FLIP ROBO TECHNOLOGIES

Worksheet Set – 4

Batch DS2301

Intern: Vishnukanth

Mail ID: vishnukh25@gmail.com

- ✓ **Machine Learning**
- ✓ **SQL**
- ✓ **Statistics**

Machine Learning

Worksheet – 4

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

Answer : c) High R-squared value for train-set & Low R-squared value for test-set

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.

Answer : b) Decision trees are highly prone to overfitting

3. Which of the following is an ensemble technique?

A) SVM B) Logistic Regression

C) Random Forest D) Decision tree

Answer : c) Random Forest

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.

Answer : b) Sensitivity (also known as True Positive Rate or Recall)

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

Answer : b) Model 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

Answer : a) Ridge and d) Lasso

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

A) Adaboost B) Decision Tree C) Random Forest D) Xgboost

Answer : B) Decision Tree

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

Answer : a) Pruning and c) Restricting the max depth of the tree.

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

Answer : b) A tree in the ensemble focuses more on the data points on which the previous tree was not performing well.

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?

The adjusted R-squared penalizes the presence of unnecessary predictors in the model by reducing the R-squared value as the number of predictors increase. The adjusted R-squared takes into account the number of predictors in the model and the sample size, so it provides a better estimate of the goodness of fit of the model compared to the R-squared. The adjusted R-squared is calculated by subtracting (1 - R^2) * (n - 1) / (n - p - 1), where n is the sample size and p is the number of predictors in the model. As the number of predictors increases, the value of (n - p - 1) decreases, and the penalization term becomes larger, causing the adjusted R-squared value to decrease, which indicates that the model is less good at fitting the data.

11. Differentiate between Ridge and Lasso Regression.

Ridge and Lasso Regression are two commonly used regularization techniques in linear regression. The main difference between them is the way they penalize the coefficients of the predictors in the model:

- Ridge Regression: In Ridge Regression, the regularization term is the sum of the squares of the coefficients of the predictors, multiplied by a regularization parameter (lambda). This has the effect of shrinking the coefficients towards zero, but it does not set any coefficients to exactly zero.
- Lasso Regression: In Lasso Regression, the regularization term is the sum of the absolute values of the coefficients of the predictors, multiplied by a regularization parameter (lambda). This has the effect of both shrinking the coefficients towards zero and setting some coefficients to exactly zero, effectively reducing the number of predictors in the model.

In summary, Ridge Regression is best suited for situations where all predictors are important and we want to reduce multicollinearity, while Lasso Regression is best suited for situations where only a subset of predictors is important and we want to perform feature selection.

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

- VIF (Variance Inflation Factor) is a measure of the amount of multicollinearity in a multiple regression model. It is calculated for each predictor in the model, and it represents the increase in the variance of the estimated coefficients when a predictor is included in the model compared to when it is excluded.
- A VIF value close to 1 indicates that the predictor is not highly correlated with any of the other predictors in the model, and a VIF value close to or greater than 10 indicates that the predictor is highly correlated with one or more of the other predictors in the model and that multicollinearity is present.
- In general, a suitable value of a VIF for a feature to be included in a regression modeling is a VIF value less than 5, although this threshold may vary depending on the specific modeling problem and the analyst's preference. If a feature has a VIF value greater than 5, it may indicate that the feature is redundant with one or more other features in the model, and it may be worth considering removing it.

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

Scaling the data before feeding it to the model is important for several reasons:

- Algorithm stability: Some algorithms, such as K-Nearest Neighbors and Neural Networks, are sensitive to the scale of the input features. Scaling the data helps to ensure that the algorithms are stable and perform well.
- Optimization convergence: Optimization algorithms used in many machine learning models, such as Gradient Descent, may converge more slowly or not at all if the input features are not scaled. Scaling the data can help the optimization algorithms converge more quickly.
- Handling of sparse data: Some algorithms, such as PCA and linear regression, perform better on data that is centered and scaled. Scaling the data can also help to handle sparse data by removing the influence of the large values and allowing the smaller values to contribute more to the model.
- Improved interpretability: When the data is scaled, the coefficients of the model are easier to interpret because they are not affected by the scale of the input features.
- Overall, scaling the data can improve the performance and interpretability of the model, making it an important step in the data pre-processing stage.

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

There are several metrics used to check the goodness of fit in linear regression:

- R-squared: It is a measure of the proportion of the variance in the dependent variable that is explained by the independent variables. A higher R-squared value indicates a better fit of the model.
- Mean Squared Error (MSE): It measures the average squared difference between the actual and predicted values. A lower MSE indicates a better fit of the model.
- Root Mean Squared Error (RMSE): It is the square root of the MSE and gives a more interpretable measure of the error in the model.
- Adjusted R-squared: It is similar to R-squared but takes into account the number of independent variables in the model. It adjusts the R-squared value to penalize the addition of unnecessary predictors.
- F-statistic: It is a measure of how well the model fits the data compared to a model with no independent variables. A high F-statistic indicates a better fit of the model.
- AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion): They are measures of the goodness of fit of the model, taking into account both the quality of the fit and the complexity of the model. Lower values of AIC and BIC indicate a better fit of the model.

These metrics can be used together to evaluate the performance of the linear regression model and determine the best fit for the data.

