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

Ans: A) 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

Ans: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

Ans:C) Zero

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

Ans: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

Ans: D) none of these

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

A) Descriptive model) BPredictive modal

C) Reinforcement learning D) All of the above

Ans:BPredictive modal

7. Lasso and Ridge regression techniques belong to _____?

A) Cross validation B) Removing outliers

C) SMOTE D) Regularization

Ans:A) Cross validation

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

A) Cross validation B) Regularization

C) Kernel D) SMOTE

Ans:B) Regularization

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

Ans:

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

A) True B) False

Ans: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

Ans: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.

Ans: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?

Ans The word regularize means to make things regular or acceptable. This is exactly why we use it for. Regularizations are techniques used to reduce the error by fitting a function appropriately on the given training set and avoid overfitting.

14. Which particular algorithms are used for regularization?

Ans: here are three main regularization techniques, namely:

Ridge Regression (L2 Norm)

Lasso (L1 Norm)

Dropout.

Techniques of Regularization

Mainly, there are two types of regularization techniques, which are given below:

Ridge Regression

Lasso Regression

Ridge Regression

☞ Ridge regression is one of the types of linear regression in which we introduce a small amount of bias, known as Ridge regression penalty so that we can get better long-term predictions.

☞ In Statistics, it is known as the L-2 norm.

☞ In this technique, the cost function is altered by adding the penalty term (shrinkage term), which multiplies the lambda with the squared weight of each individual feature. Therefore, the optimization function(cost function) becomes:

ridge Regularization

Fig. Cost Function for Ridge Regression

Image Source:

In the above equation, the penalty term regularizes the coefficients of the model, and hence ridge regression reduces the magnitudes of the coefficients that help to decrease the complexity of the model.

☞ Usage of Ridge Regression:

When we have the independent variables which are having high collinearity (problem of) between them, at that time general linear or polynomial regression will fail so to solve such problems, Ridge regression can be used.

If we have more parameters than the samples, then Ridge regression helps to solve the problems.

☞ Limitation of Ridge Regression:

Not helps in Feature Selection: It decreases the complexity of a model but does not reduce the number of independent variables since it never leads to a coefficient being zero rather only minimizes it. Hence, this technique is not good for feature selection.

Model Interpretability: Its disadvantage is model interpretability since it will shrink the coefficients for least important predictors, very close to zero but it will never make them exactly zero. In other words, the final model will include all the independent variables, also known as predictors.

Lasso Regression

☞ Lasso regression is another variant of the regularization technique used to reduce the complexity of the model. It stands for Least Absolute and Selection Operator.

☞ It is similar to the Ridge Regression except that the penalty term includes the absolute weights instead of a square of weights. Therefore, the optimization function becomes:

Lasso Regularization

Fig. Cost Function for Lasso Regression

Image Source:

☞ In statistics, it is known as the L-1 norm.

☞ In this technique, the L1 penalty has the effect of forcing some of the coefficient estimates to be exactly equal to zero which means there is a complete removal of some of the features for model evaluation when the tuning parameter λ is sufficiently large. Therefore, the lasso method also performs Feature selection and is said to yield sparse models.

☞ Limitation of Lasso Regression:

Problems with some types of Dataset: If the number of predictors is greater than the number of data points, Lasso will pick at most n predictors as non-zero, even if all predictors are relevant.

Multicollinearity Problem: If there are two or more highly collinear variables then LASSO regression selects one of them randomly which is not good for the interpretation of our model

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

Ans: 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.

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) &

Ans:%

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

A) 0.666 B) 0

C) 1 D) 0.67

Ans:A) 0.666

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

A) 36 B) 10

C) 24 D) 45

Ans:A) 36

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

A) 2 B) True

C) False D) 0

Ans:B) True

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

A) 2 B) 4

C) 0 D) 6

Ans:

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

Ans:

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

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

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

Ans: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

b) Central Mean Theorem

c) Centroid Limit Theorem

d) All of the mentioned

Ans: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

Ans: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

Ans:d) All of the mentioned

5. _____ random variables are used to model rates.

a) Empirical

b) Binomial

c) Poisson

d) All of the mentioned

Ans:c) Poisson

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

a) True

b) False

Ans: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

Ans: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

Ans:a) 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

Ans: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?

Ans: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. In graph form, normal distribution will appear as a bell curve.

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

Ans:Common Methods

Mean or Median Imputation. When data is missing at random, we can use list-wise or pair-wise deletion of the missing observations. ...

Multivariate Imputation by Chained Equations (MICE) MICE assumes that the missing data are Missing at Random (MAR). ...

Random Forest.

12. What is A/B testing?

Ans:A/B testing is a basic randomized control experiment. It is a way to compare the two versions of a variable to find out which performs better in a controlled environment. For instance, let's say you own a company and want to increase the sales of your product.

13. Is mean imputation of missing data acceptable practice?

Ans:he 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.

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?

Ans: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.

15. What are the various branches of statistics

Ans:Descriptive statistics describe what is going on in a population or data set. Inferential statistics, by contrast, allow scientists to take findings from a sample group and generalize them to a larger population. The two types of statistics have some important differences.