UNIT1

INTRODUCTION TO DATA SCIENCE

Topic 1: Defining Data Science and Big Data

By :- Shravani Shinde (Branch COMP)

DEFINING DATA SCIENCE MCQS

1. Data science is the process of diverse set of data through ?
A. organizing data
B. processing data
C. analysing data
D. All of the above
Ans : D
Explanation: Data science is the process of deriving knowledge and insights from a huge and diverse set of data through organizing, processing and analysing the data.

2. The modern conception of data science as an independent discipline is sometimes attributed to?
A. William S.
B. John McCarthy
C. Arthur Samuel
D. Satoshi Nakamoto
Ans : A
Explanation: Data science developed by William S.

3. Which of the following is not a part of data science process?
A. Discovery
B. Model Planning
C. Communication Building
D. Operationalize
Ans : C
Explanation: Communication Building is not a part of data science process.

4. Which of the following is not a application for data science?
A. Recommendation Systems
B. Image & Speech Recognition
C. Online Price Comparison
D. Privacy Checker
Ans : D
Explanation: Privacy Checker is not a application for data science

5. Raw data should be processed only one time.
A. True
B. False
C. Can be true or false
D. Can not say

Ans : B
Explanation: Raw data may only need to be processed once.

6. Which of the following step is performed by data scientist after acquiring the data?
A. Data Cleaning
B. Data Integration
C. Data Replication
D. All of the above
Ans : A
Explanation: Data cleaning, data cleaning or data scrubbing is the process of detecting and correcting (or removing) corrupt or inaccurate records from a record set, table, or database.

7. Which of the following is one of the key data science skills?
A. Statistics
B. Machine Learning
C. Data Visualization
D. All of the above
Ans : D
Explanation: Data visualization is the presentation of data in a pictorial or graphical format.

8. Which of the following is the most important language for Data Science?
A.Java
B. Ruby
C. R
D. None of the mentioned
Answer: C
Explanation: R is free software for statistical computing and analysis.

9. Which of the following is characteristic of Processed Data?
A. Data is not ready for analysis
B.All steps should be noted
C.Hard to use for data analysis
D.None of the mentioned
Answer: B
Explanation: Processing includes merging, summarizing and subsetting data.

By :- Srushti Jadhav (Branch COMP)

BIG DATA MCQS

1. According to analysts, for what can traditional IT systems provide a foundation when they're integrated with big data technologies like Hadoop?
a. Big data management and data mining
b. Data warehousing and business intelligence
c. Management of Hadoop clusters
d. Collecting and storing unstructured data
Answer: a
Explanation: Hadoop is the technology/framework which stores and process the big data on large clusters of commodity hardware.

2. What are the main components of Big Data?
a. MapReduce
b. HDFS
c. YARN
d. All of the above
Answer: d
Explanation: All of the above are the main components of Big Data

3. How many V's of Big Data
a. 2
b. 3
c. 4
d. 5
Answer : d
Explanation: Big Data was defined by the "3Vs" but now there are "5Vs" of Big Data which are Volume, Velocity, Variety, Veracity, Value

4. All of the following accurately describe Hadoop, EXCEPT _____
a. Open-source
b. Real-time
c. Java-based
d. Distributed computing approach
Ans : b
Explanation: Apache Hadoop is an open-source software framework for distributed storage and distributed processing of Big Data on clusters of commodity hardware.

5. What are the different features of Big Data Analytics?
a. Open-Source
b. Scalability
c. Data Recovery
d. All the above
Answer: d
Explanation: open source,scalability and data recovery all three are features of big data.

6.The examination of large amounts of data to see what patterns or other useful information can be found is known as
a. Data examination
b. Information analysis
c. Big data analytics
d. Data analysis
Answer : c
Explanation: The examination of large amounts of data to see what patterns or other useful information can be found is known as Big data analytics.

7. _____ has the world's largest Hadoop cluster.
a. Apple
b. Datamatics
c. Facebook
d. None of the above
Answer: c
Explanation: Facebook has many Hadoop clusters.And has a largest cluster of hadoop.

8. Facebook Tackles Big Data With _____ based on Hadoop.
a. Project Prism
b. Prism
c. Project Big
d. Project Data
Answer : a
Explanation: Prism automatically replicates and moves data wherever it's needed across a vast network of computing facilities.

Topic 2: Recognizing the different types of data

By :- Kajol Pawar (branch IT)

1. Which of the following data is put into a formula to produce a commonly? accepted result?
a) Raw
b) Processed
c) Synchronized
d) All of the mentioned
Answer: b
Explanation: Raw data (e.g. Information entered into a database) comes from direct measurements. These are converted to processed data by editing, cleaning or modifying it. Now the processed data can be analysed and formula can be thus be applied on it to get an accepted output. Synchronized data is any form of data that traverses between source and destination system for the purpose of maintaining data consistency and harmony. So, processed data is put into formula to produce acceptable output.

2. Which of the following is another name for raw data?
a) Destination data
b) Eggy data
c) Secondary data
d) Machine learning
Answer: b
Explanation: Raw data is the data obtained from any source i.e. it's a source data. Thus, we cannot call it destination data
Even though raw data may reside in secondary storage, it can't be called secondary data because secondary data is the data that is being aggregated from raw data, and does not contain original data collected from sources like survey, etc.
When raw data is collected, processed, and analysed it is called processed

data. Now, data scientists use this data to train a machine to learn automatically from past data. This is called machine learning which is unlike raw data.
So, the remaining option, eggy data is the correct answer since eggy means uncooked or unprocessed or raw.

3. Which of the following is an example of tidy data?
a) Complicated JSON from facebook API
b) Complicated JSON from twitter API
c) Unformatted excel file
d) All of the mentioned
Answer: d
Explanation: Tidy data is obtained after processing script. It is of the form of data matrix in which rows corresponds to sample individuals and columns to variables. Unformatted excel file is in the table form i.e. Matrix form. So, it is can be example of tidy data.
Java Script Object Notation that converts human readable text to attribute/value pair and array data types. (can be analoged to matrix form with key and value as this is nothing but tidy data, where variables corresponds with columns and data entries with rows)
So, option a and b are examples of tidy data. Thus, option d is correct option.

4. Which of the following is a trait of tidy data?
a) Each variable in one column
b) Each observation in different row
c) Each value must have its own cell.
d) All of the mentioned
Answer: d
Explanation: option a,b,c are the 3 rules that makes a dataset tidy. (The ith observation is placed in the ith row. The jth variable in the jth column. So, 'i*j' individual cells are formed for palcing the corresponding value. Eg. the value of jth variable of i th observation is found at the cell 'i*j'.

5. Which of the following package is used for tidy data?
a) tidyr
b) souryr
c) NumPy
d) All of the mentioned
Answer: a
Explanation: tidyr is used for tidy data with spread and gather functions. Gather takes multiple columns and gathers them into key-value pairs. Sometimes 2 variables are clumped together in one column, separate() allows you to tease them apart.
While NumPy is used for working with arrays.

6. Point out the wrong statement.
a) Tidy datasets are all alike but every messy dataset is messy in its own way.
b) Most statistical datasets are data frames made up of rows and columns.
c) Tidy datasets provide a standardized way to link the structure of a dataset with its semantics.
d) None of the mentioned.
Answer: d
Explanation: the tidy data is structured data with a defined physical layout and its semantics. So. In this tidy datasets these structure (physical layout)is linked with its semantics (by the use of key/value pairs). Statistical data is represented in the form of matrices or tables, i.e. rows and columns. Thus, option d is the answer.

7. Strange binary file generated from machines is an example of tidy data.
a) True
b) False
Answer: b
Explanation: Data sets stored in spreadsheets, such as Microsoft's excel, are tidy datasets. But, the binary files generated from machines i.e. raw data cannot be mapped into key value pair or any table form, so it is not an example of tidy data.

8. Which of the following is the most common problem with messy data?
a) Column headers are values
b) Variables are stored in both rows and columns
c) A single observational unit is stored in multiple tables
d) All of the mentioned
Answer: d
Explanation: real datasets can, and often do, violate the three precepts(
• Each variable in one column
• Each observation in different row
• Each value must have its own cell.
of tidy data in almost every way imaginable. The above option a, b and c completely violates the 3 precepts of tidy data. So, it is called messy datasets.

By :- Ranu Kumari (Branch IT)

9. Data stored already in order is
a. Structured data
b. Unstructured data
c. Both A and B
d. None
Ans: a
Explanation: By Definition of structured data

10. Examples of unstructured data is
a. Videos
b. Images
c. Name and Address of a person
d. A and B
e. All of the above
Ans: d
Explanation: We can't store data in rows and columns database.

11. Point out the correct statement
a. Data has only qualitative value
b. Data has only quantitative value
c. Data has both qualitative and quantitative values
d None of the mentioned
Answer: c
Explanation: Data has both qualitative and quantitative values
Structured data is quantitative while unstructured qualitative
In unstructured data it's difficult to gather, store, and organize in typical databases like Excel and SQL

12. Unstructured data based on
a. Character

b. Binary
c. Both
d. None
Answer: c
Explanation: Because data isn't organised in unstructured data

13. Which type of data is widely used
a. Unstructured data
b. Structured data
c. Both
d. None
Answer: a
Explanation: Unstructured data is widely used like satellite generated images, scientific data or images, social media, images, videos, text documents, PDFs etc.

14. Analysis Methods for structured data is
a. Classification, Regression and data clustering
b. Data Stacking and Data mining
c. Both A and B
d. None
Ans: a
Explanation: Data stacking and data mining is analysis method for unstructured data

15. Specialists to handle data unstructured data are
a. Business Analysts
b. Data Scientists
c. Both
d. None
Ans: b
Explanation: Unstructured data handled by data scientist as they have strong statistical knowledge, ML modelling etc.
While structures data handled by Business Analysts as they have ability to understand the data insights

Topic 3: Gaining insight into Data Science Process

By :- Vrushali Phatale (branch IT)

1)Redundant whitespaces cause error
A. True
B. False
Answer: A
Explanation: Whitespaces remain as the cleaning was not executed properly.
A whitespace in one string can cause mismatch of strings. Eg.- "FR" - "FR "
Some languages have inbuilt functions to remove the whitespaces like Python has strip() function.

2) How many values can dummy variables take?

A. 1
B. 2
C. 3
D. 4
Answer: B

Explanation: Turning variables into dummy is a data transformation that breaks a variable that has multiple classes into multiple variables each having only 2 possible values i.e. 0 (false)or 1(true).
Eg- If observation is made on monday you put 1 there and 0 elsewhere..

3) Data can be stored in _____

A) Databases
B) Datamarts
C) Data warehouses
D) Data lakes
E) All of these
Answer: E
Explanation: All these are data repositories maintained by IT professionals.
The primary goal of a database is data storage, while a data warehouse is designed for reading and analyzing that data.
A datamart is a subset of the data warehouse and geared toward serving a specific business unit.
Data warehouses and datamarts are home to preprocessed data, data lakes contains data in its natural or raw format.

4) Select the techniques to handle missing data
A) Omit the values
B) Set value to NULL
C) Modeling the value
D) Impute a value from an estimated or theoretical distribution
E) All of these
Answer: E
Explanation: These techniques are easy to perform and does not disturb the model.

5) _____ is the first step in data science process
A) Research Goal
B) Data Retrieval
C) Data Preparation
D) None
Answer: A
Explanation: The main purpose of this step is to understand what,why & how of the project.

By :- Tanvi Kathed (Branch IT)

6) A *agile* project model is an alternative to sequential process with iterations.
A) True
B) False
Answer: A
Explanation: This methodology wins more ground in IT so it is adopted by data science community.

7)There are how many steps in Data Science Process?
A) 4
B)6
C)7
D)5
Answer: B
Explanation: There are 6 steps in Data Science Process.1.Setting the research goal. 2.Retrieving Data 3.Data Preparation 4.Data Exploration. 5.Data Modeling. 6.Presentation and automation

8) Data Preparation Process consists of
A)Data Cleaning
B)Data Transformation
C)Combining Data
D)ALL of the Above
Answer:D
Explanation: Data Preparation process consists of 1st Data Cleaning 2nd Data Transformation 3rd Combining data.

9)Data exploration process consists of
A)Simple graphs
B)Merging/Joining datasets
C)Set operators
D)Creating View
Answer:A
Explanation: Data exploration process consists of Simple Graphs,Combined Graphs,Link and Brush, Nongraphical techniques.

10) Data Modelling is a process of
A)Model and variable selection
B)Model execution
C)Model Diagnostic and Model comparison
D) All Ofthe above
Answer: D
Explanation: Data modelling process consists of Model and variable selection,model execution,Model diagnostic and model comparison

Topic 4: Data Science Process: Overview, Different steps

By :- Shraddha Deshmukh (branch IT)

1)In the first step of data science process (setting the research goal) what questions must be kept in mind ?
A)Where,how,what
B)What,how,why
c)How,where
D)For,where
Answer:B
Explanation:The outcome of these questions provide a clear research goal, a good understanding of the context, well-defined deliverables, and a plan of action with a timetable.
is then best placed in a project charter.

2)State True or False :-data lakes contains data in its natural or raw format
A)True
B)False
Answer: A
Explanation: Data present in the data lakes is raw and needs to be refined

3)Data cleansing is a subprocess of the data science process that focuses on ?
A)Ignoring the errors and collecting data
B)data integrated with errors
C) removing errors in your data so your data becomes a true and consistent
D)none of the above
Answer:C
Explanation:By removing the errors we get a proper refined data which is required.

4)In Data cleansing process a good practice is to mediate data errors as _____possible ?
A)Late
B)Early
C)In the middle
D)None
answer:B
Explaination: If the errors are resolved in the early stage it gets easier to perform various operations on the collected data

5)What are the sub-step in data preperation step of data process model ?
A)Data cleaning,Data transformation,combining data.
B)data retrival,data ownership
C)data exploration
D)Retriving data
Answer:A
Explaination: For getting refined data these are the substeps under data preparation that need to be followed.

By :- Swarada Mone (branch ENTC)

6) What will happen if Exploratory Data Analysis is not done ?
A) It will not affect on the model
B) It will produce an inaccurate model
C) You can proceed to the next step
D) It is not mandatory
Answer - B
Explanation- EDA is the selection of feature variables that will be used in model development. Skipping EDA might end up choosing wrong variable .

7) Main step(s) most models consists is/are :
A) Selection of a modeling technique & variables to enter in model
B)Execution of the model
C) Diagnosis and model comparison
D) All of the above
Answer - D
Explanation - Sub-step of building the models

8) After successful analysis of the data and building a well-performing model , _____ is done.
A) Retrieving data
B) Data modeling
C) Data Preparation
D) Presentation of the data
Answer - D
Explanation - Presenting data & automating data analysis is the last process to be done

9)_____ is done using machine learning and statistical techniques to achieve project goals.
A) Setting the research goal
B) Data Modeling
C) Data Preparation
D) Data Exploration
Answer - B
Explanation - Both are used in Model execution, which is a part of Data Modeling

Topic 5: Machine Learning Definition and Relation with Data Science

By Purva Komajpillewar (branch ENTC)

1) what is true about Machine Learning?
A) Machine Learning is that field of computer science
B) ML is a type of artificial intelligence that extract patterns out of raw data by using an algorithm
C) The main focus of ML is to allow computer systems learn from experience without being explicitly programmed
D) All of the above
Answer: D
Explanation: All statement are true about ML

2) Different learning methods does not include?
A) Introduction
B) Analogy
C) Deduction
D) Memorization
Answer: A
Explanation: Different learning methods does not include the introduction

3) which of the factors affect the performance of learner system does not include?
A) Representation scheme used
B) Training scenario
C) Type of feedback
D) Good data structures
Answer: D
Explanation: Factors that affect the performance of learner system does not include good data structures

4) In language understanding the level of knowledge that does not include?
A) phonological
B) Syntactic
C) Empirical
D) Logical
Answer: c
Explanation: In language understanding,the level of knowledge that does not include empirical knowledge

5) A model of language consist of categories which does not include?
A) Language units
B) Role structure of unit
C) system constraints
D) structural units
Answer: D
Explanation : A model of language consist of the categories which does not include structural units.

By :- Rutuja Gathe (branch ENTC)

6)which of the following are one of the important steps to pre-process the text in NLP based projects?
a) Stemming b) Stop word removal c) Object standardization
A) 1&2
B) 1&3
C) 2&3
D) 1,2&3
Answer : D
Explanation : Stemming,stop word removal,object standardization is required to pre process the text in NLP

7) Which of the following is not supervised learning?
A) PCA
B) Decision Tree
C) Linear regression
D) Naive Bayesian
Answer: A
Explanation: PCA is not supervised learning

8) The action 'STACK(A, B)' of a robot arm specify to _____
A)Place block B on Block A
B) Place blocks A, B on the table in that order
C) Place blocks B, A on the table in that order
D) Place block A on block B
Answer: D
Explanation: The action 'STACK(A,B)' of a robot arm specify to Place block A on block B.

9) High entropy means that the partitions in classification are
A) Pure
B) Not pure
C) Useful
D) Useless
Answer: B
Explanation: High entropy means the partitions in classification are not pure

10)When performing regression or classification, which of the following is the correct way to preprocess the data ?
A) Normalize the data , PCA , training
B) PCA , Normalize PCA output , training
C) Normalize the data , PCA , normalize PCA output , training
D) Training , PCA , Normalization"
Answer: A
Explanation: Normalize the data , PCA , training is the correct way to preprocess the data

EXTRA MCQ

By :- Trushna Patil (branch IT)

1. Data that summarize all observations in a category are called _____ data.
a) frequency
b) summarized
c) raw
d) none of the mentioned
Answer: b
Explanation: The summary could be the sum of the observations, the number of occurrences, their mean value, and so on.

2. Which of the following is an example of raw data?
a) original swath files generated from a sonar system
b) initial time-series file of temperature values
c) a real-time GPS-encoded navigation file
d) all of the mentioned
Answer: d
Explanation: Raw data refers to data that have not been changed since acquisition.

3. Point out the correct statement.
a) Primary data is original source of data
b) Secondary data is original source of data
c) Questions are obtained after data processing steps
d) None of the Mentioned

Answer: a
Explanation: Primary data is also referred to as raw data.

4. Which of the following data is put into a formula to produce commonly accepted results?
a) Raw
b) Processed
c) Synchronized
d) All of the Mentioned
Answer: b
Explanation: Raw data came from direct measurements.

UNIT 2

STATISTICS AND PROBABILITY BASICS FOR DATA ANALYSIS

Topic 1:-Describing a Single Set of Data, Correlation

By-Pranjal Patil (IT)

1 . Which of the following implies no relationship with respect to correlation?

