MACHINE LEARNING

In Q1 to Q11, only one option is correct, choose the correct option:

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

Ans - 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

Ans – A

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

A) Positive B) Negative
C) Zero D) Undefined

Ans – 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

Ans - C

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

Ans – 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

Ans – B

7. Lasso and Ridge regression techniques belong to _________?

A) Cross validation B) Removing outliers
C) SMOTE D) Regularization

Ans – D

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

A) Cross validation B) Regularization
C) Kernel D) SMOTE

Ans - 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

Ans – A

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

A) True B) False

Ans – 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

Ans - B

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.

Ans - D

13. Explain the term regularization?

Ans -

Regularization means restricting a model to avoid overfitting by shrinking the coefficient estimates to zero. When a model suffers from overfitting, we should control the model's complexity. Technically, regularization avoids overfitting by adding a penalty to the model's loss function:

There are three commonly used regularization techniques to control the complexity of machine learning models, as follows:

- L2 regularization
- L1 regularization

L2 regularization

- Ridge regression is one of the types of linear regression in which a small amount of bias is introduced so that we can get better long-term predictions.
- Ridge regression is a regularization technique, which is used to reduce the complexity of the model. It is also called as **L2 regularization**.

L1 regularization

Lasso regression is another regularization technique to reduce the complexity of the model. It stands for **Least Absolute and Selection Operator.**

It is similar to the Ridge Regression except that the penalty term contains only the absolute weights instead of a square of weights.

Since it takes absolute values, hence, it can shrink the slope to 0, whereas Ridge Regression can only shrink it near to 0.

14. Which particular algorithms are used for regularization?

Ans -

Regularization is a technique used to prevent overfitting in machine learning models by adding a penalty term to the objective function. Two common algorithms used for regularization are:

Lasso Regression (L1 Regularization): Lasso regression adds the absolute values of the coefficients as a penalty term to the objective function. It can lead to sparse models, meaning some of the coefficients become exactly zero.

Ridge Regression (L2 Regularization): Ridge regression adds the squared values of the coefficients as a penalty term to the objective function. It tends to shrink the coefficients toward zero without necessarily making them exactly zero.

These regularization techniques are especially useful when dealing with linear regression models, and they help in controlling the complexity of the model and preventing overfitting.

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

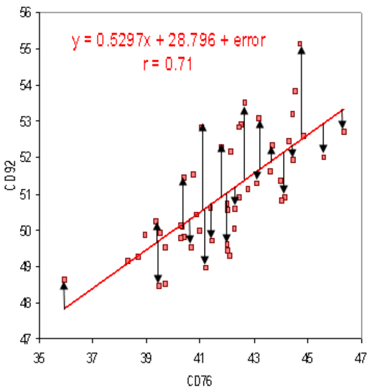
Ans –

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

Examples of the Error Term in Statistics

In econometric theory, the classical normal linear regression model (CNLRM) involves finding the best fitting linear model for observed data that shows the relationship between two variables.

For example, let's say you were running a study on the way the number of exams in a certain college affect the amount of red bull purchased from college vending machines. You could collect data which told you how many exams were given and how much red bull was purchased on a dozen or more days during the semester. This data can be plotted as a scatter plot, with exams (E^x) per given day on the x axis and red bull purchased (R^B) per given day on the y axis. Then you would look for the line $y = \beta_0 + \beta_1$ x that best fit the data.



“Best fit” here means that the error term, the distance from each point to the line, is minimized. Since the relationship between variables is probably not completely linear and because there are other factors outside the scope of our study (sales on red bull, sales on other caffeine drinks, difficult physics homework sets, etc.) the graph of the probability distribution won’t actually go through all our data points. The distance between each point and the linear graph (shown as black arrows on the above graph) is our error term. So we can write our function as $RB=\beta0 + \beta1\ Ex + \varepsilon$ where $\beta0$ and $\beta1$ are constants and ε is an (non constant) error term

PYTHON – WORKSHEET 1

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

Ans –C

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

Ans - B

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

Ans – C

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

Ans - A

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

Ans – 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

Ans – 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

Ans – 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.

