ASSIGNMENT - 4

MACHINE LEARNING

In Q1 to Q5, only one option is correct, Choose the correct option:

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

A) High R-squared value for train-set and High R-squared value for test-set.

B) Low R-squared value for train-set and High R-squared value for test-set.

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

D) None of the above

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

A) Decision trees are prone to outliers.

B) Decision trees are highly prone to overfitting.

C) Decision trees are not easy to interpret

D) None of the above.

3. Which of the following is an ensemble technique?

A) SVM B) Logistic Regression

C) Random Forest D) Decision tree

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?

A) Accuracy B) Sensitivity

C) Precision D) None of the above.

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?

A) Model A B) Model B

C) both are performing equal D) Data Insufficient

In Q6 to Q9, more than one options are correct, Choose all the correct options:

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

A) Ridge B) R-squared

C) MSE D) Lasso

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

A) Adaboost B) Decision Tree

C) Random Forest D) Xgboost.

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

A) Pruning B) L2 regularization

C) Restricting the max depth of the tree
D) All of the above

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

A) We initialize the probabilities of the distribution as 1/n, where n is the number of data-points

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

D) None of the above

Q10 to Q15 are subjective answer type questions, Answer them briefly.

10. Explain how does the adjusted R-squared penalize the presence of unnecessary predictors in the model?

ANSWER:

The adjusted R-Square compensates for the addition of variable and only increases if the new predictor anhances the model above what would be obtained by probability conversely, it will decrease when a predictor improves the model less than what is predicted by chance.

11. Differentiate between Ridge and Lasso Regression.

ANSWER:

Ridge- Ridge regression is a technique used to analyze multi-linear regression (multicollinear), also know as L2 regularization. It is Applied when predicted value are greater than the observed values.

Lasso- Lasso stands for – Least Absolute Shrinkage and selection operator. It is technique where data points are shrunk towards a central point, like the mean. Lasso is also as L1 regularization.

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

ANSWER:

A Variance inflation factor (VIF) is a measure of the amount of multicollinearity in regression analysis. Multicollinearity

exist when there is a correlation b/w multiple independent variable in a multiple regression model. This can adversely affect the regression results.

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

ANSWER:

Scaling the target value is a good idea in regression modelling; Scaling of the data makes it easy for a model to learn and understand the problem. Scaling of the data comes under the set of steps of data preprocessing when we are preforming machine learning algorithms in the data set.

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

ANSWER:

R-square, the proportion of variance in the outcome Y, explain by the covariates X, is commonly described as a measure of goodness of fit. This of course seems very reasonable, since R square measures how close the observed Y values are to the predicted (fitted) value from the model.

15. From the following confusion matrix calculate sensitivity, specificity, precision, recall and accuracy.

Actual/Predicted	True	False
True	1000	50
False	250	1200

ANSWER:

Accuracy = 1000+1200/1000+1200+50+250

=2,501.2

Precision = 1000/1000+50

=51

Sensitivity or Recall = 1000/1000+250

=251

Specificity = 1200/1200+50

=51

WORKSHEET 4 SQL

Q1 and Q2 have one or more correct answer. Choose all the correct option to answer your question.

1. Which of the following are TCL commands?

A. Commit

B. Select

C. Rollback

D. Savepoint

2. Which of the following are DDL commands?

A. Create

B. Select

C. Drop

D. Alter

Q3 to Q10 have only one correct answer. Choose the correct option to answer your question.

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

A. SELECT NULL FROM SALES;

B. SELECT NAME FROM SALES;

C. SELECT * FROM SALES WHEN PRICE = NULL;

D. SELECT # FROM SALES;

4. DCL provides commands to perform actions like

A. Change the structure of Tables

B. Insert, Update or Delete Records and Values

C. Authorizing Access and other control over Database

D. None of the above

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

A. Dates

B. Column Alias

C. String

D. All of the mentioned

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

A. ROLLBACK

B. COMMIT

C. TRUNCATE

D. DELETE

7. A subquery in an SQL Select statement is enclosed in:

A. Parenthesis - (...).

B. brackets - [...].

C. CAPITAL LETTERS.

D. braces - {...}.

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

A. FILE

B. REPORT

C. TABLE

D. FORM

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

A. Data types

B. Primary keys

C. Default values

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

B. DESC

C. There is no default value

D. None of the mentioned

Q11 to Q15 are subjective answer type questions, Answer them briefly.