Option A . Cor (X, Y) = 1

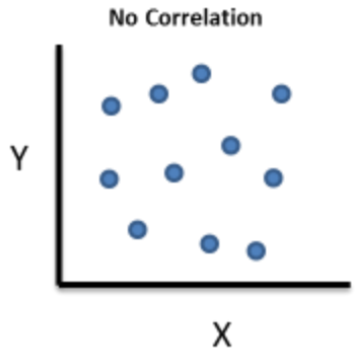
Option B . Cor (X, Y) = 0

Option C . Cor (X, Y) = 2

Option D . All of the mentioned

Answer B

Explanation **Correlation is a statistical technique that can show whether and how strongly pairs of variables are related. Correlation Coefficient values can range from +1 to -1, where +1 indicates a perfect positive relationship, -1 indicates a perfect negative relationship, and a 0 indicates no relationship exists**



2 . Identify the false claim about correlations.

Option A If two variables are correlated, this necessarily means that variation in one cause variation in the other.

Option B If two variables are causally related, they will be correlated.

Option C Sometimes correlations are products of some other, unobserved, factor.

Option D Correlations between variables can be either negative or positive.

Answer **Option A**

Explanation **If two variables are correlated, this necessarily means that variation in one cause variation in the other.**

3 . What do we call it when two variables accompany one another and move in the same direction?

Option A Positive correlation

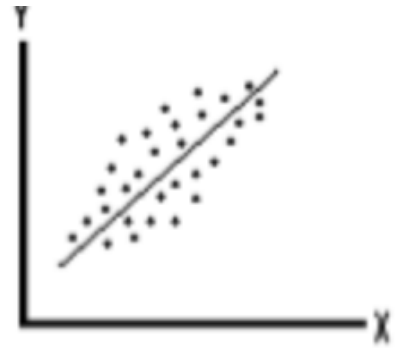
Option B Positive causation

Option C Negative correlation

Option D Negative Causation

Answer **Option A**

Explanation **A positive correlation means that the variables move in the same direction. Put another way, it means that as one variable increases so does the other, and conversely, when one variable decreases so does the other. Examples of positive correlations are the relationship between an individual's height and weight or the relationship between a person's age and number of wrinkles.**



Positive Correlation

4 . **Which of the following is most likely to be an example of a correlation where an omitted variable is responsible for the observed correlation?**

Option A A correlation between ice cream consumption and murders

Option B A correlation between butter production and the performance of the stock market

Option C A correlation between economic development and democracy

Option D This never happens, by definition

Answer **Option A**

Explanation **A correlation between ice cream consumption and murders.**

It possible that indulging in your favorite flavor of ice cream could send you on a crime spree? Or, after committing crime do you think you might decide to treat yourself to a cone? There is no question that a relationship exists between ice cream and crime It is much more likely that both ice cream sales and crime rates are related to the temperature outside. It is much more likely that both ice cream sales and crime rates are related to the temperature outside. When the temperature is warm, there are lots of people out of their houses, interacting with each other, getting annoyed with one another, and sometimes committing crimes. Also, when it is warm outside, we are more likely to seek a cool treat like ice cream. The example of ice cream and crime rates is a positive correlation because both variables increase when temperatures are warmer.

5 . **Two variables are said to be positively correlated when with _____ in the value of one variable, the value of other variable also _____**

Option A Fall, Rises

Option B Fall, falls

Option C No change, Rises

Option D Rise, Falls

Answer **Option B**

Explanation **Positive correlation implies that both the variables will move in the same direction, either both will rise or both will fall. Meditation increases level of concentration.**

6 . **If with the fall in the value of one variable the value of another variable rises in the same proportion then it is said to be**

Option A Negatively correlated

Option B Positively correlated

Option C None Of the above

Option D Both

Answer **Option A**

Explanation **Because both the variables are moving in different directions.**
Example of negative correlation to exist between someone's tiredness during the day and the number of hours they slept the previous night: the amount of sleep decreases as the feelings of tiredness increase. A student who has many absences has a decrease in grades.

7 . **Correlation coefficient is denoted by**

Option A Co

Option B R

Option C 1

Option D C

Answer **Option B**

Explanation **"r" denotes Karl Pearson's coefficient of correlation The Karl Pearson's product-moment correlation coefficient (or simply, the Pearson's correlation coefficient) is a measure of the strength of a linear association between two variables and is denoted by r or rxy(x and y being the two variables involved)**

Topic: Simpson's Paradox

By-Mansi Khopade (Comp)

1 . Simpson's Paradox occurs when

Option A No baseline risk is given, so it is (not know whether or not a high relative risk has practical importance

Option B A confounding variable rather than the explanatory variable is responsible for a change in the response variable

Option C The direction of the relationship between two variables changes when the categories of a confounding variable are taken into account

Option D The results of a test are statistically significant but are really due to chance.

Answer **C**

Explanation **Simpson's paradox occurs when some groups of data show a certain relationship in each group, but when the data is combined, that relationship is reversed**

2 . In Simpson's paradox when data of some groups is combined the relationship is _____

Option A Reversed

Option B Remains same

Option C Irrational

Option D Illogical

Answer **A**

Explanation **Simpson's paradox occurs when some groups of data show a certain relationship in each group, but when the data is combined, that relationship is reversed**

3 . Simpson's paradox can also be illustrated using the _____ dimensional vector space.

Option A 1

Option B 2

Option C 3

Option D 4

Answer **B**

Explanation **Since the Simpson's paradox can be represented in the form of a vector on the plane with coordinates and a slope i.e vector A = (p, q) with slope p/q.**

4 . It is also referred to as

Option A Yule–Simpson effect

Option B amalgamation paradox

Option C reversal paradox

Option D All of these

Answer **D**

Explanation **It is also referred to as Simpson's reversal, Yule–Simpson effect, amalgamation paradox, or reversal paradox.**

5 . The name Simpson's paradox was introduced by Colin R. Blyth in _____ year

Option A 1970

Option B 1972

Option C 1976

Option D 1969

Answer **B**

Explanation **No explanation**

Topic: Some Other Correlational Caveats

By– Komal Jha (IT)

1 . If the scatter diagram is drawn the scatter points lie on a straight line, then it indicates

Option A Skewness

Option B perfect correlation

Option C No correlation

Option D None of the above

Answer **B**

Explanation **scatter points if projected by a linear graph are perfect correlation**

2. If two variables oppose each other than the correlation will be

Option A Positive Correlation

Option B Zero correlation

Option C Perfect correlation

Option D Negative correlation

Answer **D**

Explanation **When the coefficient of correlation (r) is less than 0, it is negative. When r is -1.0, there is a perfect negative correlation**

3 . In correlation, both variables are always

Option A Random

Option B Non- Random

Option C Same

Option D None

Answer **A**

Explanation **With correlation, the X and Y variables are interchangeable. With correlation, X and Y are typically both random variables, such as height and weight or blood pressure and heart rate**

4 . Following are the elements of correlation except

Option A There should be three or more variables

Option B There should be only curvilinear relations among variables

Option C The change in the value of one affect another

Option D There should be relationship among them

Answer **B**

Explanation **It is not necessary that there must be a curvilinear relation among variables,it can be linear or non linear correlation**

5 . When the correlation is only studied between two variables it is called

Option A Simple Correlation

Option B Positive correlation

Option C Multiple Correlation

Option D Negative correlation

Answer **D**

Explanation **Because it is very easy and simple to calculate correlation between two variables**

Topic: Correlation & Causation

By- Janhavi Majge (IT)

1 . The goal of correlational research is to:

Option A assess the causal impact of one variable on another

Option B Assess busy relationships

Option C assess relationships between variables

Option D all of the above

Answer **C**

Explanation **correlation deals with the variance of two variables with respect to each other**

2 . If income and happiness are positively correlated, then a person with a low income would be predicted to be

Option A not depressed at all

Option B less depressed than a person with a high income

Option C more depressed than a person with a high income

Option D cannot make a prediction from correlational data

Answer **C**

Explanation **since both happiness and income are said to be correlated to each other they are directly proportional. hence the answer**

3 . A strength of correlational designs is that they:

Option A can demonstrate causation

Option B do not require ethics board approval

Option C can be used with variables which cannot be manipulated

Option D by a researcher are more intrusive than experimental designs

Answer **D**

Explanation **No Explanation**

4 . Correlation coefficients range from:

Option A r = -1 to r = +1

Option B r = 0 to r = +1

Option C r = -1 to r = 0

Option D r = +.5 to r = +1

Answer **A**

Explanation

5 . Research shows that the older a person is, the larger their vocabulary. This is an example of a:

Option A positive correlation

Option B negative correlation

Option C causal correlation

Option D partial correlation

Answer **A**

Explanation **the vocabulary significantly increases with the age hence a positive correlation**

Topic: Probability Dependence & Independence

By-Prachi Kaladeep (IT)

1. A roulette wheel has 38 slots – 18 red, 18 black, and 2 green. You play five games and always bet on red slots. How many games can you expect to win?

Option A	1.1165
Option B	2.3684
Option C	2.6316
Option D	4.7368

Answer **(B)**

Explanation **The probability that it would be Red in any spin is 18/38. Now, you are playing the game 5 times and all the games are independent of each other. Thus, the number of games that you can win would be 5*(18/38) = 2.3684**

2 . A jar contains 4 marbles. 3 Red & 1 white. Two marbles are drawn with replacement after each draw. What is the probability that the same colour marble is drawn twice?

Option A	7/8
Option B	9/6
Option C	5/8
Option D	9/8

Answer **(C)**

Explanation **If the marbles are of the same colour, then it will be 3/4 * 3/4 + 1/4 * 1/4 = 5/8.**

3 . The expected value or _____ of a random variable is the centre of its distribution.

Option A	Mode
Option B	Median
Option C	Mean
Option D	Bayesian inference

Answer **(C)**

Explanation **A probability model connects the data to the population using assumptions.**

4 . In a class of 30 students, approximately what is the probability that two of the students have their birthday on the same day (defined by same day and month) (assuming it's not a leap

year)? For example – Students with birthday 3rd Jan 1993 and 3rd Jan 1994 would be a favourable event.

Option A 0.49

Option B 0.52

Option C 0.696

Option D 0.35

Answer **(C)**

Explanation **The total number of combinations possible for no two persons to have the same birthday in a class of 30 is 30 * (30-1)/2 = 435. Now, there are 365 days in a year (assuming it's not a leap year). Thus, the probability of people having a different birthday would be 364/365. Now there are 870 combinations possible. Thus, the probability that no two people have the same birthday is (364/365) ^435 = 0.303. Thus, the probability that two people would have their birthdays on the same date would be 1 – 0.303 = 0.696**

5 . **A coin of diameter 1-inches is thrown on a table covered with a grid of lines each two inches apart. What is the probability that the coin lands inside a square without touching any of the lines of the grid? You can assume that the person throwing has no skill in throwing the coin and is throwing it randomly. You can assume that the person throwing has no skill in throwing the coin and is throwing it randomly.**

Option A 3/5

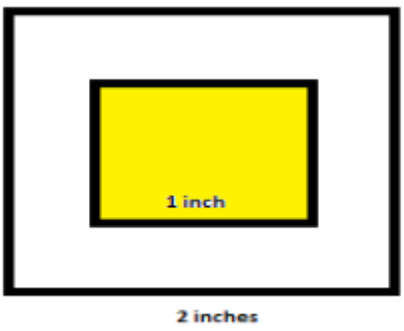
Option B 1/4

Option C 2/4

Option D 8/4

Answer **(B)**

Explanation **Think about where all the centre of the coin can be when it lands on 2 inches grid and it not touching the lines of the grid. If the yellow region is a 1-inch square and the outside square is of 2 inches. If the centre falls in the yellow region, the coin will not touch the grid line. Since the total area is 4 and the area of the yellow region is 1, the probability is ¼.**



6 . **There are a total of 8 bows of 2 each of green, yellow, orange & red. In how many ways can you select 1 bow?**

Option A 1

Option B 2

Option C 4

Option D 8

Answer **(C)**

Explanation **You can select one bow out of four different bows, so you can select one bow in four different ways.**

Topic: Conditional Probability

By Madhura Mirikar (ETC)

1 . Let E and F be events of a sample space S of an experiment, if P(S|F) = P(F|F), then value of P(F|F) is _____

Option A 0

Option B -1

Option C 1

Option D 2

Answer **Option C**

Explanation **We know that P(S|F) = P(S∩F) / P(F). (By formula for conditional probability) Which is equivalent to P(F|F) = P(F) / P(F) = 1, hence the value of P(F|F) = 1.**

2 . If E and F are two events associated with the same sample space of a random experiment, then P (E|F) is given by _____

Option A P(E∩F) / P(F), provided P(F) ≠ 0

Option B P(E∩F) / P(F), provided P(F) = 0

Option C P(E∩F) / P(F)

Option D P(E∩F) / P(E)

Answer **Option A**

Explanation **E and F are two events associated with the same sample space of a random experiment. The value of P (E|F) = (E∩F) / P(F), provided P(F) ≠ 0. We know that if P(F) = 0, then the value of P(E|F) will reach a value which is not defined hence it is wrong option. Also, P(E∩F) / P(F) and P(E∩F) / P(E) are wrong and do not equate to P(E|F).**

3 . Given that E and F are events such that P(E) = 0.6, P(F) = 0.3 and P(E∩F) = 0.2, then P(E|F)?

Option A 2/3

Option B 1/3

Option C ¾

Option D ¼

Answer **Option A**

Explanation **We know that P(E|F) = P(E∩F) / P(F). (By formula for conditional probability)**
Value of P(E∩F) is given to be 0.2 and value of P(F) is given to be 0.3.
P(E|F) = (0.2) / (0.3).
P(E|F) = 2 / 3.

4 . If P(A) = 5/13, P(B) = 7/13 and P(A∩B) = 3/13, evaluate P(A|B)

Option A 1/7

Option B 3/7

Option C 3/5

Option D 2/7

Answer **Option B**

Explanation **We know that P(A|B) = P(A∩B) / P(B). (By formula for conditional probability)**
Which is equivalent to (3/13) / (7/13), hence the value of P(A|B) = 3/7.

5 . If P(A) = 1/5, P(B) = 0, then what will be the value of P(A|B)?

Option A 0

Option B 1

Option C Not Defined

Option D 1/5

Answer Option C

Explanation We know that P(A|B) = P(A∩B) / P(B). (By formula for conditional probability)
The value of P(B) = 0 in the given question. As the value of denominator becomes 0, the value of P(A|B) becomes un-defined.

Topic: Bayles's Theorem, & Random Variables

By-Komal Singh (ETC)

1 . **Naina receives emails that consists of 18% spam of those emails. The spam filter is 93% reliable i.e., 93% of the mails it marks as spam are actually a spam and 93% of spam mails are correctly labelled as spam. If a mail marked spam by her spam filter, determine the probability that it is really spam.**

Option A 50%

Option B 84%

Option C 39%

Option D 63%

Answer A

Explanation 18% email are spam, and 82% email are not spam. Now, 18% of mail marked as spam is spam and 82% mail marked as spam are not spam. By Bayes theorem the probability that a mail marked spam is really a spam = (Probability of being spam and being detected as spam)/ (Probability of being detected as spam) = (0.18 * 0.82)/ (0.18 * 0.82) + (0.18 * 0.82) = 0.5 or 50%.

2 . **A jar containing 8 marbles of which 4 red and 4 blue marbles are there. Find the probability of getting a red given the first one was red too.**

Option A a) 4/13

Option B b) 2/11

Option C c) 3/7

Option D d) 8/15

Answer C

Explanation **Explanation: Suppose, P (A) = getting a red marble in the first turn, P (B) = getting a black marble in the second turn. P (A) = 4/8 and P (B) = 3/7 and P (A and B) = 4/8∗3/7=3/14 P(B/A) = P(AandB)P(A)= (3/14)/ (1/2) =3/7.**

3 . **A bin contains 4 red and 6 blue balls, and three balls are drawn at random. Find the probability such that both are of the same colour.**

Option A a) 10/28

Option B b) 1/5

Option C c) 1/10

Option D d) 4/7

Answer **B**

Explanation **Total no of balls = 10. Number of ways drawing 3 balls at random out of 10 = 10C3 = 120. Probability of drawing 3 balls of same colour = 4C3 + 6C3 = 24. Hence, the required probability is 24/120=1/5**

4 . **A football player makes 75% of his 5-point shots and 25% his 7-point shots. Determine the expected value for a 7-point shot of the player**

Option A a) 4.59

Option B b) 12.35

Option C c) 5.25

Option D d) 42.8

Answer **C**

Explanation **Multiply the outcome by its probability, so the expected value becomes 0.75 * 7 points = 5.25.**

5 . **A Random Variable X can take only two values, 4 and 5 such that P (4) = 0.32 and P (5) = 0.47. Determine the Variance of X.**

Option A a) 8.21

Option B b) 12

Option C c) 3.7

Option D d) 4.8

Answer **C**

Explanation **Expected Value: μ = E(X) = ∑x * P(x) = 4 × 0.32 + 5 × 0.47 = 3.63. Next find ∑x2 * P(x): ∑x2 * P(x) = 16 × 0.32 + 25 × 0.47 = 16.87. Therefore, Var(X) = ∑x2P(x) − μ2 = 16.87 − 13.17 = 3.7.**

Option A a) 3.8

Option B b) 2.9

Option C c) 4.78

Option D d) 5.32

Answer **C**

Explanation **We know that E(X) = ∑ x*P(x) = 2 × 0.45 + 4 × 0.97 = 4.78, where x= {2,4}.**

6 . A jar of pickle is picked at random using a filling process in which an automatic machine is filling pickle jars with 2.5 kg of pickle in each jar. Due to few faults in the automatic process, the weight of a jar could vary from jar to jar in the range 1.7 kg to 2.9 kg excluding the latter. Let X denote the weight of a jar of pickle selected. Find the range of X.

Option A a) $3.7 \leq X < 3.9$

Option B b) $1.6 \leq X < 3.2$

Option C c) $1.7 \leq X < 2.9$

Option D d) $1 \leq X < 5$

Answer **C**

Explanation **Possible outcomes should be $1.7 \leq X < 2.9$. That is the probable range of X for the answer.**

Topic: Continuous Distributions - Anupriya

1 . Which of these cannot be shown on the continuous distributions?

