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 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

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.

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 model
   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 an 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:

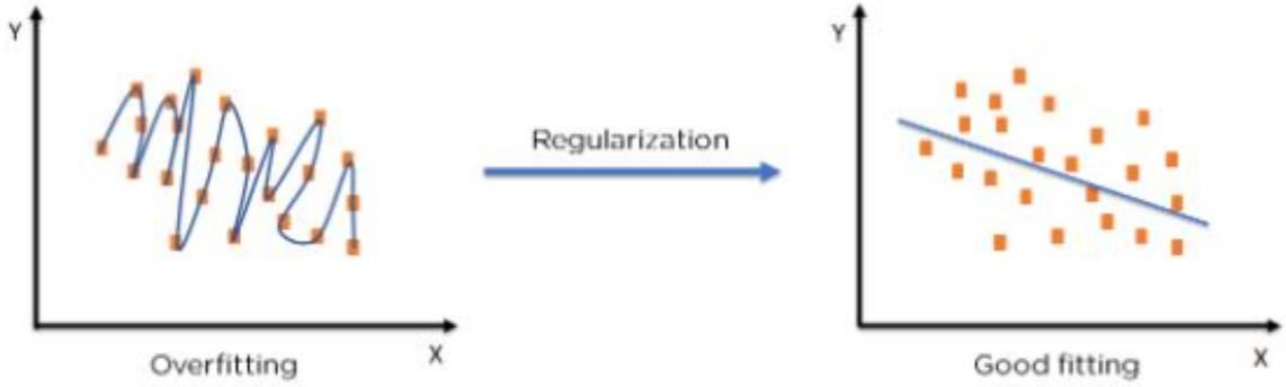
12. Which of the following is true about Normal Equation used to compute the coefficient of the Linear Regression?
    A) We don't have to choose the learning rate.
    B) It becomes slow when number of features is very large.
    C) We need to iterate.
    D) It does not make use of dependent variable.

MACHINE LEARNING

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

13. Explain the term regularization?
14. Which particular algorithms are used for regularization?
15. Explain the term error present in linear regression equation?

Ans 13. Regularization are the process to train machine learning models in such a way that prevents underfitting or overfitting and minimizes the overall loss. The diagram below shows the same in a simplified manner.



Ans 14.

Ans 15.

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. 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 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

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

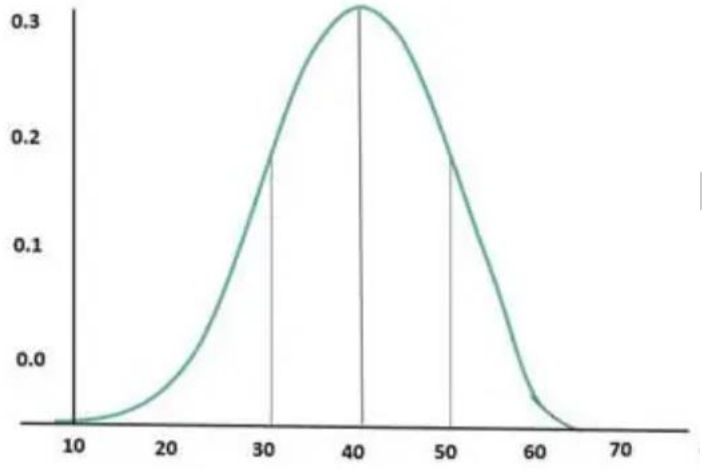
10. What do you understand by the term Normal Distribution?
11. How do you handle missing data? What imputation techniques do you recommend?
12. What is A/B testing?
13. Is mean imputation of missing data acceptable practice?
14. What is linear regression in statistics?
15. What are the various branches of statistics?

Ans 10. Normal distribution is a type of probability distribution (type of symmetrical distribution) which is also known as the Gaussian Distribution. The major property of the distribution is that its mean, median and mode are all equal. The majority of the distribution peaks at the center, which tells us that the readings near the mean are more frequent and they fall symmetrically from the mean. The formula for this distribution is:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

Where: x = value of the variable or data being examined and f(x) the probability function, μ = the mean, and σ = the standard deviation

e.g.:



Ans 15. There are two branches of statistics: Descriptive and Inferential

Descriptive Statistics: It is used to 'present' or 'show' the data in a way which is understandable. This can be done in two ways, i.e. visually using graphs and numerally using mean and other properties.

Inferential Statistics: This is the part which comes after descriptive statistics, where we make conclusions based on results of the descriptive statistics.

Ans 14. Linear regression is a type of predictive analysis where the relationship between two variables is predicted by assuming a linear relationship between the dependent variable and one or more independent variables. If we're using one independent variable, then it is simple linear regression, otherwise, its multiple linear regression.

In simple linear regression, the data is used to find the 'line of best fit' using the lease square method. Formula for simple linear regression is y = c + b*x, where y = estimated dependent variable score, c = constant, b = regression coefficient, and x = score on the independent variable.

Ans 11. There are two basic ways of handling missing values, deleting, and imputing. Deleting missing values is generally not recommended as it might remove some important data which might lead to wrong results. Imputation of missing data is basically substituting the missing values with other relevant values. There are many imputation techniques like:

a. Mean/Median/Mode imputation: the missing values are replaced with the mean/median/mode of that dataset. The outliers should be processed first before applying this technique.
b. Random sample imputation: the missing values are replaced with values taken at random from the dataset comparable to the dataset.

c. Interpolation and Extrapolation: the missing values are replaced with the interpolated or extrapolated results from the data. This is generally used for the data that is taken in a particular length of time.

Ans 13. The mean interpolation is one of the most basic ways to impute data, hence it is not always practical. There are a few drawbacks to it like:

a. Mean is sensitive to the outliers and hence that can bias the data trend.
b. Replacing the missing data with the mean takes the assumption that the missing values are not related to other variables, which is usually not the case.

Ans 12. A/B testing is a type of statistical hypothetical testing. It is a way to determine which of the two versions of a thing are better.