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

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

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

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

5. _____ random variables are used to model rates.

a) Empirical

b) Binomial

c) Poisson

d) All of the mentioned

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

a) True

b) False

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

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

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

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?

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

Actually,

1st we use Data Dropping

Using the dropna() function is the easiest way to remove observations or features with missing values from the DataFrame. Below are some techniques.

Drop observations with missing values

These three scenarios can happen when trying to remove observations from a data set:

dropna(): drops all the rows with missing values.

dropna(how = ‘all’): the rows where all the column values are missing

2. We prefers

Drop columns with missing values.

The parameter axis = 1 can be used to explicitly specify we are interested in columns rather than rows. dropna(axis = 1): drops all the columns with missing values.

3 rd we use

Replace methods. mean, mode ,median.

There are some imputers to remove null values.

This are known as Advanced technics.

1. encoder_imputer

2. get_dummies

3. odinal_encoder

4. binary_encoder

5.knn_imputer

6.iterative_imputer

Here I recomanded

iterative_imputers. & encoder_impueters.

12. What is A/B testing?

In statistical terms, A/B testing is a method of two-sample hypothesis testing. This means comparing the outcomes of two different choices (A and B) by running a controlled mini-experiment. This method is also sometimes referred to as split testing.

13. Is mean imputation of missing data acceptable practice?

True, imputing the mean preserves the mean of the observed data. So if the data are missing completely at random, the estimate of the mean remains unbiased. That’s a good thing.

Plus, by imputing the mean, you are able to keep your sample size up to the full sample size. That’s good too.

14. What is linear regression in statistics?

Once the degree of relationship between variables has been established using co-relation analysis, it is natural to delve into the nature of relationship. Regression analysis helps in determining the cause and effect relationship between variables. It is possible to predict the value of other variables (called dependent variable) if the values of independent variables can be predicted using a graphical method or the algebraic method.

Graphical Method

It involves drawing a scatter diagram with independent variable on X-axis and dependent variable on Y-axis. After that a line is drawn in such a manner that it passes through most of the distribution, with remaining points distributed almost evenly on either side of the line.

A regression line is known as the line of best fit that summarizes the general movement of data. It shows the best mean values of one variable corresponding to mean values of the other. The regression line is based on the criteria that it is a straight line that minimizes the sum of squared deviations between the predicted and observed values of the dependent variable.

Algebraic Method

Algebraic method develops two regression equations of X on Y, and Y on X.

Regression equation of Y on X
$Y=a+bX$

Where −

- Y= Dependent variable
- X = Independent variable
- a = Constant showing Y-intercept
- b = Constant showing slope of line

15. What are the various branches of statistics?

Branches of Statistics

A) Descriptive Statistics

Descriptive statistics is the first part of statistics that deals with the collection of data.

Descriptive statistics have two parts;

- Central tendency measures
- Variability measures

There are 3 methods in central tendency measures

1. Mean

Mean is a conventional method used to describe the central tendency. Typically, calculate the average of values, count all values, and then divide them with the number of available values.

Formula of Mean

m= Sum of the terms/numbers of terms

2. Median

It is the result that is in the middle of a set of values. An easy way to calculate the median is to edit the results in numerical journals and locate the result that is in the center of the distributed sample

Formula of Median

To solve the median, there are two formulas;

- **When n is odd,**

(n+1 / 2)th observation

- **When n is even,**

median=(n/2)th + (n/2+ 1)th observation / 2

3. Mode

The mode is the frequently occurring value in the given data set.

B) Inferential statistics

Different types of inferential statistics include:

- **Regression analysis:** It is a set of statistical methods used to estimate relationships between a dependent variable and one or more independent variables. It includes several variations, like linear, multiple linear, and nonlinear. The most well-known models are simple linear and multiple linear.
- **Analysis of variance (ANOVA):** ANOVA is a statistical method that distributes observed variance data into various components. A one-way ANOVA is applied for three or more data groups to gain information about the relationship between the dependent and independent variables.
- **Analysis of covariance (ANCOVA):** It is used to test categorical variables' main and interaction effects on constant dependent variables and keep control for the impact of selected other constant variables. The control variables are known as covariates.
- **Statistical significance (t-test):** It is used to determine a significant difference between the means of two groups related to particular features. A t-test studies the t-statistic, the t-distribution values, and the degree of freedom to learn the statistical significance.
- **Correlation analysis:** It is a statistical method that is used to find the relationship between two variables or datasets and discover how strong the relationship may be.