11. What is denormalization?

ANSWER:

Denormalization is the process of adding precomputed data to an otherwise normalized relational database to improve read performance of the database.

12. What is a database cursor?

ANSWER:

A database cursor is an identifier with a group of rows. It is in a sense, a pointer to the current row in a buffer.

13. What are the different types of the queries?

ANSWER:

These SQL commands are mainly categorized into five categories as:

DDL – Data Definition Language

DQL – Data Query Language

DML – Data Manipulation Language

DCL – Data Control Language

TCL – Transaction Control Language

14. Define constraint?

ANSWER:

SQL constraints are used to specify rules for the data in a table. Constraints are used to limit the type of data that can go into a table. This ensures the accuracy and reliability of the data in the table. If there is any violation b/w the constraint and the data action, the action is aborted.

15. What is auto increment?

ANSWER:

Auto-increment allows a unique number to be generated automatically when a record is inserted into a table. Often this is the primary key field that we would like to be created automatically every time a new record is inserted.

STATISTICS WORKSHEET

Q1 to Q9 have only one correct answer. Choose the correct option to answer your question.

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

a) The outcome from the roll of a die

b) The outcome of flip of a coin

c) The outcome of exam

d) All of the mentioned

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

a) Discrete

b) Non Discrete

c) Continuous

d) All of the mentioned

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

a) pdf

b) pmv

c) pmf

d) all of the mentioned

4. The expected value or _____ of a random variable is the center of its distribution.

a) mode

b) median

c) mean

d) bayesian inference

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

a) variance

b) standard deviation

c) empirical mean

d) all of the mentioned

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

a) variance

b) standard deviation

c) mode

d) none of the mentioned

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

a) 0 and 10

b) 1 and 2

c) 0 and 1

d) None of the mentioned

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

difficult statistics?

a) baggyer

b) bootstrap

c) jacknife

d) none of the mentioned

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

a) frequency

b) summarized

c) raw

d) none of the mentioned

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

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

ANSWER:

Histograms are a special kind of bar graph that shows a bar for a range of data values instead of a single value. A box plot is a data display that draws a box over a number line to show the interquartile range of the data. The ‘whislers’ of a box plot show the least and greatest values in the data set.

11. How to select metrics?

ANSWER:

Prioritize objectives, examines which metric consistently predicts their achievement, and identify which activites influence predictors, in that order.

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

ANSWER:

To assess statistical significance, you would use hypothesis testing. The null hypothesis and alternate hypothesis would be stated first. Second, you'd calculated the p-value, which is the likelihood of getting the test's observed findings if the null hypothesis is true. Finally, you would select the threshold of significance (alpha) and reject the null hypothesis if the p-value is samller than the alpha- in other words, the result is statistically significant.

13. Give examples of data that doesnot have a Gaussian distribution, nor log-normal.

ANSWER:

Any distribution of money or value will be non-Gaussian. For example: distributions of income; distributions of house prices, distributions of bets placed on a sporting event. These distributions cannot have negative values and usually have extended right hand tails.

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

ANSWER:

Income is the classic example of when to use the median instead of the mean because its distribution tends to be skewed. The median indicates that half of all incomes fall below 27581 and half are above it. For these data, the mean overstimates where most household income fall.

15. What is the Likelihood?

ANSWER:

The likehood is the probability that a particular outcomesis observed when the true value of the parameter is, equivalent to the probability mass on : it is not a probability density over the parameter. The likehood, should not be confused with, which is the posterior probability of given the data.