STATISTICS

Q1 to Q9 have only one correct answer. Choose the correct option to answer your question.

1. Which of the following is the correct formula for total variation?
a) Total Variation = Residual Variation – Regression Variation

2. Collection of exchangeable binary outcomes for the same covariate data are called outcomes.
a) random

3. How many outcomes are possible with Bernoulli trial?
c) 4

4. If Ho is true and we reject it is called
c) Standard error

5. Level of significance is also called:
a) Power of the test

6. The chance of rejecting a true hypothesis decreases when sample size is:
c) Both of them

7. Which of the following testing is concerned with making decisions using data?
a) Probability

8. What is the purpose of multiple testing in statistical inference?
c) Minimize false negatives

9. Normalized data are centred at
and have units equal to standard deviations of the original data
d) 10

Q10 and Q15 are subjective answer type questions, Answer them in your own words briefly.

10. What Is Bayes' Theorem?

Understanding Bayes' Theorem

Applications of Bayes' Theorem are widespread and not limited to the financial realm. For example, Bayes' theorem can be used to determine the accuracy of medical test results by taking into consideration how likely any given person is to have a disease and the general accuracy of the test. Bayes' theorem relies on incorporating prior probability distributions in order to generate posterior probabilities.

Prior probability, in Bayesian statistical inference, is the probability of an event occurring before new data is collected. In other words, it represents the best rational assessment of the probability of a particular outcome based on current knowledge before an experiment is performed.

Posterior probability is the revised probability of an event occurring after taking into consideration the new information. Posterior probability is calculated by updating the prior probability using Bayes' theorem. In statistical terms, the posterior probability is the probability of event A occurring given that event B has occurred.

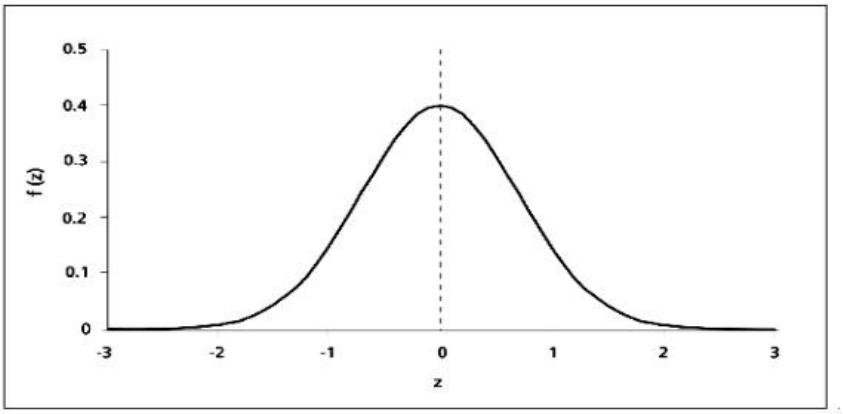
How Is Bayes' Theorem Used in Machine Learning?

Bayes Theorem provides a useful method for thinking about the relationship between a data set and a probability. In other words, the theorem says that the probability of a given hypothesis being true based on specific observed data can be stated as finding the probability of observing the data given the hypothesis multiplied by the probability of the hypothesis being true regardless of the data, divided by the probability of observing the data regardless of the hypothesis.

11. What is z-score?

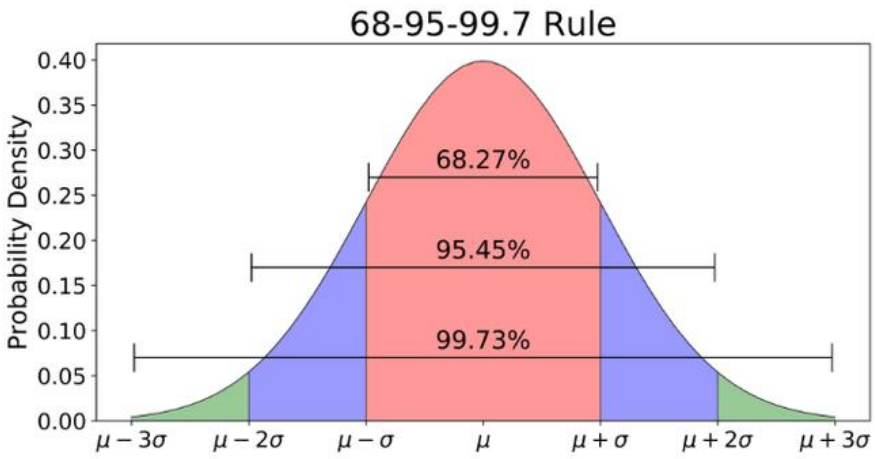
A z-score describes the position of a raw score in terms of its distance from the mean, when measured in standard deviation units. The z-score is positive if the value lies above the mean, and negative if it lies below the mean.

