Worksheet 3 Machine Learning

1. Which of the following is an application of clustering?

   Ans. d. All of the above (Biological network analysis, Market trend prediction, Topic modeling)

2. On which data type, we cannot perform cluster analysis?

   Ans. d. None

3. Netflix's movie recommendation system uses

   Ans. c. Reinforcement learning and Unsupervised learning

4. The final output of Hierarchical clustering is

   Ans. b. The tree representing how close the data points are to each other

5. Which of the step is not required for K-means clustering?

   Ans. d. None

6. Which is the following is wrong?

   Ans. c. k-nearest neighbor is same as k-means

7. Which of the following metrics, do we have for finding dissimilarity between two clusters in hierarchical clustering?

   Ans. d. 1, 2 and 3 (Single-link, Complete-link, Average-link)

8. Which of the following are true?

   Ans. a. 1 only (Clustering analysis is negatively affected by multicollinearity of features)

9. n the figure above, if you draw a horizontal line on y-axis for y=2. What will be the number of clusters formed?

   Ans. a. 2

10. For which of the following tasks might clustering be a suitable approach?

    Ans. c. Predicting whether stock price of a company will increase tomorrow.

11. Given, six points with the following attributes:

    Which of the following clustering representations and dendrogram depicts the use of MIN or Single link proximity function in hierarchical clustering:

    Ans. (A)

Worksheet 3 Machine Learning

12. Given, six points with the following attributes:

    Which of the following clustering representations and dendrogram depicts the use of MAX or Complete link proximity function in hierarchical clustering

    Ans. (B)

13. What is the importance of clustering?

    Ans. Clustering is very much important as it determines the intrinsic grouping among the unlabeled data present. There are no criteria for a good clustering. It depends on the user, what is the criteria they may use which satisfy their need

14. How can I improve my clustering performance?

    Ans. Cluster performance improved by using its methods are as follow:

    a. **Density-Based Methods:** These methods consider the clusters as the dense region having some similarities and differences from the lower dense region of the space.

    b. **Hierarchical Based Methods:** The clusters formed in this method form a tree-type structure based on the hierarchy.

    c. **Partitioning Methods:** These methods partition the objects into k clusters and each partition forms one cluster

    d. **Grid-based Methods:** In this method, the data space is formulated into a finite number of cells that form a grid-like structure.

Worksheet 3 SQL

1. Write SQL query to create table Customers.
   Ans. cursor.execute("CREATE TABLE customer_data (customer_number INT PRIMARY KEY, customer_name TEXT, phone_number INT, customer_address_line1 TEXT, customer_address_line2 TEXT, customer_city TEXT, customer_state TEXT, customer_country TEXT, customer_postalcode INT, customer_salesRepemployee_number INT, customer_creditLimit INT)")
2. Write SQL query to create table Orders.
   Ans. cursor.execute("CREATE TABLE orders (order_number INT PRIMARY KEY, order_date INT, order_requiredDate INT, order_shippedDate INT, order_status TEXT, order_comments TEXT, customer_number INT)")
3. Write SQL query to show all the columns data from the Orders Table.
   Ans. orders=cursor.execute("SELECT * FROM orders")
4. Write SQL query to show all the comments from the Orders Table.
   Ans. orders =cursor.execute("SELECT * comments FROM orders ")
5. Write a SQL query to show orderDate and Total number of orders placed on that date, from Orderstable.
   Ans. orders =cursor.execute("SELECT order_placed_date, SUM(total_order) FROM orders WHERE order_placed_date=30-12-22 GROUP BY order_placed_date")
6. Write a SQL query to show employeNumber, lastName, firstName of all the employees from employees table.
   Ans. employee= cursor.execute("SELECT employee_number, employee_lastName, employee_firstName FROM employees ")
7. Write a SQL query to show all orderNumber, customerName of the person who placed the respective order.
   Ans. orders=cursor.execute("SELECT *order_number FROM orders")
   ```
   For row in orders:
       Print(row)
   ```
8. Write a SQL query to show name of all the customers in one column and salerepemployee name inanother column
   Ans. customer=cursor.execute("SELECT customer_name , saleremploye_name FROM customer_data")
9. Write a SQL query to show Date in one column and total payment amount of the payments made on that date from the payments table
   Ans. pyt=cursor.execute("SELECT payment_date , payment_amount FROM payments WHERE payment_date=31-12-22")
10. Write a SQL query to show all the products productName, MSRP, productDescription from the products table.
    Ans. prd =cursor.execute("SELECT * product_Name, MSRP, product_Description FROM products")
11. Write a SQL query to print the productName, productDescription of the most ordered product.
    Ans. prd =cursor.execute("SELECT MAX (amount),product_Name, product_Description FROM products")
    ```
    Print("Maximum order="prd.fetchone())
    ```

