

PYTHON – WORKSHEET 1 ANSWERS

1. Which of the following operators is used to calculate remainder in a division?

Answer: C (%)

2. In python 2//3 is equal to?

Answer: B (0)

3. In python, 6<<2 is equal to?

Answer: C (24)

4. In python, 6&2 will give which of the following as output?

Answer: A (2)

5. In python, 6|2 will give which of the following as output?

Answer: D (6)

6. What does the finally keyword denotes in python?

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

7. What does raise keyword is used for in python?

Answer: A (It is used to raise an exception)

8. Which of the following is a common use case of yield keyword in python?

Answer: C (In defining a generator)

9. Which of the following are the valid variable names?

Answer: A and C (_abc and abc2)

10. Which of the following are the keywords in python?

Answer: A and B (Yield and Raise)

STATISTICS WORKSHEET-1

1. Bernoulli random variables take (only) the values 1 and 0.

Answer: A (True)

2. Which of the following theorem states that the distribution of averages of iid variables, properly normalized, becomes that of a standard normal as the sample size increases?

Answer: A (Central Limit Theorem)

3. Which of the following is incorrect with respect to use of Poisson distribution?

Answer: B (Modelling bounded count data)

4. Point out the correct statement.

Answer: D (All of the mentioned)

5. _____ random variables are used to model rates.

Answer: C (Poisson)

6. Usually replacing the standard error by its estimated value does change the CLT.

Answer: B (False)

7. Which of the following testing is concerned with making decisions using data?

Answer: B (Hypothesis)

8. Normalized data are centered at _____ and have units equal to standard deviations of the original data.

Answer: A (0)

9. Which of the following statement is incorrect with respect to outliers?

Answer: C (Outliers cannot conform to the regression relationship)

10. What do you understand by the term Normal Distribution?

Answer: Normal distribution is also known as the Gaussian distribution. It is 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.

11. How do you handle missing data? What imputation techniques do you recommend?

Answer: To handle missing data deletion methods are to eliminate missing data.

Data scientists use two data imputation techniques to handle missing data: Average imputation and common-point imputation.

Average imputation: uses the average value of the responses from other data entries to fill out missing values. However, a word of caution when using this method – it can artificially reduce the variability of the dataset.

Common-point imputation: on the other hand, is when the data scientists utilise the middle point or the most commonly chosen value.

12. What is A/B testing?

Answer: A/B testing, also known as split testing, refers to a randomized experimentation process wherein two or more versions of a variable (web page, page element, etc.) are shown to different segments of website visitors at the same time to determine which version leaves the maximum impact and drive business metrics.

13. Is mean imputation of missing data acceptable practice?

Answer: Mean imputation is typically considered terrible practice since it ignores feature correlation. Also, mean imputation decreases the variance of our data while increasing bias. As a result of the reduced variance, the model is less accurate and the confidence interval is narrower.

14. What is linear regression in statistics?

Answer: Linear regression analysis is used to predict the value of a variable based on the value of another variable. The variable we want to predict is called the dependent variable. The variable we are using to predict the other variable's value is called the independent variable.

15. What are the various branches of statistics

Answer: There are two main branches of statistics

- Inferential Statistic.

- Descriptive Statistic.

Inferential Statistics: Inferential statistics used to make inference and describe about the population. These stats are more useful when its not easy or possible to examine each member of the population.

Descriptive Statistics: Descriptive statistics are use to get a brief summary of data. You can have the summary of data in numerical or graphical form.

There are two types of Descriptive Statistics:

1. Central Tendency

a. Mean

b. Median

c. Mode

2. Dispersion of Data

a. Range

b. Variance

c. Standard Deviation

d. Skewness/Percentile

MACHINE LEARNING WORKSHEET-1

1. Which of the following methods do we use to find the best fit line for data in Linear Regression?

Answer: A (Least Square Error)

2. Which of the following statement is true about outliers in linear regression?

Answer: A (Linear regression is sensitive to outliers)

3. A line falls from left to right if a slope is _____?

Answer: B (Negative)

4. Which of the following will have symmetric relation between dependent variable and independent variable?

Answer: B (Correlation)

5. Which of the following is the reason for over fitting condition?

Answer: C (Low bias and high variance)

6. If output involves label then that model is called as:

Answer: B (Predictive modal)

7. Lasso and Ridge regression techniques belong to _____?

Answer: D (Regularization)

8. To overcome with imbalance dataset which technique can be used?

Answer: D (SMOTE)

9. The AUC Receiver Operator Characteristic (AUCROC) curve is an evaluation metric for binary classification problems. It uses _____ to make graph?

Answer: A (TPR and FPR)

10. In AUC Receiver Operator Characteristic (AUCROC) curve for the better model area under the curve should be less.

Answer: A (True)

11. Pick the feature extraction from below:

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

12. Which of the following is true about Normal Equation used to compute the coefficient of the Linear Regression?

Answer: A & B (We don't have to choose the learning rate and It becomes slow when number of features is very large)

13. Explain the term regularization?

Answer: Regularization is a method for constraining or regularizing the size of the coefficients, thus shrinking them towards zero. It reduces model variance and thus minimizes overfitting. Regularization helps to solve over fitting problem in machine learning. Regularization is nothing but adding a penalty term to the objective function and control the model complexity using that penalty term. It can be used for many machine learning algorithms.

14. Which particular algorithms are used for regularization?

Answer: Ridge Regression (L2), and Lasso (L1) are the two main techniques use for regularization.

15. Explain the term error present in linear regression equation?

Answer: An error term in statistics is a value which represents how observed data differs from actual population data. It can also be a variable which represents how a given statistical model differs from reality. The error term is often written ϵ . The error term includes everything that separates your model from actual reality. This means that it will reflect nonlinearities, unpredictable effects, measurement errors, and omitted variables.