MACHINE LEARNING

ASSIGNMENT - 6

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.

Ans-

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.

Ans-b

3. Which of the following is an ensemble technique? A) SVM B) Logistic Regression C) Random Forest D) Decision tree

Ans-a

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.

Ans-b

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

Ans-b

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

Ans-a and d

7. Which of the following is not an example of boosting technique? A) Adaboost B) Decision Tree C) Random Forest D) Xgboost.

Ans-b and c

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

Ans-a and c

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

Ans-d

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?

Ans- The adjusted R-squared is a modified version of R-squared that has been adjusted for the number of predictors in the model. The adjusted R-squared increases only if the new term improves the model more than would be expected by chance. It decreases when a predictor improves the model by less than expected by chance.

11. Differentiate between Ridge and Lasso Regression.

Ans-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.

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- the range of all features should be normalized so that each feature contributes approximately proportionately to the final distance.that is why we need to scale the data before feeding it to the train the model

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

Ans-While R Square is a relative measure of how well the model fits dependent variables, Mean Square Error is an absolute measure of the goodness for the fit. MSE is calculated by the sum of square of

prediction error which is real output minus predicted output and then divide by the number of data points.

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

Actual/Predicted	True	False
True	1000	50
False	250	1200 .

Ans- Sensitivity aka Recall (true positives / all actual positives) = TP / TP + FN.

= 1000/1000+50= 0.952

Accuracy (all correct / all) = TP + TN / TP + TN + FP + FN.

=1000+1200/1000+250+50

= 1.69

Precision (true positives / predicted positives) = TP / TP + FP.

=1000/1000+250

= 0.8

Specificity (true negatives / all actual negatives) =TN / TN + FP

=1200/1200+250

= 0.82

WORKSHEET 6 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

Ans-c and d

2. Which of the following are DDL commands? A. Create B. Select C. Drop D. Alter

Ans-a,c and d

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;

Ans-b

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

Ans-c

5. Which of the following should be enclosed in double quotes? A. Dates B. Column Alias C. String D. All of the mentioned

Ans-b

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

Ans-b

7. A subquery in an SQL Select statement is enclosed in: A. Parenthesis - (...). B. brackets - [...]. C. CAPITAL LETTERS. D. braces - {...}.

Ans-a

8. The result of a SQL SELECT statement is a :- A. FILE B. REPORT C. TABLE D. FORM

Ans- b

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

Ans-d

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

Ans-a

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

11. What is denormalization?

Ans-Denormalization is a database optimization technique in which we add redundant data to one or more tables. This can help us avoid costly joins in a relational database. ... In a traditional normalized database, we store data in separate logical tables and attempt to minimize redundant data.

12. What is a database cursor?

Ans- 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?

Ans- 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?

Ans-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 between the constraint and the data action, the action is aborted.

15. What is auto increment?

Ans- Auto-increment allows a unique number to be generated automatically when a new 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- 6

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

Ans-d

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

Ans-a

3. Which of the following function is associated with a continuous random variable? a) pdf b) pmv c) pmf d) all of the mentioned

Ans-a

4. The expected value or _____ of a random variable is the center of its distribution. a) mode b) median c) mean d) bayesian inference

Ans-c

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

Ans-a

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

Ans-a

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

Ans-c

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

Ans-b

9. Data that summarize all observations in a category are called _____ data. a) frequency b) summarized c) raw d) none of the mentioned

Ans- b

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

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

Ans- 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.

11. How to select metrics?

Ans- Choosing the right metrics

1. Good metrics are important to your company growth and objectives. Your key metrics should always be closely tied to your primary objective. ...
2. Good metrics can be improved. Good metrics measure progress, which means there needs to be room for improvement. ...
3. Good metrics inspire action.

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

Ans- **Here are the steps for calculating statistical significance:**

1. Create a null hypothesis.
2. Create an alternative hypothesis.
3. Determine the significance level.
4. Decide on the type of test you'll use.
5. Perform a power analysis to find out your sample size.
6. Calculate the standard deviation.
7. Use the standard error formula.

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

Ans-

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

Ans-

15. What is the Likelihood?

Ans- In statistics, the likelihood function measures the goodness of fit of a statistical model to a sample of data for given values of the unknown parameters.