Rakesh Chandra Maity (rakesh.chandra.maity@gmail.com)

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) $

Option C (%)

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

Option B (0)

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

Option C (24)

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

Option A (2)

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

Option 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

Option 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
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.

Option A

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

Option A & Option C

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

Option A & Option 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.

Answers Shared in GitHub

MACHINE LEARNING ASSIGNMENT

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

Option A

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

Option A

3. A line falls from left to right if a slope is _____?
A) Positive B) Negative
C) Zero D) Undefined

Option B

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

Option B

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

Option C

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

Option C

7. Lasso and Ridge regression techniques belong to _____?
A) Cross validation B) Removing outliers
C) SMOTE D) Regularization

Option B

8. To overcome with imbalance dataset which technique can be used?
A) Cross validation B) Regularization
C) Kernel D) SMOTE

Option D

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

Option C

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

Option 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

Option D

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.

Options A,B

MACHINE LEARNING

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

13. Explain the term regularization?

Answer:

In general, regularization means to make things regular or acceptable. This is exactly why we use it for applied machine learning. In the context of machine learning, regularization is the process which regularizes or shrinks the coefficients towards zero. In simple words, regularization discourages learning a more complex or flexible model, to prevent overfitting.

14. Which particular algorithms are used for regularization?

Answer:

There are mainly three regularization techniques, namely:

1. Ridge Regression (L2 Norm)
2. Lasso (L1 Norm)
3. Dropout

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

Answer:

An error term in statistics is a value which represents how observed data differs from actual population data. It can also be a variable which represents how a given statistical model differs from reality. The error term is often written ε.

he error term includes everything that separates your model from actual reality. This means that it will reflect nonlinearities, unpredictable effects, measurement errors, and omitted variables.

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

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

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

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

D) All the mentioned

5. _____ random variables are used to model rates.
a) Empirical
b) Binomial
c) Poisson
d) All of the mentioned

C) Poisson

6. 10. Usually replacing the standard error by its estimated value does change the CLT.
a) True
b) False

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

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

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

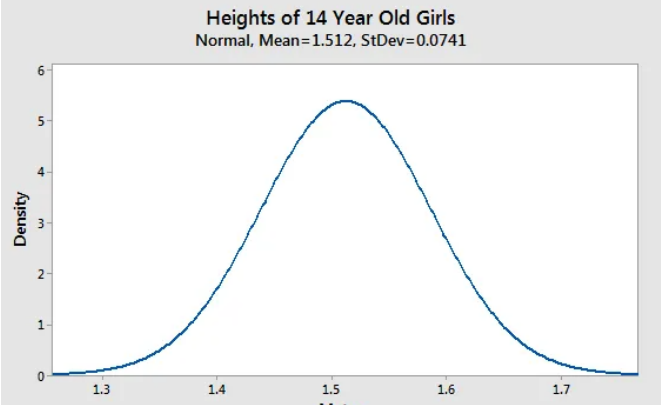
C) None of the mentioned

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

10. What do you understand by the term Normal Distribution?

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 distribution are similarly unlikely.

As with any probability distribution, the normal distribution describes how the values of a variable are distributed. It is the most important probability distribution in statistics because it fits many natural phenomena. 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.



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

With the help of different imputation techniques I will handle missing data.

Multiple imputation is more advantageous than the single imputation because it uses several complete data sets and provides both the within-imputation and between-imputation variability.

Imputation Techniques

- Complete Case Analysis(CCA):- This is a quite straightforward method of handling the Missing Data, which directly removes the rows that have missing data i.e we consider only those rows where we have complete data i.e data is not missing.
- Arbitrary Value Imputation. ...
- Frequent Category Imputation.

12. What is A/B testing?

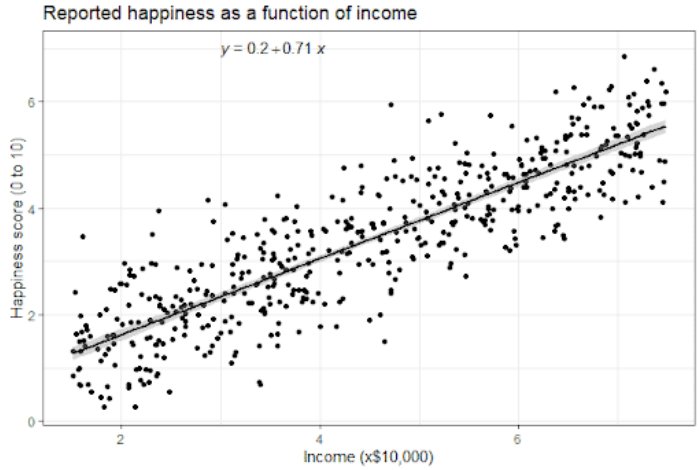
An AB test is an example of statistical hypothesis testing, a process whereby a hypothesis is made about the relationship between two data sets and those data sets are then compared against each other to determine if there is a statistically significant relationship or not

13. Is mean imputation of missing data acceptable practice?

- Bad practice in general.
- Leads to an underestimate of the standard deviation
- Distorts relationships between variables by “pulling” estimates of the correlation toward zero

14. What is linear regression in statistics?

Simple linear regression is **a regression model that estimates the relationship between one independent variable and one dependent variable using a straight line**. Both variables should be quantitative. ... Linear regression most often uses mean-square error (MSE) to calculate the error of the model.



15. What are the various branches of statistics?

The two main branches of statistics are descriptive statistics and inferential statistics. Both of these are employed in scientific analysis of data and both are equally important.

Descriptive statistics deals with the presentation and collection of data. This is usually the first part of a statistical analysis. It is usually not as simple as it sounds, and the statistician needs to be aware of designing experiments, choosing the right focus group and avoid biases that are so easy to creep into the experiment.

Inferential statistics, as the name suggests, 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. Most predictions of the future and generalizations about a population by studying a smaller sample come under the purview of inferential statistics. Most social sciences 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.