It is also known as a standard score, because it allows comparison of scores on different kinds of variables by standardizing the distribution. A standard normal distribution (SND) is a normally shaped distribution with a mean of 0 and a standard deviation (SD) of 1



Interpretation of Z-score

- An element having a z-score less than 0 represents that the element is less than the mean.
- An element having a z-score greater than 0 represents that the element is greater than the mean.
- An element having a z-score equal to 0 represents that the element is equal to the mean.
- An element having a z-score equal to 1 represents that the element is 1 standard deviation greater than the mean; a z-score equal to 2, 2 standard deviations greater than the mean, and so on.
- An element having a z-score equal to -1 represents that the element is 1 standard deviation less than the mean; a z-score equal to -2, 2 standard deviations less than the mean, and so on.
- If the number of elements in a given set is large, then about 68% of the elements have a z-score between -1 and 1; about 95% have a z-score between -2 and 2; about 99% have a z-score between -3 and 3. This is known as the Empirical Rule or the 68-95-99.7 Rule and can be demonstrated in the image below



The 68–95–99.7 Rule for a Normal Distribution

12. What is t-test?

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.

13. What is percentile?

In statistics, a percentile is a term that describes how a score compares to other scores from the same set. While there is no universal definition of percentile, it is commonly expressed as the percentage of values in a set of data scores that fall below a given value.

14. What is ANOVA?

Analysis of variance (ANOVA) is an analysis tool used in statistics that splits an observed aggregate variability found inside a data set into two parts: systematic factors and random factors. The systematic factors have a statistical influence on the given data set, while the random factors do not. Analysts use the ANOVA test to determine the influence that independent variables have on the dependent variable in a regression study.

The t- and z-test methods developed in the 20th century were used for statistical analysis until 1918, when Ronald Fisher created the analysis of variance method.12 ANOVA is also called the Fisher analysis of variance, and it is the extension of the t- and z-tests. The term became well-known in 1925, after appearing in Fisher's book, "Statistical Methods for Research Workers."3 It was employed in experimental psychology and later expanded to subjects that were more complex.

15. 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. Analysis of variance (ANOVA) is a statistical technique that is used to check if the means of two or more groups are significantly different from each other.

MACHINE LEARNING

Q1 to Q12 have only one correct answer. Choose the correct option to answer your question.

1. Which of the following is an application of clustering?
a. Biological network analysis

2. On which data type, we cannot perform cluster analysis?
d. None

3. Netflix's movie recommendation system uses
a. Supervised learning

4. The final output of Hierarchical clustering is
d. All of the above

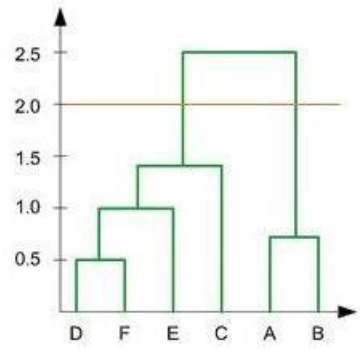
5. Which of the step is not required for K-means clustering?
c. Initial guess as to cluster centroids

6. Which is the following is wrong?
b. k-means clustering tries to group n observations into k clusters
c. k-nearest neighbour is same as k-means

7. Which of the following metrics, do we have for finding dissimilarity between two clusters in hierarchical clustering?
i.
Single-link
ii.
Complete-link
iii.
Average-link
Options:
a.1 and 2

8. Which of the following are true?
i. Clustering analysis is negatively affected by multicollinearity of features
ii. Clustering analysis is negatively affected by heteroscedasticity
Options:
c. 1 and 2

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



a. 2

10. For which of the following tasks might clustering be a suitable approach?
b.Given a database of information about your users, automatically group them into different market segments.