Option A Length dimension measurement of a box

Option B Volume measurement of the box

Option C Area measurement of the box

Option D Number of defects on the surface of the box

Answer **D**

Explanation **Continuous distributions are used to describe the variation in the values of variables which are continuous, i.e., which take values on continuous scale. Number of defects is discrete not continuous parameter.**

2 . **Which of these is a continuous distribution?**

Option A Pascal distribution

Option B Lognormal distribution

Option C Binomial distribution

Option D Hyper geometric distribution

Answer **B**

Explanation **Pascal, binomial, hyper geometric distributions are all part of discrete distribution which are used to describe variation of attributes. Lognormal distribution is a continuous distribution used to describe variation of the continuous variable.**

3 . **Which of these distributions has an appearance of bell-shaped or unimodal curve?**

Option A Lognormal distribution

Option B Normal distribution

Option C Exponential distribution

Option D Cumulative exponential distributions

Answer **B**

Explanation **Out of all continuous distributions, Normal distributions are the only distributions, which have a shape of curve as a bell. The curves of them are mostly unimodal**

4 . **Normal distribution is applied for**

Option A Continuous random distribution

Option B Discrete random variable

Option C Irregular random variable

Option D Uncertain random variable

Answer **A**

Explanation **This is the rule on which Normal distribution is defined.**

5 . **Which of the following is true about continuous distribution**

Option A A probability distribution in which the random variable X can take on any value

Option B The sum of all the probability is equal to 1

Option C Both A and B

Option D None of these

Answer **A**

Explanation **A continuous distribution has a range of values that are infinite, and X can take infinite value. Therefore, it is true about continuous distribution that X can take any value.**

6 . **Which of these are types of continuous distribution?**

Option A Triangular distribution

Option B Lognormal distribution

Option C Beta Distribution

Option D All of these

Answer **D**

Explanation **Three-point estimates are displayed using triangular and beta. standard deviations are displayed using lognormal distribution.**

Topic: The Normal Distribution

By – Sudhisha Zare (IT)

1 . **Normal Distribution is applied for _____**

Option A Continuous Random Distribution

Option B Discrete Random Variable

Option C Irregular Random Variable

Option D Uncertain Random Variable

Answer **A**

Explanation **This is the rule on which Normal distribution is defined, no details on this as of why for more knowledge on this aspect, you can refer to any book or website which speaks on the same.**

2 . **The shape of the Normal Curve is _____**

Option A Bell Shaped

Option B Flat

Option C Circular

Option D Spiked

Answer **A**

Explanation **Due to the nature of the Probability Mass function, a bell-shaped curve is obtained.**

3 . **For a standard normal variate, the value of mean is?**

Option A ∞

Option B 1

Option C 0

Option D not defined

Answer **C**

Explanation **For a normal variate, if its mean = 0 and standard deviation = 1, then it's called as Standard Normal Variate. Here, the converse is asked.**

4 . **The area under a standard normal curve is?**

Option A 0

Option B 1

Option C ∞

Option D not defined

Answer **B**

Explanation

For any probability distribution, the sum of all probabilities is 1. Area under normal curve refers to sum of all probabilities.

5 . **Normal Distribution is also known as _____**

Option A Cauchy's Distribution

Option B Laplacian Distribution

Option C Gaussian Distribution

Option D Lagrangian Distribution

Answer **C**

Explanation **Named after the one who proposed it. For further details, refer to books or internet.**

6 . **In Standard normal distribution, the value of mode is _____**

Option A 2

Option B 1

Option C 0

Option D Not fixed

Answer **C**

Explanation **In a standard normal distribution, the value of mean is 0 and in normal distribution mean and mode coincide.**

7 . **The shape of the normal curve depends on its _____**

Option A Mean deviation

Option B Standard deviation

Option C Quartile deviation

Option D Correlation

Answer **B**

Explanation **This can be seen in the pdf of normal distribution where standard deviation is a variable.**

8 . The value of constant 'e' appearing in normal distribution is _____

Option A 2.5185

Option B 2.7836

Option C 2.1783

Option D 2.7183

Answer **D**

Explanation **This is a standard constant.**

Topic: The Central Limit Theorem

By – Shreya Jamsandekar (Comp).

1 . The Central Limit Theorem says that the sampling distribution of the sample mean is approximately normal if

Option A All possible samples are selected

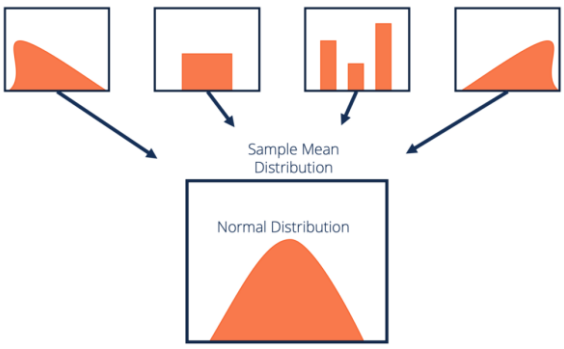
Option B Sample size is large

Option C Standard error of the sampling distribution is small

Option D None of the above

Answer **C**

Explanation **As per the definition of Central limit theorem**



2 . What does the central limit theorem state?

Option A if the sample size increases sampling distribution must approach normal distribution

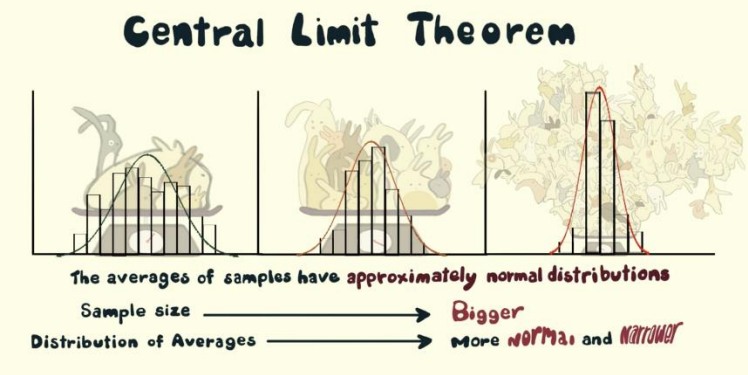
Option B if the sample size decreases then the sample distribution must approach normal distribution

Option C if the sample size increases then the sampling distribution much approach an exponential distribution

Option D if the sample size decreases then the sampling distribution much approach an exponential distribution

Answer **A**

Explanation **The central limit theorem states that if the sample size increases sampling distribution must approach normal distribution. Generally, a sample size more than 30 us considered as large enough.**



3 . The difference between the sample value expected and the estimates value of the parameter is called as?

Option A Bias

Option B Error

Option C Contradiction

Option D Difference

Answer **A**

Explanation **The difference between the expected sample value and the estimated value of parameter is called as bias. A sample used to estimate a parameter is unbiased if the mean of its sampling distribution is exactly equal to the true value of the parameter being estimated.**

4 . **In which of the following types of sampling the information is carried out under the opinion of an expert?**

Option A quota sampling

Option B convenience sampling

Option C purposive sampling

Option D judgement sampling

Answer **D**

Explanation **In judgement sampling is carried under an opinion of an expert. The judgement sampling often results in a bias because of the variance in the expert opinion.**

5 . **Which of the following is a subset of population?**

Option A Distribution

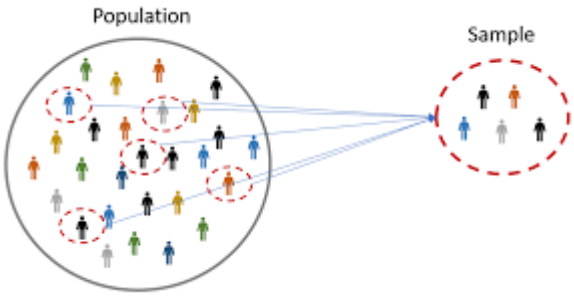
Option B Sample

Option C Set

Option D Data

Answer **B**

Explanation **In sampling distribution, we take a subset of population which is called as a sample. The main advantage of this sample is to reduce the variability present in the statistics.**



6 . **The sampling error is defined as?**

Option A difference between population and parameter

Option B difference between sample and parameter

Option C difference between population and sample

Option D difference between parameter and sample

Answer **C**

Explanation **In sampling distribution, the sampling error is defined as the difference between population and the sample. Sampling error can be reduced by increasing the sample size.**

7 . Any population which we want to study is referred as?

Option A standard population

Option B final population

Option C infinite population

Option D target population

Answer **D**

Explanation **In sampling distribution, we take a part of a population under study which is called as target population. Target population is also called as a sample.**

Done By-Group 4

Team Leader- Shreya Jamsandekar

Team Members-

1. **Manasi Khopade - COMP**
2. **Komal Jha - IT**
3. **Janhavi Majge - IT**
4. **Prachi Kaladeep-IT**
5. **Sudhisha Zare-IT**
6. **Pranjal Patil - IT**
7. **Madhura Mirikar - ENTC**
8. **Komal Singh - ENTC**
9. **Anupriya kumari-ENTC**

UNIT - 3

Data Analysis in Depth

1 . DATA ANALYSIS THEORY AND METHODS (general introduction)

By :- Pranali Jamdade (branch IT)

1) Which of the following is good way of performing experiments in data science ?

a) Measure variability

b) Generalize to the problem

c) Have Replication

d) All of the Mentioned

Answer: Option D

Explanation:

Experiments on causal relationships investigate the effect of one or more variables on one or more outcome variables.

2) Which of the following approach should be used to ask Data Analysis question ?

a) Find only one solution for particular problem

b) Find out the question which is to be answered

c) Find out answer from dataset without asking question

d) None of the mentioned

Answer: Option B

Explanation:

Data analysis has multiple facets and approaches.

3) Which of the following step is performed by data scientist after acquiring the data ?

a) Data Cleansing

b) Data Integration

c) Data Replication

d) All of the Mentioned

Answer: Option A

Explanation:

Data cleansing, data cleaning or data scrubbing is the process of detecting and correcting (or removing) corrupt or inaccurate records from a record set, table, or database.

4) Which of the following systems record changes to a file over time ?

a) Record Control

b) Version Control

c) Forecast Control

d) None of the mentioned

Answer: Option B

Explanation:

Version control is also known as revision control.

5) Which of the following command allows you to update the repository ?

a) push

b) pop

c) update

d) None of the Mentioned

Answer: Option A

Explanation:

The git branch command is your general-purpose branch administration tool.

6) Which of the following technique comes under practical machine learning ?

a) Bagging

b) Boosting

c) Forecasting

d) None of the Mentioned

Answer: Option B

Explanation:

Boosting is an approach to machine learning based on the idea of creating a highly accurate predictor

7) Which of the following uses data on some object to predict values for other object ?

a) Inferential

b) Exploratory

c) Predictive

d) None of the Mentioned

Answer: Option C

Explanation:

A prediction is a forecast, but not only about the weather.

8) Which of the following relationship are usually identified as average effects ?

a) Descriptive
b) Causal
c) Predictive
d) None of the Mentioned

Answer: Option B

Explanation:

A correlation is a measure or degree of relationship between two variables.

9) Which of the following is a revision control system ?

a) Git
b) NumPy
c) Slidify
d) None of the mentioned

Answer: Option A

Explanation:

Git is a free and open source distributed version control system designed to handle everything from small to very large projects with speed and efficiency.

2 . CLUSTERING OVERVIEW AND INTRODUCTION

By :- Swamini Sontakke (branch IT)

1. The goal of clustering is to-

a) Divide the data points into groups
b) Classify the data points in different classes
c) Predict the output values of input data points
d) All of the above

Answer : A

Explanation : Clustering is the task of dividing data points into a number of groups such that data points in the same groups are similar, i. e. to seggregate groups with similar traits and forming a cluster

2. Clustering is

a) Supervised Learning
b) Unsupervised Learning
c) Reinforcement Learning
d) None

Answer : B

Explanation :Clustering is an example of Unsupervised Learning where we draw reference from datasets having inputs which are not labelles.

3. Which of the following is a bad characteristic of a dataset for clustering analysis-

a) Data Points with Outliers
b) Data point with different densities
c) Data points with non-convex shape
d) All of the above

Answer : D

Explanation :All above are bad characteristics of datasets for clustering analysis.

4. For Clustering we dont require

a) Labeled Data
b) Unlabeled Data
c) Numerical Data
d) Categorical Data

Answer : A

Explanation :Clustering is an example of Unsupervised learning and hence uses unlabelled data. So labelled data is not required.

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

a) Biological Network Analysis
b) Market Trend Prediction
c) Topic Modeling
d) All of the above

Answer : D

Explanation :All above are applications of Clustering.

6. On which data type, we can not perform cluster analysis?

a) Time Series Data
b) Text Data
c) Multimedia Data
d) None

Answer : D

Explanation :Clustering can be performed on interval ordinal and categorical data.

7. Indicate which is/are a method of clustering

a) linkage method
b) split and merge
c) both a and b
d) neither a nor b

Answer : C

Explaination : Both Linkage Method and Split and Merge are methods of clustering.

8. K means and K-medioids are example of which type of clustering method?

a) Hierarchical

b) partition

c) probabilistic

d) None of the above.

Answer : B

Explaination :Both K-means and K-medioida are example of Partitioning clustering algorithm.

3. K-MEANS CLUSTERING ALGORITHM

By :- Yutika Vora (branch IT)

1. Which of the following is required by K-means clustering?

a) defined distance metric

b) number of clusters

c) initial guess as to cluster centroids

d) None of the mentioned

Answer: D

Explanation: K-means clustering follows partitioning approach.

2. Point out the wrong statement.

a) k-means clustering is a method of vector quantization

b) k-means clustering aims to partition n observations into k clusters

c) k-nearest neighbor is same as k-means

d) none of the mentioned

Answer: C

Explanation: k-nearest neighbor has nothing to do with k-means.

3. The k-means algorithm...

a) always converges to a clustering that minimizes the mean-square vector-representative distance

b) can converge to different final clustering, depending on initial choice of representatives

c) is typically done by hand, using paper and pencil

d) should only be attempted by trained professionals

Answer : B

Explaination : On data that does have a clustering structure, the number of iterations until convergence is often small, and results only improve slightly after the first dozen

4. The choice of k, the number of clusters to partition a set of data into,...

a) is a personal choice that shouldn't be discussed in public

b) depends on why you are clustering the data

c) should always be as large as your computer system can handle

d) has maximum 10

Answer : B

Explaination : The number of clusters to be formed always depends on the Dataset and why the clusters are to be formed and no other factor is considered.

4. Which of the following statements about the K-means algorithm are correct?

a) The K-means algorithm is not sensitive to outliers.

b) For different initializations, the K-means algorithm will definitely give the same clustering results.

c) The centroids in the K-means algorithm may not be any observed data points.

d) The K-means algorithm can detect non-convex clusters.

Answer : C

Explaination : A centroid is a data point that represents the center of the cluster (the mean), and it might not necessarily be a member of the dataset.

5. K means and K-medioids are example of which type of clustering method?

a) Hierarchical

b) partition

c) probabilistic

d) None of the above.

Answer: B

Explaination : Both the k-means and k-medoids algorithms are partitional (breaking the dataset up into groups). K-means attempts to minimize the total squared error, while k-medoids minimizes the sum of dissimilarities between points labeled to be in a cluster and a point designated as the center of that cluster. In contrast to the k-means algorithm, k-medoids chooses datapoints as centers (medoids or exemplars).

6. K Means is which type of learning algorithm?

a) Supervised

b) Unsupervised

c) Reinforcement

d) None of the above

Answer : B

Explaination : K-means clustering is a type of unsupervised learning, which is used when you have unlabeled data (i.e., data without defined categories or groups).

7. Which of the following algorithm has similarity with K Means?

a) Support Vector Machine

b) Logistic Regression

c) K-NN

d) Linear Regression

Answer : C

Explaination : Both KNN and Kmeans are clustering algorithms

8.K-means algorithm can be used for which of the following?

a) Cluster analysis

b) Feature learning

c) Both of the above

d) None of the above

Answer : C

Explaination : We use Kmeans to form clusters i.e grouping the data on basis of categories, which further helps in future learning.

9. Which of the following forms key step of K-Means clustering algorithm?

a) Assigning data to one of the clusters

b) Recompute the centroid

c) Both of the above

d) None of the above

Answer : C

Explaination : The algorithm works in a way where data is assigned to a particular cluster and then further the centroid is recomputed.

10. K-Means squared error function is related with which of the following?

a) Manhattan distance

b) Hamming distance

c) Euclidean distance

d) Minkowski distance

Answer : C

Explaination : k-means clustering minimizes within-cluster variances (squared Euclidean distances), but not regular Euclidean distances, which would be the more difficult problem: the mean optimizes squared errors, whereas only the geometric median minimizes Euclidean distances.

4 . DETERMINING OF CLUSTERS

By:- Krupali R. Bhoir (Branch IT)

1. Which of the following method is used for finding optimal of cluster in K-Mean algorithm?

a) Elbow method

b) Manhattan method

c) Ecludian mehthod

d) All of the above

Answer : A

Explanation:only elbow method is used for finding the optimal number of clusters. The elbow method looks at the percentage of variance explained as a function of the number of clusters: One should choose a number of clusters so that adding another cluster doesn't give much better modeling of the data.

2. Hierarchical clustering should be primarily used for exploration.

a) True

b) False

c) Both

d) None

Answer: A

Explaination : Hierarchical clustering is deterministic.

3. K-means is not deterministic and it also consists of number of iterations.

a) True

b) False

c) Both

d) None

Answer: A

Explaination : K-means clustering produces the final estimate of cluster centroids.

4. In which of the following cases will K-Means clustering fail to give good results?

1. Data points with outliers
2. Data points with different densities
3. Data points with round shapes
4. Data points with non-convex shapes

a) 1 and 2

b) 2 and 3

c) 2 and 4

d) 1, 2 and 4

Answer: D

Explaination : K-Means clustering algorithm fails to give good results when the data contains outliers, the density spread of data points across the data space is different and the data points follow non-convex shapes.

5. What is true about K-Mean Clustering?

1. K-means is extremely sensitive to cluster center initializations
2. Bad initialization can lead to Poor convergence speed
3. Bad initialization can lead to bad overall clustering

a) 1 and 3

b) 2 and 3

c) 1 and 2

d) 1, 2 and 3

Answer: D

Explaination : All three of the given statements are true. K-means is extremely sensitive to cluster center initialization. Also, bad initialization can lead to Poor convergence speed as well as bad overall clustering.

6. Which of the following are true?

1. Clustering analysis is negatively affected by multicollinearity of features
2. Clustering analysis is negatively affected by heteroscedasticity

a) 1 only

b) 2 only

c) 1 and 2

d) None

Answer: A

Explaination : Clustering analysis is not negatively affected by heteroscedasticity but the results are negatively impacted by multicollinearity of features/ variables used in clustering as the correlated feature/ variable will carry extra weight on the distance calculation than desired.

