MACHINE LEARNING

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

Q1-Which of the following method do we use to find the best fit line for data in Linear Regression?

Ans.- (a**)** Lear Square Error

Q2- Which of the following statement is true about outliers in linear regression?

Ans. (a**)**Linear regression is sensitive to outliers

Q3-A line falls from left to right if a slope is_______?

Ans-(b) Negative

Q4-Which of the following will have symmetric relation between independent variable and dependent variable?

Ans-(b) Correlation

Q5- Which of the following is the reason for over fitting condition?

Ans.-(d)none of these

Q6-If output involves label then that model is known as________

Ans-(a**) descriptive model**

Q7-LASSO and RIDGE regression technique belongs to_________

Ans- (d) Regularization

Q8-To overcome with imbalance dataset which technique can be used?

Ans-(d) SMOTE

Q9-The AUC Receiver Operator Characteristics (AUCROC) curve is an evaluation metric for binary classification problems. It uses _________ to make graph?

Ans-(c) sensitivity and specificity

Q10- In AUC Receiver Operator Characteristics (AUCROC) curve for the better model area under the curve should be less ?

Ans- (b) False

Q11-Pick the feature extraction from below:

Ans-B) Apply PCA to project high dimensional data

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

Q12-which of the following is true about Normal Equation used to compute the coefficient of the Linear Regression?

Ans-opt. 1,2 and 3

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

Q13-Explain the term regularization.

A-when we use regression model to train some data there is a good chance that the model will overfit the given training data set. Regularization helps this overfitting problem by restricting the degrees of freedom of a given equation i.e. simply reducing the number of degrees of a polynomial function by reducing their corresponding weights.

In a linear equation we don't want huge weights/coefficients as a small change in weight can make a large difference for the dependent variable. So regularization constraints the weights of such features to avoid overfitting.

To regularize the model a shrinkage penalty is added to the cost of function. Let's see different types of regularization in regression:

- LASSO
- RIDGE
- ELASTICENT

Q14-which particular algorithm is used for regularization?

A-The algorithms which are used for regularizations are :

- LASSO
- RIDGE

1. LASSO :-

LASSO(Least Absolute Shrinkage and Selection Operator) Regression L1 form

LASSO regression penalizes the model based on the sum of magnitude of the coefficients. The regularization term is given by

Regularization = λ*$\sum|\beta j|$
Where, λ=the shrinkage factor

2. RIDGE :-

RIDGE REGRESSION(L2 Form)

Ridge regression penalizes the model selection based on the sum of the squares of magnitude of the coefficients. The regularization term is given by

Regularization=λ*$\sum|\beta^2 j|$
Where, λ=the shrinkage factor

Q15-Explain the term error present in linear regression equation.

Ans.-An error term is a residual variable produced by a statistical or mathematical model, which is created when the model does not fully represent the actual relationship between the independent variables and the dependent variables . As a result of this incomplete relationship , the error term is the amount at which equation may differ during empirical analysis.

Understanding an Error term

An error term represents the margin of error within a statistical mode; 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 result. The regression line is used as a point of analysis when attempting to determine the correlation between one independent variable and one dependent variable.

What do Error term tell us?

Within a linear regression model tracking a stock's price over time, the error term is the difference between the expected price at a particular time and the price that was actually observed. In instances where the price is exactly what was anticipated at a particular time , the price will fall on the trend line and the error term will be zero.

PYTHON

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

Q1 Which of the following operators is used to calculate remainder in a division?

Ans-(c) %

Q2. In python 2//3 is equal to?

Ans-(b) 0

Q3. In python, 6<< 2 is equal to

Ans-(c) 24

Q4. In python, 6&2 will give which of the following as output?

Ans-(a) 2

Q5. In python, 6|2 will give which of the following as output?

Ans-(d) 6

Q6. What does the finally keyword denotes in python?

Ans-(c) the finally block will be executed no matter the try block raises the error or not

Q7. What does raise keyword is used for in python?

Ans-(a) it is used to raise an exception

Q8. Which of the following is a common use case of yield keyword in python?

Ans-(D) in for loop.

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

Q9. Which of the following are the valid variable names?

Ans-(a and c) _abc and abc2

Q10. Which of the following are the keywords in python?

Ans-(a and b) yield and raise

STATISTICS

Q1 to Q9 have only one correct answer. Choose the correct option to answer your question

Q1. Bernoulli random variables take (only) the values 1 and 0

Ans-(a) true

Q2. 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?

Ans-(a) Central Limit Theorem

Q3. Which of the following is incorrect with respect to use of Poisson distribution?

Ans-(b) Modeling Bounded count data

Q4. Point out the correct statement.

Ans-(d) All of the mentioned

Q5_______ random variables are used to model rates.

Ans-(c) Poisson

Q6. 10. Usually replacing the standard error by its estimated value does change the CLT

Ans-(b) False

Q7. 1. Which of the following testing is concerned with making decisions using data?

