WORKSHEET 1

1. Which of the following operators is used to calculate remainder in a division?

ANS. %

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

ANS.1

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

Ans.24

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

Ans.2

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

ans. 6

5. What does the finally keyword denotes in python?

ans. the finally block will be executed no matter if the try block raises an error or not.

6. What does raise keyword is used for in python?

Ans. Raise an exception

8. Which of the following is a common use case of yield keyword in python?

Ans, in defining generator

9. Which of the following are the valid variable names?

Ans. 1ab, abc2

10. Which of the following are the keywords in python?

Ans. yield raise look-in

WORKSHEET-2

1. The computational complexity of linear regression is:

Ans, $O(n)$

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

Ans. Polynomial regression

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

Ans.lasso

5.Which gradient descent algorithm always gives optimal solution?

Ans, A) Stochastic Gradient Descent B) Mini-Batch Gradient Descent C) Batch Gradient Descent.

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

ans, true

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

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

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.

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

11. Which of the following sentence is false regarding regression?) It relates inputs to outputs.

B) It is used for prediction.

C) It discovers causal relationship.

SUBJECTIVE ANSWERS

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

Ans.)You could use batch gradient descent, stochastic gradient descent, or mini-batch gradient descent. SGD and MBGD would work the best because neither of them need to load the entire dataset into memory in order to take 1 step of gradient descent. Batch would be ok with the caveat that you have enough memory to load all the data.

The normal equations method would not be a good choice because it is computationally inefficient. The main cause of the computational complexity comes from inverse operation on an (n x n) matrix. O n2 . 4 to O n3.

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

Ans) The normal equations method does not require normalizing the features, so it remains unaffected by features in the training set having very different scales.

Feature scaling is required for the various gradient descent algorithms. Feature scaling will help gradient descent converge quicker.

WORKSHEET STATISTICS

1. Bernoulli random variables take (only) the values 1 and 0.

Ans. 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?

Ans. a) Central Limit Theorem b) Central Mean Theorem c) Centroid Limit Theorem

3. Which of the following is incorrect with respect to use of Poisson distribution?

Ans. a) Modeling event/time data b) Modeling bounded count data c) Modeling contingency tables

4. Point out the correct statement.

a) The exponent of a normally distributed random variables follows what is called the log-normal distribution

5. _____ random variables are used to model rates.

Ans. Poission.

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

Ans.false

7.Which of the following testing is concerned with making decisions using data?

Ans. Hypothesis

8.Normalized data are centered at_____and have units equal to standard deviations of the original data.

Ans. 0

9.Which of the following statement is incorrect with respect to outliers?

Ans. c) Outliers cannot conform to the regression relationship

SUBJECTIVE ANSWERS (STATISTICS)

1. 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

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

Ans.

1. Mean or Median Imputation. When data is missing at random, we can use list-wise or pair-wise deletion of the missing observations. ...
2. Multivariate Imputation by Chained Equations (MICE) MICE assumes that the missing data are Missing at Random (MAR). ...
3. Random Forest.

3.What is A/B testing?

Ans. A/B Testing also known as split testing, refers to a randomized experimentation process wherein two or more versions of a variable (web page, page element, etc.) are shown to different se gments of website visitors at the same time to determine which version leaves the maximum impact and drive business metrics.

4. What is linear regression in statistics?

The linear regression model describes the dependent variable with a straight line that is defined by the equation $Y = a + b \times X$, where a is the y-intersect of the line, and b is its slope. ... The regression line enables one to predict the value of the dependent variable Y from that of the independent variable X.

5.What are the various branches of statistics?

Ans.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 for the student of statistics.