Assignment 1- Machine Learning

1. Which of the following methods do we use to find the best fit line for data in Linear Regression?

Answer: Least Square Error

2. Which of the following statement is true about outliers in linear regression?

Answer: Linear Regression is sensitive to outliers.

3. A line falls from left to right if a slope is _____?

Answer: zero.

4. Which of the following will have symmetric relation between dependent variable and independent variable?

Answer: Correlation.

5. Which of the following is the reason for over fitting condition?

Answer: High Bias and High Variance.

6. If output involves label then that model is called as:

Answer: Predictive Model.

7. Lasso and Ridge regression techniques belong to _____?

Answer: Regularization

8. To overcome with imbalance dataset which technique can be used?

Answer: Regularization

9. The AUC Receiver Operator Characteristic (AUCROC) curve is an evaluation metric for binary classification problems. It uses _____ to make graph?

Answer: TPR and FPR

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

Answer: False

11. Pick the feature extraction from below:

Answer: Apply PCA to project high dimensional data

12. Which of the following is true about Normal Equation used to compute the coefficient of the Linear Regression?

Answer- A) We don't have to choose the learning rate.

B) It becomes slow when number of features is very large.

13. Often in machine learning while predicting a label there is an issue of Overfitting (when the model takes too many features thereby resulting in high accuracy during the training phase and less accuracy during testing phase). In order to avoid such overfitted model, regularization techniques are used which imposes some penalty over the data and brings down the weight. There are two Regularization method: Lasso and Ridge.

14. There are two types of algorithms used in regularization: Lasso (L1 Technique) and Ridge (L2 Technique)

15. In Linear Regression algorithm, an error is often termed as the difference between actual data and predicted data in the model. Error is also known as residue in machine learning and often a best fitted line in linear regression is formed keeping the difference between actual data and predicted data (error) the least.

    It can be also expressed as r= y-(mx+b)
    R= Residue or Error
    Y= Predicted value
    M= Co-efficient
    X= Input Data (Feature)
    B= Intercept.

Assignment:2- PYTHON – WORKSHEET 1

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

Answer: %

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

Answer: 0

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

Answer: 24

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

Answers: 2

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

Answer: 6

6. What does the finally keyword denotes in python?

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

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

Answer: It is used to raise an exception

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

Answer: in defining a generator

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

Answer: _abc and abc2

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

Answer: yield and raise

STATISTICS WORKSHEET-1

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

Answer: 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?

Answers: Central Limit Theorem

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

Answers: a) Modeling event/time data b) Modeling bounded count data

4. Point out the correct statement.

Answers: 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.

5. _____ random variables are used to model rates.

Answers: a) Empirical b) Binomial

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

Answers: True

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

Answers: Hypothesis

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

Answers: 0

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

Answers: 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

10. Normal distribution is a distribution curve which is more symmetrical near the mean that is there is more data near the mean and there is zero skewness. A normal distribution curve is often identified as a bell-shaped curve.

11. Missing data can be handled by either i) **deleting the missing data** from a data set by using dropna() method or del() method ii) We can also use **imputation method** where we use the fillna(), mean() or mode()

12. A/B Testing is a method of comparing two webpages against each other to improve business and reach more people.

13. Yes, mean imputation of missing data is acceptable practice.

14. Linear regression is a method of predicting the value of a variable (Dependent variable) based on one or more continuous variable (Independent Variable).

15. There are three branches of statistics: Data collection, descriptive statistics and inferential statistics.