ASSIGNMENT – 39

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 D) Both A and B

Least Square Error and Maximum Likelihood Methods are used.

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) Linear regression is sensitive to outliers

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

A) Positive B) Negative

C) Zero D) Undefined

Answer If a slope is Negative

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 C) Both of them

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) Low bias and high variance

6. 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 B) 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 A) 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 False

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 B) Apply PCA to project high dimensional data

In 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.

ASSIGNMENT – 39

MACHINE LEARNING

Q13 and Q15 are subjective answer type questions, Answer them briefly.

13. Explain the term regularization?

Answer Regularization in machine learning is a technique used to prevent overfitting. It is applied where a model becomes too complex and fits the training data too closely leading to poor generalization on new or unseen data.

It helps to achieve a more balanced model that performs well on both the training and test datasets by avoiding the memorization of noise in the training data. Common regularization methods include
a. L1 regularization (Lasso)

b. L2 regularization (Ridge)

c. elastic net

14. Which particular algorithms are used for regularization?

Answer Regularization is a concept applied to various machine learning algorithms to prevent overfitting. Common algorithms used for Regularization are:

1. Linear Regression:

- L1 Regularization (Lasso Regression)

- L2 Regularization (Ridge Regression)

2. Logistic Regression:

- L1 Regularization

- L2 Regularization

3. Support Vector Machines (SVM)

- L2 Regularization

4. Decision Trees and Random Forests:

- Regularization techniques are typically embedded in tree pruning methods.

6. Elastic Net:

- Combines both L1 and L2 regularization and is applicable to linear regression problems.

The specific choice of regularization method depends on the characteristics of the problem and the desired properties of the model.

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

Answer

In the context of linear regression, the term "error" refers to the difference between the predicted values generated by the regression model and the actual observed values in the dataset. This difference can also be referred as "residuals."

The linear regression equation models the relationship between the independent variable and the dependent variable as a linear combination of coefficients. Mathematically, it can be represented as:

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n}$$

The error term is the difference between the predicted values and the actual values. In an ideal scenario, the model would make perfect predictions and the error would be zero for every data point.

But in reality, due to factors such as noise and complexity in real-world phenomena, there is usually some level of error present in the predictions.

The goal of linear regression is to minimize the sum of squared errors (SSE) or residuals which can be achieved through methods like least squares to find the optimal values for the coefficients that minimize the difference between predicted and actual values.

WORKSHEET

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 **Modulo %**

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

A) 0.666 B) 0

C) 1 D) 0.67

Answer B) 0 (// means Floor Division in Python)

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

A) 36 B) 10

C) 24 D) 45

Answer C) 24 (“<<” operator means left-shift operator in python. We get the binary value 11000 equal to 24 in decimal)

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

A) 2 B) True

C) False D) 0

Answer D) 0 (“&” operator is bitwise AND. 6 is 110 and 2 is 010. Performing 6&2, we get 000 equal to 0)

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

A) 2 B) 4

C) 0 D) 6

Answer D) 6 (“|” operator is bitwise OR operator in python. 6 is 110 and 2 is 010.Performing 6|2 gives 110 which is 6 in decimal)

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) 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?

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) 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 C) in defining a generator

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 A) _abc C) abc2

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

A) yield B) raise

C) look-in D) all of the above

Answer A) yield B) 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.

WORKSHEET

STATISTICS WORKSHEET-1

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 a) 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 b) Modeling bounded count data

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 c) The square of a standard normal random variable follows what is called chi-squared distribution

5. _____ random variables are used to model rates.

a) Empirical

b) Binomial

c) Poisson

d) All of the mentioned

Answer Poisson

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

a) True

b) False

Answer True

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

Answer b) Hypothesis

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

Answer Normalized data are centered at 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 **c**) Outliers cannot conform to the regression relationship

WORKSHEET

Q10and Q15 are subjective answer type questions, Answer them in your own words briefly.

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

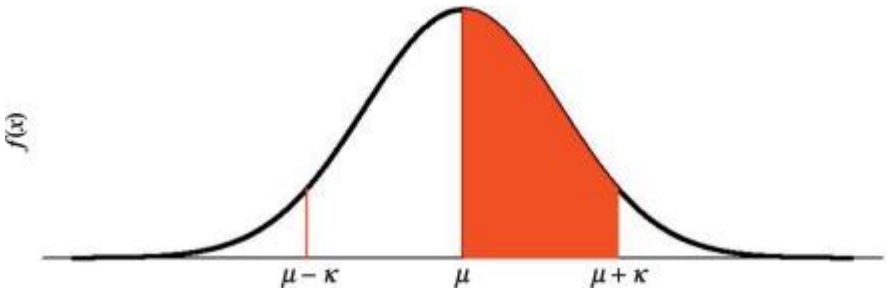
Answer A normal distribution is also known as a "Gaussian distribution". It is a symmetric probability distribution that is characterized by a bell-shaped curve. In a normal distribution:

1. Symmetry: The distribution is symmetric around its mean, which is also its median and mode. The curve is bell-shaped.

2. Mean, Median, and Mode Equality: The mean (average), median, and mode of a normal distribution are equal and located at the center of the distribution.