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:

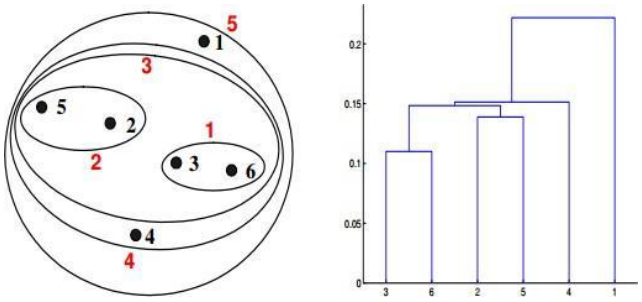
point	x coordinate	y coordinate
p1	0.4005	0.5306
p2	0.2148	0.3854
p3	0.3457	0.3156
p4	0.2652	0.1875
p5	0.0789	0.4139
p6	0.4548	0.3022

Table : X-Y coordinates of six points.

	p1	p2	p3	p4	p5	p6
p1	0.0000	0.2357	0.2218	0.3688	0.3421	0.2347
p2	0.2357	0.0000	0.1483	0.2042	0.1388	0.2540
p3	0.2218	0.1483	0.0000	0.1513	0.2843	0.1100
p4	0.3688	0.2042	0.1513	0.0000	0.2932	0.2216
p5	0.3421	0.1388	0.2843	0.2932	0.0000	0.3921
p6	0.2347	0.2540	0.1100	0.2216	0.3921	0.0000

Table : Distance Matrix for Six Points

a.



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.

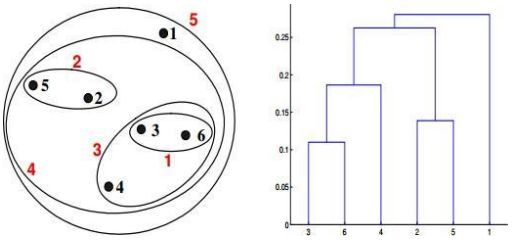
point	x coordinate	y coordinate
p1	0.4005	0.5306
p2	0.2148	0.3854
p3	0.3457	0.3156
p4	0.2652	0.1875
p5	0.0789	0.4139
p6	0.4548	0.3022

Table : X-Y coordinates of six points.

	p1	p2	p3	p4	p5	p6
p1	0.0000	0.2357	0.2218	0.3688	0.3421	0.2347
p2	0.2357	0.0000	0.1483	0.2042	0.1388	0.2540
p3	0.2218	0.1483	0.0000	0.1513	0.2843	0.1100
p4	0.3688	0.2042	0.1513	0.0000	0.2932	0.2216
p5	0.3421	0.1388	0.2843	0.2932	0.0000	0.3921
p6	0.2347	0.2540	0.1100	0.2216	0.3921	0.0000

Table : Distance Matrix for Six Points

c.



Q13 to Q14 are subjective answers type questions, Answers them in their own words briefly

13. What is the importance of clustering?

Clustering methods (like Hierarchical method, Partitioning, Density-based method, Model-based clustering, and Grid-based model) help in grouping the 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.

Importance of Clustering Methods

Having clustering methods helps in restarting the local search procedure and remove the inefficiency. In addition, clustering helps to determine the internal structure of the data.

1. This clustering analysis has been used for model analysis, vector region of attraction.
2. 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.
3. Clustering quality depends on the methods and the identification of hidden patterns.
4. They play a wide role in applications like marketing economic research and weblogs to identify similarity measures, Image processing, and spatial research.
5. They are used in outlier detections to detect credit card fraudulence.

14. How can I improve my clustering performance?

Monitoring cluster coordination at national and sub-national level is necessary to ensure that clusters are:

- efficient and effective coordination mechanisms
- fulfil the core cluster functions
- support efficient delivery of relevant services
- meet the needs of cluster members and demonstrate accountability to affected people

Monitoring also ensures that the architecture of coordination responds to changes in the context and in coordination needs

Cluster Coordination Performance Monitoring (CCPM)

Cluster Coordination Performance Monitoring (CCPM) is a self-assessment exercise. Clusters assess their performance against the six core cluster functions and accountability to affected populations. It is a country-led process, supported globally. Ideally, it is carried out by all clusters/sectors at the same time but can be implemented on demand by individual clusters. The process enables all cluster partners and coordinators to identify strengths and weaknesses of performance and paths to improvement.

The CCPM should ideally be implemented by all clusters three to six months after the onset of an emergency and annually thereafter. In protracted crises, the recommendation is, for all clusters, to complete a CCPM annually.

