MACHINE LEARNING

1. **In which of the following you can say that the model is overfitting?**

Ans. : C) High R-squared value for train-set and Low R-squared value for test-set.

2. **Which among the following is a disadvantage of decision trees?**

Ans. : B) Decision trees are highly prone to overfitting.

3. **Which of the following is an ensemble technique?**

Ans. : C) Ensemble Technique

4. **Suppose you are building a classification model for detection of a fatal disease where detection of the disease is most important. In this case which of the following metrics you would focus on?**

Ans: A) Accuracy

5. **The value of AUC (Area under Curve) value for ROC curve of model A is 0.70 and of model B is 0.85. Which of these two models is doing better job in classification?**

Ans: B) Model B

6. **Which of the following are the regularization technique in Linear Regression?**

Ans: A) Ridge D) Lasso

7. **Which of the following is not an example of boosting technique?**

Ans: B) Decision Tree

8. **Which of the techniques are used for regularization of Decision Trees?**

Ans: A) Pruning B) L2 regularization

9. **Which of the following statements is true regarding the Adaboost technique?**

Ans: B) A tree in the ensemble focuses more on the data points on which the previous tree was not performing well C) It is example of bagging technique

11.Differentiate between Ridge and Lasso Regression.

Ans. Ridge Regression : In Ridge regression, we add a penalty term which is equal to the square of the coefficient. The *L2* term is equal to the square of the magnitude of the coefficients. We also add a coefficient lambda to control that penalty term. In this case if lambda is zero then the equation is the basic OLS else if lambda then it will add a constraint to the coefficient. As we increase the value of lambda this constraint causes the value of the coefficient to tend towards zero. This leads to both low variance (as some coefficient leads to negligible effect on prediction) and low bias (minimization of coefficient reduce the dependency of prediction on a particular variable).

Lasso Regression : Lasso regression stands for Least Absolute Shrinkage and Selection Operator. It adds penalty term to the cost function. This term is the absolute sum of the coefficients. As the value of coefficients increases from *0* this term penalizes, cause model, to decrease the value of coefficients in order to reduce loss. The difference between ridge and lasso regression is that it tends to make coefficients to absolute zero as compared to Ridge which never sets the value of coefficient to absolute zero.

12. What is VIF? What is the suitable value of a VIF for a feature to be included in a regression modelling?

Ans: Variance inflation factor (VIF) is a measure of the amount of multicollinearity in a set of multiple regression variables. Mathematically, the VIF for a regression model variable is equal to the ratio of the overall model variance to the variance of a model that includes only that single independent variable.

There are some guidelines we can use to determine whether our VIFs are in an acceptable range. A rule of thumb commonly used in practice is if a VIF is > 10, you have high multicollinearity. In our case, with values around 1, we are in good shape, and can proceed with our regression.

13. Why do we need to scale the data before feeding it to the train the model?

Ans : To ensure that the gradient descent moves smoothly towards the minima and that the steps for gradient descent are updated at the same rate for all the features, we scale the data before feeding it to the model. Having features on a similar scale can help the gradient descent converge more quickly towards the minima.

14. What are the different metrics which are used to check the goodness of fit in linear regression?

Ans: Mean Absolute Error, Root Mean Squared Error (also called Standard Error of the Regression), Relative Absolute Error and the Relative Squared Error, Coefficient of Determination or R Squared (R^2).

15. Sol:

Sensitivity : (1000/1250)*100 = 80 = Recall

Speficity: (1200/1250)* 100 = 96

Precision: (1000/1050)*100 = 95.23

Accuracy: (1250/2500) *100 = 50

SQL

1. Which of the following are TCL commands?

Ans : C. Rollback D. Savepoint

2. Which of the following are DDL commands?

Ans: A. Create C. Drop D. Alter

3. Which of the following is a legal expression in SQL?

B. SELECT NAME FROM SALES;

4. DCL provides commands to perform actions like

C. Authorizing Access and other control over Database

5. Which of the following should be enclosed in double quotes?

C. String

6. Which of the following command makes the updates performed by the transaction permanent in the database?

B. COMMIT

7. A subquery in an SQL Select statement is enclosed in: A. Parenthesis - (...).

8. The result of a SQL SELECT statement is a :- C. TABLE

9. Which of the following do you need to consider when you make a table in a SQLs

D. All of the mentioned

10. If you don't specify ASC and DESC after a SQL ORDER BY clause, the following is used by___? A. ASC

11. What is denormalization?

Sol : Denormalization is a database optimization technique where we add redundant data in the database to combine multiple table data into one so that it can be queried quickly.