7. What is the minimum no. of variables/ features required to perform clustering?

a) 0

b) 1

c) 2

d) 3

Answer: B

Explaination :At least a single variable is required to perform clustering analysis. Clustering analysis with a single variable can be visualized with the help of a histogram.

8. Which of the following algorithm is most sensitive to outliers?

a) K-means clustering algorithm

b) K-medians clustering algorithm

c) K-modes clustering algorithm

d) K-medoids clustering algorithm

Answer: A

Explaination : Out of all the options, K-Means clustering algorithm is most sensitive to outliers as it uses the mean of cluster data points to find the cluster center.

9. Point out the correct statement.

a) The choice of an appropriate metric will influence the shape of the clusters

b) Hierarchical clustering is also called HCA

c) the merges and splits are determined in a greedy manner

d) All of the above

Answer: D

Explaination : Some elements may be close to one another according to one distance and farther away according to another.

10. Which of the following clustering requires merging approach?

a) Partitional

b) Hierarchical

c) Naive Bayes

d) None of the above

Answer: B

Explaination : Hierarchical clustering requires a defined distance as well.

5 . ASSOCIATION RULES GENERAL INFORMATION

By: RajeshwariChillarge(Branch IT)

1. Association rules provide information in the form of "if-then" statements.

a) True

b) False

c) Both

d) None

Answer : A

Explanation: Association rules provide information of this type in the form of “if-then” statements. These rules are computed from the data and, unlike the if-then rules of logic, association rules are probabilistic in nature.

2. What do you mean by support(A)?

a)Total number of transactions containing A

b)Total number of transactions not containing A

c)Number of transactions containing A / Total number of transactions

d)Number of transactions not containing A / Total number of transactions

Answer : C

Explaination: Support(A)=Number of transactions in which A appears/Total number of transactions

3. Which of the following is/are interestingness measures for association rules?

a) recall

b) lift

c) accuracy

d) compactness

Answer : B

Explaination: None

4. Transaction ID Items bought.
1 Tea, Cake, Cold Drink.
2 Tea, Coffee, Cold Drink
3 Eggs, Tea, Cold Drink
4 Cake, Milk, Eggs
5 Cake, Coffee, Cold Drink, Milk, Eggs
From the above given data(Table 1),what is the value of support that if a person buy Tea, also buy Cake

a) 10%

b) 20%

c) 30%

d) 15%

Answer : C

Explaination: Support that if a person buy Tea, also buy Cake : 1 / 5 = 0.2 = 20%,Why 1? because Tea and Cake occur together only in 1 transaction,Why 5?because we have a total of 5 transactions

5. Transaction ID Items bought.
1 Tea, Cake, Cold Drink.
2 Tea, Coffee, Cold Drink
3 Eggs, Tea, Cold Drink
4 Cake, Milk, Eggs
5 Cake, Coffee, Cold Drink, Milk, Eggs
From the above given data(Table 1),value of absolute support and relative support of Coffee respectively is

a) 3 and 0.6

b) 2 and 0.4

c) 1and 0.2

d) 1 and 0.1

Answer : B

Explaination: Absolute Support is the absolute number of transactions which contains an itemset.
Relative support is the relative number of transactions which contains an itemset relative to the total transactions.
Formula:Total number of transactions containing an itemset X / Total number of transactions
Therefore,absolute support of coffee = 2 as we have two transactions of coffee and relative support for the same =2/5=0.4(5 because we have total 5 transactions)

6. Confidence is the conditional probability that a randomly selected transaction will include all the items in the consequent given that the transaction includes all the items in the antecedent.

a) True

b) False

c) Both

d) None

Answer : A

Explaination: Confidence can be interpreted as the likelihood of purchasing both the products A and B. Confidence is calculated as the number of transactions that include both A and B divided by the number of transactions includes only product A.

7. If an item set 'XYZ' is a frequent item set, then all subsets of that frequent item set are

a) Undefined

b) Not frequent

c) Frequent

d) Can not say

Answer : C

Explaination: All the subsets of frequent itemsets are frequent.

8. Which of the following is direct application of frequent itemset mining?

a) Social Network Analysis

b) Market Basket Analysis

c) Outlier Detection

d) Intrusion Detection

Answer : B

Explaination: Market Basket Analysis is one of the application of frequent itemset mining

6. APRIORI ALGORITHM

By :- Jagtap Rutuja Vijay(Branch Computer)

1.The number of iterations in apriori _____

a. increases with the size of the data

b.decreases with the increase in size of the data

c. increases with the size of the maximum frequent set

d. decreases with increase in size of the maximum frequent set

Answer:C

Explanation:The repeated itemset mining from data streams is an important data mining and that focuses on looking at sequences of actions or events.

2.An algorithm called____is used to generate the candidate item sets for each pass after the first.

A. apriori.

B. apriori-gen.

C. sampling.

D. partition.

Answer: B

Explanation:Apriori uses a "bottom up" approach, where frequent subsets are extended one item at a time (a step known as candidate generation), and groups of candidates are

3.Significant Bottleneck in the Apriori algorithm is Select one:

a. Finding frequent itemsets

b. Pruning

c. Candidate generation

d. Number of iterations

Answer: C

Explanation:Apriori uses a "bottom up" approach, where frequent subsets are extended one item at a time (a step known as candidate generation, and groups of candidates are t

4.____is the most well known association rule algorithm and is used in most commercial products.

A. Apriori algorithm.

B. Partition algorithm.

C. Distributed algorithm.

D. Pincer-search algorithm.

Answer: A

Explanation:It is by far the most well-known association rule algorithm. This technique uses the property that any subset of a large itemset must be a large itemset.

5.The apriori property means

(a) If a set cannot pass a test, all of its supersets will fail the same test as well

(b) To improve the efficiency the level-wise generation of frequent item sets

(c) If a set can pass a test, all of its supersets will fail the same test as well

(d) To decrease the efficiency the level-wise generation of frequent item sets

Answer: B

Explanation:The apriori property means to improve the efficiency the level-wise generation of frequent item sets

6.In Apriori algorithm, if 1 item-sets are 100, then the number of candidate 2 item-sets are

a. 100

b. 4950

c.200

d.5000

Answer:B

Explanation : -Association rule

7.Evaluation Of Association Rules

By :- Madhavi Gaikwad (Branch Computer)

1.What is the relation between a candidate and frequent itemsets?

a) A candidate itemset is always a frequent itemset

b) A frequent itemset must be a candidate itemset

c) No relation between these two

d) Strong relation with transactions

Answer : B

Explaination :- A frequent itemset is an itemset whose support is greater than some user-specified minimum support (denoted Lk, where k is the size of the itemset)
A candidate itemset is a potentially frequent itemset (denoted Ck, where k is the size of the itemset)

2. What is association rule mining?

a) Same as frequent itemset mining

b) Finding of strong association rules using frequent itemsets

c) Using association to analyze correlation rules

d) Finding Itemsets for future trends

Answer : B

Explaination :- Association rule mining is the data mining process of finding the rules that may govern associations and causal objects between sets of items. So in a given transaction with multiple items, it tries to find the rules that govern how or why such items are often bought together.

3.For the question given below consider the data Transactions :

I1, I2, I3, I4, I5, I6
I7, I2, I3, I4, I5, I6
I1, I8, I4, I5
I1, I9, I10, I4, I6
I10, I2, I4, I11, I5
With support as 0.6 find all frequent itemsets?

a) <I1>, <I2>, <I4>, <I5>, <I6>, <I1, I4>, <I2, I4>, <I2, I5>, <I4, I5>, <I4, I6>, <I2, I4, I5>

b)<I2>, <I4>, <I5>, <I2, I4>, <I2, I5>, <I4, I5>, <I2, I4, I5>

c)<I11>, <I4>, <I5>, <I6>, <I1, I4>, <I5, I4>, <I11, I5>, <I4, I6>, <I2, I4, I5>

d)<I1>, <I4>, <I5>, <I6>

Answer : A

Explaination :- As there are 5 data sets and min support is given as 0.6 so we have to find an itemsets with frequency greater than or equal to 3(i.e 5X0.6)

4.Frequency of occurrence of an itemset is called as _____

a) Support

b) Confidence

c) Support Count

d) Rules

Answer : C

Explaination :- Defination of Support count.

5.What will happen if support is reduced?

a)Number of frequent itemsets remains the same

b)Some itemsets will add to the current set of frequent itemsets

c)Some itemsets will become infrequent while others will become frequent

d)Can not say

Answer:B

Explaination:frequent itemset mining may generate a huge number of frequent itemsets, especially when the min_sup threshold is set low or when there exist long patterns in the data set.

6.An itemset whose support is greater than or equal to a minimum support threshold is _____

a)Itemset

b)Frequent Itemset

c)Infrequent items

d)Threshold values

Answer:B

Explaination:Defination of Frequent Itemset

7.How do you calculate Confidence(A -> B)?

a. Support(A∩B) / Support (A)

b. Support(A∩B) / Support (B)

c. Support(AUB) / Support (A)

d. Support(AUB) / Support (B)

Answer:A

Explaination:Defination of Confidence

8. Regression (Overview and introduction)

By :- Roshni Patil (branch Comp)

1. A relationship where the flow of the data points is best represented by a curve is called:

a) Linear Relationship

b) Nonlinear relationship

c) Linear positive

d) Linear negative

Answer : Option A

Explaination :

A relationship where the flow of the data point is best represented by the curve is called linear relationship.

2.It is possible that two regression coefficients have:

a) Opposite signs

b) Same signs

c) No sign

d) Difficult to tell

Answer : Option B

Explaination :

The regression coefficients must have the same sign. i.e., both must be positive or both must be negative.

3. In simple regression equation, the numbers of variables involved are:

a) 0

b) 1

c) 2

d) 3

Answer : Option C

Explaination :

In simple regression equation, the number of variables involved are :Two Simple linear regression is a statistical method that allows us to summarize and study relationships betw

4. If the value of any regression coefficient is zero, then two variables are:

a) Qualitative

b) Correlation

c) Dependent

d) Independent

Answer : Option D

Explaination :

If the correlation coefficient of two variables is zero, it signifies that there is no linear relationship between the variables. However, this is only for a linear relationship; it is possible

5. If one regression coefficient is greater than one, then other will he:

a) More than one

b) Equal to one

c) Less than one

d) Equal to minus one

Answer : Option C

Explaination :

Both the regression coefficients (bxy,byx) must have the same sign. i.e., if one of them is positive other should positive or if one of them is negative other should be negative. If or

6. The dependent variable is also called:

a) Regression

b) Regressand

c) Continuous variable

d) Independent

Answer : Option B

Explaination :

A linear regression line has an equation of the form Y=a+bX where Y is called dependent variable or response or regressand X is called independent variable or predictors or exp

7. Regression coefficient is independent of:

a) Units of measurement

b) Scale and origin

c) Both (a) and (b)

d) None of them

Answer : Option C

Explaination :

The regression coefficient is the measure of the relationship between two or more variables. The regression coefficient is independent of the measurement units, scale and origin

9.LINEAR REGRESSION METHOD

By : Janhavi Sharma (Branch Comp)

1. Linear Regression is a supervised machine learning algorithm.

A)True

B)False

C)Both

D)None

Answer: A

Explanation: Yes, Linear regression is a supervised learning algorithm because it uses true labels for training. Supervised learning algorithm should have input variable (x) and a

2. Which of the following methods do we use to find the best fit line for data in Linear Regression?

A) Least Square Error

B) Maximum Likelihood

C) Logarithmic Loss

D) Both A and B

Answer:A

Explanation: In linear regression, we try to minimize the least square errors of the model to identify the line of best fit.

3.Which of the following is true about Residuals ?

A) Lower is better

B) Higher is better

C) A or B depend on the situation

D) None of these

Answer: A

Explanation: Residuals refer to the error values of the model. Therefore lower residuals are desired.

4. How many coefficients do you need to estimate in a simple linear regression model (One independent variable)?

a) 1

b) 2

c) 3

d) 4

Answer: b

Explanation: In simple linear regression, there is one independent variable so 2 coefficients (Y=a+bx+error).

5. Function used for linear regression in R is _____

a) lm(formula, data)

b) lr(formula, data)

c) lrm(formula, data)

d) regression.linear(formula, data)

Answer: a

Explanation: lm(formula, data) refers to a linear model in which formula is the object of the class “formula”, representing the relation between variables. Now this formula is o
advertisement

6. Which of the following evaluation metrics can be used to evaluate a model while modeling a continuous output variable?

A) AUC-ROC

B) Accuracy

C) Logloss

D) Mean-Squared-Error

Answer: (D)

Explaination: Since linear regression gives output as continuous values, so in such case we use mean squared error metric to evaluate the model performance. Remaining optic

7.Suppose that we have N independent variables (X1,X2… Xn) and dependent variable is Y. Now Imagine that you are applying linear regression by fitting the best fit line u

You found that correlation coefficient for one of it's variable(Say X1) with Y is -0.95.

Which of the following is true for X1?

A) Relation between the X1 and Y is weak
B) Relation between the X1 and Y is strong
C) Relation between the X1 and Y is neutral
D) Correlation can't judge the relationship

Answer: (B)

Explaination:The absolute value of the correlation coefficient denotes the strength of the relationship. Since absolute correlation is very high it means that the relationship is s

8.True-False: It is possible to design a Linear regression algorithm using a neural network?

A) TRUE
B) FALSE

Answer: (A)

Explaination: True. A Neural network can be used as a universal approximator, so it can definitely implement a linear regression algorithm.

10. CLASSIFICATION AND OVERVIEW

By : Nikita Barve (Branch COMP)

1.VarImp is a wrapper around the evimp function in the_____ package

A.numpy

B.earth

c) plot

d) none of the mentioned

Answer: b

Explanation: The earth package is an implementation of Jerome Friedman's Multivariate Adaptive Regression Splines.

2. Point out the wrong statement.

a) The trapezoidal rule is used to compute the area under the ROC curve

b) For regression, the relationship between each predictor and the outcome is evaluated

c) An argument, para, is used to pick the model fitting technique

d) All of the mentioned

Answer: c

Explanation: An argument, nonpara, is used to pick the model fitting technique.

3. Which of the following curve analysis is conducted on each predictor for classification?

a) NOC

b) ROC

c) COC

d) All of the mentioned

Answer: b

Explanation: For two class problems, a series of cutoffs is applied to the predictor data to predict the class.

4. Which of the following function tracks the changes in model statistics?

a) varImp

b) varImpTrack

c) findTrack

d) none of the mentioned

Answer: a

Explanation: GCV change value can also be tracked.

5. Which of the following model include a backwards elimination feature selection routine?

a) MCV

b) MARS

c) MCRS

d) All of the mentioned

Answer: b

Explanation: MARS stands for Multivariate Adaptive Regression Splines.

6. The advantage of using a model-based approach is that is more closely tied to the model performance.

a) True

b) False

Answer: a

Explanation: Model-based approach is able to incorporate the correlation structure between the predictors into the importance calculation.

7. Which of the following argument is used to set importance values?

a) scale

b) set

c) value

d) all of the mentioned

Answer: a

Explanation: All measures of importance are scaled to have a maximum value of 100.

8.Who introduced MARS?

a) Jerome H.Friedman

b)Travis Oliphant

c)Jonh D

d)J. Oliphant

Answe: a

Explanation: Jerome H. Friedman introduced MARS in 1991

11. NAIVE BAYE'S CLASSIFIER

By :- Vedanti Chinchmalatpure (branch ENTC)

1. How many terms are required for building a bayes model?

a) 1

b) 2

c) 3

d) 4

Answer: c

Explanation: The three required terms are a conditional probability and two unconditional probability.

2. Where is Naive Bayes Used?

a) Face Recognition

b)Weather Prediction

c) Medical Diagnosis

d) All of the mentioned

Answer: d

Explanation: You can use Naive Bayes for all the mentioned options

3. What does the bayesian network provides?

a) Complete description of the domain

b) Partial description of the domain

c) Complete description of the problem

d) None of the mentioned

Answer: a

Explanation: A Bayesian network provides a complete description of the domain.

4.Bayes theorem is given by

a) P(A|B)= [P(B|A) * P(A)] / P(B)

b) P(B/A)= [P(B|A) * P(A)] / P(B)

c) P(A/B)= [P(B|A)]/P(A) P(B)

d) None of the mentioned

Answer: a

Explanation: The Bayes theorem gives us the conditional probability of event A, given that event B has occurred. In this case, the first coin toss will be B and the second coin toss

5. With the help of a _____ Google News recognizes whether the news is political, world news, and so on.

a) Naive Bayes classifier

b) Bayes theroem

c) Partial distribution

d) All of the mentioned

Answer: a

Explanation: News recognitoin is one of the aplication of Naive Bayes classifier.

6. Advantages of Naive Bayes Classifier are:

a) It is simple and easy to implement

b)It handles both continuous and discrete data

c)It doesn’t require as much training data

d) All of the mentioned

Answer: a

Explanation: The following are some of the benefits of the Naive Bayes classifier:

It is simple and easy to implement
It doesn’t require as much training data
It handles both continuous and discrete data
It is highly scalable with the number of predictors and data points
It is fast and can be used to make real-time predictions
It is not sensitive to irrelevant features

7.Text classification is one of the most popular applications of

a) Naive Bayes classifier

b) Bayes theroem

c) Partial distribution

d) All of the mentioned

Answer: a

Explanation:Text classification is one of the aplication of Naive Bayes classifier.

UNIT 4

ADVANCE DATA ANALYSIS MEANS

Topic 1: WHAT IS DECISION TREE?CREATING A DECISION TREE

By :- Arati Ratnaparkhi (Branch COMP)

1. **A _____ is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility.**
   A. Decision tree
   B. Graphs
   C. Trees
   D. Neural Networks
   Ans : A
   Explanation: Refer the definition of Decision tree.

2. **Decision Tree is a display of an algorithm**.
   A. True
   B. False
   Ans : A
   Explanation: it is one way to display algorithm that only contains conditional control statements.

3. **What is Decision Tree?**
   A. Flow-Chart
   B. Structure in which internal node represents test on an attribute, each branch represents outcome of test and each leaf node represents class label
   C. Both a and b
   Ans : C
   Explanation: Refer the definition of Decision tree.

4. **Decision Trees can be used for Classification Tasks.**
   A. True
   B. False
   Ans :A
   Explanation: Decision tree is a type of supervised learning algorithm that can be used in both regression and classification problems.