Structured Querying Language

Worksheet – 4

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

Answer : a) Commit , c) Rollback, d) Savepoint

2. Which of the following are DDL commands?

A. Create

B. Select

C. Drop

D. Alter

Answer : a) Create , c) Select, 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;

Answer : b) SELECT NAME 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

Answer : c) Authorizing Access and other control over Database

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

A. Dates B. Column Alias C. String D. All of the mentioned

Answer : c) String

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

Answer : b) COMMIT

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

A. Parenthesis - (...). B. brackets - [...].

C. CAPITAL LETTERS. D. braces - {...}.

Answer : b) Parenthesis - (...)

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

A. FILE B. REPORT C. TABLE D. FORM

Answer : c) table

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

Answer : 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

Answer : a) ASC

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

11. What is denormalization?

Denormalization is the process of adding redundant data to a database schema, in order to improve query performance and reduce the complexity of data relationships. This is often done by breaking down the relationships between tables and repeating data in multiple places, making the database design less normalized and more optimized for specific types of queries.

12. What is a database cursor?

A database cursor is a control structure that enables traversal over the rows in a result set of a database query. It allows processing of each row individually, and is used for fetching data from the database in a more efficient manner than fetching all the data in a single query and processing it in the application. It acts as a pointer to the current row in the result set, and is used to manage the context of the row being processed. Cursors are commonly used in stored procedures and other database programming constructs.

13. What are the different types of the queries?

- Select query
- Insert query
- Update query
- Delete query
- Merge query
- Alter query
- Create query
- Drop query
- Truncate query
- Index query.

14. Define constraint?

A constraint in SQL is a rule that is used to limit the values that can be stored in a column or a set of columns of a table. Constraints are used to enforce data integrity and maintain the correctness of the data in the database. Examples of constraints include primary key, foreign key, unique, not null, and check constraints. Constraints help ensure that the data stored in the database is accurate, consistent, and free from undesirable values. They also help enforce referential integrity between related tables.

15. What is auto increment?

Auto increment is a feature in SQL that allows a unique number to be generated automatically for a specified column in a table, whenever a new record is inserted into the table. The auto increment column is usually used as the primary key for the table, and its values are guaranteed to be unique for each record. The auto increment feature is supported by most relational database management systems, and it eliminates the need for manual value assignment for the primary key column, improving the efficiency and accuracy of the data management process. The auto increment value is usually assigned by the database management system and increments by a set amount each time a new record is added to the table.

STATISTICS

Worksheet – 4

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

Answer : 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

Answer : a) Discrete

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

a) pdf

b) pmv

c) pmf

d) all of the mentioned

Answer : a) pdf (Probability Density Function).

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

a) mode

b) median

c) mean

d) bayesian inference

Answer : c) mean

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

Answer : c) empirical mean

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

Answer : a) variance

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

Answer : c) 0 and 1

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

Answer : b) bootstrap

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

a) frequency

b) summarized

c) raw

d) none of the mentioned

Answer : a) frequency

Q10 and Q15 are subjective answer type questions, Answer them in your own words briefly.

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

A boxplot displays summary statistics of a distribution, including the median, quartiles, and outliers. A histogram displays the distribution of a set of continuous data by dividing the data into a set of intervals and plotting the frequency of observations within each interval.

11. How to select metrics?

- Define the objectives: Clearly define what you want to measure and what are the goals of your project.
- Identify relevant data: Collect data that is relevant to your objectives and goals.
- Choose appropriate metrics: Select metrics that accurately reflect the objectives and goals.
- Ensure the metrics are actionable: Choose metrics that are easy to understand and can be used to make decisions.
- Consider context: The metrics that are relevant in one context may not be relevant in another, consider the context in which the metrics will be used.
- Refine and iterate: Regularly review and refine the metrics to ensure they continue to accurately reflect the objectives and goals.

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

To assess the statistical significance of an insight, you can use hypothesis testing. The steps are:

- Formulate the null and alternative hypotheses: The null hypothesis represents the status quo, while the alternative hypothesis represents the new insight.
- Choose a significance level: The significance level (α) determines the level of risk you are willing to take in rejecting the null hypothesis when it is actually true.
- Determine the test statistic: Based on the data and the hypothesis, calculate a test statistic.
- Determine the p-value: The p-value is the probability of observing the test statistic if the null hypothesis is true.
- Make a decision: If the p-value is less than the significance level, reject the null hypothesis and accept the alternative hypothesis. This means the insight is statistically significant.
- Report the results: Clearly and concisely report the results, including the p-value and the significance level.

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

Examples of data that doesn't have a Gaussian (Normal) distribution nor log-normal:

- Bernoulli distribution: A discrete distribution with two possible outcomes (e.g. success or failure).
- Poisson distribution: A discrete distribution used to model the number of events over a fixed interval of time or space.
- Exponential distribution: A continuous distribution used to model the time between events in a Poisson process.
- Weibull distribution: A continuous distribution used to model the time to failure of a system or component.
- Uniform distribution: A continuous distribution with a constant probability over a defined interval.

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

An example where the median is a better measure than the mean is when the data has extreme outliers or is skewed. In such cases, the mean may be greatly influenced by the outliers, leading to a misrepresentation of the central tendency of the data. The median, on the other hand, is less sensitive to outliers and more accurately represents the central tendency in these cases. For example, in the case of income data, the mean may be greatly influenced by a small number of individuals with extremely high incomes, while the median represents the middle value and is a more representative measure of the typical income in the population.

15. What is the Likelihood?

Likelihood is a function that describes the probability of observing a specific set of data, given a set of parameters for a statistical model. It represents the fit of the model to the data, and is used in maximum likelihood estimation to find the parameter values that maximize the likelihood of the data given the model. The likelihood function is a measure of how well the model represents the data and is used to compare different models and to make inferences about population parameters.