Glenn Nigel Ebenezer- Worksheet1 –Internship 31

Assigned on 14-08-2022 04:56:26 pm **Last Date** Sunday, 21-08-2022, at 11:59 PM

I Python Worksheet-1

MCQs 1-10

Python Worksheet1	
Q#	Choice
1	C
2	B
3	C
4	A
5	D
6	C
7	A
8	C
9	A,B,C
10	A,B

Q11 to Q13 in python.

Link: https://github.com/glennnigel/internship/blob/main/Python%20Worksheet1-%20Q11-Q15.ipynb

II STATISTICS WORKSHEET-1

MCQ 1-9

STATISTICS WORKSHEET-1	
Q#	Choice
1	A
2	A
3	B
4	D
5	C
6	B
7	A
8	A
9	C

10. What do you understand by the term Normal Distribution?

A: Starting with an example, the distribution of marks among students in a class, while there maybe a few toppers, and a few students who have underperformed, there will be a rather large chuck of students who have scored average, some below and above average, if these results were plotted on a graph, marks against students or their student ids. One would find a bell curve. Where the data also known as samples are distributed in a shape conforming a bell curve, where there is ideally a symmetry between the RHS and LHS

Hence, Normal distribution, also known as the Gaussian distribution or Bell 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.

11. How do you handle missing data? What imputation techniques do you recommend?

A: Just like having incorrect data, having missing data can also negatively impact data analysis. It is vital that the data is handled properly to ensure correct analysis and its inferences. Data imputation is one such procedure - it is the process of filling in missing values based on other data. Although not authentic data it is the next best thing.

There are several methods to Data Imputation -

- deleting entire observations containing one or more unknown values if we have enough observations/samples to go by
- Replacing unknown values with the most frequent values
- Replacing unknown values by exploring correlations
- Replacing unknown values by exploring similarities between cases

Data Imputation works best when the right kind of imputation as listed above is used. This ofcourse is left to the judgement of the Data Scientist. Different situations require one or more or of these imputation techniques. The quality of the analysis and decision is purely dependant on Data scientists' judgement in handling missing data.

12. What is A/B testing?

A: Starting with an example I use a lot hailing from the Audio industry, if I were to make changes to a track, id apply an effect, like an EQ (equaliser) where I have randomly boosted the lower frequency, based on this, I compare the track with the effect "B" with the track with the effect on BYPASS as "A", comparing A/B, it's easy to make a decision as to which is better and tweak if necessary for optimal results.

Similarly, in context with Data science, A/B testing is a random control experiment. It is a way to compare the two versions of a variable to find out which performs better in a controlled environment.

13. Is mean imputation of missing data acceptable practice?

A: While it is an imputation technique, it can be detrimental to the study, as it merely maintains the existing overall mean, and does very little to keep the estimation real. While infact over/underestimates the standanrd deviation. Hence, while it is used, it can be bad practise in general, but can be used when the risk and stakes are very low.

14. What is linear regression in statistics?

A: When data prediction/estimation is required, estimation of coefficients of the linear variable is made involving one or more independent variables to best predict the value of the dependant variable. Where the variable we want to predict is called the dependent variable. And the variable we use to predict the other variable's value is called the independent variable. It employs a method of fitting the dependant variable in a straight line such that it reduces the difference between the predicted and actual values.

15. What are the various branches of statistics?

The Following are some of the branches in the field of statistics.

- Econometric
- Actuarial
- Psychometrics
- Physics Statistics
- Population Statistics
- Official Statistics
- Biostatistics
- Industrial Statistics
- Computing statistics

III MACHINE LEARNING – Assignment 39

MCQ 1-11

MACHINE LEARNING	
Assignment-39	
Q#	Choice
1	A
2	A
3	B
4	B
5	A
6	A
7	D
8	D
9	A
10	B
11	B
12	A,B,C

13. Explain the term regularization?

A: 'Regularisation' as obvious as it seems, is used to regularise data and make it acceptable to train. It is done by reducing the error margin by fitting an appropriate function on training data, such that no over fitting is done.

14. Which particular algorithms are used for regularization?

A:

- Ridge Regression
- Dropout
- Lasso regression
- Regularisation

15. Explain the term error present in linear regression equation?

A: Considering the Linear Regression model has been given it will give us an expected value for a certain set of features in data. The difference between the expected and the actual value is defined on some exogenous factor, this exogenous factor is often termed as error term.