PYTHON WORSHEET 1

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. A) _abc , C) abc2
10. A) yield , B) raise

STATISTICS WORKSHEET 1

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. NORMAL DISTRIBUTION: [GAUSSIAN]

    Normal Distribution is a special type of bell-shaped density curve. It describes the tendency for data to cluster the central valve which is the population mean(μ) located in the middle of the curve. It is symmetric about its mean. For example, Exam scores, height of the adults, weight etc follows ND. Some data points will be below the mean and some of them above.

    The following are the parameters to determine the ND. μ characterises the position of the ND, σ(Population Standard deviation) characterises the spread of the ND. Larger the deviation more the spread out of the distribution would be.
    The total area under the curve is 1. The data is symmetrically distributed with no skew
    The empirical rule or 68-95-99.7 rule tells where the most of the values lie in the ND.
    Around 68% of the values within 1 SD from the mean.
    Around 95% of the values within 2 SD from the mean.
    Around 99.7% of the values within 3 SD from the mean.

11. HANDLING MISSING DATA AND IMPUTATION TECHNIQUES:

    Missing data from the source is unavoidable. Most of the time the data missing from the source would be because of item non response which is people unwilling to answer some questions. Due to the missing data the statistical power of the analysis can reduce which can impact the validity of the results. The best way to handle the situation is to make possible plans to minimise the damage.

Imputation is the process of substituting an estimated value in place of the missing value and analysing the entire data set as if the estimated values were the observed ones. The following are some of the techniques to impute a value,

- AVERAGE IMPUTATION:
  This uses the average value of the responses from other data entries to fill out the missing values. This can reduce the variability.
- COMMON POINT IMPUTATION:
  This uses the most commonly used value for substitution. This takes the MEAN if the data is numeric and not skewed or MEDIAN if the data is numeric and skewed or MOST FREQUENT which uses the most frequent used value if the data is string or numeric of the observed values which in most of the cases gives the same value.
- IMPUTATION WITH CONSTANT VALUE:
  It replaces the missing value with zero or any constant value.
- HOT DECK IMPUTATION:
  Choosing a random variable from the sample which has the comparable on the other factors.
- COLD DECK IMPUTATION:
  This chooses the value deliberately from an individual with similar values on other values but without random variance.
- REGRESSION IMPUTATION:
  Regressing the missing variable to get a predicted value. Relying on the anticipated value keeps the association between the variable and the imputed model but not the variability around them.
- STOCHASTIC REGRESSION IMPUTATION:
  This uses the predicted value of regression and the random value. Majority of multiple imputation is based on this.
- K_NEAREST NEIGHBOUR IMPUTATION:
  This helps to impute the missing data by finding the closest neighbours using the Euclidean distance metric to the observation with the missing data and imputing them based on non-missing values in the neighbours. This requires normalising data as it is a distance based imputer.

12. A/B TESTING:

This refers to the experiments where two or more variations of the same webpage are compared against each other by displaying them to real-time visitors to find which is better for the given goal. This can also be used for emails, apps, popups and much more. It is an analytical method for making decisions that estimates population parameters based on sample statistics. This optimise the web marketing strategies that allows the decision makers to use the best design for a website by looking the analytical results using two alternatives A and B.

People served with two designs and based on their activity which is collected through the web analytics the statistical tests are applied to find out which has the better efficacy. This can be measured using the Discrete or binomial metrics like Click through rate, conversion rate, bounce rate and the Continuous or the non-binomial kike average revenue per user, average session duration. The testing starts by making a claim i.e Hypothesis. Followed by launching the test to gather the statistical evidence to accept or reject the hypothesis.

The final data Shows whether the hypothesis is correct or incorrect. The null hypothesis states the default position to be tested or the situation as it is assumed to be now. The Alternative hypothesis challenges the null hypothesis and is a basically a hypothesis that researcher believes to be true.

13. MEAN IMPUTATION OF MISSING DATA:

    Mean imputation is considered to be avoidable practice as it ignores the feature correlation. It does not preserve the relationship among the variables. If the missing data is completely at random it will not bias the parameter estimate. This also leads to the underestimate of standard errors. It decreases the variance of the data with increasing bias. This makes the model less accurate. The alternative option would be multiple imputation.

14. LINEAR REGRESSION:

    This is the commonly used predictive analysis. It examines two things.

    1) predict the value of the variable(dependent) based on the value of the other variable(independent)

    2) which of the variables are significant predictor of outcome variable and in what way they do indicate by magnitude and sign of beta estimates impact the outcome variable.

    The simplest form of regression equation is y=c+b*x

    Where y=estimated dependent variable score, c=constant, b=regression co efficient, x=score on independent variable.

    This regression fits the straight line or the surface that minimises the discrepancies between the predicted and the actual output values.

    The three major uses would be determining the strength of the predictors, trend forecasting, forecasting an effect.

15. BRANCHES OF STATISTICS:

    The two main branches would be Descriptive and inferential statistics.

    Descriptive focuses on collecting summarising and presenting the data. This can be either visually presented like graphs, charts etc or numerically like averages.

    Inferential focuses on analysing the sample data to draw conclusions about a population. In this we start with the hypothesis to see whether the data is consistent with that hypothesis.

MACHINE LEARNING WORKSHEET 1

1. A) Least Square error
2. A)Linear regression is sensitive to outliers
3. B) Negative
4. B) Correlation
5. C) Low bias and high variance
6. B) Predictive model
7. D) Regularization
8. A) Cross validation
9. A) TPR and FPR
10. B) False
11. B) Apply PCA to project high dimensional data
12. A) We don't have to choose the learning rate
    B) It becomes slow when number of features is very large.

13. REGULARIZATION:

    It is a technique used to reduce the errors by fitting the function appropriately on the given set and avoid overfitting. Overfitting is a phenomenon that occurs in the machine learning model is constraint to training set and not able to perform well on unseen data. This technique discourages the learning a more complex or flexible model. This converts the complex model to a simpler one to avoid over fitting and shrinks the co efficient for lesser computational cost.
14. ALGORITHMS FOR REGULARIZATION:

    The commonly used regularization techniques are

- L1 regularization
- L2 regularization
- Elastic net regularization

    The regression model that uses the L1 regularization technique is called as LASSO (Least Absolute Shrinkage and Selection operator). This adds "absolute value of magnitude" of coefficient as penalty term to loss function(L).

    The model that uses the L2 regularization is called RIDGE regression. This adds" squared magnitude" of coefficient as penalty term to loss function(L). During regularization the output doesn't change only the loss function changes. A standard least squares model tends to have some variance in it. It significantly reduces the variance of the model without substantial increase in its bias.

    Elastic net technique linearly combines L1 and L2 penalties of the LASSO and RIDGE respectively.
15. TERM ERROR IN THE LINEAR REGRESSION:

    The error term is the value which represents how the observed data differs from actual population data. It refers to the sum of variations within the regression line. It can also be a variable which represents how a given statistical model differs from reality. This means that it reflects the nonlinearities, unpredictable effects, measurement errors, omitted variables. This is usually the remainder residual or the disturbance term denoted by *e*, ε or u.

The error term and residual go synonymously but the difference is that the error term represents the way the observed data differs from the actual population and the residual represents the observed data differs from the sample population data.