MACHINE LEARNING_ASSIGNMENT 1

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

(HERE CORRECT ANSWERS IS SELECTED WITH GREEN COLOUR-)

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.

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

13. **Explain the term regularization?**

Sometimes what happens is that our Machine Learning model performs well on the training dataset but does not perform well on test dataset. It means the model is not able to predict the output or target column for the unseen data, hence the model is called an Overfitted model. **Regularizations** are techniques used to reduce the error by fitting a function appropriately on the given training set and avoid overfitting so able to make a Generalise model.

14. **Which particular algorithms are used for regularization?**

1. **Lasso Regression**: - This regularization technique is used to reduce the complexity of the model, it nullify or give zero importance to those features which were not contributing to label.

* LASSO stands for Least Absolute and Selection Operator

* Lasso also known as L1-norm

2. **Ridge regression**: - In this techniques we introduce a small amount of bias, known as Ridge regression penalty so that we can get better long-term predictions. It gives very little importance to those features which were not contributing to label but not Zero.

* it is known as L2 norm

3.**Elastic-Net regression** :- it is also a regularisation technique but very less popular

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

The error is the difference between the actual value and the predicted value. The error term of a regression equation represents all of the variation in the dependent variable not explained by the weighted independent variables.

A regression equation is the formula for a straight line- in this case, the best-fit line through a scatterplot of data. If there were no error, all the data points would be located on the regression line ; to the extent they are not represents error; this is what the error term summarizes

---

STATISTICS ASSIGNMENT- 1

Q1 to Q9 have only one correct answer. Choose the correct option to answer your question.(correct answers marked with green-)

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. 10. 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 centred 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?

* 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 taper off equally in both directions. Extreme values in both tails of the distributions are similarly unlikely.

*The normal distribution describes how the value are distributed and it is also known as Gaussian distribution.

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

***** Missing data is an inevitable part of the process and is a huge problem for data analysis because it distorts findings. So to handle missing data imputation is used. There are several imputation techniques but here are some recommendation: -

A) Mean imputation – it uses the average value of the responses from the other data entries to fill out missing values.

B) Common point imputation – it uses the middle point or the most commonly chosen value

12. What is A/B testing?

An A/B testing is an example of statistical hypothesis testing, a process whereby a hypothesis is made about the relationship between two datasets and those datasets are then compared against each other to determine if there is a statistically significant relationship or not

13. Is mean imputation of missing data acceptable practice?

Mean imputation is typically considered terrible practice since it ignores feature correlation. 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?

Linear regression is used to predict the value of a variable based on the value of another variable. It estimates the relationship between one independent variable and one dependent variable using a straight line. Both variable should be quantitative.

Regression equation => y= c+b*x

Where y= estimated dependent variable score, c= constant, b= regression coefficient, and x = score on the independent variable.

15. What are the various branches of statistics

There are the real branches of statistics

1) descriptive statistics and 2) inferential statistics

Descriptive analysis : - It deals with the presentation and collection of data. This is usually the first part of a statistical analysis.

Inferential statistics: - It involves drawing the right conclusions from the statistical analysis that has been performed using descriptive statistics. In the end, it is the inferences that make studies important and this aspect is dealt with in inferential statistics.

Python Assignment 1

Q1 to Q8 have only one correct answer. Choose the correct option to answer your question.(correct answers marked with green colour)

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 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

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

- Next questions are in the jupyter notebook

_____ (Thank you)_____