The process generally involves the following four steps and outputs:

Step 1 - Planning

- The HCT should initiate discussions and agree on the broad parameters for the implementation of the process.
- The Inter-cluster Coordination Group should then look in more detail at how to roll-out the process: deciding on issues such as whether only national or also sub-national levels will be involved; agreeing the timeline; and allocating roles and responsibilities for the process.
- Individual clusters meet with their partners to explain the purpose and clarify the different steps and processes involved (A template presentation is available to facilitate this meeting).

Step 2 - CCPM Survey

- The cluster coordinator completes a cluster description survey (online).
- The cluster coordinator and cluster partners each complete separate (online) feedback questionnaires (20-30 minutes).
- Global Clusters use an automated system to compile survey data and produce a Cluster Description Report with information on the cluster' structures and on the availability of key outputs linked to the cluster functions. The system also produces a Preliminary Cluster Coordination Performance Report which includes a colour coded analysis of the six core functions and on accountability to affected population. This preliminary performance analysis is a snap shot, and primarily serves to focus the discussion with partners to agree on an action plan for strengthening the cluster's performance.

Step 3 - Cluster Analysis and Action Planning

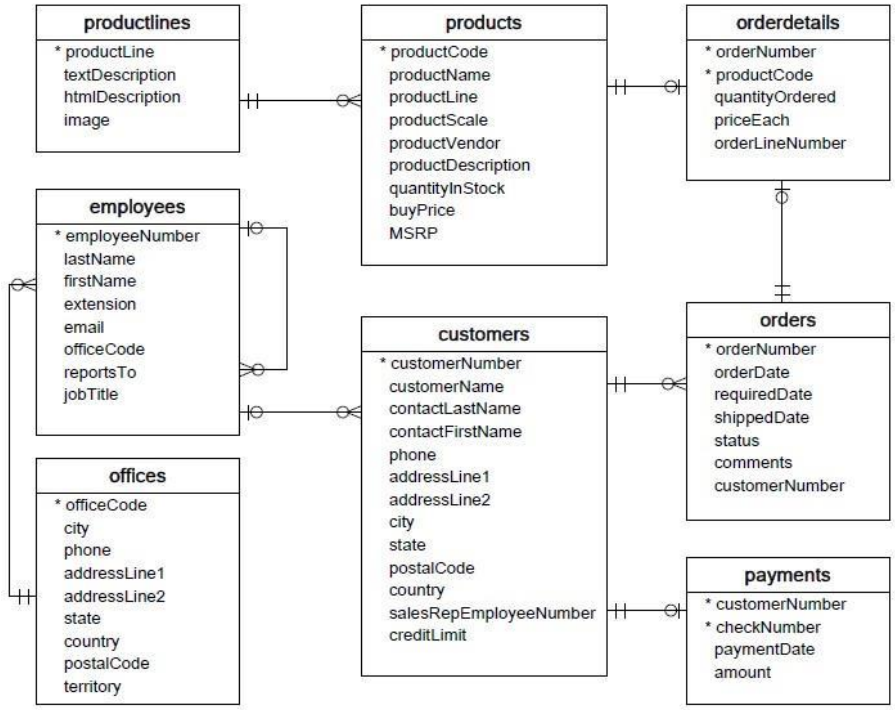
- In a half or full day workshop, each cluster discusses the cluster description and the survey results (and any related questions), identifies mitigating factors and explanations of performance and agreeing on specific actions that could be taken to improve performance if necessary. The Performance Report and an Action Plan are then finalized with the additional information and shared with stakeholders.

Step 4 - Follow-up and Monitoring

- The Inter-Cluster Coordination Group reviews the final Cluster Coordination Performance Reports and Action Plans and identifies common weaknesses across clusters that need to be addressed systematically.
- The reports and action plans are presented to the HCT to agree which actions require their support and to Global Clusters to identify individual cluster support requirements.
- Each cluster monitors implementation of its Action Plan at regular intervals, reporting to the HCT on progress.

SQL

Refer the following ERD and answer all the questions in this worksheet. You have to write the queries using mysql for the required Operation.



Customers: stores customer's data.
Products: stores a list of scale model cars.
ProductLines: stores a list of product line categories.
Orders: stores sales orders placed by customers.
OrderDetails: stores sales order line items for each sales order.
Payments: stores payments made by customers based on their accounts.
Employees: stores all employee information as well as the organization structure such as who reports to whom.
Offices: stores sales office data.