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

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

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

A) Positive B) Negative

C) Zero D) Undefined

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

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

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

7. Lasso and Ridge regression techniques belong to _____?

A) Cross validation B) Removing outliers

C) SMOTE D) Regularization

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

A) Cross validation B) Regularization

C) Kernel 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 B) Sensitivity and precision

C) Sensitivity and Specificity D) Recall and precision

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

A) True 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.

C) We need to iterate.

D) It does not make use of dependent variable.

MACHINE LEARNING

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

13. Explain the term regularization?

While training a machine learning model, the model can easily be overfitted or under fitted.

To avoid this, we use regularization in machine learning to properly fit a model onto our test set. Regularization techniques help reduce the chance of overfitting and help us get an optimal model.

Regularization refers to techniques that are used to calibrate machine learning models in order to minimize the adjusted loss function and prevent overfitting or underfitting.

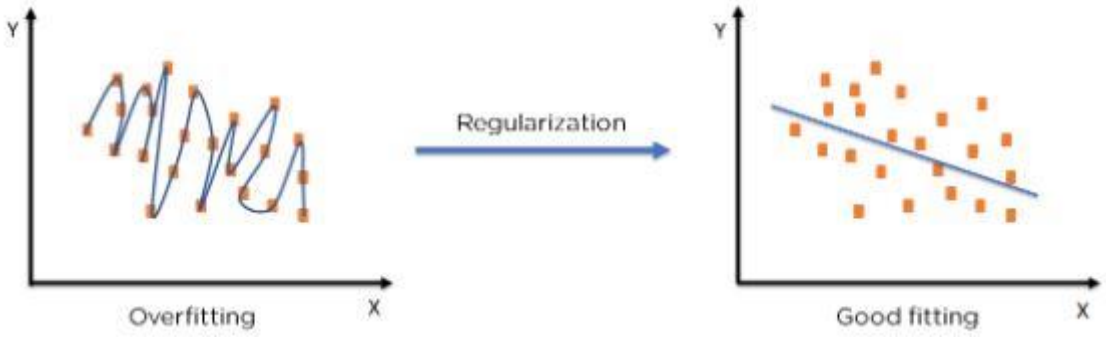

Figure 5: Regularization on an over-fitted model

Using Regularization, we can fit our machine learning model appropriately on a given test set and hence reduce the errors in it.

Regularization Techniques

There are two main types of regularization techniques:

1. Ridge Regularization (L2 Regularization)

2. Lasso Regularization (L1 Regularization

14. Which particular algorithms are used for regularization?

Ridge Regularization :

Also known as Ridge Regression, it modifies the over-fitted or under fitted models by adding the penalty equivalent to the sum of the squares of the magnitude of coefficients.

This means that the mathematical function representing our machine learning model is minimized and coefficients are calculated. The magnitude of coefficients is squared and added. Ridge Regression performs regularization by shrinking the coefficients present. The function depicted below shows the cost function of ridge regression :

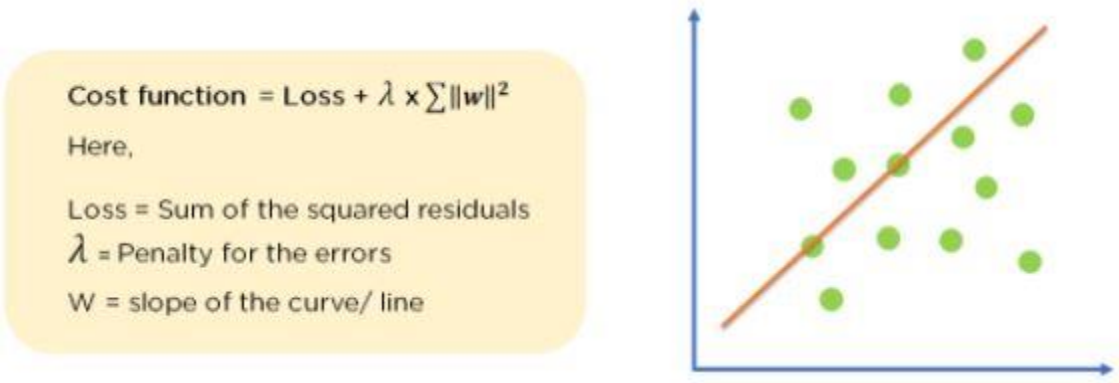

Figure 7: Cost Function of Ridge Regression

In the cost function, the penalty term is represented by Lambda λ. By changing the values of the penalty function, we are controlling the penalty term. The higher the penalty, it reduces the magnitude of coefficients. It shrinks the parameters. Therefore, it is used to prevent multicollinearity, and it reduces the model complexity by coefficient shrinkage.