3. 68-95-99.7 Rule: Approximately 68% of the data falls within one standard deviation of the mean, 95% within two standard deviations, and 99.7% within three standard deviations.

4. Parameters: The distribution is defined by two parameters - the mean (μ) and the standard deviation (σ).



The normal distribution is a fundamental concept in statistics and is often used in various fields to model and analyze random variables. For Example - such as heights, weights, and measurement errors tend to follow a normal distribution making it an essential statistical tool.

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

Answer Handling missing data is crucial in data analysis and machine learning to avoid biased or inaccurate results. Here are some common techniques for handling missing data:

1. Deletion:

- Listwise Deletion: Removing entire rows with missing values.

2. Imputation:

Techniques used for handling missing data:

a. Mean, Median, or Mode Imputation: Replace missing values with the mean, median, or mode of the observed values in that variable.

b. Forward Fill or Backward Fill: Propagate the last observed value forward or use the next observed value backward to fill missing entries.

c. K-Nearest Neighbors (KNN) Imputation: Predict missing values based on the values of their k-nearest neighbors in the feature space.

d. Multiple Imputation: Generate multiple datasets with imputed values and analyze each separately considering the variability introduced by imputation.

3. Prediction Models:

- Use machine learning models to predict missing values based on other features.

The choice of imputation technique depends on the nature of data, the extent of missingness, and understanding the underlying patterns causing the missing data.

12. What is A/B testing?

Answer A/B testing is also known as split testing. It is a statistical method used to compare two versions (A and B) of a variable, like a webpage, app, email etc. to determine which one performs better. It analyses the differences in their responses or outcomes.

Some features of the A/B testing process:

1. Setup: Identify a metric you want to improve. Create two versions (A and B) of the element you want to test. Version A is often the existing version while version B includes the changes you want to test.

2. Randomization: Users are randomly assigned to either version A or B. This helps ensure that any differences in performance can be attributed to the changes made.

3. Implementation: Show version A to one group and version B to another and collect relevant data on user interactions.

4. Analysis: Compare the performance of A and B using statistical methods to determine if there's a significant difference in the measured metric. For example Use chi-square tests.

5. Conclusion: Based on the analysis, decide whether the changes in version B had a statistically significant impact on the chosen metric. If so, you may choose to implement the better-performing version.

A/B testing is widely used in marketing, product development and user experience optimization to make data-driven decisions about design, content or features.

13. Is mean imputation of missing data acceptable practice?

Answer Mean imputation is where missing values are replaced with the mean of the observed values in a variable. It is a commonly used method for handling missing data. However, its acceptability depends on the context and characteristics of the data. Here are some considerations:

Advantages:

1. Simplicity: Mean imputation is simple to implement and computationally efficient.

2. Preservation of Sample Size

Concerns:

1. Bias:

2. Distortion of Variability

3. Impact on Relationships

4. Not Suitable for Categorical Data: Mean imputation is inappropriate for categorical data because it applies only to numerical variables.

Alternatives:

1. Consider Imputation Methods: k-nearest neighbors imputation

2. Understand Data Missingness:

Mean imputation is a simple approach, its appropriateness depends on the assumptions made about the missing data mechanism and the nature of the data.

14. What is linear regression in statistics?

Answer Linear regression is a statistical method used to model the relationship between a dependent variable (denoted as Y) and one or more independent variables (denoted as X). The goal

is to find the linear relationship that best predicts the values of the dependent variable based on the given independent variables.

The general form of a simple linear regression model with one independent variable is:

$y= a0+a1x+ \varepsilon$

Here,

Y= Dependent Variable (Target Variable)

X= Independent Variable (predictor Variable)

a0= intercept of the line

a1 = Linear regression coefficient

ε = random error

Linear regression is widely used in various fields including economics, finance, biology for predictive modelling and understanding the relationships between variables.

15. What are the various branches of statistics?

Answer Statistics is a broad field with several branches working on different aspects of data such as analysis, interpretation, and inference.

Some of the various branches of Statistics are:

1. Descriptive Statistics
2. Inferential Statistics
3. Social Statistics
4. Business Statistics
5. Spatial Statistics
6. Bayesian Statistics
7. Environmental Statistics
8. Statistical Computing