MACHINE LEARNING

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

Answer – 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

Answer - A

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

Answer - 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

Answer – 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

Answer – 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

Answer – B

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

Answer – D

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

Answer – 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

Answer – A

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

Answer – B

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

Answer – A

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.

Answer – A, B

13. Explain the term regularization?

Answer - **Regularization** is a technique use in machine learning to reduce errors of overfitting and improve the generalization performance by adding extra and relevant data to the model. Regularization works by adding a penalty or complexity term to the complex model.

14. Which particular algorithms are used for regularization?

Answer - Regularization uses many algorithms, we particularly use two type of regression algorithms to prevent overfitting.

1) **Ridge regression** – it's a regularization technique where a tolerable amount of bias is introduced to obtain better long-term predictions, this method adds the square of the coefficients as a penalty to the loss function. The penalty term acts to reduce the magnitude of the parameter estimates making the model less sensitive to the training data.

2) **Lasso regression** – it's a regularization technique short for least absolute shrinkage and selection operator, its used to prevent overfitting in models with a large number of parameters by encouraging the selection of important features and adding a penalty of the absolute value of the coefficients to the loss function. This makes the model less sensitive to the training data.

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

Answer - The error term captures the unobserved factors in the relationship between the independent and dependent variables. It's the difference between Orignal answer and the Machine answer.

STATISTICS WORKSHEET-1

1) Bernoulli random variables take (only) the values 1 and 0.
a) True
b) False

Answer – A

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

Answer - A

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

Answer - 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

Answer – C

5) **_____ random variables are used to model rates.**
a) Empirical
b) Binomial

c) Poisson
d) All of the mentioned

Answer – C

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

Answer – B

7) **Which of the following testing is concerned with making decisions using data?**
a) Probability
b) Hypothesis
c) Causal
d) None of the mentioned

Answer – B

8) **Normalized data are centered at_____and have units equal to standard deviations of the original data.**
a) 0
b) 5
c) 1
d) 10

Answer – A

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

Answer - C

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

Answer - The normal distribution is a statistical distribution that is symmetrical around its mean, and is bell-shaped. Also known as a Gaussian distribution, it's a type of continuous probability distribution in which most data points cluster toward the middle of the range which indicates that values near the mean occur more frequently than the values that are farther away from the mean.

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

Answer - Handling missing data usually depends on the type and nature of data, the amount of data missing and the reason for missing data, in case of few missed data, we just put N/A i.e. null in its place, however with large number of missing data can affect the prediction model so many imputation techniques are used to challenge that

a) Mean, median and mode imputation – Missing values are replaced with the mean, median or mode of the observed values for that variable. Outliers presence makes this technique not suitable.
b) Linear regression imputation – Using a regression model to predict missing values based on other variables.
c) K-Nearest Neigbours (KNN) imputation – Data is filled by the values of their k-nearest neighbours.

12. What is A/B testing?

Answer - A/B testing also known as split testing, is a statistical method for compairing two variables, A & B against each other and statistically analysed to determine which variation performs according to the goal.

13. Is mean imputation of missing data acceptable practice?

Answer – Mean imputation means the missing values are replaced with the mean of the observed values for the variable, Mean imputation is a simple and easy to implement technique which is actually an effective way to substitute missing values, given that there are not much outliers present.

14. What is linear regression in statistics?

Answer - Linear regression is a statistical method used to model the relationship between a dependent variable and independent variables. Its form is in the simple linear regression equation ($Y= a + bx + e$) , where Y is dependent variable, a is intercept, b is slope or coefficient, x is the independent variable and e is the error.

15. What are the various branches of statistics?

Answer – Statistics has mainly 2 branches, Descriptive statistics and Inferential statistics. Descriptive statistics has two branches (a) Central tendency which has mean, median, mode and (b) Dispersion of data which has range, variance, standard deviation, percentile, skew etc.

Inferential statistics has hypothesis testing, z score testing, t test, regression test, chi-square test etc.

_____THE END_____