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) $
Ans:- C) %

2. In python 2//3 is equal to?
A) 0.666 B) 0 C) 1 D) 0.67
Ans:- B) 0

3. In python, 6<<2 is equal to?
A) 36 B) 10 C) 24 D) 45
Ans:- C) 24

4. In python, 6&2 will give which of the following as output?
A) 2 B) True C) False D) 0
Ans:- a) 2

5. In python, 6|2 will give which of the following as output?
A) 2 B) 4 C) 0 D) 6
Ans:- 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
Ans:- C)

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
Ans:- A)

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.
Ans:- C)

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
Ans:- A) & C)

10. Which of the following are the keywords in python?
A) yield B) raise C) look-in D) all of the above
Ans:- A) & B)

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)

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)

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)

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)

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)

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:- The normal distribution is the most important probability distribution in statistics for independent , randomly generated variables because it fits many natural phenomena symmetrically toward either extreme. For example, heights, blood pressure, measurement error, and IQ scores follow the normal distribution. It is also known as the Gaussian distribution and the bell curve.

In normal distribution the mean, mode and median are all the same.

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

Ans:- Missing data in a dataset can either be drop those rows or columns, or decide to replace them with another value.

In pandas there are two very useful methods: isnull() and dropna() that will help you find columns of data with missing data and drop those values. If I want to fill the invalid values with a placeholder value (for example, 0), I could use the fillna() method.

There are many options to consider when it comes to the imputation of missing values.

1. Imputing by a constant value that has meaning within the domain, like 0, which is different from values.
2. For the Continuous variable, the most common methods used are mean and median imputation. In some cases, these metrics cannot be used as a proxy because of high variance in the variable which contains missing values. In such cases, we can use regression to model the missing variable and predict the same.
3. For the Categorical variable, imputing with the mode is generally used. But if there is a class imbalance in the variable, imputing with mode will increase the problem further. So in such cases, we can use a classification model to predict the missing classes.

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 example- suppose you own a company and want to increase the sale of your product. Here, either you can use random experiments, or you can apply scientific and statistical methods. A/B testing is one of the most prominent and widely used statistical tools. In this scenario, you may divide the products into two parts – A and B. Here A will remain unchanged while you make significant changes in B's packaging. Now, on the basis of the response from customer groups who used A and B respectively, you try to decide which is performing better.

It is a hypothetical testing methodology for making decisions that estimate population parameters based on sample statistics. The population refers to all the customers buying your product, while the sample refers to the number of customers that participated in the test.

13. Is mean imputation of missing data acceptable practice?

Ans:- In general it's a bad practice. Missing data arises in almost all serious statistical analysis. They all are just considerations that we plan to make but couldn't. In datasets, missing values could be replaced with ?, nan, N/A, blank cell or sometimes, inf, -inf. Mean imputation is the easiest way to replace each missing value with the mean of the observed values for that variable. But this strategy can severely distort the distribution in the variable, leading to issues with measures, and also leads to underestimated standard deviation. Additionally, mean imputation distorts relationships between variables by dragging estimates of the correlation towards zero.

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 prediction an outcome (dependent) variable?
2. Which variables in particular are significant predictors 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 = m*x + c, where y = dependent variable, c = is the y-intercept, m = slope of the line, and x = independent variable.

   There are many names for a regression's dependent variable. It may be called an outcome variable, criterion variable, endogenous variable, or regressand. The independent variable can be called exogenous variables, or regressors.
   Three major uses for regression analysis are :-
   1. Determining the strength of predictors
   2. Forecasting an effect
   3. Trend forecasting

      Types of Linear regression
      1. Simple linear regression
      2. Multiple linear regression
      3. Logistic regression
      4. Ordinal regression
      5. Multinomial regression
      6. Discriminant analysis

15. What are the various branches of statistics?

Ans:-There two branches of statistics are Descriptive statistics and Inferential statistics

1. Descriptive Statistics:- In involves the presentation and collection of data. It's the first phase of the statistical analysis. It's not simple as it seems with its definition, as it involves the experiments, choosing the right samples and avoid being biases so that the experiment must be easy to conduct.

