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

   Answer: Least Square Error

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: Linear regression is sensitive to outliers

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

4. 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: Correlation

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: Low bias and high variance

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

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

   Answer: Regularization

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

Answer: SMOTE

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: TPR and FPR

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

A) **True** B) False

Answer: True

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 In

Answer: Construction bag of words from a email

Q12, more than one options are correct, choose all the correct options:

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) We don't have to choose the learning rate. B) It becomes slow when number of features is very large.

Q13 and Q15 are subjective answer type questions,

Answer them briefly.

13. Explain the term regularization?

Answer: Regularization is a set of methods for reducing overfitting in machine learning models. Typically, regularization trades a marginal decrease in training accuracy for an increase in generalizability.Regularization encompasses a range of techniques to correct for overfitting in machine learning models. such, regularization is a method for increasing a model's generalizability—that is, it's ability to produce accurate predictions on new datasets.1 Regularization provides this increased generalizability at the sake of increased training error. In other words, regularization methods typically lead to less accurate predictions on training data but more accurate predictions on test data.

Regularization differs from optimization. Essentially, the former increases model generalizability while the latter increases model training accuracy. Both are important concepts in machine learning and data science.There are many forms of regularization. Anything in the way of a complete guide requires a

much longer book-length treatment. Nevertheless, this article provides an overview of the theory necessary to understand regularization's purpose in machine learning as well as a survey of several popular regularization techniques. This concession of increased training error for decreased testing error is known as bias-variance trade-off. Bias-variance tradeoff is a well-known problem in machine learning. It's necessary to first define "bias" and "variance." To put it briefly:

Bias measures the average difference between predicted values and true values. As bias increases, a model predicts less accurately on a training dataset. High bias refers to high error in training.
Variance measures the difference between predictions across various realizations of a given model. As variance increases, a model predicts less accurately on unseen data. High variance refers to high error during testing and validation.
Bias and variance thus inversely represent model accuracy on training and test sets respectively.2 Obviously, developers aim to reduce both model bias and variance. Simultaneous reduction in both is not always possible, resulting in the need for regularization. Regularization decreases model variance at the cost of increased bias.

14. Which particular algorithms are used for regularization?

Answer:Regularization Algorithms

Ridge regression – Its purpose is to overcome problems such as data overfitting and multicollinearity in data. When there is considerable collinearity (the existence of near-linear connections among the independent variables) among the feature variables, a typical linear or polynomial regression model will fail. Ridge Regression adjusts the variables by a modest squared bias factor. The feature variable coefficients are pulled away from this rigidity by such a squared bias factor, providing a little bit of bias into the model but considerably lowering variation.

Ridge is an excellent way to prevent overfitting.

Use regularization to solve overfitting and feature selection if you have a model with a high number of features in the dataset and want to prevent making the model too complicated.However, the ridge has one major drawback: the final model has all N characteristics.Ridge regression decreases the two coefficients towards each other when the variables are highly linked. Lasso is torn between the two and prefers one over the other.One never knows which variable will be chosen depending on the situation. Elastic-net is a hybrid of the two that tries to shrink while still doing the sparse selection.

LASSO – It simply penalizes large coefficients, in contrast to Ridge Regression. When the hyperparameter is big enough, Lasso has the effect of driving certain coefficient estimations to be absolutely zero. As a result, Lasso conducts variable selection, resulting in models that are significantly easier to read than Ridge Regression models. In a nutshell, it's about lowering variability and increasing the accuracy of linear regression models.If we have a large number of features, LASSO works effectively for feature selection.

It reduces coefficients to zero and if a set of predictors is highly associated, lasso selects one and reduces the others to zero.

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

Answer: Understanding an Error Term

An error term represents the margin of error within a statistical model; it refers to the sum of the deviations within the regression line, which provides an explanation for the difference between the theoretical value of the model and the actual observed results. The regression line is used as a point of analysis when attempting to determine the correlation between one independent variable and one dependent variable.

Error Term Use in a Formula

An error term essentially means that the model is not completely accurate and results in differing results during real-world applications. For example, assume there is a multiple linear regression function that takes the following form:

$Y=\alpha X+\beta\rho+\epsilon$where:$\alpha$,=Constant parameters$X$,$\rho$=Independent variables$\epsilon$=Error term
Y=*αX*+*βρ*+*ϵ***where:***α*,*β*=Constant parameters*X*,*ρ*=Independent variables*ϵ*=Error term

When the actual Y differs from the expected or predicted Y in the model during an empirical test, then the error term does not equal 0, which means there are other factors that influence Y.

PYTHON – WORKSHEET 1

Q1 to Q8 have only one correct answer. Choose the correct option to answer your question.

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

A) # B) & C) % D) $

Answer: %

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

A) 0.666 B) **0** C) 1 D) 0.67

Answer: 0

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

A) 36 B) 10 C) **24** D) 45

Answer: 24

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

A) **2** B) True C) False D) 0

Answer: 2

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

A) 2 B) 4 C) 0 D) **6**

Answer: 6

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: It encloses the lines of code which will be executed if any error occurs while executing the lines of code in the try block

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: It is used to raise an exception

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: in defining an iterator

Q9 and Q10 have multiple correct answers. Choose all the correct options to answer your question.

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

A) **_abc** B) 1abc C) **abc2** D) None of the above

Answer: _abc, abc2

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

A) **yield** B) **raise** C) look-in D) all of the above

Answer: yield, raise

Q11 to Q15 are programming questions.

Answer them in Jupyter Notebook.

