*a** alternative hypothesis testing.

13.

Somehow, imputation of mean to replace the missing data can be done as this has been a technique engaged in a wider scale. However, it comes with a lot of problems or difficulties. Because of the biasness of the mean-imputed variable variance, the downward display from the variance of the un-imputed variable will affect the dataset since the bias will equally affect standard errors, confidence intervals, inferential statistics and many more.

However, imputation of mean to replace missing values, need to be considered critically based on the sample nature. For instance, this can be done where:

1. in a dataset you have variables that does not contain extreme values, then the use of **mean imputation** can be employed.
2. Using the **Mean Variance** where the dataset contains extreme values.
3. The use of **Mode importation** where in a dataset the there is a repetition of values.

If these three approaches are engaged, there’s likelihood that errors can be reduced to the bearable level when replacing missing values.

14.

In statistics, Simple linear regression is the use of a statistical technique to make an estimate in a relationship involving two quantifiable variables (x and y) for a possibleoutcome**.** In observing the week or strong relationship between these two-variables knowing fully well that one is dependent on another, for instance:

1. The amount of salary end versus happiness gain.
2. the number of persons that will come to the stadium to watch a match on a raining day.
3. Air travel versus cost
4. Number of employments in and the job vacancies.
5. Attention to patients versus the number of available doctors.

Therefore, the use simple linear regression can answer the above scenario with possible outcome.

15.

The two major branches of statistics are Descriptive and Inferential Statistics. Descriptive Statistics is the presentation and collection of data, and statistical analysis of the data. Inferential statistics studies a statistical sample and, through analysis, can say something about the population from which the sample is drown from.

When we talk about the various branches of statistics, we are referring to the branches of statistics. To say this right, statistics has two main branches which are:

1. ***Descriptive statistics:*** that have to do with the collection and organization of data, summarize collected data and measure the mean, standard deviation, rates count, percentages and range. In a simpler form, it will describe the data but will not generalize it beyond the data itself.

2. ***Inferential statistics***: on the other hand takes care of organization of data from sample state to population level. It’s in the inferential level that hypotheses testing is done to ascertain the level of relationships that exist between variables at the same time making predictions for possible outcome. Here, the involvement of numerical methods is engaged to proof that a research data can support a hypotheses or not which will explain the outcome as true or an occurrence that is due to chance.