5. **Choose from the following that are Decision Tree nodes**?
   A. Decision Nodes
   B. End Nodes

C. Chance Nodes
D. All of the mentioned
Ans : D
Explanation: A decision tree consists of three types of nodes: Decision nodes – typically represented by squares. Chance nodes – typically represented by circles. End nodes – typically represented by triangles.

6. **Decision Nodes are represented by _____**
A. Disks
B. Squares
C. Circles
D. Triangles
Ans : B
Explanation: Decision Nodes are represented by square.

By :- Siddhi Jadhav (Branch ENTC)

7.End Nodes are represented by _____
A. Disks
B. Squares
C. Circles
D. Triangles
Ans : D
Explanation: End Nodes are represented by triangles.

8.Which of the following are the advantage/s of Decision Trees?
A. Possible Scenarios can be added
B. Use a white box model, If given result is provided by a model
C. Worst, best and expected values can be determined for different scenarios
D. All of the mentioned
Ans : d
Explanation: None.

9. In the general case, imagine that we have d binary features, and we want to count the number of features with value 1. How many leaf nodes would a decision tree need to represent this function?
A. 2^1 leaf nodes
B. 2^d leaf nodes
C. 2^{d-1} leaf nodes
D. 2^d-1 leaf nodes
Ans : B
Explanation: We need 2^d leaf nodes. For example, for one feature we have two leaf nodes at maximum (either 0 or 1). For 2 features we have four leaf nodes and so on.

10. What is the biggest weakness of decision trees compared to logistic regression classifiers?

A. Decision trees are more likely to overfit the data
B. Decision trees are more likely to underfit the data
C. Decision trees do not assume independence of the input features
D. None of the mentioned

Ans : A

Explanation: Decision trees are more likely to overfit the data since they can split on many different combination of features whereas in logistic regression we associate only one parameter with each feature.

11. Chance Nodes are represented by _____.

A. Disks
B. Squares
C. Circles
D. Triangles

Ans : C

Explanation: A chance node is represented by a circle, shows a probability of a certain result.

12.Decision trees are an algorithm for which machine learning task?

A. clustering
B. classification
C. regression
D. Both b & c

Ans : D

Explanation: Decision trees are non parametric supervised learning methods used for both classification & regression tasks.The goal is to create a model that predicts the value of a target variable.

13. Which error metric is most appropriate for evaluating a {0,1} classification task?

A. worst-case error
B. sum of squares error
C. Entropy
D. precision and recall

Ans : D

Explanation:The F1 score is a number between 0 and 1 and is the harmonic mean of precision and recall.

TOPIC 2: ENTROPY AND THE OF A PARTITION

By :- Rutuja Tarapure (Branch COMP)

1. **What is entropy?**

A. A measure of the randomness in the information being processed.
B. A conclusion drawn from a set of information.
C. A relationship determined from a set of information.
D. None of the other Ans s

Ans : A

Explanation: Entropy is a measure of the randomness in the information being processed. The higher the entropy, the harder it is to draw any conclusions from that information.

2. **High entropy means that the partitions in classification are**

A. pure
B. not pure
C. useful
D. useless

Ans : B

Explanation:Entropy is a measure of the randomness in the information being processed. The higher the entropy, the harder it is to draw any conclusions from that information.
It is a measure of disorder or purity or unpredictability or uncertainty.

3. **High entropy is an indication of:**

A. That there is an observable relationship in the information
B. Predicting coin toss results
C. The randomness in the information being processed
D. How easy it is to draw conclusions from the information

Ans :C

Explanation:Higher entropy indicates higher uncertainty and a more chaotic system.

4. **.Entropy in information theory is directly analogous to the entropy in**

A. thermodynamics
B. biochemistry
C. zoology
D. physics

Ans :A

Explanation:Entropy is a scientific concept, as well as a measurable physical property that is most commonly associated with a state of disorder, randomness, or uncertainty.

Entropy, the measure of a system's thermal energy per unit temperature that is unavailable for doing useful work in thermodynamics

5. **The concept of information entropy was introduced by**

A. John Tukey
B. PeterNaur
C. Claude Shannon
D. none of the above

Ans :C

6. **Negentropy(negative entropy) is used as a measure of**
    A. distribution
    B. distance to normality.
    C. uncertainty
    D. none of the above

    Ans : B

    Explanation: Negentropy is a negation of entropy. It measures the difference in entropy between a given distribution and the Gaussian distribution with the same mean and variance

By:-Meenal Kore(Branch ENTC)

1. **Entropy of N random variables is the _____ of the entropy of individual random variable.**
    A. Sum
    B. Product
    C. Sum of squares
    D. Average

    Ans : A

    Explanation: Entropy of N random variables is the sum of the entropy of individual random variable.

2. **Average energy per bit is given by**
    A. average energy symbol/log2 M
    B. average energy symbol * log2 M
    C. log2 M/ Average energy symbol
    D. none of the mentioned

    Ans : A

    Explanation: Average energy per bit is given by average energy symbol divided by log2 M.

3. **Entropy of a random variable is**
   A. 0
   B. 1
   C. Infinite
   D. Cannot be determine

   Ans : C

   Explanation: Entropy of a random variable is also infinity.

4. **Entropy is**
   A. Amplitude of signal
   B. Information in a signal
   C. Average information per message
   D. All of the above

   Ans : C

   Explanation: entropy is average information per message so we can calculate uncertainty

5. **The equation of entropy**

   A. $H(X)= – \Sigma\ p(x)\ log2p(x)$
   B. $H(X)= \Sigma\ p(x)\ log2p(x)$
   C. $H(X)= \Sigma\ -logp(x)$
   D. $H(X)=\Sigma\ p(-x)\ log\ p(x)$

   Ans : A

   Explanation: Where 'P' is simply the frequentist probability of an element/class 'i' in our data. It is the entropy formula for an event X with n possible outcomes and probabilities p_1, …, p_n.
   the units are bits (based on the formula using log base 2). The intuition is entropy is equal to the number of bits you need to communicate the outcome of a certain draw.

6. **Properties of entropy are**

   A. Uniform distributions with more outcomes have more uncertainty
   B. Uncertainty is additive for independent events
   C. it is always nonnegative
   D. none of the above

   Ans : A & B

   Explanation: a)distribution is uniform when all of the outcomes have the same probability
   A good measure of uncertainty achieves its highest values for uniform distributions.
   The uncertainty associated with both event should be the sum of the individual uncertainties: H(X,Y)=H(X)+H(Y)
   Remaining option is the property of negative entropy

Topic 3: RANDOM FORESTS NEURAL NETWORKS : PERCEPTRONS

By : Kokate Neha (Branch ENTC)

1 . What is perceptron ?

A. a single layer feed-forward neural network with pre-processing
B. an auto-associative neural network
C. a double layer auto-associative neural network
D. a neural network that contains feedback

Ans : A

Explanation : The perceptron is a single layer feed-forward neural network. It is not an auto-associative network because it has no feedback and is not a multiple layer neural network because the pre-processing stage is not made of neurons.

2. Which of the following is true for neural networks?
(i) The training time depends on the size of the network.
(ii) Neural networks can be simulated on a conventional computer.
(iii) Artificial neurons are identical in operation to biological ones.
A. All of the mentioned
B. (ii) is true
C. (i) and (ii) are true
D. None of the mentioned

Ans : C

Explanation : The training time depends on the size of the network; the number of neuron is greater and therefore the number of possible 'states' is increased. Neural networks can be simulated on a conventional computer but the main advantage of neural networks – parallel execution – is lost. Artificial neurons are not identical in operation to the biological ones.

3. Which is true for neural networks?
A. It has set of nodes and connections
B. Each node computes it's weighted input

C. Node could be in excited state or non-excited state
D. All of the mentioned

Ans : D

Explanation : All mentioned are the characteristics of neural network.

4. Neural Networks are complex _____ with many parameters.
A. Linear Functions
B. Nonlinear Functions
C. Discrete Functions
D. Exponential Functions

Ans : A

Explanation : Neural networks are complex linear functions with many parameters.

5. A perceptron adds up all the weighted inputs it receives, and if it exceeds a certain value, it outputs a 1, otherwise it just outputs a 0.
A. True
B. False
C. Sometimes – it can also output intermediate values as well
D. Can't say

Ans : A

Explanation : Yes the perceptron works like that.

6. Which of the following is an application of NN (Neural Network)?
A. Sales forecasting
B. Data validation
C. Risk management
D. All of the mentioned

Ans : D

Explanation : All mentioned options are applications of Neural Network.

7. Which of the following is correct with respect to random forest?

A. Random forest are difficult to interpret but often very accurate

B. Random forest are easy to interpret but often very accurate

C. Random forest are difficult to interpret but very less accurate

D. None of the mentioned

Ans : A

Explanation : Random forest is top performing algorithm in prediction.

Topic 4: FEED-FORWARD NEURAL NETWORK

By :- Apurva Dahiphalkar (Branch IT)

1.What are the advantages of neural networks over conventional computers?

(i) They have the ability to learn by example

(ii) They are more fault tolerant

(iii)They are more suited for real time operation due to their high 'computational' rates

A. (i) and (ii) are true

B. (i) and (iii) are true

C. Only (i)

D. All of the mentioned

Ans: D

Explanation: Neural networks learn by example. They are more fault tolerant because they are always able to respond and small changes in input do not normally cause a change in output. Because of their parallel architecture, high computational rates are achieved.

2.In which ANN, loops are allowed?

A. FeedForward ANN

B. FeedBack ANN

C. Both A and B

D. None of the Above

Ans : B

Explanation: FeedBack ANN loops are allowed. They are used in content addressable memories.

3.How many types of Artificial Neural Networks?

A. 2

B. 3

C. 4

D. 5

Ans : A

Explanation: There are two Artificial Neural Network topologies : FeedForward and Feedback.

4.Which of the following is not an Machine Learning strategies in ANNs?

A. Unsupervised Learning

B. Reinforcement Learning

C. Supreme Learning

D. Supervised Learning

Ans : C

Explanation: Supreme Learning is not an Machine Learning strategies in ANNs

5.In FeedForward ANN, information flow is ___.

A. unidirectional

B. bidirectional

C. multidirectional

D. All of the above

Ans : A

Explanation: FeedForward ANN the information flow is unidirectional.

6.What is full form of ANNs?

A. Artificial Neural Node

B. AI Neural Networks

C. Artificial Neural Networks

D. Artificial Neural numbers

Ans : C

Explanation: Artificial Neural Networks is the full form of ANNs.

7.The first artificial neural network was invented in ___.

A. 1957

B. 1958

C. 1959

D. 1960

Ans : B

Explanation: The first artificial neural network was invented in 1958.

8.Which of the following is an Applications of Neural Networks?

A. Automotive

B. Aerospace

C. Electronics

D. All of the above

Ans : D

Explanation: All above are application of Neural Networks.

9.What is the advantage of basis function over multilayer feedforward neural networks?

A. training of basis function is faster than MLFFNN

B. training of basis function is slower than MLFFNN

C. storing in basis function is faster than MLFFNN

D. none of the mentioned

Ans: A

Explanation: The main advantage of basis function is that the training of basis function is faster than MLFFNN.

TOPIC 5: BACKPROPAGATION

By :- Priyanaka Dhasade (Branch IT)

1. What is the objective of backpropagation algorithm?
A. to develop learning algorithm for multilayer feedforward neural network
B. to develop learning algorithm for single layer feedforward neural network
C. to develop learning algorithm for multilayer feedforward neural network, so that network can be trained to capture the mapping implicitly
D. none of the mentioned
Ans: C

Explanation: The objective of backpropagation algorithm is to to develop learning algorithm for multilayer feedforward neural network, so that network can be trained to capture the mapping implicitly.

2.The backpropagation law is also known as generalized delta rule, is it true?
A. yes
B. no

Ans: A
Explanation: Because it fulfils the basic condition of delta rule.

3.What is true regarding backpropagation rule?
A. it is also called generalized delta rule
B. error in output is propagated backwards only to determine weight updates
C. there is no feedback of signal at nay stage

D. all of the mentioned

Ans: D

Explanation: These all statements defines backpropagation algorithm.

4.There is feedback in final stage of backpropagation algorithm?

A. yes

B. no

Ans: B

Explanation: No feedback is involved at any stage as it is a feedforward neural network.

5.What is true regarding backpropagation rule?

A. it is a feedback neural network

B. actual output is determined by computing the outputs of units for each hidden layer

C. hidden layers output is not all important, they are only meant for supporting input and output layers

D. none of the mentioned

Ans: B

Explanation: In backpropagation rule, actual output is determined by computing the outputs of units for each hidden layer.

By: - Priti Dhokte (Branch IT)

6.What is meant by generalized in statement "backpropagation is a generalized delta rule" ?

A.because delta rule can be extended to hidden layer units

B. because delta is applied to only input and output layers, thus making it more simple and generalized

C. it has no significance

D. none of the mentioned

Ans : A

Explanation: The term generalized is used because delta rule could be extended to hidden layer units.

7.What are the general tasks that are performed with backpropagation algorithm?

A. pattern mapping

B. function approximation

C. prediction

D. all of the mentioned

Ans : D

Explanation: These all are the tasks that can be performed with backpropagation algorithm in general.

8.Does backpropagation learning is based on gradient descent along error surface?

A. yes

B. no

C. cannot be said

D. it depends on gradient descent but not error surface

Ans : A

Explanation: Weight adjustment is proportional to negative gradient of error with respect to weight

9. What are general limitations of back propagation rule?

A. local minima problem

B. slow convergence

C. scaling

D. all of the mentioned

Ans : D

Explanation: These all are limitations of backpropagation algorithm in general.

10.How can learning process be stopped in backpropagation rule?

A. there is convergence involved

B. no heuristic criteria exist

C. on basis of average gradient value

D. none of the mentioned

Ans : C

Explanation: If average gradient value fall below a preset threshold value, the process may be stopped.

Topic6:-Example:defeating A Captcha Mapreduce: Why Mapreduce?

By:- Manisha Lakhe(Branch IT)

1. Point out the correct statement.

A. MapReduce tries to place the data and the compute as close as possible

B. Map Task in MapReduce is performed using the Mapper() function

C. Reduce Task in MapReduce is performed using the Map() function

D. All of the mentioned

Ans : A

Explanation: This feature of MapReduce is “Data Locality”.

2. _____ part of the MapReduce is responsible for processing one or more chunks of data and producing the output results.

A. Maptask

B. Mapper

C. Task execution

D. All of the mentioned

Ans : A

Explanation: Map Task in MapReduce is performed using the Map() function.

3. _____ function is responsible for consolidating the results produced by each of the Map() functions/tasks.

A. Reduce

B. Map

C. Reducer

D. All of the mentioned

Ans : A

Explanation: Reduce function collates the work and resolves the results.

4. _____ maps input key/value pairs to a set of intermediate key/value pairs.

A. Mapper

B. Reducer

C. Both Mapper and Reducer

D. None of the mentioned

Ans : A

Explanation: Maps are the individual tasks that transform input records into intermediate records.

5. The number of maps is usually driven by the total size of _____

A. inputs

B. outputs

C. tasks

D. None of the mentioned

Ans : A

Explanation: Total size of inputs means the total number of blocks of the input files.

6. Running a _____ program involves running mapping tasks on many or all of the nodes in our cluster.

A. MapReduce

B. Map

C. Reducer

D. All of the mentioned

Ans : A

Explanation: In some applications, component tasks need to create and/or write to side-files, which differ from the actual job-output files.

7. Which of these are not a captcha solving service

A. DeathbyCAPTCHA

B. 2Captcha

C. Kolotibablo

D. All of the above

Ans : D

8. Google released an upgraded version of its reCAPTCHA called

A.CAPTCHA reformed

B. Invisible reCAPTCHA

C. 2ReCAPTCHA

D. None

Ans: B

Explanation :In March of 2017, Google released an upgraded version of its reCAPTCHA called “Invisible reCAPTCHA.” Unlike “no CAPTCHA reCAPTCHA,” which required all users to click the infamous “I’m not a Robot” button, Invisible reCAPTCHA allows known human users to pass through while only serving a reCAPTCHA image challenge to suspicious users.

9. Is Google InvinsiblereCAPTCHA really invinsible ?

A. True
B. False

Ans: B

Explanation: The fact that Invisible reCAPTCHA can be bypassed isn’t because there was a fatal flaw in the design of the newer CAPTCHA. It’s that any reverse Turing test is inherently beatable when the pass conditions are known.

10. Once the attacker receives the CAPTCHA solution from captcha solving services, Invisible reCAPTCHA can be defeated via automation of

A. JavaScript action that calls a function to supply the solved token with the page form submit
B. HTML code change directly in the webpage to substitute a snippet of normal CAPTCHA code with the solved token input.
C. All of the above
D. None of the above

Ans: C

Explanation: When the pass Conditions are known, the CAPTCHA can be defeated using both the solutions.

By: - Priyanka Desai (Branch IT)

1)_____ can best be described as a programming model used to develop applications that can process massive amount of data.

A. Oozie
B. mahout
C. MapReduce
D. All of the mentioned

Ans : C

Explanation:

MapReduce is a programming model and an associated implementation for processing and generating large data sets with a parallel, distributed algorithm.

2)_____ is general-purpose computing model and runtime system for distributed data analytics.

A. Oozie
B. MapReduce
C. mahout

D. All of the mentioned

Ans : B

Explanation:

MapReduce provides a flexible and scalable foundation for analytics,from traditional reporting to leading-edge machine learning algorithms.

3)MapReduce process executes in how many stages.

A. 3
B. 4
C. 5
D. 6

Ans : A

Explanation:

MapReduce program executes in three stages, namely map stage, shuffle stage, and reduce stage.

4)Which among the following is the programming model designed for processing large volumes of data in parallel by dividing the work into a set of independent tasks.

A. MapReduce
B. HDFS
C. Pig
D. None of the mentioned

Ans : A

Explanation:

It is a programming model designed for processing huge volumes of data (both ... in parallel by dividing the work into a set of independent sub-work (tasks)

5)The most common type of CAPTCHA (displayed as Version 1.0) was first invented in _____by two groups working in parallel.

A. 1996
B. 1997
C. 1998
D. None of the mentioned.

Ans : B

Explanation:
Captcha Verification is presented with a picture of words or characters, and the respondent must correctly type out those characters in order to proceed invented in 1997.

.

6)What is full form of CAPTCHA

A. Computer Automated per tuning cell and Hard Apart Computer Automated per tuning cell and Hard Apart
B. Completely Automated Public Turing test to tell Computers
and Humans Apart Completely Automated Public Turing test to tell Computers
and Humans Apart
C. Community Action Planning Council test to tell computers and human Apart. Community Action Planning Council test to tell computers and human Apart.
D. None of the mentioned. None of the mentioned.