Ans - C

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

Ans – A & C

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

Ans – B

STATISTICS WORKSHEET-1

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

a) True

b) False

Ans – 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

Ans – 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

And – 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

Ans – C

5. ______ random variables are used to model rates.

a) Empirical
b) Binomial
c) Poisson
d) All of the mentioned

Ans – C

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

a) True
b) False

Ans – B

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

a) Probability
b) Hypothesis
c) Causal
d) None of the mentioned

Ans – B

8. 4. Normalized data are centered at______and have units equal to standard deviations of the original data.

a) 0
b) 5
c) 1
d) 10

Ans – 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

Ans – C

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

Ans -

Normal distribution, the most common distribution function for independent, randomly generated variables. Its familiar bell-shaped curve is ubiquitous in statistical reports, from survey analysis and quality control to resource allocation.
The graph of the normal distribution is characterized by two parameters: the mean, or average, which is the maximum of the graph and about which the graph is always symmetric; and the standard deviation, which determines the amount of dispersion away from the mean. A small standard deviation (compared with the mean) produces a steep graph, whereas a large standard deviation (again compared with the mean) produces a

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

Ans –

One of the most common problems I have faced in Data Cleaning/Exploratory Analysis is handling the missing values. Firstly, understand that there is NO good way to deal with missing data. I have come across different solutions for data imputation depending on the kind of problem — Time series Analysis, ML, Regression etc. and it is difficult to provide a general solution. I am attempting to summarize the most commonly used methods and trying to find a structural solution.

Imputation Techniques

Before jumping to the methods of data imputation, we have to understand the reason why data goes missing.

1. **Missing at Random :** Missing at random means that the propensity for a data point to be missing is not related to the missing data, but it is related to some of the observed data

2. **Matrix Factorization Techniques:** Use methods like Singular Value Decomposition (SVD) or Principal Component Analysis (PCA) to fill in missing values based on the patterns in the observed data.

3. **K-Nearest Neighbors (KNN) Imputation**: Estimate missing values based on the values of their k-nearest neighbors. This method considers similarity between instances to impute missing values.

12. What is A/B testing?

Ans –

A/B testing is an experimental method in which two versions of anything are contrasted to see which is **“better” or more effective**.
This is often done in marketing when two different types of content—whether it be **email copy, a display ad, a call-to-action (CTA)** on a web page, or any other marketing asset—are being compared. This is usually done before launching any product in the market so that the company can get better results.
This also helps in comparing the performance of **two or more variants of emails** and then selecting the best among them based on the result given by the audience. So, now without waiting any time let’s move forward and take a look that what is A/B testing:

13. Is mean imputation of missing data acceptable practice?

Ans

Mean imputation, where missing values are replaced with the mean of the observed (non-missing) values for that variable, is a simple and quick method for handling missing data. While it is widely used due to its simplicity, there are both advantages and limitations to mean imputation, and its acceptability depends on the context and characteristics of the data. Here are some considerations:
Advantages:

1. **Simplicity:** Mean imputation is straightforward and easy to implement, making it a quick solution for handling missing data, especially in situations where time or resources are limited.
2. **Preservation of Sample Size:** Mean imputation retains the original sample size, ensuring that the analysis is conducted on the maximum available data.

14. What is linear regression in statistics?

Ans -

Linear regression is a data analysis technique that predicts the value of unknown data by using another related and known data value. It mathematically models the unknown or dependent variable and the known or independent variable as a linear equation. For instance, suppose that you have data about your expenses and income for last year. Linear regression techniques analyze this data and determine that your expenses are half your income. They then calculate an unknown future expense by halving a future known income.

15. What are the various branches of statistics?

Ans –

Statistics is a broad field that encompasses various branches, each focusing on different aspects of data collection, analysis, interpretation, and presentation. Here are some major branches of statistics:

Descriptive Statistics

Descriptive statistics deals with the presentation and collection of data. This is usually the first part of a statistical analysis. It is usually not as simple as it sounds, and the statistician needs to be aware of designing experiments, choosing the right focus group and avoid biases that are so easy to creep into the experiment.

Different areas of study require different kinds of analysis using descriptive statistics. For example, a physicist studying turbulence in the laboratory needs the average quantities that vary over small intervals of time. The nature of this problem requires that physical quantities be averaged from a host of data collected through the experiment.

Inferential Statistics

Inferential statistics, as the name suggests, involves drawing the right conclusions from the statistical analysis that has been performed using descriptive statistics. In the end, it is the inferences that make studies important and this aspect is dealt with in inferential statistics.

Most predictions of the future and generalizations about a population by studying a smaller sample come under the purview of inferential statistics. Most social sciences experiments deal with studying a small sample population that helps determine how the population in general behaves. By designing the right experiment, the researcher is able to draw conclusions relevant to his study.