11. Write a python program to find the factorial of a number.

12. Write a python program to find whether a number is prime or composite.

13. Write a python program to check whether a given string is palindrome or not.

14. Write a Python program to get the third side of right-angled triangle from two given sides.

15. Write a python program to print the frequency of each of the characters present in a given string.

Answers for above 5 questions are solved in jupyter notebook

WORKSHEET STATISTICS

Q1 to Q9 have only one correct answer.

Choose the correct option to answer your question.

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

a) **True** b) False

Answer: 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?

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

Answer: Central Limit Theorem

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: All of the mentioned

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: The exponent of a normally distributed random variables follows what is called the log-normal distribution

5. _____ random variables are used to model rates.

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

Answer: Poisson

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

a) True b) **False**

Answer: False

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: Hypothesis

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: 0

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: Outliers can be the result of spurious or real processes

Q10and Q15 are subjective answer type questions,

Answer them in your own words briefly.

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

Answer: Normal distribution, also known as the Gaussian distribution, 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. The normal distribution appears as a "bell curve" when graphed.

Normal distributions have key characteristics that are easy to spot in graphs:

- The mean, median and mode are exactly the same.
- The distribution is symmetric about the mean—half the values fall below the mean and half above the mean.
- The distribution can be described by two values: the mean and the standard deviation.

$z = (X - \mu) / \sigma$

where X is a normal random variable, μ is the mean of X, and σ is the standard deviation of X. You can also find the normal distribution formula here. In probability theory, the normal or Gaussian distribution is a very common continuous probability distribution.

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

Answer: Missing values are a common issue in machine learning. This occurs when a particular variable lacks data points, resulting in incomplete information and potentially harming the accuracy and dependability of your models. It is essential to address missing values efficiently to ensure strong and impartial results in your machine-learning projects. In this article, we will see How to Handle Missing Values in Datasets in Machine Learning.

Missing values can pose a significant challenge in data analysis, as they can:

- **Reduce the sample size:** This can decrease the accuracy and reliability of your analysis.
- **Introduce bias:** If the missing data is not handled properly, it can bias the results of your analysis.
- **Make it difficult to perform certain analyses:** Some statistical techniques require complete data for all variables, making them inapplicable when missing values are present

12. What is A/B testing?

Answer: A/B testing compares two versions of an app or webpage to identify the better performer. It's a method that helps you make decisions based on real data rather than just guessing. It compares options to learn what customers prefer. You can test website/app layouts, email subject lines, product designs, CTA button text, colors, etc.

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 drives business metrics.

Essentially, A/B testing eliminates all the guesswork out of website optimization and enables experience optimizers to make data-backed decisions. In A/B testing, A refers to 'control' or the original testing variable. Whereas B refers to 'variation' or a new version of the original testing variable. The version that moves your business metric(s) in the positive direction is known as the 'winner.' Implementing the changes of this winning variation on your tested page(s) / element(s) can help optimize your website and increase business ROI. The metrics for conversion are unique to each website. For instance, in the case of eCommerce, it may be the sale of the products. Meanwhile, for B2B, it may be the generation of qualified leads.

A/B testing is one of the components of the overarching process of Conversion Rate Optimization (CRO), using which you can gather both qualitative and quantitative user insights. You can further use this collected data to understand user behavior, engagement rate, pain

points, and even satisfaction with website features, including new features, revamped page sections, etc. If you're not A/B testing your website, you're surely losing out on a lot of potential business revenue.

13. Is mean imputation of missing data acceptable practice?

Answer: The process of replacing null values in a data collection with the data's mean is known as mean imputation.

Mean imputation is typically considered terrible practice since it ignores feature correlation. Consider the following scenario: we have a table with age and fitness scores, and an eight-year-old has a missing fitness score. If we average the fitness scores of people between the ages of 15 and 80, the eighty-year-old will appear to have a significantly greater fitness level than he actually does.

In a single imputation method the missing data are filled by some means and the resulting completed data set is used for inference. Mean imputation (MI) is one such method in which the mean of the observed values for each variable is computed and the missing values for that variable are imputed by this mean.

Second, 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 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.

This form of analysis estimates the coefficients of the linear equation, involving one or more independent variables that best predict the value of the dependent variable. Linear regression fits a straight line or surface that minimizes the discrepancies between predicted and actual output values. There are simple linear regression calculators that use a "least squares" method to discover the best-fit line for a set of paired data. You then estimate the value of X (dependent variable) from Y (independent variable). Linear-regression models are relatively simple and provide an easy-to-interpret mathematical formula that can generate predictions. Linear regression can be applied to various areas in business and academic study.

You'll find that linear regression is used in everything from biological, behavioral, environmental and social sciences to business. Linear-regression models have become a proven way to scientifically and reliably predict the future. Because linear regression is a long-established statistical procedure, the properties of linear-regression models are well understood and can be trained very quickly.

15. What are the various branches of statistics?

Answer: The two main branches of statistics are descriptive statistics and inferential statistics. Both of these are employed in scientific analysis of data and both are equally important for the 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. Home >
- Descriptive Statistics and Inferential Statistics
  Every student of statistics should know about the different branches of statistics to correctly understand statistics from a more holistic point of view. Often, the kind of job or work one is involved in hides the other aspects of statistics, but it is very important to know the overall idea behind statistical analysis to fully appreciate its importance and beauty.

  The two main branches of statistics are descriptive statistics and inferential statistics. Both of these are employed in scientific analysis of data and both are equally important for the student of statistics.
  Descriptive StatisticsDescriptive 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 StatisticsInferential 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.