Machine learning worksheet 2

Q .1) A

Q .2) D

Q .3) A

Q .4) A

Q .5) B

Q .6) B

Q .7) A

Q .8) D

Q .9) A

Q .10) D

Q .11) D

Q.12) IS K sensitive to outliers?

The K-means clustering algorithm is sensitive to outliers, because a mean is easily influenced by extreme values. K-medoids clustering is a variant of K-means that is more robust to noises and outliers.

Q.13) why is K means better?

Relatively simple to implement.

Scales to large data sets.

Guarantees convergence.

Can warm-start the positions of centroids.

Easily adapts to new examples.

Generalizes to clusters of different shapes and sizes, such as elliptical clusters.

Q.14) is k means deterministic algorithm?

One of the significant drawbacks of K-Means is its non-deterministic nature. K-Means starts with a random set of data points as initial centroids. This random selection influences the quality of the resulting clusters. Besides, each run of the algorithm for the same dataset may yield a different output.

SQL WORKSHEET 2

Q .1) D

Q .2) C

Q .3) D

Q .4) C

Q .5) B

Q .6) B

Q .7) A

Q .8) D

Q .9) A

Q .10) B

Q .11) B

Q .12) D

Q .13) A

Q .14) B & C

Q .15) A& B

STATISTICS WORKSHEET 2

Q .1) B

Q .2) C

Q .3) D

Q .4) B

Q .5) B

Q .6) B

Q .7) A

Q .8) B

Q .9) D

Q .10) A

Q .11) C

Q .12) D

Q .13) D

Q .14) A

Q .15) D

Machine learning worksheet 3

Q .1) D

Q .2) D

Q .3) C

Q .4) B

Q .5) D

Q .6) C

Q .7) D

Q .8) A

Q .9) A

Q .10) A

Q .11) A

Q .12) B

Q.13) what is the importance of clustering?

They can cluster different customer types into one group based on different factors, such as purchasing patterns. The factors analysed through clustering can have a big impact on sales and customer satisfaction, making it an invaluable tool to boost revenue, cut costs, or sometimes even both.

Q.14) how can I improve my clustering performance?

Graph-based clustering performance can easily be improved by applying ICA(Independent Component Analysis) blind source separation during the graph Laplacian embedding step. Applying unsupervised feature learning to input data using either RICA or SFT, improves clustering performance.

SQL WORKSHEET 3

Q .1) The SQL CREATE TABLE statement for the customers table is:

Cursor.execute(“CREATE TABLE customer_data(customer_no INT PRIMARY KEY,customer_name TEXT,contact_lastname TEXT, contact_firstname TEXT,phone INT,addressline1 TEXT, addressline2 TEXT,city TEXT,state TEXT,postal_code INT,country TEXT,salesrepemployee_no INT,credit_limit INT”))

Q .2)

Cursor.execute(“CREATE TABLE orders(order_no INT PRIMARY KEY,order_date INT,required_date INT, shipped_date INT,status TEXT,comments TEXT,customer_no INT”))

Q .3) SELECT order_no,order_date,reuuired_date,shipped_date,status,comments,customer_no FROM orders;

Q .4) SELECT * FROM orders;

Q .5) SELECT COUNT(*) FROM orders WHERE ord_date='2022-08-18';

Q .6) SELECT employee_no, first_name, last_name FROM employees

Q .7) SELECT a.cust_name, b.ord_no,FROM customer a LEFT OUTER JOIN orders b ON a.customer_id=b.customer_id order by b.ord_date;

Q .8) SELECT customer_name,salesrepemployee _name FROM customers LEFT OUTER JOIN employees;

Q .9) SELECT payment_date,amount FROM payments;

Q .10) SELECT product_name,MSRP,product_discription FROM products;

Q .11)
```
SELECT p.`product_id`, p.`name`, SUM(o.`quantity`) AS quantity
FROM `Order_Detail` AS o
   INNER JOIN `Product` AS p
   ON o.`product_id` = p.`product_id`
GROUP BY o.`product_id`
ORDER BY SUM(o.`quantity`) DESC, p.`name` ASC
LIMIT 3
```

STATISTICS WORKSHEET 3

Q .1) B

Q .2) C

Q .3) A

Q .4) A

Q .5) B

Q .6) B

Q .7) B

Q .8) D

Q .9) A

Q .10) what is Baye's theorem?

Baye's theorem is one of the most popular machine learning concepts that helps to calculate the probability of occurring one event with uncertain knowledge while other one has already occurred.

Bayes' theorem can be derived using product rule and conditional probability of event X with known event Y:

- According to the product rule we can express as the probability of event X with known event Y as follows;

1. P(X ? Y)= P(X|Y) P(Y) {equation 1}

- Further, the probability of event Y with known event X:

1. P(X ? Y)= P(Y|X) P(X) {equation 2}

Mathematically, Bayes theorem can be expressed by combining both equations on right hand side. We will get:

$$P(X|Y) = \frac{P(Y|X).P(X)}{P(Y)}$$

Here, both events X and Y are independent events which means probability of outcome of both events does not depends one another.

The above equation is called as Bayes Rule or Bayes Theorem.

- P(X|Y) is called as **posterior**, which we need to calculate. It is defined as updated probability after considering the evidence.
- P(Y|X) is called the likelihood. It is the probability of evidence when hypothesis is true.
- P(X) is called the **prior probability**, probability of hypothesis before considering the evidence
- P(Y) is called marginal probability. It is defined as the probability of evidence under any consideration.

Hence, Bayes Theorem can be written as:

posterior = likelihood * prior / evidence

Q .11) what is z-score ?

Simply put, a z-score (also called a standard score) gives you an idea of how far from the mean a data point is. But more technically it's a measure of how many standard deviations below or above the population mean a raw score is. A z-score can be placed on a normal distribution curve.

Q .12) what is t-test?

The t test estimates the true difference between two group means using the ratio of the difference in group means over the pooled standard error of both groups. You can calculate it manually using a formula, or use statistical analysis software.

Q .13) what is percentile?

A percentile is a comparison score between a particular score and the scores of the rest of a group. It shows the percentage of scores that a particular score surpassed. For example, if you score 75 points on a test, and are ranked in the 85 th percentile, it means that the score 75 is higher than 85% of the scores.

Q .14) what is ANOVA?

ANOVA is to test for differences among the means of the population by examining the amount of variation within each sample, relative to the amount of variation between the samples. Analyzing variance tests the hypothesis that the means of two or more populations are equal.

Q .14) how can ANOVA help?

ANOVA is helpful for testing three or more variables. It is similar to multiple two-sample t-tests. However, it results in fewer type I errors and is appropriate for a range of issues. ANOVA groups differences by comparing the means of each group and includes spreading out the variance into diverse sources.