PYTHON WORKSHEET

1. Which of the following operators is used to calculate remainder in a division? A) # B) & C) % D) $

   Answer: C

2. In python 2//3 is equal to? A) 0.666 B) 0 C) 1 D) 0.67

   Answer: B

3. In python, 6<<2 Is equal to? A) 36 B)10 C) 24 D) 45

   Answer: C

4. In python, 6&2 will give which of the following as output? A) 2 B) True C) False D) 0

   Answer: A

5. In python, 6|2 will give which of the following as output? A) 2 B) 4 C) 0 D) 6

   Answer: D

6. What does the finally keyword denotes in python? A) It is used to mark the end of the code B) It encloses the lines of code which will be executed if any error occurs while executing the lines of code in the try block. C) the finally block will be executed no matter if the try block raises an error or not. D) None of the above

   Answer: C

7. What does raise keyword is used for in python? A) It is used to raise an exception. B) It is used to define lambda function C) it's not a keyword in python. D) None of the above

   Answer: A

8. Which of the following is a common use case of yield keyword in python? A) in defining an iterator B) while defining a lambda function C) in defining a generator D) in for loop.

   Answer: C

9. Which of the following are the valid variable names? A) _abc B) 1abc C) abc2 D) None of the above

   Answer: A and C

10. Which of the following are the keywords in python? A) yield B) raise C) look-in D) all of the above

    Answer: A and B

STATISTICS WORKSHEET-1

1. Bernoulli random variables take (only) the values 1 and 0. a) True b) False

   Answer: a

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? a) Central Limit Theorem b) Central Mean Theorem c) Centroid Limit Theorem d) All of the mentioned

   Answer: a

3. Which of the following is incorrect with respect to use of Poisson distribution? a) Modeling event/time data b) Modeling bounded count data c) Modeling contingency tables d) All of the mentioned

   Answer: b

4. Point out the correct statement. a) The exponent of a normally distributed random variables follows what is called the log- normal distribution b) Sums of normally distributed random variables are again normally distributed even if the variables are dependent c) The square of a standard normal random variable follows what is called chi-squared distribution d) All of the mentioned

   Answer: d

5. _____ random variables are used to model rates. a) Empirical b) Binomial c) Poisson d) All of the mentioned

   Answer: c

6. Usually replacing the standard error by its estimated value does change the CLT. a) True b) False

   Answer: a

7. Which of the following testing is concerned with making decisions using data? a) Probability b) Hypothesis c) Causal d) None of the mentioned

   Answer:b

8. Normalized data are centered at______and have units equal to standard deviations of the original data. a) 0 b) 5 c) 1 d) 10

   Answer: a

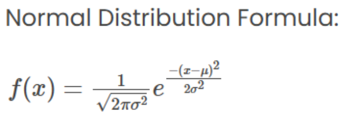
9. Which of the following statement is incorrect with respect to outliers? a) Outliers can have varying degrees of influence b) Outliers can be the result of spurious or real processes c) Outliers cannot conform to the regression relationship d) None of the mentioned

   Answer:c

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

The Normal Distribution, also known as the Gaussian distribution, is a fundamental concept in statistics and probability theory. It is characterized by a bell-shaped curve that is symmetric around its mean.

Probability Density Function: The PDF of the normal distribution is given by:



Normal Distribution Formula:

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

where x is a random variable, μ is the mean, σ2 is the variance, π equal to 3.14159, and e is the base of the natural logarithm (approximately equal to 2.71828).

1. **68-95-99.7 Rule**: In a normal distribution:

   68% of the population is within 1 standard deviation of the mean.
   95% of the population is within 2 standard deviation of the mean.
   99.7% of the population is within 3 standard deviation of the mean.

2. **Applications**: The normal distribution is widely used in statistics, science, and engineering due to its symmetry.

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

Handling missing data is crucial in data analysis to avoid bias and loss of information. Some common techniques for handling missing data include:

- **Mean/Median/Mode Imputation**: Replace missing values with the mean, median, or mode of the non-missing values in that column.
- **Forward Fill or Backward Fill**: Propagate the last observed value forward or the next observed value backward.
- **Multiple Imputation**: Generate multiple plausible values for each missing value to account for uncertainty.
- **Prediction Models**: Use regression or machine learning models to predict missing values based on other variables.

12. What is A/B testing?

A/B testing (or split testing) is a method of comparing two versions of a webpage, app feature to determine which one performs better. It involves randomly assigning users to two groups: A and B where each group experiences a different version (A or B) of the test variable. Statistical analysis is then used to determine if there is a statistically significant difference in performance between the two versions, often focusing on metrics such as conversion rates, click-through rates, or user engagement.

Example:

- Product Descriptions: Testing different product descriptions or formats (e.g., bullet points vs. paragraph style) to see which leads to more purchases.
- Checkout Process: Testing variations in the checkout process (e.g., single-page checkout vs. multi-step process) to identify the most user-friendly option.

