Worksheet_set_1

MCQ PART

First assignment:-Machine learning	Second assignment:-statistics	Third assignment:-python
Ans1) A	Ans1) A	Ans1) C
Ans2) A	Ans2) A	Ans2) B
ANS3) B	Ans3) B	Ans3) C
Ans4) C	Ans4) D	Ans4) A
Ans5) C	Ans5) C	Ans5) D
Ans6) B	Ans6) B	Ans6) C
Ans7) D	Ans7) B	Ans7) A
Ans8) D	Ans8) A	Ans8) C
Ans9) A	Ans9) C	Ans9) A, B and C
Ans10) B		Ans10) A and B
Ans11) B		
Ans12) A,B and C		

Theory part

First assignment:-machine learning

Ans13) Regularization means the dataset which is used by the machine to learn for the process of training and testing is overfitted or underfitted ,to overcome the such problem ,the data scientist has to give some additional information to the dataset for the making the complex model into simpler one and also to shrink the complexities of independent .It is a technique for overcoming the problem of overfitting and underfitting.Overfitting is where the machine tries to learn every single data thus it has a high variance and underfitting is where the machine not able to learn but give random output thus it has high bias.

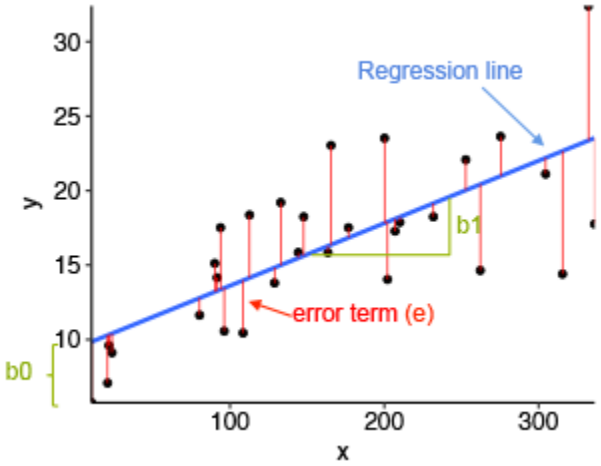
Ans14) The particular algorithms used for regularization are:-

1)ridge regression means to analyse the data which is highly related to each other or multi collinearity.its add a l2 penalty to its loss as equal to the coefficients magnitude.it shrinks the coefficients to reduce the complexity and multi coliinearity of model.

2) lasso regression means it is regression analysis which performs both selecting of feature and regularization for making better accuracy prediction.its add a penalty of l1 to its loss as equal to magnitude of coefficients.it also converts the coefficients of less important feature to zero and also shrinks the rest coefficient for the model complexity.

3)elastic net regression means it is a combination of both lasso and ridge regression such as l1 and l2.

Ans15) Error is the difference between the actual value and Predicted value and the goal is to reduce this difference.



The vertical distance between the data point and the regression line is known as error or residual**.** Each data point has one residual and the sum of all the differences is known as the Sum of Residuals/Errors.

Mathematical Approach:

Residual/Error = Actual values – Predicted Values

Sum of Residuals/Errors = Sum(Actual- Predicted Values)

Square of Sum of Residuals/Errors = (Sum(Actual- Predicted Values))^2

Second assignment:-statistics

Ans10) Normal distribution is known as the Gaussian distribution as it is probability distribution that is symmetric to the mean.the data which is near to mean is more frequent than the data far away from mean.the graph form is like bell shape. the proportion of outcomes of a particular kind in a large number of independent repetitions of an experiment in which the probabilities remain constant from trial to trial

Ans11) The best way to handle the missing data is :-

To delete the rows or columns which have missing values but that missing values should not be much important and delete the missing data is also suitable when there is a large amount of data with you on a particular matter.

Try to fill those missing values with the predicted values on the basis of data.

Try to use the mean average and median of the following row or column which has missing values and thus put the average mean & median of it in the missing values.

The techniques are:-

Dropping the values with null

Dropping features with high nullity

Implementing of model such as bagging, boosting, voting classifier etc

Implementing of average mean and median values to missing data

Ans12) A/B testing is the way of machine to learn about the data input and output provided in the dataset.The machine is giving a percentage of the dataset to learn about input and the output given accordingly by input values.the dataset is divided into the proportion in which some of the data is given for training purpose and rest is given for testing purpose.the A/B testing is the way in which the machine learns and thus performs testing part according to the learned part .in other words the machine is given the part of data to learn in which there is both input and output ,after learning process the machine has to perform the testing part in which the machine is given the only input of the data and the machine gives output according to the predicted output which he learned in the training phase.

Ans13) yes ,mean imputation of missing data is acceptable practice as sometimes there is less amount of data with the data scientist to examine and there are some missing values too in the data which is an obstacle for the data scientist to interpret the best outcome of the data.the data scientist can remove or delete the rows or columns containing missing values in the data but there is less amount of data and less amount of data means the missing values are important for the accuracy so the scientist will come to solution of calculating the average mean of the rows or columns containing missing values and thus fill the missing values with the value got from mean .this solution will keep the data accuracy and thus help for better examine of the data by the data scientist.

Ans14) in statistics , the linear regression is the graphical presentation of the relationship of the dependent variables and independent variables.the linear regression is the best modelling technique for the verifying the relation among the dependent and independent variables.its help to know about the nature of independent variables and dependent variables such as how much the dependent variables depend on the independent variables.the graphical presentation of the linear regression is line from getting left to right in upward direction .the best fit line is the line touches the maximum data possible.

Ans15) There are only two branches of statistics that are:-

1) Descriptive statistics:-this branch focuses on collecting, summarizing and presenting a set of data. it deals with presentation and collection of data. It is not simple, needs to be aware of designing experiments, choosing the right focus group and avoid biases that are so easy to create problem into the experiment.

2) Inferential statistics:-this branch analyses the data and make conclusions on it. It involves taking the right conclusions from the statistical analysis that has been performed using descriptive statistics. In the end, it is the inferences that make studies important and this aspect is dealt with in inferential statistics.