WORKSHEET 7 SQL

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

1. The primary key is selected from the

Candidate keys

2. Which is/are correct statements about primary key of a table?

Primary keys cannot contain NULL values

A table can have only one primary key with single or multiple fields

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

3. Which SQL command is used to insert a row in a table?

Insert

4. Which one of the following sorts rows in SQL?

ORDERBY

5. The SQL statement that queries or reads data from a table is

SELECT

6. Which normal form is considered adequate for relational database design?

3NF

7. SQL can be used to

All of the above can be done by SQL

8. SQL query and modification commands make up

DML

9. The result of a SQL SELECT statement is a(n).

Table

10. Second normal form should meet all the rules for

1 NF

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

11. What are joins in SQL?

Join in DBMS is a binary operation which allows to combine join product and selection in one single statement.

The goal of creating a join condition is that it helps to combine the data from two or more DBMS tables.

The tables in DBMS are associated using the primary key and foreign keys.

A SQL Join statement is used to combine data or rows from two or more tables based on a common field between them.

12. What are the different types of joins in SQL?

Different types of Joins are:

INNER JOIN: The INNER JOIN keyword selects all rows from both the tables as long as the condition satisfies.

This keyword will create the result-set by combining all rows from both the tables where the condition satisfies

LEFT JOIN: This join returns all the rows of the table on the left side of the join and matching rows for the table on the right side of join.

The rows for which there is no matching row on right side, the result-set will contain null. LEFT JOIN is also known as LEFT OUTER JOIN.

RIGHT JOIN:RIGHT JOIN is similar to LEFT JOIN. This join returns all the rows of the table on the right side of the join and matching rows for the table on the left side of join. The rows for which there is no matching row on left side, the result-set will contain null. RIGHT JOIN is also known as RIGHT OUTER JOIN.

FULL JOIN: Creates the result-set by combining result of both LEFT JOIN and RIGHT JOIN. The result-set will contain all the rows from both the tables.

The rows for which there is no matching, the result-set will contain NULL values

13. What is SQL Server?

SQL SERVER is a relational database management system (RDBMS) developed by Microsoft. It is primarily designed and developed to compete with MySQL and Oracle database.

SQL Server supports ANSI SQL, which is the standard SQL (Structured Query Language) language. However, SQL Server comes with its own implementation of the SQL language, T-SQL

SQL Server works exclusively on Windows environment for more than 20 years. In 2016, Microsoft made it available on Linux.

SQL Server 2017 became generally available in October 2016 that ran on both Windows and Linux.

Usage of SQL Server

To create databases.

To maintain databases.

To analyze the data through SQL Server Analysis Services (SSAS).

To generate reports through SQL Server Reporting Services (SSRS).

To carry out ETL operations through SQL Server Integration Services (SSIS).

14. What is primary key in SQL?

The PRIMARY KEY constraint uniquely identifies each record in a table.

A table can have only one primary key, which may consist of one single or of multiple fields

Each table can have only one SQL Primary Key.

All the values are unique and Primary key SQL value can uniquely identify each row.

The system will not allow inserting a row with SQL Server Primary Key which already exists in the table.

Primary Key cannot be NULL.

15. What is ETL in SQL?

ETL stands for Extract, Transform and Load, which is a process used to collect data from various sources,

transform the data depending on business rules/needs and load the data into a destination database.

The need to use ETL arises from the fact that in modern computing business data resides in multiple locations and in many incompatible formats.

Extract, Transform and Load

Extract – The first step in the ETL process is extracting the data from various sources.

Each of the source systems may store its data in completely different format from the rest.

The sources are usually flat files or RDBMS, but almost any data storage can be used as a source for an ETL process.

Transform – Once the data has been extracted and converted in the expected format, it's time for the next step in the ETL process, which is transforming the data according to set of business rules. The data transformation may include various operations including but not limited to filtering, sorting, aggregating,

joining data, cleaning data, generating calculated data based on existing values, validating data, etc.

Load – The final ETL step involves loading the transformed data into the destination target, which might be a database or data warehouse.

ETL tools, includes IBM (IBM InfoSphere DataStage), Oracle (Oracle Warehouse Builder) and of course Microsoft with their SQL Server Integration Services (SSIS) included in certain editions of Microsoft SQL Server 2005 and 2008.

MACHINE LEARNING

1. Which of the following in sk-learn library is used for hyper parameter tuning?

All of the above

2. In which of the below ensemble techniques trees are trained in parallel?

Random forest

3. In machine learning, if in the below line of code:

sklearn.svm.SVC (C=1.0, kernel='rbf', degree=3)

we increasing the C hyper parameter, what will happen?

The regularization will decrease

4. Check the below line of code and answer the following questions:

sklearn.tree.DecisionTreeClassifier(*criterion='gini',splitter='best',max_depth=None, min_samples_split=2)

Which of the following is true regarding max_depth hyper parameter?

Both A & B

5. Which of the following is true regarding Random Forests?

It's an ensemble of weak learners.

6. What can be the disadvantage if the learning rate is very high in gradient descent?

Both of them

7. As the model complexity increases, what will happen?

Bias will decrease, Variance increase

8. Suppose I have a linear regression model which is performing as follows:

Train accuracy=0.95 and Test accuracy=0.75

Model is over fitting

<u>Q9 to Q15 are subjective answer type questions, Answer them briefly.</u>

9. Suppose we have a dataset which have two classes A and B.

The percentage of class A is 40% and percentage of class B is 60%. Calculate the Gini index and entropy of the dataset.

10. What are the advantages of Random Forests over Decision Tree?

Higher resolution in the feature space Trees are unpruned. While a single decision tree like CART is often pruned, a random forest tree is fully grown and unpruned, and so, naturally, the feature space is split into more and smaller regions.

