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

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

A) Linear regression is sensitive to outliers

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

B) Negative

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

B) Correlation

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

C) Low bias and high variance

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

B) Predictive modal

7. Lasso and Ridge regression techniques belong to _____?

D) Regularization

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

D) 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

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

B) 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

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.

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

13. Explain the term regularization?

Regularization is one of the most important concepts of machine learning. It is a technique to prevent the model from overfitting by adding extra information to it. It is a form of regression that shrinks the coefficients estimates towards zero. In other words, this technique forces us not to learn a more complex or flexible model, to avoid the problem of overfitting.

In the regularization technique we reduce the magnitude of the independent variables by keeping the same number of variables.

Following are the two techniques of the regularization;

1. Ridge Regression
2. Lasso Regression
3. Dropout

14. Which particular algorithms are used for regularization?

There are three particular algorithms which are used for regularization;

1. Ridge Regression
2. Lasso Regression

1. Ridge Regression → Ridge regression is a regularization technique to perform well where there is a high variance as compare to the data which is used to train the model. Ridge regression is used to avoid the overfitting of data. By using the regularization technique it penalize the model for high variance and create new fit line and gives the better output.

2.Lasso Regression → Lasso regression is mostly similar to the ridge regression but there is a one important difference between them. In Ridge regression the magnitude of coefficient is almost zero but in Lasso regression the magnitude of coefficient is exactly zero.

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

An error term is a value which shows that how observed data is differ from the actual data. An error term includes everything that separates your model from actual reality.

PYTHON

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?

C) %

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

A) 0.66

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

C) 24

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

A) 2

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

D) 6

6. What does the finally keyword denotes in python?

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.

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

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

C) abc2

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

A) yield

B) raise

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

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

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

b) Modeling bounded count data

4. Point out the correct statement.

d) All of the mentioned

5. _____ random variables are used to model rates.

c) Poisson

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

b) False

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

b) Hypothesis

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

a) 0

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

c) Outliers cannot conform to the regression relationship

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

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

The normal distribution also known as the Gaussian distribution in statistics for independent random variable. The normal distribution is continuous probability distribution that is symmetrical around its mean.

As with any probability distribution, the normal distribution describes how the values of a variable are distributed. It is the most important probability distribution in statistics because it accurately describes the distribution of values for many natural phenomena.

Normal distributions are symmetrical but not all symmetrical distributions are normal. In normal distribution data is distributed symmetrically and when plotted on graph , the data follows a bell shape.

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

Following are the imputation techniques for handling the missing data;

1. Delete the rows
2. Replace with the most frequent value
3. Apply classifier algorithm to Predict
4. Apply Unsupervised Machine learning

As I write the above techniques are basics and more useful techniques to handle the missing data.

By using these various techniques we can handle the missing data. According to first technique (delete the row) we have to delete the missing data row, but sometimes by using this technique the important data should be deleted, so we avoid to use this technique. Similarally by using the Replace most frequent value technique, the data will be imbalanced.

The remaining two techniques, Apply classifier algorithm to predict and Apply Unsupervised Machine learning. By applying these technique we can used the different classifier algorithm to predict the missing data. By applying Unsupervised Machine learning we can handle the missing data by grouping technique.

12. What is A/B testing?

A/B testing is a way to compare two version of something to figure out which performs better. While its most often associated with website and app. A/B testing is also known as split testing. A/B testing provides the most benefits when it operates continuously. A regular flow of test can deliver a stream of recommendations on how to fine-tune performance. And continuous testing is possible because the available options for testing are nearly unlimited.

A/B testing can be used to evaluate just about any digital marketing assets;

Ex. emails, website pages, advertisement, mobile apps, etc.

A/B testing plays an important role in campaign management since it helps determine what is and isn't working. A/B testing can help you to see which element of your marketing strategy has the biggest impact, and which one needs to be dropped altogether.

13. Is mean imputation of missing data acceptable practice?

Mean is nothing but the sum of the numbers divided by the count of the numbers. In machine learning we used the mean data for the numerical features not for the categorical ones. There is a lots of disadvantages to use the mean imputation, so it is not a good idea to handle the missing data by using mean imputation technique.

14. What is linear regression in statistics?

Linear regression is the most basic and commonly used predictive analysis. The variable you want to predict is called dependent variable and the variable you are using to predict the another variable's value is called the independent variable. Regression estimates are used to describe data and to explain the relationship.

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.

There are simple linear regression calculators that uses a "least squares" method to discover the best fit line for a set of paired data. Estimate the value of Y (dependent variable) from X (independent variable).

Ex. $Y = a + bX$

Where, Y = Dependent variable

X = Independent variable.

15. What are the various branches of statistics?

Statistics have mainly two branches;

1. Descriptive Statistics
2. Inferential Statistics

In Descriptive Statistics we analyze the data, summarize the data and organizing the data in the form of number and graph. Mean, median and mode are also included in descriptive statistics. Barplot, histogram, pie-chart are some of the examples of descriptive statistics.

Inferential Statistics involves using a sample to draw conclusion about a population. Hypothesis testing, chi square testing are the examples of inferential statistics.