Ans : B

Explanation:

CAPTCHAs were created in response to bots (software agents) that automatically fill in Web forms as if they were individual users.

7)A common method is to use a CAPTCHA solving service, which utilizes ____-cost human labor in developing countries to solve CAPTCHA images.

A. low
B. high
C. large
D. None of the mentioned

Ans : A

Explanation:
Cybercriminals subscribe to a service for CAPTCHA solutions, which streamline into their automation tools via APIs, populating the Ans s on the target website so low human labour is needed.

8)Which of the following is a layer of CNN ?

A. Logistic layer
B. Convo layer
C. Pooling layer
D. All of the mentioned

Ans : D

Explanation:

There are five different layers in CNN:-Input layer,Convolayer,Poolinglayer,Fully connected(FC) layer,Softmax/logistic layer,Output layer

9)Layers in CNNs are special as they are organized in _____ dimensions.

A. 1
B. 2
C. 3
D. None of the mentioned

Ans : C

Explanation:

The CNN Layers are width, height and depth so they are organized in 3 dimensions.

10)To use 2Captcha as a service, a customer (i.e., an attacker) integrates the_____ into her attack to create a digital supply chain.

A. 1Cpatcha API
B. 2Captcha API
C. 3Captcha API
D. None of the mentioned

Ans : B

Explanation:

The process of solving captchas with 2captcha get your API key from your account settings page. Each user is given a unique authentication token, we call it API key.

Topic 7: - Examples Like Word Count and Matrix Multiplication

By: - Yukta Patil (Branch IT)

1. **In a word count query using MapReduce, what does the map function do?**
   A. It sorts the words alphabetically and returns a list of the most frequently used words.
   B. It returns a list with each document as a key and the number of words in it as the value. The master JobTracker sends map and reduce functions to the same machines or nodes in a cluster.
   C. It creates a list with each word as a key and every occurrence as value 1.
   D. It creates a list with each word as a key and the number of occurrences as the value.

Ans: C

2. **The MapReduce algorithm contains two important tasks, namely _____.**
   A. mapped, reduce
   B. mapping, Reduction
   C. Map, Reduction
   D. Map, Reduce

Ans: D

Explanation: The MapReduce algorithm contains two important tasks, namely Map and Reduce.

3. **In how many stages the MapReduce program executes?**

A. 2
B. 3
C. 4
D. 5

Ans: B

Explanation: MapReduce program executes in three stages, namely map stage, shuffle stage, and reduce stage.

4. **Although the Hadoop framework is implemented in Java, MapReduce applications need not be written in ____**

A. Java
B. C
C. C#
D. None of the mentioned

Ans: A

Explanation: Hadoop Pipes is a SWIG- compatible C++ API to implement MapReduce applications (non JNITM based).

5. **____ is a utility which allows users to create and run jobs with any executables as the mapper and/or the reducer.**

A. Hadoop Strdata
B. Hadoop Streaming
C. Hadoop Stream
D. None of the mentioned

Ans: B

Explanation: Hadoop streaming is one of the most important utilities in the Apache Hadoop distribution.

UNIT 5

BASICS OF DATA VISUALIZATION

Topic 1: - Challenges and advantages of data visualization

By:-Gayatri gadas (Branch-IT)

1) What is true about Data Visualization?

A. Data Visualization is used to communicate information clearly and efficiently to users by the usage of information graphics such as tables and charts.
B. Data Visualization helps users in analyzing a large amount of data in a simpler way.
C. Data Visualization makes complex data more accessible, understandable, and usable.
D. All of the above

Answer:-D

2) Data can be visualized using?

A. graphs
B. charts
C. maps
D. All of the above

Answer: - D

3) Data visualization is also an element of the broader _____.

A. deliver presentation architecture
B. data presentation architecture
C. dataset presentation architecture
D. data process architecture

Answer:- B

4) Which method shows hierarchical data in a nested format?

A. Treemaps
B. Scatter plots
C. Population pyramids
D. Area charts

Answer:- A

5) Which is used to inference for 1 proportion using normal approx?
A. fisher.test()
B. chisq.test()
C. Lm.test()
D. prop.test()
Answer:-D

6) Which is used to find the factor congruence coefficients?
A. factor.mosaicplot
B. factor.xyplot
C. factor.congruence
D. factor.cumsum
Answer:-C

7) Which of the following is tool for checking normality?
A. qqline()
B. qline()
C. anova()
D. lm()
Answer:-A

Topic 2: - Data visualization and options

By:-Dipali Salunke (Branch-Computer)

1) Which of the following is false?
A. data visualization include the ability to absorb information quickly
B. Data visualization is another form of visual art
C. Data visualization decrease the insights and take solwer decisions
D. None Of the above
Answer:- C
Explanation: Data visualization decrease the insights and take slower decisions is false statement.

2) Common use cases for data visualization include?
A. Politics
B. Sales and marketing
C. Healthcare
D. All of the above
Answer:- D
Explanation: All option are Common use cases for data visualization

3) Which of the following plots are often used for checking randomness in time series?

A. Autocausation
B. Autorank
C. Autocorrelation
D. None of the above

Answer:- C

Explanation: If the time series is random, such autocorrelations should be near zero for any and all time-lag separations.

4) Which are pros of data visualization?

A. It can be accessed quickly by a wider audience.
B. It can misrepresent information
C. It can be distracting
D. None Of the above

Answer :- A

Explanation: Pros of data visualization : it can be accessed quickly by a wider audience.

5) Which are cons of data visualization?

A. It conveys a lot of information in a small space.
B. It makes your report more visually appealing.
C. visual data is distorted or excessively used.
D. None Of the above

Answer :- C

Explanation: It can be distracting : if the visual data is distorted or excessively used

6) Which of the intricate techniques is not used for data visualization?

A. Bullet Graphs
B. Bubble Clouds
C. Fever Maps
D. Heat Maps

Answer : C

Explanation: Fever Maps is not is not used for data visualization instead of that Fever charts is used

7) Which one of the following is most basic and commonly used techniques?

A. Line charts
B. Scatter plots
C. Population pyramids
D. Area charts

Answer : A

Explanation: Line charts. This is one of the most basic and common techniques used. Line charts display how variables can change over time.

Topic 3: - Dashboard types

By:-Anuja Bhosale (Branch-IT)

1)An ____ dashboard is a reporting tool that is used to analyze large volumes of data to allow users to investigate trends, predict outcomes, and discover insights.
a)Operational
b)Strategic
c)Analytical
d)Data

Answer-c)Analytical.
Explanation-Analytical dashboards are more common within business intelligence tools because they are typically developed and designed by data analysts.

2)An ___ dashboard is a reporting tool that is used to monitor business processes that frequently change and to track current performance of key metrics and KPIs.
a)Strategic
b)Operational
c)Interactive
d)Analytical

Answer-b)Operational.
Explanation- Compared to other types of dashboards, the data updates very frequently, sometimes even on a minute-by-minute basis.

3)A ____ dashboard is a reporting tool used to monitor the status of key performance indicators (KPIs), and are typically used by executives.
a)Operational
b)Analytical
c)Strategic
d)Data.

Answer-c)Strategic.
Explanation-A strategic dashboard is a reporting tool used to monitor the status of key performance indicators (KPIs), and are typically used by executives.

4)The data behind a strategic dashboard updates on a recurring basis, but at more frequent intervals than an analytical dashboard.
a)True
b)False

Answer-b)False
Explanation-The data behind a strategic dashboard updates on a recurring basis, but at less frequent intervals than an operational dashboard.

5)What is the mean by KPIs?
a)key performance indicators
b)key process indicators
c)key program indicators
d)key positions indicators

Answer:- a
Explanation-A Key Performance Indicator is a measurable value that demonstrates how effectively a company is achieving key business objectives.

6)Operational dashboards are designed to be viewed ___ time throughout the day.
a)one
b)multiple
c)two
d)None of the mentioned.
Answer:- b
Explanation-In operational dashboards data updates very frequently, sometimes even on a minute-by-minute basis.

7)Which are the types of business dashboards?
a)Operational
b)Strategic
c)Analytical
d)All of the above.
Answer:-d
Explanation- There are three types of business dashboards from top to bottom: strategic dashboard, analytical dashboard, and operational dashboard.

Topic 4: - Data dashboard

By:- Neha Pandhare (Branch-IT)

1. A data dashboard is an ___ that visually tracks, analyzes and display key performance indicators (KPI).
A. Information management tool
B. Business information tool
C. Data visualization tool
D. Open source dashboard tool
Answer: A

2.Data Dashboard provide an object view of ___ and serve as an effective foundation for further dialogue
A. Testing matrics
B. Source code matrics
C. Performance matrics
D. Development matrics
Answer: C

3.___ are helpful performance monitors for department manager and front line worker.
A. Analytical dashboard
B. Operational dashboard
C. Central dashboard
D. Strategic dashboard
Answer: B

4. Dashboard are capable of correlate data from different source in a ___ if the user so chooses
A. Multiple visualization
B.double visualization
C. Single visualization
D. None
Answer: C

5.____are typically designed to help decision makers, executive and senior leaders.
A. Analytical dashboard
B. Central dashboard
C. Operation dashboard
D. Strategic dashboard
Answer: A

6. Monitoring multiple KPI's and metrics on one ___ , user can make adjustment to their business pratices in real time.
A. KPI dashboard
B.Central dashboard
C. Dynamic dashboard
D. Excel dashboard
Answer: B

7. A dashboard is a ___ used to display data visualization in a way that is immediately understood
A. Business intelligence tool
B. Information management tool
C. Data visualization tool
D. Open source dashboard tool
Answer: A

Topic 5: - Dashboard design and Principles

By:-Muskan Chawan (Branch-IT)

1. What combines the outward manifestation of the computer-based system , coupled with all supporting information that describe system syntax and semantics?
a) mental image
b) interface design
c) system image
d) interface validation

Answer: c
Explanation: When the system image and the system perception are coincident, users generally feel comfortable with the software and use it effectively.

2. What establishes the profile of end-users of the system?
a) design model
b) user's model
c) mental image
d) system image

Answer: b
Explanation: To build an effective user interface, all design should begin with an understanding of the intended users, including their profiles of their age, physical abilities, education, etc.

3. What incorporates data, architectural, interface, and procedural representations of the software?
a) design model
b) user's model
c) mental image
d) system image

Answer: a
Explanation: The requirements specification may establish certain constraints that help to define the user of the system, but the interface design is often only incidental to the design model.

4. A software engineer designs the user interface by applying an iterative process that draws on predefined design principles.
a) True
b) False

Answer: a
Explanation: The statement is true.

5. A software might allow a user to interact via
a) keyboard commands
b) mouse movement
c) voice recognition commands
d) all of the mentioned

Answer: d
Explanation: All the mentioned input mediums are available today.

6. When users are involved in complex tasks, the demand on ____ can be significant.
a) short-term memory
b) shortcuts
c) objects that appear on the screen
d) all of the mentioned

Answer: a
Explanation: The interface should be designed to reduce the requirement to remember past actions and results.

Topic 6: - Display of dashboard media

By:-Rohini sawant (Branch-IT)

1) For creating variable size bins we use__.
A)sets
B)Groups
C) calculated fields
D)table calculations
Answer: C
Explanation: calculated fields allows you to create new data from data that already exists in your data source.

2)____feature allows the user to know more comprehensive dashboard information related to particular element.
A)drill-down
B)click-to-filter
C)time interval widget
D)none of the above
Answer: A
Explanation: using drill down we can get comprehensive dashboard information related to particular element, variable or a key performance indicator without overcrowding the overall design.

3)Which of the of following would be the right choice, when we want to compare items in the same category?
A)line charts
B)Bar charts
C)pie charts
D)none
Answer: B
Explanation: Ex. Rainfall in a single month

4)Gauge will display the latest value that is running in the board.
A)True
B)False
Answer: A
Explanation:Ex. Speedometer in a car shows only the current speed of the car.

5)Which medium is used to import data into dashboard from data warehouse?
A)web services
B)Excel files
C)CSV files
D)All of the above
Answer:D
Explanation: Data can be fetched or imported from Excel files,web services as well as CSV files which are created using R.

6)A good reason to use a bullet graph is___.
A)analysing the trend for a time period
B)comparing the actual against the target sales
C)adding data to bins and calculating count measures
D) displaying the sales growth for a particular year
Answer : B

Topic 7: - Types of data visualization

By:-Shrutika Shinde (Branch-IT)

1.Of the following which are types of data visualization.
a. catter plots.
b. Polar area diagram
c. Time series sequences.
d. All of the above
Answer:- d

2.A graph of plotted points that show the relationship between two sets of data.
a. Scatter plot
b. Histogram
c. Cartogram
d. none of the above
Answer: a
Cartogram: A cartogram distorts the geometry or space of a map to convey the information of an alternative variable.
Histogram : It uses rectangles with the heights proportional to the count and the widths equal to range of small interval.

3.Which of the following is true about Column chart?
a. Easy to read and understand
b.One data set can be changed without affecting others
c. With too many categories, it can become a bit too cluttered
d. Both a and b
Answer. Both a and b

4 . Which of the following are temporal
Visualizations?
a. Connected Scatter Plot
b. Polar Area Diagram
c .Time Series
d. All of thr above
Answer: d

5.Which of the following are Multidimensional Visualizations?
a. Pie Chart
b. Histogram
c. Scatter plot
d. All of the above
Answer: d

6.What are Network Data Visualizations?
a. Alluvial Diagram
b. Node -Link Diagram
c. Matrix
d. All of the above
Answer : d

7.Which of the following is true about Bar Chart?
a. You want to compare two or more values in the same category
b.You want to understand how multiple similar data sets relate to each other.
c. The category you're visualizing only has one value associated with it.
d. Both a and b.
Answer : d.

Topic 8: - Charts and histograms

By:- Pooja Sharma (Branch-E&tc)

1)Which of the following is appropriate to graph a single continuous variable?

a. Waffle chart
b. Histogram
c. Bar chart
d. Pie chart

Ans: b

2)mosaic plot is used when graphing

a. the relationship between two continuous variables.
b. the relationship between one continuous and one categorical variable.
c. the relationship between two categorical variables.
d. data that are not normally distributed by group.

Answer: c

3)Density plots, histograms, and boxplots can all be used to

a. examine frequencies in categories of a factor.
b. examine the relationship between two categorical variables.
c. determine whether two continuous variables are related.
d. examine the distribution of a continuous variable.

Answer: d

4)which of the following statement make a mosaic plot?
a) histogram()
b) mosaicplot()
c) bar()
d) which.max(x)

Answer: b
Explanation: histogram() is lattice command for producing a histogram.

5)Which of the following is used to view dataset in a spreadsheet-type format ?
a) Disp()
b) View()
c) Seq()
d) lm()

Answer: b
Explanation: seq() make arithmetic progression vector.

6)Data can be visualized using?

a) graphs
b) charts
c) maps
d) All of the above

Answer : D
Explanation: Data visualization is a graphical representation of quantitative information and data by using visual elements like graphs, charts, and maps.

Topic 9: - Need of data visualization techniques and its advantages

By:- Shaurya Raina (Branch-IT)

1. Which are pros of data visualization?
A. It can be accessed quickly by a wider audience.
B. It can misrepresent information
C. It can be distracting
D. None Of the above
Answer-a
Explanation: Pros of data visualization : it can be accessed quickly by a wider audience.

2.Which are cons of data visualization?
A. It conveys a lot of information in a small space.
B. It makes your report more visually appealing.
C. visual data is distorted or excessively used.
D. None Of the above
Answer- C

Explanation- It can be distracting : if the visual data is distorted or excessively used.

3.Which of the intricate techniques is not used for data visualization?
A. Bullet Graphs
B. Bubble Clouds
C. Fever Maps
D. Heat Maps
Answer-C
Explanation-Fever Maps is not is not used for data visualization instead of that Fever charts is used.

4.Which one of the following is most basic and commonly used techniques?
A. Line charts
B. Scatter plots
C. Population pyramids
D. Area charts
Answers- a
Explanation-Line charts. This is one of the most basic and common techniques used. Line charts display how variables can change over time
5.Which is used to query and edit graphical settings?
A. anova()
B. par()
C. plot()
D. cum()
Answer – b
Explanation-Explanation: par() is used to query and edit graphical settings.

6.Which of the following method make vector of repeated values?
A. rep()
B. data()
C. view()
D. read()
Answer-b
Explanation-data() load (often into a data.frame) built-in dataset.

7.____is used for density plots?
a) par
b) lm
c) kde
d) C
Answer- C
Explanation: kde is used for density plots.

Topic 10: - Stream line & static measures

By:-Shubhda Ghone (Branch-E&tc)

1. Statistical inference is the process of drawing formal conclusions from data.
a) True
b) False
Answer: a
Explanation: Statistical inference requires navigating the set of assumptions and tools.

2. The expected value or ___ of a random variable is the center of its distribution.
a) mode
b) median
c) mean
d) bayesian inference
Answer: c
Explanation: A probability model connects the data to the population using assumptions.

3. Point out the correct statement.
a) Some cumulative distribution function F is non-decreasing and right-continuous
b) Every cumulative distribution function F is decreasing and right-continuous
c) Every cumulative distribution function F is increasing and left-continuous
d) None of the mentioned
Answer: d
Explanation: Every cumulative distribution function F is non-decreasing and right-continuous.

4. Cumulative distribution functions are used to specify the distribution of multivariate random variables.
a) True
b) False
Answer: a
Explanation: In the case of a continuous distribution, it gives the area under the probability density function from minus infinity to x.

Topic 11: - Plot, graphs and networks

By:- Shraddha More (Branch-IT)

1. Which of the following gave rise to need of graphs in data analysis?
a) Data visualization
b) Communicating results
c) Decision making
d) All of the mentioned
Answer: d
Explanation: A picture can tell better story than data

2.Point out the correct statement.
a) coplots are one dimensional data graph
b) Exploratory graphs are made quickly
c) Exploratory graphs are made relatively less in number
d) All of the mentioned
Answer: a
Explanation: coplot is used for two dimensional representation

3.Which of the following graph can be used for simple summarization of data?
a) Scatterplot
b) Overlaying
c) Barplot
d) All of the mentioned
Answer: c
Explanation: A bar chart or bar graph is a chart that presents Grouped data with rectangular bars with lengths proportional to the values that they represent.

