WORKSHEET – 1 **ASSIGNMENT – 39**

MACHINE LEARNING

Q1: A) (Least square error)

Q2: B) (Linear regression is sensitive to outliers)

Q3: B) (Negative)

Q4: B) (Correlation)

Q5: D) (none of these)

Q6: B) (Predictive modal)

Q7: D) (Regularization

Q8: A) (Cross Validation)

Q9: A) (TPR and FPR)

Q10: B) (False)

Q11: B) (Apply PCA to project high dimensional data)

Q12: B) (It becomes slow when number of features is very large)

C) (We need to iterate)

Q13: The word regularize means Regular or acceptable. Regularization is most commonly used and important technique in Machine Leaning. This technique is used to prevent the error by fitting a function appropriately on Data set to avoid Model from Over fitting. Over fitting occurs when a model fits against its training data. Then, the model cannot perform correctly against test data. when we train our model for long time it start to learn noise or irrelevant information this unable to generalize with the test data and we will not be able to get the correct prediction.

Q14: Algorithms used for Regularization:

1. LASSO Regression (L1 Form)

2. RIDGE Regression (L2 Form)

LASSO (Least Absolute Shrinkage and Selection Operator):

- LASSO removes unnecessary features which are not important for the model.
- Basically it checks the relation between every feature with the label. If no relation found it nullify or remove that feature (for give no weightage to that feature).
- It also acts as feature selection (it select only important feature).

RIDGE :

- RIDGE is used if the model is to remove multicollinearity.
- It doesn't remove or completely nullify the unimportant feature. it just give liess importance to that feature which is not important for prediction.

Q15: Error variable produce by model, when model is not able to establish relation between independent variable and dependent variable. The amount of error term can change the equation analysis. Term error specifically mean that our model is not working properly, that its not accurate and give us different results when we execute it in real time applications. The equation for Linear Regression is :

$$Y = a + b * X + e$$

Where,

a = intercept

b = slope of the line

e = error term

when Actual data (Y) is differ from predicted data (Y) in model during execution of model and error term is not equal to 0, That means Predicted data is getting influenced by other factors.

PYTHON – WORKSHEET 1

Q1: C) (%)

Q2: B) (0)

Q3: C) (24)

Q4: A) (2)

Q5: D) (6)

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

Q7: C) it's not keyword in python.

Q8: C) in defining a generator

Q9: C) abc2

A) _abc

Q10: A) yield

C)look-in

STATISTICS WORKSHEET -1

Q1: a) True

Q2: a) Central Limit Theorem

Q3: a) Modeling event/time data

Q4: b) Sum of normally distributed random variables are again normally distributed even if the variables are dependent.

Q5: c) Poisson

Q6: b) False

Q7: b) Hypothesis

Q8: a) 0

Q9: a) outliers can have varying degrees of influence.

Q10: The normal distribution describes how the values of a variable are distributed. It is the most important probability distribution in statistics because it fits many natural phenomena. For example, heights, blood pressure, measurement error, and IQ scores follow the normal distribution. It is also known as the Gaussian distribution and the bell curve. As with any probability distribution, the parameters for the normal distribution define its shape and probabilities entirely. The normal distribution has two parameters, the mean and standard deviation..

Q11: Missing data can skew anything for data scientists, from economic analysis to clinical trials. After all, any analysis is only as good as the data. A data scientist doesn't want to produce biased estimates that lead to invalid results. The concept of missing data is implied in the name: it's data that is not captured for a variable for the observation in question. Missing data reduces the statistical power of the analysis, which can distort the validity of the results. Lets understand why data is missing.

- Missing at Random (MAR)
- Missing completely at Random (MCAR)
- Missing not at Random (MNAR)

The Imputation techniques are:

1.Mean, Median, Mode - This is one of the most common methods of imputing values when dealing with missing data. In cases where there are a small number of missing observations

2. Time Series Specific method - Another option is to use time-series specific methods when appropriate to impute data. There are four type of time series specific methods –

- No trend or seasonality

- Trend, but no seasonality,
- Seasonality, but no trend,
- Both trend and seasonality

Q12: A/B testing is a basic randomized control experiment. It is a way to compare the two versions of a variable to find out which performs better in a controlled environment. It is a hypothetical testing methodology for making decisions that estimate population parameters based on sample statistics. We make hypothesis testing it means that A hypothesis is a tentative insight into the natural world; a concept that is not yet verified but if true would explain certain facts or phenomena.

There are two types of Hypothesis :

1. Null Hypothesis – (H_0): The null hypothesis is the one that states that sample observations result purely from chance. From an A/B test perspective, the null hypothesis states that there is no difference between the control and variant groups

2 Alternative Hypothesis : The alternative hypothesis challenges the null hypothesis and is basically a hypothesis that the researcher believes to be true. The alternative hypothesis is what you might hope that your A/B test will prove to be true

Once we are ready with our null and alternative hypothesis, the next step is to decide the group of customers that will participate in the test. Then we randomly select the sample from the population is called random sampling. It is a technique where each sample in a population has an equal chance of being chosen. Random sampling is important in hypothesis testing because it eliminates sampling bias, and it's important to eliminate bias because you want the results of your A/B test to be representative of the entire population rather than the sample itself. After that we conduct the test to calculate daily conversion rates.

Q13. The process of replacing null values in a data collection with the data's mean is known as mean imputation. Mean imputation is typically considered terrible practice since it ignores feature correlation. Mean imputation decreases the variance of our data while increasing bias. As a result of the reduced variance, the model is less accurate.

Q14: Linear regression is a basic and commonly used type of predictive analysis. The overall idea of regression is to examine two things:

(1) Does a set of predictor variables do a good job in predicting an outcome (dependent) variable.

(2) Which variables in particular are significant predictors of the outcome variable, and in what way do they indicated by the magnitude and sign of the beta estimates–impact the outcome variable.

These regression estimates are used to explain the relationship between one dependent variable and one or more independent variables. Equation used for Regression is :

$$\mathbf{y = c + b*x}$$

where,
y = estimated dependent variable score
c = constant
b = regression coefficient
x = score on the independent variable

Q15: There are three real branches of statistics:

1) Data collection
2) Descriptive statistics
3) Inferential statistics

Data Collection : Data collection is all about how the actual data is collected. Sometimes, data is harder to collect. if you are collecting data, you need to be careful where you get it from You can't realistically ask everyone in the whole country (the population), so you have to choose a representative sample of people. The words population and sample are used in general in statistics. The population is the entire set of data, and a sample is a (hopefully representative) subset of the population.

Descriptive statistics : Descriptive statistics is the part of statistics that deals with presenting the data we have. This can take two basic forms – presenting aspects of the data either visually or numerically. The basic aim of descriptive statistics is to 'present the data' in an understandable way. If you simply write down every piece of data, it means little to someone who sees it; it needs to be summarized. Instead, of that you are presented with visual charts.

Inferential statistics: Inferential statistics is the aspect that deals with making conclusions about the data. it take the data you have and make an 'inference' or 'conclusion' from it.