1. Write SQL query to create table **Customers.**
Solution:

```
CREATE TABLE customers
(
customerNumber int,
customerName varchar(255),
customerFirstName varchar(255),
customerLastName varchar(255),
phone int,
addressLine1 varchar(255),
addressLine2 varchar(255),
city varchar(255),
state varchar(255),
postalcode int,
country varchar(255),
salesRepEmployeeNumber int,
creditLimt int
);
```

2. Write SQL query to create table **Orders.**
Solution:

```
CREATE TABLE Orders
(
```

```
orderNumber int,
orderDate int,
requiredDate int,
shippedDate int,
status varchar(255),
comments varchar(255),
customerNumber int
);
```

3. Write SQL query to show all the columns data from the **Orders** Table.
Solution:

```
select * from orders
select orderNumber,orderDate,requiredDate,shippedDate,status,comments,customerNumber
from orders
SELECT COLUMN_NAME FROM INFORMATION_SCHEMA.COLUMNS WHERE
TABLE_NAME='orders'
```

4. Write SQL query to show all the comments from the **Orders**Table.
Solution:

```
COMMENT ON TABLE ORDERS
IS 'Orders Information';
SELECT * FROM USER__TAB_COMMENTS
WHERE TABLE_NAME='orders'
```

5. Write a SQL query to show orderDate and Total number of orders placed on that date, from Order**s**table.
Solution:

```
SELECT orderDate(order_placed_date),
COUNT(order_id) AS num_orders,
SUM(order_total) AS daily_total
FROM orders
WHERE orders_placed_date>=date_sub(current_date,INTERVAL 31 DAY)
GROUP BY date(order_placed_date)
```

6. Write a SQL query to show employeNumber, lastName, firstName of all the employees from **employees** table.
Solution:

```
CREATE TABLE employees
(
EmployeNumber int,
LastName varchar(255),
FirstName varchar(255)
);
```

7. Write a SQL query to show all orderNumber, customerName of the person who placed the respective order.
Solution:

```
SELECT orders.orderNumber,
 Customer.customerName
FROM orders,customers
WHERE orders.customerNumber=customer.customerNumber
```

8. Write a SQL query to show name of all the customers in one column and salerepemployee name in another column.
Solution:

```
   SELECT customer.customerName,
 Salesman.salesrepemployeename
FROM customer,salesman
WHERE salesman.salesrepemployeename=customer. Salesrepemployeename
```

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.
Solution:

```
SELECT Date "Date",
COUNT(*) "Total Payments"
FROM Date
GROUP BY paymentDate;
```

10. Write a SQL query to show all the products productName, MSRP, productDescription from the **products** table.
Solution:

```
CREATE TABLE products
(
productCode int,
productName varchar(255),
producrline int,
productScale varchar(255),
productVendor varchar(255),
productDescription varchar(255),
quantityinStock int,
buyPrice int,
MSRP int,
);
```

11. Write a SQL query to print the productName, productDescription of the most ordered product.
Solution:

```
SELECT products
 p.Name
 FROM products.productCode
 INNER JOIN Production.Product p
 ON sod.Productcode = p.Productcode
 GROUP BY p.Name
 ORDER BY COUNT(*) DESC
```

12. Write a SQL query to print the city name where maximum number of orders were placed.
Solution:

```
SELECT cityName, COUNT(DISTINCT ord_no),
MAX(purch_amt)
FROM orders
GROUP BY cityName
ORDER BY DESC;
```

13. Write a SQL query to get the name of the state having maximum number of customers.
Solution:

```
SELECT customer_id, COUNT(DISTINCT ord_no),
MAX(purch_amt)
FROM orders
GROUP BY customer_id
ORDER BY DESC;
```

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.
Solution:

```
SELECT E1.Enum AS Employees,
E2.EFName AS EmployeesFullName
```

```
FROM Employee E1 JOIN EMP E2
ON E1.EFName=E2.Enum
```

15. Write a SQL query to print the orderNumber, customer Name and total amount paid by the customer for that order (quantityOrdered × priceEach).

Solution:

```
SELECT a.cust_name,a.city, b.ord_no,
b.ord_date,b.purch_amt AS "Order Amount",
c.name,c.commission
FROM customer a
LEFT OUTER JOIN orders b
ON a.customer_id=b.customer_id
LEFT OUTER JOIN salesman c
ON c.salesman_id=b.salesman_id;
```