WORKSHEET SET 1

STATISTICS WORKSHEET

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

10.The normal distribution is one of the most important probability distributions in statistics and is used in many different fields such as physics, engineering, and finance.

The probability density function for a normal distribution is given by:

$f(x) = (1/\sigma\sqrt{(2\pi)}) * e^{(-(x-\mu)^2/(2\sigma^2))}$

where μ is the mean of the distribution and σ is its standard deviation.

The normal distribution is often used to model naturally occurring phenomena such as the heights of individuals in a population, the scores on a test, or the weights of objects produced by a manufacturing process. This is because many of these phenomena tend to be distributed in a way that is close to normal.

The normal distribution has some important properties, such as the fact that about 68% of the data falls within one standard deviation of the mean, about 95% of the data falls within two standard deviations of the mean, and almost all of the data falls within three standard deviations of the mean. These properties make the normal distribution useful for making predictions and for setting confidence intervals.

---

11. In machine learning, managing missing data can be done in a number of ways. Among the most popular methods are:

- Validate input data before feeding into ML model; Discard data instances with missing values: One of the simplest ways to handle missing data is to discard any data instances that have missing values. However, this method can lead to a loss of information and may not be feasible if there are many missing values.
- Predicted value imputation: Using a regression model to forecast the missing values based on other dataset properties is known as predicted value imputation.
- Distribution-based imputation: This method involves using the distribution of the non-missing values to estimate the missing values.
- Unique value imputation: This method involves replacing the missing values with a unique value such as 0 or -1.
- Reduced feature models: This method involves creating a reduced feature set that does not include any features with missing values.
- Mean substitution: In this imputation technique goal is to replace missing data with statistical estimates of the missing values. Mean, Median or Mode can be used as imputation value.

According to me, among all these techniques which technique is best to use depends on the problem and dataset one is working with.

---

12. A/B testing is a testing used for experience research. It is also known as bucket testing, split-run testing, or split testing. A/B tests entail a randomized experiment with two variants i.e. A and B, while it can also be used to multiple variants of the same variable. It incorporates the use of statistical hypothesis testing.

In A/B testing, A is referred to as ‘control’ or the original testing variable where as B refers to ‘variation’ or a new version of the original testing variable. The "winner" is the version that causes your company metric(s) to change for the better.

---

13. A popular technique for dealing with missing data is mean imputation. It entails using the mean of the observed values for that variable to fill in any missing values. When there are few missing data

points, mean imputation can be effective since it is straightforward to implement. It has certain drawbacks, though, and can result in inaccurate population parameter estimations. The fact that mean imputation lessens data variability is one of its key drawbacks. This may result in an overestimation of statistical significance and an underestimating of standard errors. Additionally, the assumption made by mean imputation—which is frequently false in practice—is that the missing data are missing entirely at random (MCAR). In conclusion, mean imputation may be appropriate when there are few missing data and the data are MCAR. When these suppositions are broken, it has some drawbacks and can result in inaccurate population parameter estimates.

---

14. A machine learning algorithm and statistical modelling method called linear regression uses one or more independent variables to predict a target value. It is employed to ascertain the connection between forecasted outcomes and factors. Utilising linear functions with parameters deduced from the data, the relationship is modelled. Simple linear regression occurs when there is just one independent variable; multiple linear regression occurs when there are numerous independent variables1. By using the results of a second variable, we may predict the scores on a first variable using simple linear regression. The criteria variable, often known as Y, is the variable we are projecting. The variable on which we are basing our forecasts is referred to as the predictor variable and is referred to as X.

---

15. The study of data gathering, analysis, interpretation, presentation, and organisation is known as statistics. In other words, gathering and summarising data is a mathematical discipline. Additionally, statistics might be considered a subfield of applied mathematics. However, uncertainty and variation are two crucial and fundamental concepts in statistics. Only statistical analysis can determine the uncertainty and variation in many sectors. The probability, which is a key concept in statistics, essentially determines these uncertainties.

Some of the main branches of statistics include:

- Descriptive statistics: This branch of statistics deals with the collection, analysis, and interpretation of data. It involves summarizing data using measures such as mean, median, mode, and standard deviation.

- Inferential statistics: This branch of statistics deals with making predictions or inferences about a population based on a sample of data. It involves hypothesis testing and estimation.
- Biostatistics: This branch of statistics deals with the application of statistical methods to problems in biology and medicine.
- Business statistics: This branch of statistics deals with the application of statistical methods to problems in business and economics.
- Social statistics: This branch of statistics deals with the application of statistical methods to problems in social sciences such as sociology and psychology.

MACHINE LEARNING ASSIGNMENT

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

---

13. Regularization is a technique used in machine learning to reduce overfitting by adding a penalty term to the loss function.

To prevent the coefficients from increasing to huge values, the penalty term is included in the loss function. To achieve this, a regularisation term that penalises high coefficients is added to the loss function.

The three regularisation methods that are frequently employed to limit the complexity of machine learning models are
1. L2 regularisation,
2. L1 regularisation, and
3. Elastic Net

---

14. There are several algorithms used for regularization in machine learning. The most common ones are:

- Ridge Regression (L2 Norm)
- Lasso (L1 Norm)
- Elastic Net Regression (combination of Ridge and Lasso regression)

Any method requiring weight parameters, including neural nets, can be implemented using Ridge and Lasso. Any type of neural network, such as an ANN, DNN, CNN, or RNN, uses dropout largely to limit the learning.

---

15. The difference between the dependent variable's actual value (y) and predicted value () is known as the error term in linear regression. The residual is another name for the error term.

The mean-square error (MSE) is the formula most frequently used to determine a linear regression model's error. MSE is determined by measuring the difference between the observed and predicted y-values at each value of x, squaring each of these distances, and then figuring out the mean of each squared distance.

PYTHON WORKSHEET

1.C

2.B

3.C

4.A

5.D

6.C

7.A

8.C

9. A and C

10. A and B

Q.11 to Q.15 are answered in jupyter notebook.