Python – Worksheet 1

Questions & Answers

Q1. Option C - %

Q2. Option B – 0

Q3. Option C - 24

Q4. Option A - 2

Q5. Option D - 6

Q6. Option C - the finally block will be executed no matter if the try block raises an error or not.

Q7. Option A - It is used to raise an exception.

Q8. Option C - in defining a generator

Q9. Option A, B, C - _abc,1abc, abc2

Q10. Option D – all of the above

Machine Learning

Q1. Option D – Both A & B

Q2. Option A -Linear regression is sensitive to outliers

Q3. Option B - Negative

Q4. Option B - Correlation

Q5. Option C - Low bias and high variance

Q6. Option D – all of the above

Q7. Option D - Regularization

Q8. Option D - SMOTE

Q9. Option A - TPR and FPR

Q10. Option B - False

Q11. Option B - Apply PCA to project high dimensional data

Q12. Option A, B, C - We don't have to choose the learning rate.

It becomes slow when number of features is very large.

We need to iterate.

Q13. Regularization is a term used to calibrate the machine learning models in order to minimize the adjusted loss function and prevent it from the overfitting and under fitting. Using Regularization, we can fit our machine learning model appropriately on a given test set and hence reduce the errors in it.

Q14. There are mainly three types of regularization algorithms:

1. LASSO
2. Ridge
3. ElasticNet

Lasso Regularization: It modifies the over-fitted or under-fitted models by adding the penalty equivalent to the sum of the absolute values of coefficients. Lasso regression also performs coefficient minimization, but instead of squaring the magnitudes of the coefficients, it takes the true values of coefficients. This means that the coefficient sum can also be 0, because of the presence of negative coefficients.

Consider the cost function for the Lasso:

Cost function = Loss + λ x $\Sigma||w||$

Where Loss=sum of squared residuals

λ = Penalty error

w= slope of curve/line

We can control the coefficient values by controlling the penalty terms. LASSO regression converts coefficients of less important features to zero, which indeed helps in feature selection, and it shrinks the coefficients of remaining features to reduce the model complexity, hence avoiding overfitting.

Ridge Regularization: It modifies the over-fitted or under fitted models by adding the penalty equivalent to the sum of the squares of the magnitude of coefficients. This means that the mathematical function representing our machine learning model is minimized and coefficients are calculated. The magnitude of coefficients is squared and added. Ridge Regression performs regularization by shrinking the coefficients present.

The Cost function for Ridge:

Cost function = Loss + λ x $\Sigma||w||^2$

Where Loss=sum of squared residuals

λ = Penalty error

w= slope of curve/line

Ridge regression shrinks the coefficients as it helps to reduce the model complexity and multi collinearity.

ElasticNet: It is a regularized method which linearly combines the penalties of LASSO & Ridge methods respectively.

Cost function will be the:

Cost function = Loss + λ x $\Sigma||w||$ + λ x $\Sigma||w||^2$

Where Loss=sum of squared residuals

λ = Penalty error

w= slope of curve/line

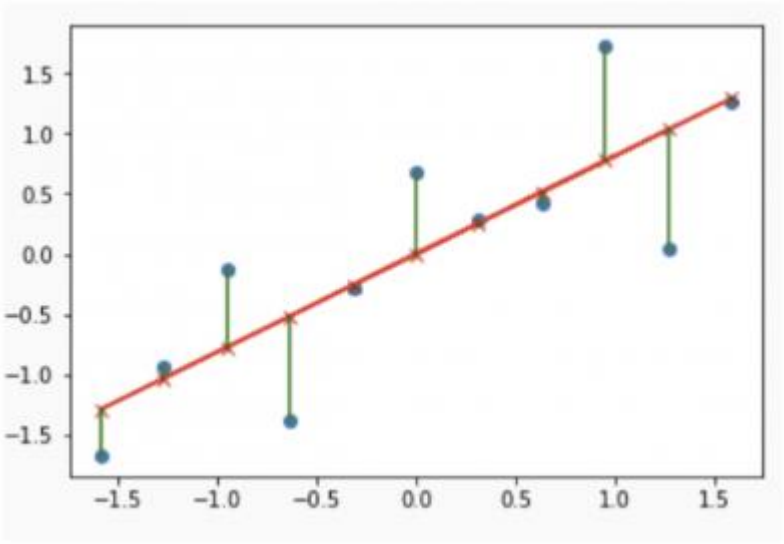
Q15. In a Linear regression model the relationship between the variables is a linear equation written as:

$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$

Where, Yi = Dependent Variable, β_0 = Population Y-intercept, β_1= Population slope, X_i = Independent Variable, ε_i = Random error

An error term in statistics is a value which represents how observed data differs from actual population data. It can also be a variable which represents how a given statistical model differs from reality.