13. Is mean imputation of missing data acceptable practice?

Mean imputation involves replacing missing values with the mean of the observed values for that variable. While it is a simple and commonly used method, it has limitations. Mean imputation can distort the distribution and variance of the data, It can also underestimate standard errors and correlations between variables. Therefore, while mean imputation is convenient, it should be used cautiously based on the specific dataset in hand.

14. What is linear regression in statistics?

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.

A linear regression line equation is written in the form of: **Y = a + bX** where X is the independent variable and plotted along the x-axis Y is the dependent variable and plotted along the y-axis.The slope of the line is b, and a is the intercept (the value of y when x = 0).

15. What are the various branches of statistics?

Statistics is broadly divided into two major branches, each focusing on different aspects of data analysis and inference:

- **Descriptive Statistics**: Involves summarizing and describing data using measures such as mean, median, mode, variance, and graphical methods.

- **Inferential Statistics**: Deals with making inferences and predictions about a population based on sample data, using techniques such as hypothesis testing and confidence intervals.

MACHINE LEARNING

1. Which of the following methods do we use to find the best fit line for data in Linear Regression? A) Least Square Error B) Maximum Likelihood C) Logarithmic Loss D) Both A and B

   Answer: D

2. Which of the following statement is true about outliers in linear regression? A) Linear regression is sensitive to outliers B) linear regression is not sensitive to outliers C) Can't say D) none of these

   Answer: A

3. A line falls from left to right if a slope is ______? A) Positive B) Negative C) Zero D) Undefined

   Answer: B

4. Which of the following will have symmetric relation between dependent variable and independent variable? A) Regression B) Correlation C) Both of them D) None of these

   Answer: B

5. Which of the following is the reason for over fitting condition? A) High bias and high variance B) Low bias and low variance C) Low bias and high variance D) none of these

   Answer: C

6. If output involves label then that model is called as: A) Descriptive model B) Predictive modal C) Reinforcement learning D) All of the above

   Answer: B

7. Lasso and Ridge regression techniques belong to _________? A) Cross validation B) Removing outliers C) SMOTE D) Regularization

   Answer: D

8. To overcome with imbalance dataset which technique can be used? A) Cross validation B) Regularization C) Kernel D) SMOTE

   Answer: D

9. The AUC Receiver Operator Characteristic (AUCROC) curve is an evaluation metric for binary classification problems. It uses _____ to make graph? A) TPR and FPR B) Sensitivity and precision C) Sensitivity and Specificity D) Recall and precision

   Answer: A

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

    Answer: B

11. Pick the feature extraction from below: A) Construction bag of words from a email B) Apply PCA to project high dimensional data C) Removing stop words D) Forward selection

    Answer: A

12. Which of the following is true about Normal Equation used to compute the coefficient of the Linear Regression? A) We don't have to choose the learning rate. B) It becomes slow when number of features is very large. C) We need to iterate. D) It does not make use of dependent variable.

    Answer: A,B

13. Explain the term regularization?

Regularization in machine learning serves as a method to forestall a model from overfitting. Overfitting transpires when a model not only discerns the inherent pattern within the training data but also incorporates the noise, potentially leading to subpar performance on fresh, unobserved data. The employment of regularization aids in mitigating this issue by augmenting a penalty to the loss function employed for model training.The primary goal of regularization is to reduce the model's complexity to make it more generalizable to new data, thus improving its performance on unseen datasets.

14. Which particular algorithms are used for regularization?

Here are some of the commonly used algorithms that incorporate regularization:

1. **Linear Regression**:
   - **Ridge Regression (L2 Regularization)**: Adds a penalty term proportional to the square of the coefficients to the loss function.
   - **Lasso Regression (L1 Regularization)**: Adds a penalty term proportional to the absolute value of the coefficients to the loss function, promoting sparsity.
2. **Logistic Regression**:
   - Similar to linear regression, logistic regression can also use L2 (Ridge) or L1 (Lasso) regularization techniques to penalize large coefficients.

3. **Support Vector Machines (SVM)**:
    - SVMs can use regularization through the tuning of the regularization parameter CCC. A smaller CCC value increases regularization strength, similar to Ridge regression.

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

In linear regression, the term "error" refers to the difference between the observed values of the dependent variable and the values predicted by the linear regression model. This concept is fundamental to understanding how well the model fits the data and to assessing its predictive accuracy.

The general form of a linear regression equation for a single independent variable XXX predicting a dependent variable y

$y=a0+a1x+e$

where:

- y is the observed value of the dependent variable (response variable),
- x is the observed value of the independent variable (feature),
- a0,a1 are the coefficients (parameters) of the linear regression model that need to be estimated,
- e is the error term.

The error term e represents the difference between the actual observed values of y and the values predicted by the linear regression model