12. What is a database cursor?

Sol: A database cursor is an identifier associated with a group of rows. It is, in a sense, a pointer to the current row in a buffer. You must use a cursor in the following cases: Statements that return more than one row of data from the database server: A SELECT statement requires a select cursor.

13. What are the different types of the queries?

Sol: Five types of SQL queries are 1) Data **Definition** Language (DDL) 2) Data Manipulation Language (DML) 3) Data Control Language(DCL) 4) Transaction Control Language(TCL) and, 5) Data Query Language (DQL)

14. Define constraint?

Sol : A **constraint** is a limitation that you place on the data that users can enter into a column or group of columns. A **constraint** is part of the table **definition**; you can **implement constraints** when you create the table or later.

15. What is auto increment?

Sol: Auto Increment is a function that operates on numeric data types. It automatically generates sequential numeric values every time that a record is inserted into a table for a field defined as auto increment.

STATISTICS

1**. Which of the following can be considered as random variable?**

Ans : d) All of the mentioned

2. Which of the following random variable that take on only a countable number of possibilities?

Ans : a) Discrete

3. Which of the following function is associated with a continuous random variable?

Ans: a) pdf

4. The expected value or _______ of a random variable is the centre of its distribution.

Ans: b) median

5. Which of the following of a random variable is not a measure of spread?

c) empirical mean

6. The _________ of the Chi-squared distribution is twice the degrees of freedom.

b) standard deviation

7. The beta distribution is the default prior for parameters between ____________

c) 0 and 1

8. Which of the following tool is used for constructing confidence intervals and calculating standard errors for difficult statistics?

b) bootstrap

9. Data that summarize all observations in a category are called __________ data.

b) summarized

10. What is the difference between a boxplot and histogram?

Sol : Histograms and box plots are graphical representations for the frequency of numeric data values. ... Histograms are preferred to determine the underlying probability distribution of a data. Box plots on the other hand are more useful when comparing between several data sets.

12. How do you assess the statistical significance of an insight?

Ans : Statistical significance can be accessed using hypothesis testing:
– Stating a null hypothesis which is usually the opposite of what we wish to test (classifiers A and B perform equivalently, Treatment A is equal of treatment B)
– Then, we choose a suitable statistical test and statistics used to reject the null hypothesis
– Also, we choose a critical region for the statistics to lie in that is extreme enough for the null hypothesis to be rejected (p-value)
– We calculate the observed test statistics from the data and check whether it lies in the critical region
Common tests:
– One sample Z test
– Two-sample Z test
– One sample t-test
– paired t-test
– Two sample pooled equal variances t-test
– Two sample unpooled unequal variances t-test and unequal sample sizes (Welch's t-test)
– Chi-squared test for variances
– Chi-squared test for goodness of fit
– Anova (for instance: are the two regression models equals? F-test)
– Regression F-test (i.e: is at least one of the predictor useful in predicting the response?)

13. Give examples of data that does not have a Gaussian distribution, nor log-normal.

Sol: The distribution of numbers that show up on the top of a fair die after a large number of throws. Each number from 1 to 6 will occur with approximately equal frequency. Increasing the number of throws will not tend to produce a bell-shaped histogram, in fact the fractional occurrence will approach a constant 1/6 over the possible numbers.

14. Give an example where the median is a better measure than the mean.

Sol : The more skewed the distribution, the greater the difference between the median and mean, and the greater emphasis should be placed on using the median as opposed to the mean.

15. What is the Likelihood

Sol: Likelihood is the value of a continuous probability density function. Its primary use is in parameter estimation. We often select the parameters that give maximum likelihood to your observations.