Ans-(b) Hypothesis

Q8. 4. Normalized data are centered at______and have units equal to standard deviations of the original data.

Ans-(a) 0

Q9. Which of the following statement is incorrect with respect to outliers?

Ans-(c) Outliers cannot conform to the regression relationship.

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

Q10-what do u understand by normal distribution?

A-The normal distribution also known as the Gaussian distribution, is the most important probability distribution in statistics for independent , random variables. Most people recognized its familiar bell shaped curve in statistical reports.

The normal distribution is a continuous probability distribution that is symmetrical around its mean, most of the observations cluster around the central peak, and the probabilities for values further away from the mean tapper off equally in both directions. Extreme values in both tails of the distribution are similarly unlikely. While the normal distribution are normal. For example , the Student's , Cauchy and logistic distributions are symmetric.

As with any probabilities 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 phenomena. Characteristics that are the sum of many independent processes frequently follow normal distributions. For example heights, blood pressure, measurement error, and IQ scores follow the normal distribution.

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

Ans- Missing data can skew anything for data scientist , from economic to clinical trials. After all any analysis is only good as the data. A data scientist does not want to produce biased estimates that lead to invalid results. The concept of missing data is implied in the name: it is data that is not captured for a variable for the observation in question. Missing data reduces the statistical power of the analysis , which can discomfort the validity of the result.

When dealing with the missing data , data scientist can use two primary methods to solve the error:

- Imputation
- Removal of the data

The imputation method develops reasonable guesses for missing data. It is most useful when the percentage of missing data is low. If the portion of missing data is too high, the result lack natural variation that could result in an effective model.

The other option is removing data, when dealing with the data that is missing at random , related data can be deleted to reduce bias. Removing data may not be the best option if there are not enough observation to result in reliable analysis. In some situations, observation of specific events or factors may required.

There are some imputation technique which can be used to handle missing data

- Mean Imputation
- Regression Imputation
- Single or Multiple imputation………….etc.

Q12. What is A/B testing?

Ans-A/B testing is one of the most popular controlled experiments used to optimize web marketing strategies. It allows decision makers to choose the best design for a website by looking at the analytics results obtained with two possible alternative A and B.

To understand what A/B testing is about, let's consider two alternative design: A and B. visitors of a website are randomly served with one of the two. Then , data about their activity is collected by web analytics. Given this data, one can apply statistical test to determine whether one of the two designs has better efficacy.

Overview

A/B testing is a shorthand for a simple randomized controlled experiments, in which two samples (A and B) of a single vectorvariable are compared. These values are similar except for one variation which might affect a user's behaviour. A/B tests are widely considered the simplest form of controlled experiments. However, by adding more variants to the test, it is complexity grows.

Q 13. Is mean imputation of missing data acceptable practice?

Ans-True, imputing the mean preserves the mean of the observed data. So if the data are missing completely at random, the estimates of the mean remains unbiased. That is a good thing ……Since most research studies are interested in the relationship among variables, mean imputation is not a good solution.

Q 14. What is linear regression in statistics?

Ans- Linear Regression is a basic and commonly used type of predictive analysis. The overall idea of regression is to examine two things: (1) does a set of predictor variables do a good job in predicting an outcome(dependent) variable? (2) Which variable in particular are significant predictor of the outcome variable, and in what way do they- indicated by the magnitude and sign of the beta estimates – impact the outcome variable? These regression estimates are used to explain the relationship between one dependent variable and one or more independent variables. The simplest form of the regression equation with one dependent and one independent variable is defined by the formula

$$Y = c+b*x,$$

Where, y= estimated dependent variable score

C= constant

B= regression

X= score on the independent variable.

Types of linear regression:-

- Simple linear regression
- Multiple linear regression
- Logistic regression
- Ordinal regression
- Multinominal regression
- Discriminant regression

Q15. What are the various branches of statistics?

Ans- Types of statistics:-

- ➢ Descriptive statistics
- ➢ Inferential statistics

1. **Descriptive statistics**

In this type of statistics, the data is summarized through the given observations. The summarization is one form of a sample of population using parameter such as the mean or standard deviation.

Descriptive statistics is a way to organize, represent and describe a collection of data using tables, graphs and summary measures. For example, the collection of people in a city using the internet or using Television.

Descriptive statistics are also categorised into four different categories:

- Measure of frequency
- Measure of dispersion
- Measure of central tendency
- Measure of position

2. Inferential statistics

This type of statistics is used to interpret the meaning of Descriptive statistics. That means once the data has been collected , analysed and summarized then we use these stats to describe the meaning of the collected data. Or we can say , it is used to draw conclusions from the data that depends on random variations such as observational errors, sampling variation etc.

Inferential statistics is a method that allows us to use information collected from a sample to make decisions, predictions or inferences to give statements that goes beyond the available data or information. For example deriving estimates from hypothetical research.