Consider the graph illustrated below which represents Linear regression :

Cost function = Loss + λ x∑‖w‖^2

For Linear Regression line, let's consider two points that are on the line,

Loss = 0 (considering the two points on the line)

λ= 1 , w = 1.4 Then, Cost function = 0 + 1 x 1.42 = 1.96

For Ridge Regression, let's assume,

Loss = 0.32 + 0.22 = 0.13

λ = 1 , w = 0.7, Then, Cost function = 0.13 + 1 x 0.72 = 0.62

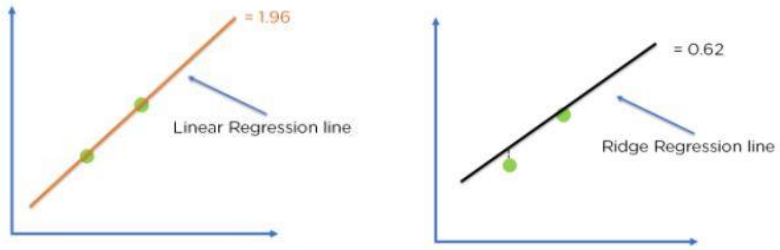

Figure 8: Linear regression mode Figure 9: Ridge regression model

Comparing the two models, with all data points, we can see that the Ridge regression line fits the model more accurately than the linear regression line.

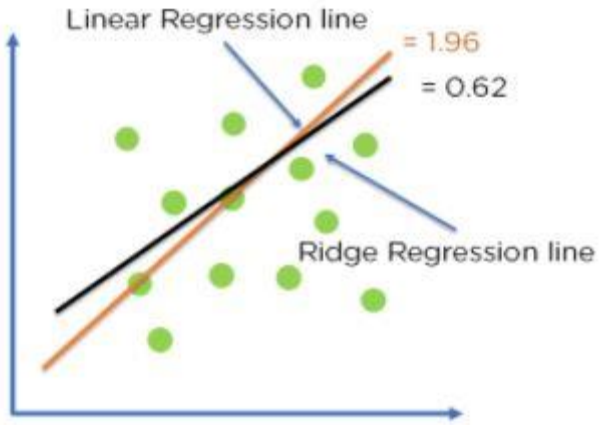

Figure 10: Optimization of model fit using Ridge Regression

Lasso Regression (L1 form)

It modifies the over-fitted or under-fitted models by adding the penalty equivalent to the sum of the absolute values of coefficients.

Lasso regression also performs coefficient minimization, but instead of squaring the magnitudes of the coefficients, it takes the true values of coefficients. This means that the coefficient sum can also be 0, because of the presence of negative coefficients. Consider the cost function for Lasso regression :

Cost function = Loss + λ x $\sum\|w\|$

For Linear Regression line, let's assume,

Loss = 0 (considering the two points on the line)

λ = 1 w = 1.4 Then, Cost function = 0 + 1 x 1.4 = 1.4

For Ridge Regression, let's assume,

Loss = 0.32 + 0.12 = 0.1

λ = 1 w = 0.7 Then, Cost function = 0.1 + 1 x 0.7 = 0.8

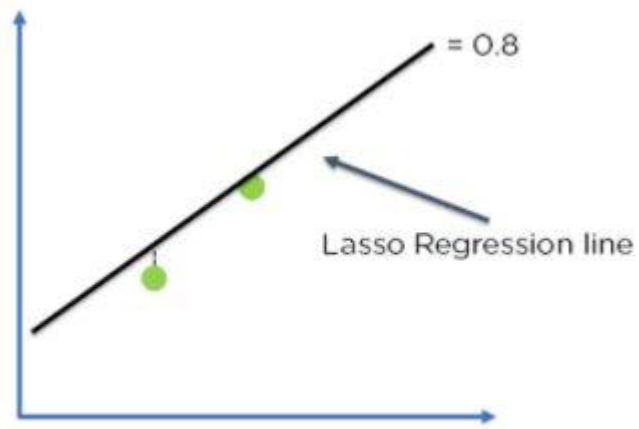

Figure 13: Lasso Regression

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

An error incomplete relationship, the error term is the amount at which the equation may differ during empirical analysis.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.

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) $

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

A) 0.666 B) 0

C) 1 D) 0.67

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

A) 36 B) 10

C) 24 D) 45

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

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

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

A) 2 B) 4

C) 0 D) 6

6. What does the finally keyword notes 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

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

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.

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

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

A) yield B) raise

C) look-in D) all of the above