PYTHON WORKSHEET - 1

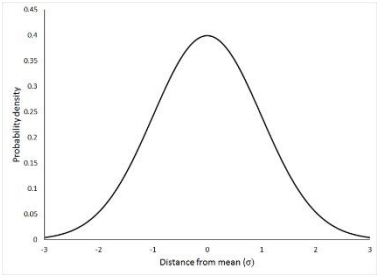
1. (C) %
2. (B) 0
3. (C) 24
4. (A) 2
5. (D) 6
6. (C) the finally block will be executed no matter if the try block raises an error or not
7. (A) It is used to raise an exception
8. (C) In defining a generator
9. (A) _abc and (C) abc2
10. (A) yield and (B) raise

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

Subjective answers:

10. A normal distribution has a bell-shaped curve which is a continuous probability distribution for independent random variables which is symmetric around its mean. It has a peak in the middle portion and the probability values decreases left and right.



It is useful to describe distribution of values for many natural phenomena such as height/salaries/IQ scores/grades, etc. of individuals from a group of people.

For a normal distribution, mean = median = mode

11. To handle missing data, two main methods can be used:
    a) Imputation
    b) Removing the data

Removing data may not be the best option because if we do not have enough data, removing some data related to the missing data will not result in reliable result. So, by imputation, we can develop a reasonable guess for the missing data.

But first, we need to find out why the data is missing, so that we can make a good guess for the missing data. Some imputation techniques would be:

a) Finding Mean or Median of the existing observations.
b) Linear Interpolation
c) Time Series Specific Methods

12. A/B testing is a statistical method to find out which performs better if we compare two variables. Let's assume that we have a product A and we want to know how it will have a success in the market. So, we need some other product so that customers can choose which one to buy. So, let's make a small change in the product A and call the changed one B (which is originally the same as A but some small changes are made).

    If product B performs better in the market, then that means we can easily know if the product A should be continued to be manufactured or not.

13. When there are null/missing values in the data collected, one can replace the null values with the mean of the data. This is called mean imputation.

    This is not an acceptable practice because it means that we ignore the correlation present in the data. Let's take an example of a data in which we have student names and marks. Suppose we are missing a data for one student who scores 3 out of 100, and we replace his mark with the mean mark (say 46 out of 100), then clearly, we increase the bias while decreasing the variance. This will reduce the accuracy of the model.

14. Linear regression is a model which tries to make a relationship between two variables using a linear equation. It tries to fit linear equation to observed data. A linear regression line has an equation of a straight line,

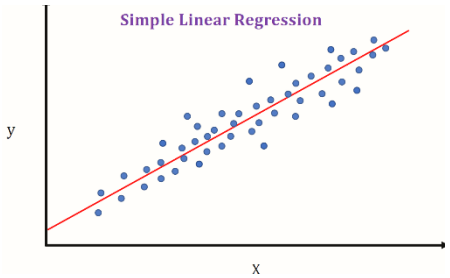
$$y = mx + c.$$

$$where\ m\ is\ the\ slope, and\ c\ is\ the\ intercept.$$

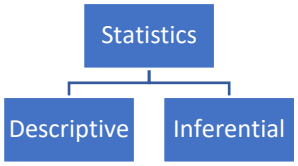
x is an explanatory variable and y is the dependent variable?

This line may not connect all the data but gives us a rough idea about the direction of the data, whether it shows increasing or decreasing trend. One may use scatter plot and tries to fit linear equation to it.

If slope is negative, that means it shows decreasing trend and if slope is positive, then the data has an increasing trend. The relationship between the two variables could be weak, strong or no relation at all. The figure below shows increasing trend.



15. The branches of statistics are described below:



Descriptive Statistics: It is a type of statistics in which one can describe a data without applying any statistical tools. For example, we can just collect raw data like marks of students in a class, opinion of people in a small constituency for an election. Using these data, we can present the data collected in an intuitive way so that it is easy to understand instead of showing the raw data by using plots, histogram, charts, etc.

Inferential Statistics: Now, if the data is too big or we have a problem in collecting all the data, then we can take sample data randomly and infer the result from it. That means these sample data will represent the result for the whole data. We can just find the average of the data which will be applicable to all the data from where we have selected our samples.

Suppose we want to know the exit poll in an election and we have to know the opinion of all the population. We cannot ask each and every person for all the constituencies since it would take a very long time, and the data would be too big. So, instead we can use inferential statistics by collecting random samples from each constituency. That means we can ask only few random persons from each constituency and take the mean of the data. This will reduce our effort and save us big time.

Machine Learning

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. (A) Descriptive model
7. (D) Regularization
8. (D) SMOTE
9. (C) Sensitivity and Specificity
10. (B) False
11. (B) Apply PCA to project high dimension data
12. (A) We don't have to choose the learning rate
    (B) It becomes slow when number of features is very large
    (D) It does not make use of dependent variable

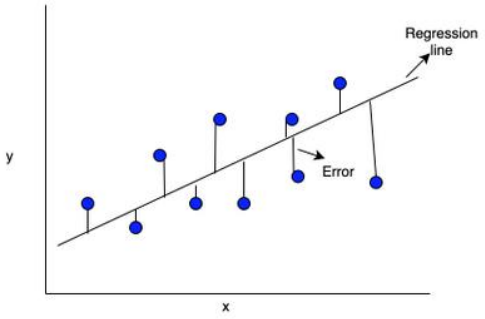
Subjective answers:

13. Regularization is a machine learning technique which use a method to fit the function appropriately to reduce errors that can be present due to overfitting of the data by reducing variance in the model. It is a technique to train the machine to learn rather than just memorizing.

If the model works accurately on training data but gives inaccurate results for unseen data, then the machine more or less just memorizes the data. So, regularization techniques help the machine to learn from the data and make it able to provide more accurate result even for unseen dataset.

14. Algorithms used for regularization are:
    (a) L1 regularization
    (b) L2 regularization
    (c) Dropout

15. The difference between the expected value and the actual or observed value is called error in linear regression. In the figure, the distance between each point and the linear graph is the error term. Let's consider an equation for linear regression as given below



$$Y = \alpha X + \beta\delta + \epsilon$$

Where α, β = constant parameter

X, δ = Independent variables

ϵ = error term