Machine Learning Assignment 3

1. d
2. d
3. c
4. b
5. d
6. c
7. d
8. a
9. a
10. b
11. a
12. b
13. Clustering methods help in grouping data points into clusters using the different techniques are used to pick the appropriate result for the problem, these clustering techniques helps in grouping the data points into similar categories, and each of these subcategories is further divided into subcategories to assist the exploration of the queries output.
    - Utilizing clustering techniques allows one to restart the local search process and eliminate inefficiencies. Clustering also aids in identifying the internal structure of the data.
    - Clustering helps in understanding the natural grouping in a dataset. Their purpose is to make sense to partition the data into some group of logical groupings.
    - Clustering quality depends on the methods and the identification of hidden patterns.
    - They play a wide role in applications like marketing economic research and weblogs to identify similarity measures, Image processing, and spatial research.
    - They aid in the detection of credit card fraud through outlier detections.

14. An unsupervised machine learning technique called clustering seeks to divide data into different groups, or clusters. Hierarchical, density-based, and similarity-based types are just a few of the variations. Each also has a few distinct algorithms connected to it. Feature engineering is one of the most challenging aspects of any machine learning method, and it can be particularly challenging with clustering because it is tough to determine what best divides your data into distinct but related groups. Similar objects should be grouped together and dissimilar objects should be in different clusters, according to the similarity-based clustering guiding principle. This is the same objective as the majority of traditional clustering techniques. A metric must be provided when using similarity-based clustering to establish how similar two objects are. This similarity measure is based on distance, and while multiple distance metrics may be used, it typically yields a number in the range [0,1], with 0 denoting no resemblance and 1 denoting equivalence. We must employ a weighted Euclidean distance function in order to gauge the value of feature weights.

SQL Worksheet

1. CREATE TABLE customers(
   customerNumber int not null auto_increment primary key,
   customerName varchar(255),
   contactLastName varchar(255),
   contactFirstName varchar(255),
   phone varchar(15),
   addressLine1 varchar(255),
   addressLine2 varchar(255),
   city varchar(255),
   state varchar(255),
   postalCode varchar(255),
   country varchar(100),
   employeeNumber int not null,
   creditLimit decimal(15, 2)
   on update cascade
   on delete restrict)

2. CREATE TABLE orders(
   orderNumber int auto_increment not null primary key,
   orderDate date,
   requiredDate date,
   shippedDate date,
   statuses text,
   comments text,
   customerNumber int not null,
   on update cascade
   on delete restrict)
3. SELECT * FROM orders
4. SELECT comments FROM orders
5. SELECT orderDate,COUNT(orderDate) FROM orders GROUP BY(orderDate)
6. SELECT employeeNumber, lastName, firstName FROM employees
7. SELECT orderNumber.orders, customerName.customers
   FROM orders,customers
   WHERE orders.customerNumber=customers.customerNumber
8. SELECT customerName.customers, firstName.employees, lastName.employees
   FROM customers,employees
   WHERE employeeNumber.employees=salesRepEmployeeNumber.customers
9. SELECT paymentDate, SUM(amount) FROM payments GROUP BY(paymentDate)
10. SELECT productName,MSRP,productDescription FROM products
11. SELECT productName,productDescription FROM products,orderdetails
    ORDER BY (quantityOrdered.orderdetails) DESC
    WHERE productCode.products=productCode.orderdetails
12.
13. SELECT state, COUNT(state) FROM customers
    GROUP BY (state)
    ORDER BY (state) DESC

14. SELECT employeeNumber, firstName||lastName FROM employees
15. SELECT orderNumber.orders, customerName.customers,
    quantityOrdered.orderdetails*priceEach.orderdetails
    FROM orders,customers,orderdetails
    WHERE orders.customerNumber=customers.customerNumber
    AND orders.orderNumber=orderdetails.orderNumber

Statistics Worksheet

1. b
2. c
3. a
4. a
5. b
6. b
7. b
8. d
9. a
10. Thomas Bayes, an English statistician, philosopher, and Presbyterian pastor, introduced the Bayes theorem in the seventeenth century. In machine learning applications, such as classification tasks, conditional probability is calculated using the Bayesian technique, a key component of the Bayes theorem. Additionally, a condensed version of the Bayes theorem (Naive Bayes classification) is employed to shorten calculation times and lower typical project costs. The Bayes theorem is frequently referred to as the Bayes rule or Bayes Law. The Bayes theorem aids in calculating an event's probability using random knowledge. It is employed to determine the likelihood of one event occurring while another has already happened. The condition probability and marginal probability can be related using this method well.
11. Simply explained, a z-score, also known as a standard score, provides a sense of how far a data point is from the mean. Technically speaking, however, it's a measurement of how many standard deviations a raw score is from or above the population mean.
12. A statistical test called a t test is employed to compare the means of two groups. It is frequently employed in hypothesis testing to establish whether a procedure or treatment truly affects the population of interest or whether two groups differ from one another.
13. Percentile is characterised as the value below which a predetermined percentage of scores fall. Although percentage and percentile are sometimes misconstrued, they are two distinct ideas. The percentiles are the values below which a specific proportion of the data in a data collection is discovered, whereas the percentage is used to indicate fractions of a whole.
14. ANOVA is a statistical analysis technique that divides systematic components from random factors to account for the observed aggregate variability within a data set. The presented data set is statistically affected by the systematic factors but not by the

random ones. The ANOVA test is used by analysts to evaluate the impact of independent factors on the dependent variable in a regression analysis.

15. To determine whether there are any appreciable variations in the means of your independent variables, a one-way ANOVA can be useful. We may start to identify which independent variable is related to the dependent variable and start to grasp what is motivating that behaviour when we recognise how each independent variable's mean differs from the others.