Python –Worksheet 1 Answers:

1) C-%, 2) B-0

3) C-24 4) A-2

5) D-6,

6) C-the finally block will be executed no matter if the try block raises an error or not

7) A-It is used to raise an exception.

8) C-in defining a generator

9) B-1abc

10) D-All of the above

Question 9 is wrongly framed. Instead of valid variable names, it should be invalid variable names.

Statistics Worksheet 1 Answers:

1)a)True

2)a)Central Limit Theorem

3)b)Modeling Bounded Count data

4)d) All of the mentioned

5)c)Poisson

6)b)False

7)b)Hypothesis

8)a)0

9)c)Outliers cannot conform to the Regression relationship

10) The **normal distribution** reflects the various values taken by many real life variables like the heights & weights of people or the marks of students in a large class. In all these cases, a large number of observations are found to be clustered around the **mean value** and their frequency drops sharply as we move away from the mean in either direction. For example, if the mean height of an adult in a city is 6 feet then a large number of adults will have heights around 6 feet. Relatively a few adults will have heights of 5 feet or 7 feet.

11)Missing data is an issue that we come across whenever we handle datasets. It is represented by Nan value in python. We can replace the numerical & categorical missing data with either mean(numeric), median and mode(category). We can remove the complete row or column that contains the missing data. We can use pairwise deletion, listwise deletion. We can also use K-Nearest Neighbor Imputation.

12)**A/B testing** is a way to understand the users' response to two different versions of a single variable. For example we can make a slight change(change in only a button) in a website and see how the website get response from the visitors. For coming at a conclusion about which of the two forms are received better than the other, we can use tools in statistics like hypothesis testing or two sample hypothesis testing. It is also known as **bucket testing** or split-run testing. It makes use of random experiment and the sampling should be unbiased.

13)**Mean imputation** of missing data is a good solution because it is very simple, the sample size remains to the fullest, although it has some shortcomings as well. The mean imputation is not so good for unbiased estimates of relationships. If the missing data is large, the relationship of the variables in the dataset gets either overestimated or underestimated. Any statistic that uses mean imputed data will have a standard error that is too low. One can make Type-I error because the pvalue gets lowered.

14)**Linear regression** in statistics is a process of finding the best fit line that shows the optimal relation among the dependent variable & one or more independent variables. We try to plot a line that has the least error. Once we do that we try to predict the values for the dependent (output)variable for certain values of the independent(input) variables. Our model predicts the output values based on the input values that match with reference to the line.

The equation is: 1)y=ax+b+e(Single independent variable-x, e-error, a-slope, b-Y-intercept)

2)y=m+ax1+bx2+cx3+e(Multiple independent variables-x1, x2, x3, e-error)

15) There are basically two branches of statistics.

a)Descriptive Statistics: It is used to describe the statistical characteristics of a dataset/sample. It uses measures of central tendency (Mean, Median, Mode, Percentile) and also dispersion(standard deviation, Variance, Range).

b) Inferential Statistics: If the population is too large for coming at certain statistical inferences, we use sampling and draw conclusions by using certain methods on those samples. This technique is known as inferential statistics. Examples are Z-score method, Hypothesis Testing, chi-square test, F-test, t-test, ANOVA-Test, MANOVA-Test. The population data has parameter while sample data has statistic.

Machine Learning-Worksheet 1 Answers

1)A)Least square Error

2)A)Linear Regression is sensitive to outliers.

3)B)Negative

4)B)Correlation

5)C)Low Bias & High Variance

6)B)Predictive Model

7)D)Regularization

8)D)SMOTE

9)C)Sensitivity & Specificity

10)B)False

11)A) Construction bag of words from a email

12)B) It becomes slow when number of features is very large.

13)In linear regression, possibility of overfitting is there. It produces error in the test output. To avoid overfitting, we generalize the Linear regression model by adding a new term that is called penalty in the model. This whole process of penalizing to avoid overfitting of a Linear regression model is known as regularization.

14)There are two algorithms used for regularization. A) Lasso Regression(L1), B)Ridge Regression(L2)

Lasso: We find out the error term(Cost function) in any regression model. Although, we try to keep it around zero, to avoid overfitting, we add a penalty in the error. In lasso regression the penalty is nothing but a term 'Lambda' multiplied by the sum of the absolute values of the slopes of the input variables present in the model. Lambda is usually taken equal to 1. It can be increased or decreased also.

Ridge: In ridge regression, the term 'Lambda' is multiplied by the sum of the squares of the slopes of the input variables present in the model.

15)Linear Regression is a technique of predicting the value of Dependent Variable(Y) given the value(s) of Independent Variable(s).

The Linear regression equation is: Y=a+bX+e. Here Y is the output, X is the Input variable, a is the Y-Intercept, b is the slope of the Linear regression Line and 'e' is the error term. There are two kinds of errors which are 'Mean square Error' & 'Root Mean Square Error'.

Mean square Error is nothing but the mean of the square of the differences between the actual outputs & the model calculated outputs. If we take square root of it, it is called root mean square error. We try to make sure that the error is as close to zero as possible. The Linear regression model is termed more(or Perfect) effective if the error value is close or equal to zero, the accuracy of predicting the Dependent Variable is very high.