4. Spinning plots can be used for two dimensional data.
a) True
b) False

Answer: a
Explanation: There are many ways to create a 3D spinning plot as well

5. Point out the correct combination with regards to kind keyword for graph plotting.
a) 'hist' for histogram
b) 'box' for boxplot
c) 'area' for area plots
d) all of the mentioned

Answer: d
Explanation: The kind keyword argument of plot() accepts a handful of values for plots other than the default Line plot

6. Which of the following plots are used to check if a data set or time series is random?
a) Lag
b) Random
c) Lead
d) None of the mentioned
Answer: a
Explanation: Random data should not exhibit any structure in the lag plot.

Topic 12: - Hierarchies

By:-Divyangi Kolhe (Branch-IT)

1.What is purpose of GetDescendant method in the following code?
DECLARE @parent HierarchyId = HierarchyId::GetRoot()
INSERT INTO H (Node,ID,Name) VALUES (@parent.GetDescendant(NULL,NULL),2,'Johnny')

a) Takes 2 arguments
b) Takes 3 arguments
c) Takes 4 arguments
d) All of the mentioned
Answer: a
Explanation: GetDescendant method takes 2 arguments indicating the left and right nodes on the child level respectively.

2.Which of the following code will not throw an error?
a)DECLARE @child HierarchyId = (SELECT Node FROM H WHERE Name = 'S1')
SELECT * FROM H WHERE Node = @child.GetAnces(2)
b)DECLARE @child HierarchyId = (SELECT Node FROM H WHERE Name = 'S1')
SELECT * FROM H WHERE Node = @child.GetAncestor(2)
c) DECLARE @child HierarchyId = (SELECT Node FROM H WHERE Name = 'S1')
SELECT * FROM H WHERE Node = @child.Ancestor(2)
d) None of the mentioned

Answer: b
Explanation: GetAncestors function returns the ancestors of a specified node in the specified level.

3.Which of the following function will be used in the following code for moving nodes?
DECLARE @newParent HierarchyId = (SELECT Node FROM H WHERE name = 'Johnny')
UPDATE H SET Node = Node.____(Node.GetAncestor(1),@newParent)
WHERE Name = 'S1'
a) GetReparentedVal
b) GetReparentedValue
c) GetValue
d) None of the mentioned

Answer: b
Explanation: The GetReparentedValue function is used to move the nodes to different locations.

4.Which of the following code snippet insert the top level manager 'Jeff Brown' as hierarchy root?

a) INSERT INTO Employees
VALUES (1, 'Jeff Brown', NULL, hierarchyid::Root());
b) INSERT INTO Employees
VALUES (1, 'Jeff Brown', NULL, hierarchyid::GET());

c) INSERT INTO Employees
VALUES (1, 'Jeff Brown', NULL, hierarchyid::GetRoot());
d) INSERT INTO Employees
VALUES (1, 'Jeff Brown', NULL, hierarchy::GetRoot());

Answer: c
Explanation: HIERARCHYID data type provides compact storage and convenient methods to manipulate hierarchies.

5.Point out the correct statement.
a) HierarchyID data type maps the data as a hashmap, so when traversing the binary tree structure
b) In real scenarios, you always need to create indexes using Hierarchy ID data type
c) In HierarchyID, we create indexes in order to make the traversal efficient
d) None of the mentioned

Answer: c
Explanation: In order to make the query execution efficient, we create indexes.

6.Which of the following is invalid code associated with hierarchical data type?

a)
```
CREATE TABLE H
(
Node HierarchyID PRIMARY KEY CLUSTERED,
NodeLevel AS Node.GetLevel(),
ID INT UNIQUE NOT NULL,
Name VARCHAR(50) NOT NULL
 )
```

b)
```
CREATE TABLE H
 (
Node HierarchyID PRIMARY KEY NON CLUSTERED,
NodeLevel AS Node.GetLevel(),
ID INT UNIQUE NOT NULL,
Name VARCHAR(50) NOT NULL
 )
```

c)
```
CREATE TABLE H (
Node HierarchyID FOREIGN KEY CLUSTERED,
NodeLevel AS Node.GetLevel(),
ID INT UNIQUE NOT NULL,
Name VARCHAR(50) NOT NULL
)
```

d) All of the mentioned

Answer: c
Explanation: Node is the column which has the HierarchyID type, NodeLevel is a calculated column which has the level of a particular node. ID and Name are custom columns for additional information.

7.Which of the code deletes node H using hierarchical data type?
a) DELETE FROM H WHERE Name = ‘Steve’
b) DROP FROM H WHERE Name = ‘Steve’
c) DELETE H WHERE Name = ‘Steve’
d) All of the mentioned

Answer: a
Explanation: Deleting a node does not automatically delete the child nodes, this would result in orphaned children.

Done by :- Group 6

Team leader :- Pooja Sharma

Team members:-

1. **Gayatri gadas - IT**
2. **dipali Rajendra salunke - comp**
3. **Anuja bhosale - IT**
4. **Neha Pandhbare- IT**
5. **Muskan Chavan - IT**
6. **Rohini Sawant-IT**
7. **Shrutika Shinde-IT**
8. **Pooja Sharma- E&tc**
9. **Shaurya Raina-IT**
10. **Shubhada ghone-E&tc**
11. **Shraddha more-IT**
12. **Divyangi kolhe- IT**

UNIT 6

DATA VISUALIZATION OF MULTIDIMENSIONAL DATA

By - Aishwarya Muneshwar (branch - IT)

1.Which of the following is the oldest database model?

a.Relational

b.Hierarchical

c.Physical

d.Network

Answer: (d).Network

Explanation:

The network model is a database model conceived as a flexible way of representing objects and their their relationships.

2.SET concept is used in :

a.Network Model

b.Hierarchical Model

c.Relational Model

d.None of these

Answer: (a).Network Model

Explanation:

Records contain fields which need hierarchical organization. Sets are used to define one-to-many relationships between records that contain one owner, many members.

3.The conceptual model is

a.dependent on hardware

b.dependent on software

c.dependent on both hardware and software

d.independent of both hardware and software

Answer: (d).independent of both hardware and software

Explanation:

It does not depend on the DBMS software used to implement the model. It does not depend on the hardware used in the implementation of the model.

4.Which of the following is record based logical model?

a.Network Model

b.Object oriented model

c.E-R Model

d.None of these

Answer: (a).Network Model

Explanation:

Record based logical models are used in describing data at the logical and view levels.

Data in the network model are represented by collections of record and relationships among data are represented by links, which can be viewed as pointers.

5.Which of the following is example of Object based logical model ?

a.Entity Relationship Model

b.Hierarchical Model

c.Relational Model

d.Network Model

Answer: (a).Entity Relationship Model

Explanation:The object-based models use the concepts of entities or objects and relationships among them rather than the implementation-based concepts, such as records, used in the record-based models. Object-based logical models provide flexible structuring capabilities and allow data constraints to be specified explicitly.

6. What is true about data mining?

a) Data Mining is defined as the procedure of extracting information from huge sets of data

b) Data mining also involves other processes such as Data Cleaning, Data Integration, Data Transformation

c) Data mining is the procedure of mining knowledge from data.

d) All of the above

Answer : d)

Explanation: Data Mining is defined as extracting information from huge sets of data. In other words, we can say that data mining is theprocedure of mining knowledge from data. The information or knowledge extracted so that it can be used.

7. Data Mining System Classification consists of?

a) Database Technology

b) Machine Learning

c) Information Science

d) All of the above

Answer : d)

Explanation: A data mining system can be classified according to the following criteria : Database Technology, Statistics,

Machine Learning, Information Science, Visualization, Other Disciplines

8. Which of the following is not a data mining task?

Select one:

a) Feature Subset Selection

b) Associtation

c) Regression

d) Sequential Pattern Discovery

Answer: a) Feature Subset Detection

Explanation: There are a number of data mining tasks such as classification, prediction, time-series analysis,

association, clustering, summarization etc. All these tasks are either predictive data mining tasks or descriptive

data mining tasks. A data mining system can execute one or more of the above specified tasks as part of data mining.

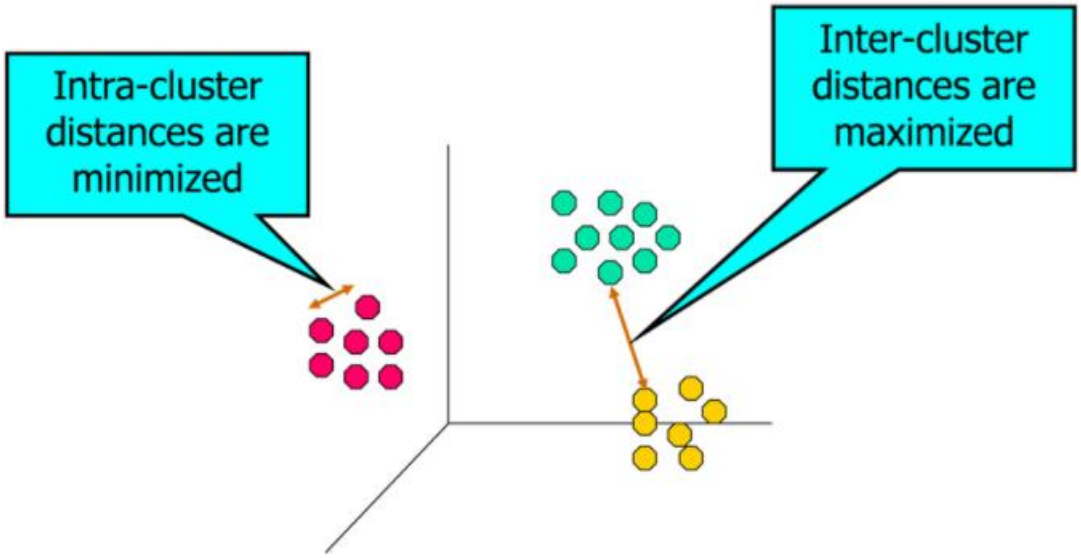
BY – Netra Hinge (branch - IT)

1. **A good clustering is one having :**

A. Low inter-cluster distance and low intra-cluster distance
B. Low inter-cluster distance and high intra-cluster distance
C. High inter-cluster distance and low intra-cluster distance
D. High inter-cluster distance and high intra-cluster distance

Answer: High inter-cluster distance and low intra-cluster distance

Explanation:

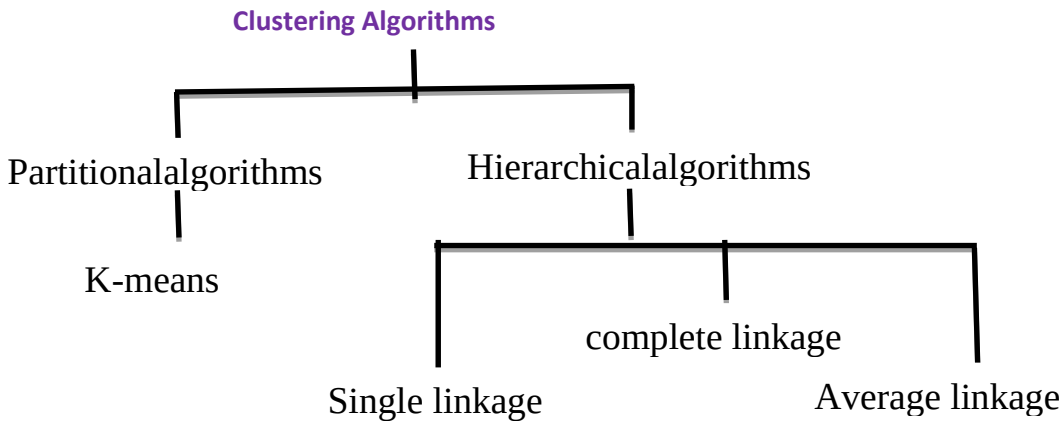


2. **Which of following is a Partitional clustering algorithm?**

A. Single linkage clustering
B. K-means clustering
C. Complete linkage clustering
D. None of above

Answer: K means clustering

Explanation:

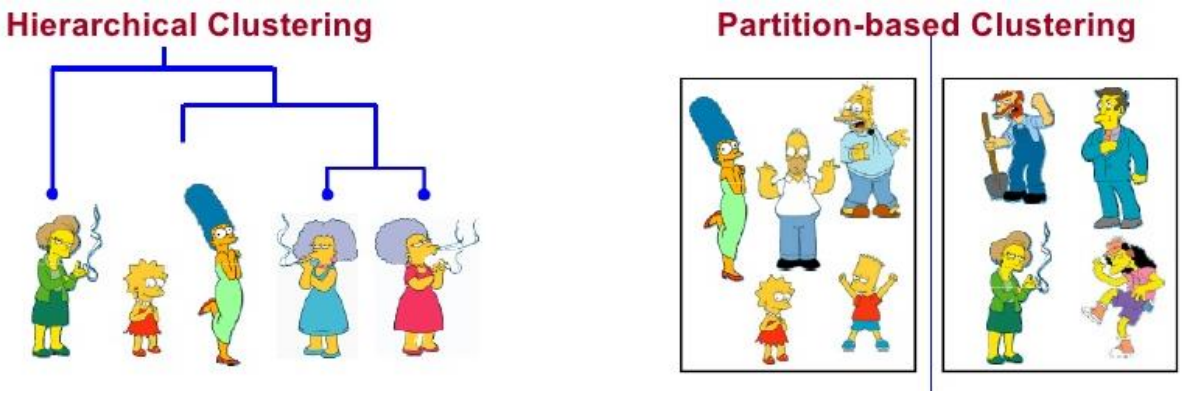


3. Which of following clustering algorithm uses dendogram?

A. Complete linkage clustering
B. K-means clustering
C. Single linkage clustering
D. None of above

Answer: complete linkage clustering

Explanation: Hierarchical clustering algorithms uses dendogram



4. Which of following clustering algorithm uses minimal spanning tree?

E. Complete linkage clustering
F. K-means clustering
G. Single linkage clustering
H. None of above

Answer : Single linkage clustering

Explanation:Single linkage clustering uses minimal spanning tree .

5. Distance between two clusters in single linkage clustering is defined as:

A. Distance between the closest pair of points between clusters
B. Distance between the furthest pair of points between clusters
C. Distance between the most centrally located pair of points between clusters
D. None of above

Answer: Distance between the closest pair of points between clusters

Explanation: Closeness between two clusters is measured by :

1.Distance between centroids of two cluster in average linkage cluster.

2. Distance between the closest pair of points between clusters in single linkage cluster

3. Distance between the furthest pair of points between clusters in complete linkage cluster

6. **Distance between two clusters in complete linkage clustering is defined as:**

A. Distance between the closest pair of points between clusters
B. Distance between the furthest pair of points between clusters
C. Distance between the most centrally located pair of points between clusters
D. None of above

Answer: Distance between the furthest pair of points between clusters

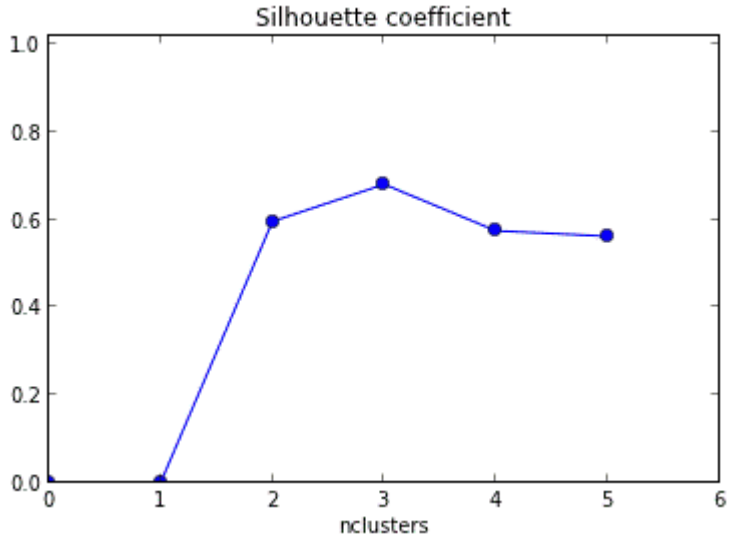
Explanation:Closeness between two clusters is measured by :

1.Distance between centroids of two cluster in average linkage cluster.

2. Distance between the closest pair of points between clusters in single linkage cluster
3. Distance between the furthest pair of points between clusters in complete linkage cluster

By – Nisha Ankam (branch - ENTC)

1.What should be the best choice of no. of clusters based on the following results:



A. 1

B. 2

C. 3

D. 4

Solution: (C)

The silhouette coefficient is a measure of how similar an object is to its own cluster compared to other clusters. Number of clusters for which silhouette coefficient is highest represents the best choice of the number of clusters

2.Which of the following method is used for finding optimal of cluster in K-Mean algorithm?

A. Elbow method

B. Manhattan method

C. Ecludian mehthod

D. All of the above

E. None of these

Solution: (A)

Out of the given options, only elbow method is used for finding the optimal number of clusters. The elbow method looks at the percentage of variance explained as a function of the number of clusters: One should choose a number of clusters so that adding another cluster doesn't give much better modeling of the data.

3. Which of the following algorithm is most sensitive to outliers?

A. K-means clustering algorithm

B. K-medians clustering algorithm

C. K-modes clustering algorithm

D. K-medoids clustering algorithm

Solution: (A)

Out of all the options, K-Means clustering algorithm is most sensitive to outliers as it uses the mean of cluster data points to find the cluster center.

4. **Which of the following can act as possible termination conditions in K-Means?**

1. **For a fixed number of iterations.**

2. **Assignment of observations to clusters does not change between iterations. Except for cases with a bad local minimum.**
3. **Centroids do not change between successive iterations.**
4. **Terminate when RSS falls below a threshold.**

Options:

A. 1, 3 and 4

B. 1, 2 and 3

C. 1, 2 and 4

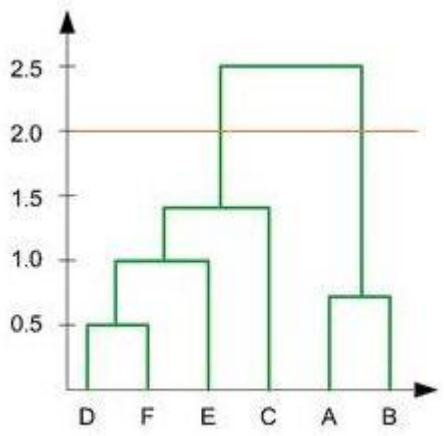
D. All of the above

Solution: (D)

All four conditions can be used as possible termination condition in K-Means clustering:

1. This condition limits the runtime of the clustering algorithm, but in some cases the quality of the clustering will be poor because of an insufficient number of iterations.
2. Except for cases with a bad local minimum, this produces a good clustering, but runtimes may be unacceptably long.
3. This also ensures that the algorithm has converged at the minima.
4. Terminate when RSS falls below a threshold. This criterion ensures that the clustering is of a desired quality after termination. Practically, it's a good practice to combine it with a bound on the number of iterations to guarantee termination.