Here the difference between each observation and the best fitted line is known as the random error or error term that created when we try to best fit the model for the best learning and better predictions.



Statistics Worksheet 1

Q1. Option A - True

Q2. Option A – Central Limit Theorem

Q3. Option B - Modeling bounded count data

Q4. Option D – All of the mentioned

Q5. Option D – All of the mentioned

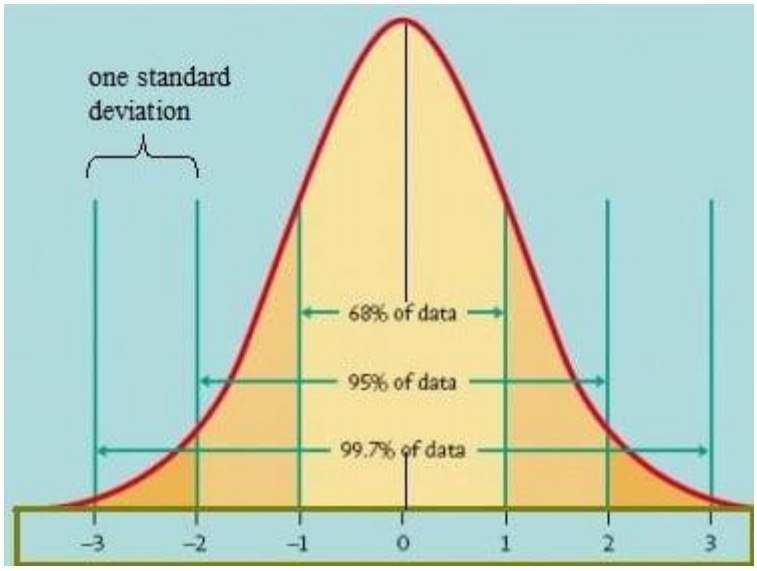
Q6. Option B - False

Q7. Option B - Hypothesis

Q8. Option A - 0

Q9. Option C - Outliers cannot conform to the regression relationship

Q10. Normal Distribution is the most widely known and used of all the distributions. It means where the mean, median and mode of the given data is the same i.e. mean=median=mode in the graphical distribution of the given data. The data is centred at zero and have units equal distance in the standard deviation. All the data is equally distributed around the mean.



Q11. Normally we don't have the complete data and there are some missing values in the data. We will firstly check that how many rows are there with the missing values, if there are less rows we will go with the method of complete deletion of the rows with missing values. But if the number of rows is more we would definitely not go with this as it results in the loss of our data so we will use the mean or median imputation method to fill the missing values.
I would recommend to use the mean or median imputation method to handle the missing data in case of numerical missing data and mode imputation data when there is categorical data.

Q12. A/B testing is basic randomized control experiment. It is basically uses to compare the two versions of a variable to check which one performs better in a controlled environment. It is hypothetical testing method for making decisions that estimate population parameters on the basis of sample statistics where population refers to whole and sample refers to some of them with which the experiment has done to predict.

Q13. No, mean imputation has its own disadvantages:

- Mean imputation reduces the variance of the imputed variables.
- Mean imputation shrinks standard errors, which invalidates most hypothesis tests and the calculation of confidence interval.
- Mean imputation does not preserve relationships between variables such as correlations.

Q14. Linear regression is used to predict the value a variable based on an another variable. If we use variable x to draw conclusions concerning a variable y, then x will be the independent or explanatory variable and y will be the dependent or response variable. If the two variables are related linearly it can be summarized by a straight line given by y=a + bx.

We will choose the value of a & b in such a way that the corresponding linear line will be the best fit for the given data and the criteria for the best fit in regression analysis is sum of the squared differences between the data points and the line itself.

Q15. There are mainly two major branches in statistics:

1. Descriptive Statistics – It mostly focuses on the central tendency, variability and distribution of the sample data. Central tendency means the estimate of characteristics like the mean, median and mode of a typical sample or population. Variability refers how much there is difference among the elements of sample which is depicted by range, standard deviation and variance. Distribution refers to the shape of the plotted data whether it is symmetric or skewed, outliers present on not in the given sample or population.
2. Inferential Statistics - It allows you to make predictions (inferences) from the data. With inferential statistics, you take data from samples and make generalizations about a population. Inferential statistics have two main uses:
    - making estimates about populations (for example, the mean SAT score of all 11th graders in the US).
    - testing hypotheses to draw conclusions about populations (for example, the relationship between SAT scores and family income).