Worksheet 3 SQL

12. Write a SQL query to print the city name where maximum number of orders were placed.
    Ans. ord=cursor.execute(“ SELECT MAX(placed_order), city FROM orders”)
    print(“Maximum order city =”prd.fetchone())
13. Write a SQL query to get the name of the state having maximum number of customers.
    Ans. cust=cursor.execute(“SELECT MAX (customer_number), Customer_state FROM customer_data”)
    Print(“Maximum customer state=”, cust.fetchone())
14. Write a SQL query to print the employee number in one column and Full name of the employee in the second column for all the employees.
    Ans. Emp=cursor.execute(“SELECT employee_number, employee_namee FROM employees”)
    For row in emp:
    Print(row)
15. Write a SQL query to print the orderNumber, customer Name and total amount paid by the customer for that order (quantityOrdered × priceEach)

    Ans. ord=cursor.execute(“SELECT orders.order_number,orders.customer_name, order_details.quantity_ordered, order_details.priceEach FROM orders INNER JOIN order_details ON orders.order_number=order_deatils.order+number”)
    For row in ord:
    Print(row)

Worksheet 3 Statistics

1. Which of the following is the correct formula for total variation?
   Ans. b) Total Variation = Residual Variation + Regression Variation
2. Collection of exchangeable binary outcomes for the same covariate data are called outcomes.
   Ans. c) Binomial
3. How many outcomes are possible with Bernoulli trial?
   Ans. a) 2
4. If Ho is true and we reject it is called
   Ans. a) Type-I error
5. Level of significance is also called:
   Ans. c) Level of confidence
6. The chance of rejecting a true hypothesis decreases when sample size is:
   Ans. b) Increase
7. Which of the following testing is concerned with making decisions using data?
   Ans. b) Hypothesis
8. What is the purpose of multiple testing in statistical inference?
   Ans. d) All of the mentioned
9. Normalized data are centered at and have units equal to standard deviations of the original data
   Ans. a) 0
10. What Is Bayes' Theorem?
    Ans. Bayes' theorem describes the probability of occurrence of an event related to any condition. It is also considered for the case of conditional probability. Bayes theorem is also known as the formula for the probability of "causes".
11. What is z-score?
    Ans. Z-score is also known as standard score gives us an idea of how far a data point is from the mean. It indicates how many standard deviations an element is from the mean. Hence, Z-Score is measured in terms of standard deviation from the mean.
12. What is t-test?
    Ans. A **t test** is a statistical test that is used to compare the means of two groups. It is often used in hypothesis testing to determine whether a process or treatment actually has an effect on the population of interest, or whether two groups are different from one another. There are two types:
    a. One sample
    b. Two sample
13. In statistics, percentiles are used to understand and interpret data. Then the percentile of a set of data is the value at which n percent of the data is below it. In everyday life, percentiles are used to understand values such as test scores, health indicators, and other measurements.
14. What is ANOVA?

Worksheet 3 Statistics

Ans. An ANOVA (Analysis of variance) test is a type of statistical test used to determine if there is a statistically significant difference between two or more categorical groups by testing for differences of means using variance.

Another Key part of ANOVA is that it splits the independent variable into 2 or more group.

15. How can ANOVA help?

Ans ANOVA helps in these ways:

- It separates observed variance data into different components to use for additional tests.
- A one-way ANOVA is used for three or more groups of data, to gain information about the relationship between the dependent and independent variables.
- If no true variance exists between the groups, the ANOVA's F-ratio should equal close to 1.