

ASSIGNMENT-7

MACHINE LEARNING

Q1. D

Q2. A

Q3. A

Q4. B

Q5. A

Q6. A

Q7. B

Q8. C

Q9.

The Gini coefficient, also known as the Gini index, is the statistical measure used to measure the income distribution among the country's population, i.e., it helps measure the income inequality of the country's population.

A=40, b=60

Gini index = $a/a+b$

= $40/(40+60)$

=0.4

Q10.

The random forest algorithm provides a higher level of accuracy in predicting outcomes over the decision tree algorithm. It can handle large datasets efficiently. A random forest produces good predictions that can be understood easily.

Q11.

the scaling is used for making data points generalized so that the distance between them will be lower. Two common techniques for scaling numerical input variables are normalization and standardization.

Q12.

- 1) We can use fixed learning rate during training without worrying about learning rate decay.
- 2) It has straight trajectory towards the minimum and it is guaranteed to converge in theory to the global minimum if the loss function is convex and to a local minimum if the loss function is not convex.

Q13.

Accuracy is not a good metric for imbalanced datasets.

This model would receive a very good accuracy score as it predicted correctly for the majority of observations, but this hides the true performance of the model which is objectively not good as it only predicts for one class.

Q14. What is “f-score” metric? Write its mathematical formula.

The F-score, also called the F1-score, is a measure of a model’s accuracy on a dataset. It is used to evaluate binary classification systems, which classify examples into ‘positive’ or ‘negative’.

$$F_1 = \frac{2}{\frac{1}{\text{recall}} + \frac{1}{\text{precision}}} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

$$= \frac{\text{tp}}{\text{tp} + \frac{1}{2}(\text{fp} + \text{fn})}$$

Q15. What is the difference between fit(), transform() and fit_transform()?

The fit() method will allow us to get the parameters of the scaling function. The transform() method will transform the dataset to proceed with further data analysis steps. The fit_transform() method will determine the parameters and transform the dataset.

WORKSHEET 7 SQL

Q1. A & B

Q2. B & C

Q3. C

Q4. C

Q5. C

Q6. C

Q7. C

Q8. B

Q9. B

Q10. A

Q11.

JOIN statement is used to combine rows from two or more table, based on related column between them.

Q12.

INNER JOIN, LEFT OUTER JOIN and RIGHT OUTER JOIN

Q13.

SQL Server is a relational database management system (RDBMS). This server is available as a separate program and responsible for handling all the database instruction, statements, or commands.

Q14.

Primary Key is used to define uniqueness of column in table. There should be only one primary key in table. Comma can be use for assigning multiple primary key for columns.

Q15.

ETL, which stands for “extract, transform, load,” are the three processes that, in combination, move data from one database, multiple databases, or other sources to a unified repository—typically a data warehouse.

STATISTICS WORKSHEET-7

Q1. B

Q2. D

Q3. C

Q4. B

Q5. None of these

Q6. B

Q7. C

Q8. B

Q9. B

Q10. A

Q11. C

Q12. A

Q13. D

Q14. D

A15. B