Internship 28

PYTHON WORKSHEET 1

1. C
2. B
3. C
4. A
5. D
6. C
7. A
8. C
9. A & C
10. A & B

11. to 15. = Jupyter Notebook

STATISTIC WORKSHEET 1

1. A
2. A
3. B
4. D
5. C
6. B
7. B
8. A
9. C

10. **NORMAL DISTRIBUTION**: - There are various types of distribution in statistic but most commonly used distribution is Normal Distribution. Normal Distribution is always normal irrespective of sample size. It is the proper term of a probability distribution that is symmetric about the mean, showing that data near the mean are more frequent in occurrence than data far from the mean. It is also known as Gaussian distribution.

    In a Normal Distribution, the mean is zero. In a graph form, normal distribution will appear as a bell curve. It says Mean, Median and Mode are equal.

11. With the help of Imputers, we can handle missing data. Imputer is nothing nut replacing NaN with some meaningful data. Following are the Imputation techniques : -

    - KNN Imputer
    - Iterative Imputer

    KNN IMPUTER: -It is a scikit-learn class used to fill out or predict the missing values in a dataset. It takes the average of three nearest neighbors and that value fill in NaN. It is more useful method which works on the basic approach of the KNN algorithm rather than the native approach of filling all the vales with the mean or median.

    ITERATIVE IMPUTER: - It refers to a process where each feature is modeled as a function of the other features. Each feature is imputed sequentially, one after the other, allowing prior imputed values to be used as part of a model in predicating subsequent features.

12. **A/B TESTING**: - An AB test is an example of statistical hypothesis testing, a process whereby a hypothesis is made about the relationship between two data sets and those data sets are then compared against each other to determine if there is a statistically significant relationship or not.

13. Yes, the mean imputation of missing data acceptable. Imputing the mean preserves the mean of the observed data. So if the data are missing completely at random, the estimate of the mean remains unbiased. Plus, by imputing the mean, you are able to keep your sample size up to the full sample size. That's good too. This is the original logic involved in mean imputation. If all you are doing is estimating means (which is rarely the point of research studies), and if the data are missing completely at random, mean imputation will not bias your parameter estimate.

14. **LINEAR REGRESSION**: - Linear regression analysis is used to predict the value of a variable based on the value of another variable. The variable you want to predict is called the dependent variable. The variable you are using to predict the other variable's value is called the independent variable. Linear regression is commonly used for predictive analysis and modeling. For example, it can be used to quantify the relative impacts of age, gender, and diet (the predictor variables) on height (the outcome variable

15. There are two real branches of statistics:

- Descriptive Statistics
- Inferential Statistics

Descriptive Statistics: - Descriptive statistics are used to describe the characteristics or features of a dataset. The term descriptive statistics can be used to describe both individual quantitative observations (also known as 'summary statistics') as well as the overall process of obtaining insights from these data. We can use descriptive statistics to describe both an entire population and an individual sample. Because they are merely explanatory, descriptive statistics are not heavily concerned with the differences between the two types of data.

Inferential Statistics: - It is focus of making generalizations about a larger population based on a representative sample of that population. Because inferential statistics focuses on making predictions (rather than stating facts) its results are usually in the form of a probability.
The accuracy of inferential statistics relies heavily on the sample data being both accurate and representative of the larger population. To do this involves obtaining a random sample. If you've ever read news coverage of scientific studies, you'll have come across the term before. The implication is always that random sampling means better results. On the flipside, results that are based on biased or non-random samples are usually thrown out. Random sampling is very important for carrying out inferential techniques, but it is not always straightforward.

MACHINE LEARNING

1. A
2. A
3. B
4. B
5. D
6. B
7. D
8. D
9. A
10. B
11. C
12. A & B

13. **Regularization**: - With the help of regularization we check whether our model is over fitted or not. Regularization is a technique used to reduce the errors by fitting the function appropriately on the given training set and avoid over fitting. The word regularizes means to make things regular or acceptable. This is exactly why we use it for. Based on the approach use to overcome over fitting, we can classify the regularization techniques into three categories:

- LASSO Regression (L1 FORM)
- RIDGE Regression (L2 FORM)
- ELASTICNET Regression (This is less popular).

14. **Regularization Algorithms:** Following are algorithms are used for regularization:-

I. LASSO Regression (L1 FORM)
II. RIDGE Regression (L2 FORM)
III. ELASTICNET Regression (This is less popular).

LASSO REGRESSION: - Least Absolute Shrinkage and Selection Operator Regression penalize the model, based on the sum of magnitude of the coefficients. It also acts as features selection. Out of four features it will select the meaningful feature.

RIDGE REGRESSION: - Ridge Regression is also penalizing the model. Based on the sum of squares of magnitude of the coefficients. Ridge regression is a regularization technique, which is used to reduce the complexity of the model. It is also called as L2 regularization. In this technique, the cost function is altered by adding the penalty term to it. The amount of bias added to the model is called Ridge Regression penalty.

15. **Error in Linear Regression Equation**: - The error term is the difference between the expected price at a particular time and the price that was actually observed.

    Linear regression most often uses mean-square error (MSE) to calculate the error of the model. MSE is calculated by:

    - Measuring the distance of the observed y-values from the predicted y-values at each value of x;
    - Squaring each of these distances;
    - Calculating the mean of each of the squared distances.

    Linear regression fits a line to the data by finding the regression coefficient those results in the smallest MSE.