Trees are diverse. Each random forest tree is learned on a random sample, and at each node, a random set of features are considered for splitting.

Both mechanisms create diversity among the trees.

Handling Over fitting

A single decision tree needs pruning to avoid overfitting. The following shows the decision boundary from an unpruned tree.

The boundary is smoother but makes obvious mistakes (overfitting).

11. What is the need of scaling all numerical features in a dataset? Name any two techniques used for scaling.

Feature scaling in machine learning is one of the most critical steps during the pre-processing of data before creating a machine learning model.

Scaling can make a difference between a weak machine learning model and a better one.

It refers to putting the values in the same range or same scale so that no variable is dominated by the other.

Normalization is a scaling technique in which values are shifted and rescaled so that they end up ranging between 0 and 1. It is also known as Min-Max scaling.

Standardization is another scaling technique where the values are centered around the mean with a unit standard deviation.

This means that the mean of the attribute becomes zero and the resultant distribution has a unit standard deviation.

12. Write down some advantages which scaling provides in optimization using gradient descent algorithm.

Main Advantages:

Faster than Batch version because it goes through a lot less examples than Batch.

Randomly selecting examples will help avoid redundant examples or examples that are very similar that don't contribute much to the learning..

13. In case of a highly imbalanced dataset for a classification problem, is accuracy a good metric to measure the performance of the model. If not, why?

Accuracy can be a useful measure if we have the same amount of samples per class but if we have an imbalanced set of samples accuracy isn't useful at all.

Even more so, a test can have a high accuracy but actually perform worse than a test with a lower accuracy.

The F-Measure is a popular metric for imbalanced classification. The Fbeta-measure measure is an abstraction of the F-measure where the balance of precisionand recall in the calculation of the harmonic mean is controlled by a coefficient called beta.

Precision metric tells us how many predicted samples are relevant i.e. our mistakes into classifying sample as a correct one if it's not true.

This metric is a good choice for the imbalanced classification scenario. The range of F1 is in [0, 1], where 1 is perfect classification and 0 is total failure

14. 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.

The F-score is a way of combining the precision and recall of the model, and it is defined as the harmonic mean of the model's precision and recall.

The F1 Score is the 2*((precision*recall)/(precision+recall)). It is also called the F Score or the F Measure. Put another way, the F1 score conveys the balance between the precision and the recall.

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

fit() :

In the fit() method, where we use the required formula and perform the calculation on the feature values of input data and fit this calculation to the transformer.

For applying the fit() method we have to use .fit() in front of the transformer object.

transform() :

For changing the data we probably do transform, in the transform() method, where we apply the calculations that we have calculated in fit() to every data point in feature F. We have to use .transform() in front of a fit object because we transform the fit calculations.

fit_transform():

This fit_transform() method is basically the combination of fit method and transform method, it is equivalent to fit().transform().

This method performs fit and transform on the input data at a single time and converts the data points. If we use fit and transform separate when we need both then it will decrease the efficiency of the model so we use fit_transform() which will do both the work.

STATISTICS WORKSHEET-7

1.A die is thrown 1402 times. The frequencies for the outcomes 1, 2, 3, 4, 5 and 6 are given in the following table:

Find the probability of getting 6 as outcome:

0.135

2.A telephone directory page has 400 telephone numbers. The frequency distribution of their unit place digit

(for example, in the number 25827689, the unit place digit is 9 is given in table below:

What will be the probability of getting a digit with unit place digit odd number that is 1, 3,5,7,9?

0.53

3.A tyre manufacturing company which keeps a record of the distance covered before a tyre needed to be replaced. The table below shows the results of 1100 cases.

If we buy a new tyre of this company, what is the probability that the tyre will last more than 9000 miles?

.745

4. Please refer to the case and table given in the question No. 3 and determine what is the probability that if we buy a new tyre then

it will last in the interval [4000-14000] miles?

0.577

5. We have a box containing cards numbered from 0 to 9. We draw a card randomly from the box.

If it is told to you that the card drawn is greater than 4 what is the probability that the card is odd?

0.5

6. We have a box containing cards numbered from 1 to 8. We draw a card randomly from the box.

If it is told to you that the card drawn is less than 4 what is the probability that the card is even?

0.33

7. A die is thrown twice and the sum of the numbers appearing is observed to be

What is the conditional probability that the number 6 has appeared at least on one of the die?

0.33

8. Consider the experiment of tossing a coin. If the coin shows tail, toss it again but if it shows head, then throw a die.

Find the conditional probability of the event that 'the die shows a number greater than 4' given that 'there is at least one Head'.

0.22

9. There are three persons Evan, Ross and Michelle. These people lined up randomly for a picture.

What is the probability of Ross being at one of the ends of the line?

0.66

10. Let us make an assumption that each born child is equally likely to be a boy or a girl. Now suppose, if a family has two children, what is the conditional probability that both are girls given that at least one of them is a girl?

0.33

11. Consider the same case as in the question no. 10. It is given that elder child is a boy.What is the conditional probability that both children are boys?

0.5

12. We toss a coin. If we get head, we toss a coin again and if we get tail we throw a die. What is the probability of getting a number greater than 4 on die?

0.34

13. We toss a coin. If we get head, we toss a coin again and if we get tail we throw a die. What is the probability of getting an odd number on die?

0.2

14. Suppose we throw two dice together. What is the conditional probability of getting sum of two numbers found on the two die after throwing is less than 4, provided that the two numbers found on the two die are different?

0.24

15.A box contains three coins: two regular coins and one fake two-headed coin, you pick a coin at random and toss it. What is the probability that it lands heads up?

3/4