Various areas requires separate analysing using this method.

For example:- The average length of the books of statistics, The variation in the weight of the rice packet from shop, the reading of an experiment conducted in the lab.

2. Inferential Statistics:- It involves outlining the right conclusion from the statistical analysis that is performed using the descriptive statistics. In the end, it is the deduction that makes studies important and this aspect is dealt with in the inferential statistics. Almost all the predictions of the future and generalizations about a population by studying a smaller sample come under the purview of inferential statistics. Most social science experiments deal with studying a small sample population that helps determine how the population in general behaves. By designing the right experiment, the researcher is able to draw conclusions relevant to his study.

   For example:- we might stand in a mall and ask a sample of 100 people if they like shopping at Lewis, the mean IIT JEE score by class 12 students in India, Election commission asking about the fairness of the voting procedure in India.

MACHINE LEARNING

In Q1 to Q8, only one option is correct, Choose the correct option:

1. The computational complexity of linear regression is:

A) $O(n2.4)$ B) $O(n)$ C) $O(n2)$ D) $O(n3)$

Ans:- B)

2. Which of the following can be used to fit non-linear data?

A) Lasso Regression B) Logistic Regression

C) Polynomial Regression D) Ridge Regression

Ans:- C)

3. Which of the following can be used to optimize the cost function of Linear Regression?

A) Entropy B) Gradient Descent C) Pasting D) None of the above.

Ans:- B)

4. Which of the following method does not have closed form solution for its coefficients?

A) extrapolation B) Ridge C) Lasso D) Elastic Nets

Ans:- C)

5. Which gradient descent algorithm always gives optimal solution?

A) Stochastic Gradient Descent B) Mini-Batch Gradient Descent

C) Batch Gradient Descent D) All of the above

Ans:- A)

6. Generalization error measures how well a model performs on training data.

A) True B) False

Ans:- A) True

7. The cost function of linear regression can be given as $J(w0,w1)= 12m\Sigma(w0+ w1x(i)- y(i))2mi=1$. The half term at start is due to:

A) scaling cost function by half makes gradient descent converge faster.

B) presence of half makes it easy to do grid search.

C) it does not matter whether half is there or not.

D) None of the above.

Ans:- C)

8. 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:- B)

In Q9 to Q11, more than one options are correct, Choose all the correct options:

9. 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 are very large.

C) We need to iterate.

D) It does not make use of dependent variable.

Ans:- A), B), C)

10. Which of the following statement/s are true if we generated data with the help of polynomial features with 5 degrees of freedom which perfectly fits the data?

A) Linear Regression will have high bias and low variance.

B) Linear Regression will have low bias and high variance.

C) Polynomial with degree 5 will have low bias and high variance.

D) Polynomial with degree 5 will have high bias and low variance.

Ans:- A)

11. Which of the following sentence is false regarding regression?

A) It relates inputs to outputs.

B) It is used for prediction.

C) It discovers causal relationship.

D) No inference can be made from regression line. Ans:- D)

Q12 and Q13 are subjective answer type questions, Answer them briefly.

12. Which Linear Regression training algorithm can we use if we have a training set with millions of features?

Ans:- since there are lots of features, but we cannot use Normal Equations because it will be very expensive and computational complexity grows quickly with the number of features, instead we can use Gradient Descent which is best used when the parameters cannot be calculated analytically(e.g. using linear algebra) and must be searched for by an optimization algorithm.

Gradient descent is a first-order iterative optimization algorithm for finding a local minimum of a different function. The idea is to take repeated steps in the opposite direction of the gradient (or approximate gradient) of the function at the current point, because this is the direction of steepest descent.

13. Which algorithms will not suffer or might suffer, if the features in training set have very different scales?

Ans:- The Gradient Descent suffers from features of different scales, because the model will take a longer time to reach the global maximum. We can always scale the features to eliminate this problem. It is important to highlight that the normal equation will work just fine without scaling.