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



A. 1

B. 2

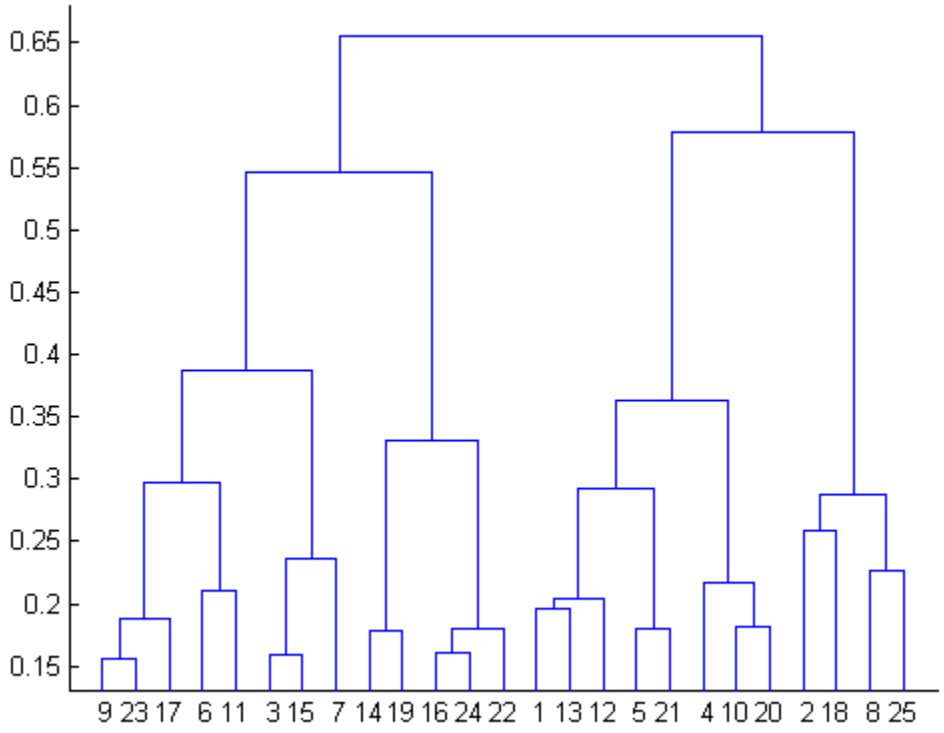
C. 3

D. 4

Solution: (B)

Since the number of vertical lines intersecting the red horizontal line at y=2 in the dendrogram are 2, therefore, two clusters will be formed.

6. What is the most appropriate no. of clusters for the data points represented by the following dendrogram:



A. 2

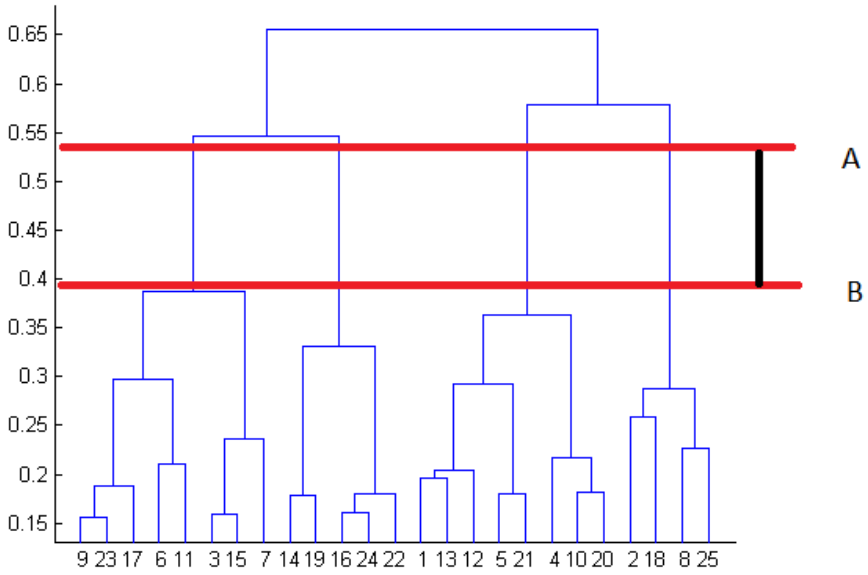
B. 4

C. 6

D. 8

Solution: (B)

The decision of the no. of clusters that can best depict different groups can be chosen by observing the dendrogram. The best choice of the no. of clusters is the no. of vertical lines in the dendrogram cut by a horizontal line that can transverse the maximum distance vertically without intersecting a cluster.



In the above example, the best choice of no. of clusters will be 4 as the red horizontal line in the dendrogram below covers maximum vertical distance AB.

By – Ishika Bagdiya(branch - ENTC)

1.The process of viewing the cross-tab (Single dimensional) with a fixed value of one attribute is

a) Slicing

b) Dicing

c) Pivoting

d) Both Slicing and Dicing

Answer : a

Explanation: The slice operation selects one particular dimension from a given cube and provides a new sub-cube. Dice selects two or more dimensions from a given cube and provides a new sub-cube.

2.OLAP stands for

a) Online analytical processing

b) Online analysis processing

c) Online transaction processing

d) Online aggregate processing

Answer : a

Explanation: OLAP is the manipulation of information to support decision making.

3.Data modeling technique used for data marts is

a) Dimensional modeling

b) ER – model

c) Extended ER – model

d) Physical model

e) Logical model

Answer : a

Explanation: A Data modeling technique used for data marts is Dimensional modeling.

4.The Synonym for data mining is

a) Data warehouse

b) Knowledge discovery in database

c) ETL

d) Business intelligence

e) OLAP

Answer : b

Explanation: The synonym for data mining is Knowledge discovery in Database.

5.Most common kind of queries in a data warehouse

a) Inside-out queries

b) Outside-in queries

c) Browse queries

d) Range queries

e) All (a), (b), (c) and (d) above

Answer : a

Explanation: The Most common kind of queries in a data warehouse is Inside-out queries.

Inside - out query

Outside – in query

6.Multiple Regression means

a) Data are modeled using a straight line

b) Data are modeled using a curve line

c) Extension of linear regression involving only one predicator value

d) Extension of linear regression involving more than one predicator value

e) All (a), (b), (c) and (d) above

Answer : d

Explanation: Multiple Regression means extension of linear regression involving more than one predicator value.

By – Megha Argade (branch - ENTC)

1)Which method shows hierarchical data in a nested format?

A. Treemaps

B. Scatter plots

C. Population pyramids

D. Area charts

Ans : A

Explanation: Treemaps are best used when multiple categories are present, and the goal is to compare different parts of a whole.

2)Which of the following plots are often used for checking randomness in time series?

A. Autocausation

B. Autorank

C. Autocorrelation

D. None of the above

Ans : C

Explanation: If the time series is random, such autocorrelations should be near zero for any and all time-lag separations.

3)Which are pros of data visualization?

A. It can be accessed quickly by a wider audience.

B. It can misrepresent information

C. It can be distracting

D. None Of the above

Ans : A

Explanation: Pros of data visualization : it can be accessed quickly by a wider audience.

4)Which are cons of data visualization?

A. It conveys a lot of information in a small space.

B. It makes your report more visually appealing.

C. visual data is distorted or excessively used.

D. None Of the above

Ans : C

Explanation: It can be distracting : if the visual data is distorted or excessively used.

5)Which of the intricate techniques is not used for data visualization?

A. Bullet Graphs

B. Bubble Clouds

C. Fever Maps

D. Heat Maps

Ans : C

Explanation: Fever Maps is not is not used for data visualization instead of that Fever charts is used.

6)Which of the following gave rise to need of graphs in data analysis?

a) Data visualization

b) Communicating results

c) Decision making

d) All of the mentioned

Answer: d

Explanation: A picture can tell better story than data.

By – Swapnali Kaldhone (branch - comp)

1)[True or False]Data warehouses and OLAP tools are based on Multidimensional data model.

True

False

Answer: A

Explanation : Multidimensional data model represents data in the form of data cubes. It allows Data warehouse & OLAP tools to model and view the data in multiple dimensions .

2) The most popularly used dimensionality reduction algorithm is Principal Component Analysis (PCA). Which of the following is/are true about PCA?

1 . PCA is an unsupervised method

2 . It searches for the directions that data have the largest variance

3 . Maximum number of principal components <= number of features

4 . All principal components are orthogonal to each other

A. 1 and 2

B. 2 & 3

C. 1,3 & 4

D. All of the above

Answer : D

Explanation: options are self explanatory.

3) [True or False] PCA can be used for projecting and visualizing data in lower dimensions.

A. True

B. False

Answer : A

Explanation : Sometimes it is very useful to plot the data in lower dimensions. We can take the first 2 principal components and then visualize the data using scatter plot.

4) Which of the following algorithm is most sensitive to outliers?

A. K-means clustering algorithm

B. K-medians clustering algorithm

C. K-modes clustering algorithm

D. K-medoids clustering algorithm

Answer : A

Explanation : Out of all the options, K-Means clustering algorithm is most sensitive to outliers as it uses the mean of cluster data points to find the cluster centre.

5) Which of the following clustering algorithms suffers from the problem of convergence at local optima?

1 . K- Means clustering algorithm

2 . Agglomerative clustering algorithm

3 . Expectation-Maximization clustering algorithm

4 . Diverse clustering algorithm

A. 1 only

B. 2 and 3

C. 2 and 4

D. 1 and 3

E. All of the above

Answer : D

Explanation: of the options given, only K-Means clustering algorithm and EM clustering algorithm has the drawback of converging at local minima.

6) What is the minimum no. of variables/ features required to perform clustering?

A. 0

B. 1

C. 2

D. 3

Answer : B

Explanation : At least a single variable is required to perform clustering analysis. Clustering analysis with a single variable can be visualized with the help of a histogram

By – Anusuiya Parihar (branch - ENTC)

1. The generalization of cross-tab which is represented visually is ____ which is also called as data cube.

a) Two dimensional cube

b) Multidimensional cube

c) N-dimensional cube

d) Cuboid

Answer: a

Explanation: Each cell in the cube is identified for the values for the three dimensional attributes.

2. The operation of moving from finer-granularity data to a coarser granularity (by means of aggregation) is called a ____

a) Rollup

b) Drill down

c) Dicing

d) Pivoting

Answer: a

Explanation: The opposite operation—that of moving from coarser-granularity data to finer-granularity data—is called a drill down.

3. Data visualization is also an element of the broader _____.

a) Deliver presentation architecture

b)Data presentation architecture

c) Dataset presentation architecture

d)Data process architecture

Ans : b

Explanation: Data visualization is also an element of the broader data presentation architecture (DPA) discipline, which aims to identify, locate, manipulate, format and deliver data in the most efficient way possible.

4. Which of the following is required by K-means clustering?

a) defined distance metric

b) number of clusters

c) initial guess as to cluster centroids

d) all of the mentioned

Answer: d

Explanation: K-means clustering follows partitioning approach.

5. K-means is not deterministic and it also consists of number of iterations.

a) True

b) False

Answer: a

Explanation: K-means clustering produces the final estimate of cluster centroids.

6. PCA works better if there is?

A linear structure in the data

If the data lies on a curved surface and not on a flat surface

If variables are scaled in the same unit

A. 1 and 2

B. 2 and 3

C. 1 and 3

D. 1 ,2 and 3

Answer: c

Explanation: Option C is correct

By – aayushi savaldekar (branch - ENTC)

1.Which of the following is not a level of data abstraction?

a.Physical Level

b.Critical Level

c.Logical Level

d.View Level

Answer -critical level

Explanation-There are mainly three levels of data abstraction: Internal Level: Actual PHYSICAL storage structure and access paths. Conceptual or Logical Level: Structure and constraints for the entire database. External or View level: Describes various user views

2.Which of the following is a Data Model?

a.Entity-Relationship model

b.Relational data model

c.Object-Based data model

d. All of the above

Answer- All the above

Explanation- Basically this all are the types of data model

3.A logical description of some portion of database that is required by a user to perform task is called

a.System View

b.User View

c.Logical View

d. Data View

Answer – User view

Explanation- DBMS is used to create and maintain the data base

4. Data warehouse architecture is based on ……..

A) DBMS

B) RDBMS

C) Sybase

D) SQL Server

Answer – RDBMS

Explanation- The Data Warehouse is based on RDBMS server which is a central information repository that is surrounded by some key data warehousing components to make the entire environment functional , manageable and accessible

5.Data warehouse contains ……. Data that is never found in the operational environment.

A) normalized

B) informational

C) summary

D) denormalized

Answer – Summary

Explanation- Data Warehouse is a relational database that is designed for query and analysis rather than for transaction processing .It usually contains historical data derived from transaction data , but it can include data from other sources.

6…….. are designed to overcome any limitations placed on the warehouse by the nature of the relational data model.

A) Operational database

B) Relational database

C) Multidimensional database

D) Data repository

ANSWERS: Multidimensional database

Explanation – A multidimensional database is a type of database that is optimized for data warehouse and online analytical processing applications. Conceptually a multidimensional database uses the idea of a data cube to represent the dimensions of data available to a user.

By - Varsha Dhope (branch - IT)

1)Which of the following techniques would perform better for reducing dimensions of a data set?

A. Removing columns which have too many missing values

B. Removing columns which have high variance in data

C. Removing columns with dissimilar data trends

D. None of these

Answer: A

Explanation:If columns have too many missing values, (say 99%) then we can remove such columns.

2)What happens when you get features in lower dimensions using PCA?

1.The features will still have interpretability

2.The features will lose interpretability

3.The features must carry all information present in data

4.The features may not carry all information present in data

A. 1 and 3

B. 1 and 4

C. 2 and 3

D. 2 and 4

Answer:D

Explanation:When you get the features in lower dimensions then you will lose some information of data most of the times and you won't be able to interpret the lower dimension data.

3)Which of the following option(s) is / are true?

1.You need to initialize parameters in PCA

2.You don't need to initialize parameters in PCA

3.PCA can be trapped into local minima problem

4.PCA can't be trapped into local minima problem

A. 1 and 3

B. 1 and 4

C. 2 and 3

D. 2 and 4

Answer:D

Explanation:PCA is a deterministic algorithm which doesn't have parameters to initialize and it doesn't have local minima problem like most of the machine learning algorithms has.

4)Under which condition SVD and PCA produce the same projection result?

A. When data has zero median

B. When data has zero mean

C. Both are always same

D. None of these

Answer:B

Explanation:When the data has a zero mean vector, otherwise you have to center the data first before taking SVD.

5)Which of the following algorithms cannot be used for reducing the dimensionality of data?

A. t-SNE

B. PCA

C. LDA False

D. None of these

Answer:D

Explanation:All of the algorithms are the example of dimensionality reduction algorithm.

6)Dimensionality reduction algorithms are one of the possible ways to reduce the computation time required to build a model.

A. TRUE

B. FALSE

Answer:A

Explanation:Reducing the dimension of data will take less time to train a model.

By – Vaishnavi Dasare (Branch - COMP)

1. Which of the following combination is incorrect?
a) Continuous – euclidean distance
b) Continuous – correlation similarity
c) Binary – manhattan distance
d) None of the mentioned
Answer: d
Explanation: You should choose a distance/similarity that makes sense for your problem
.

2.Which of the following function is used for k-means clustering?
a) k-means
b) k-mean
c) heatmap
d) none of the mentioned
Answer: a
Explanation: K-means requires a number of clusters.

3. Imagine, you have 1000 input features and 1 target feature in a machine learning problem. You have to select 100 most important features based on the relationship between input features and the target features.Do you think, this is an example of dimensionality reduction?

A. Yes

B. No

Solution: (A)

Explaination : Dimensionality reduction is the process of reducing the number of random variables under consideration, by obtaining a set of principal variables.

4.Suppose we are using dimensionality reduction as pre-processing technique, i.e, instead of using all the features, we reduce the data to k dimensions with PCA. And then use these PCA projections as our features. Which of the following statement is correct?
A. Higher 'k' means more regularization
B. Higher 'k' means less regularization
C. Can't Say
Answer:B
Explaination:Higher k would lead to less smoothening as we would be able to preserve more characteristics in data, hence less regularization.

5. What will happen when eigenvalues are roughly equal?

A. PCA will perform outstandingly

B. PCA will perform badly

C. Can't Say

D.None of above

Solution: (B)

When all eigen vectors are same in such case you won't be able to select the principal components because in that case all principal components are equal.

6. What is true about Data Visualization?

A. Data Visualization is used to communicate information clearly and efficiently to users by the usage of information graphics such as tables and charts.

B. Data Visualization helps users in analyzing a large amount of data in a simpler way.

C. Data Visualization makes complex data more accessible, understandable, and usable.

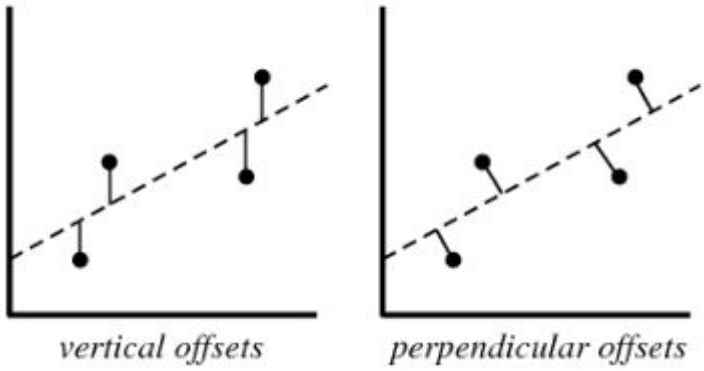
D. All of the above

Ans : D

Explanation: Data Visualization is used to communicate information clearly and efficiently to users by the usage of information graphics such as tables and charts. It helps users in analyzing a large amount of data in a simpler way. It makes complex data more accessible, understandable, and usable.

By - Sanskruti Dhamale (branch - IT)

1. Which of the following offset, do we consider in PCA?



A. Vertical offset

B. Perpendicular offset

C. Both

D. None of these

Answer - B

Explaination : We always consider residual as vertical offsets. Perpendicular offset are useful in case of PCA

2. Which of the following is finally produced by Hierarchical Clustering?

a) final estimate of cluster centroids

b) tree showing how close things are to each other

c) assignment of each point to clusters

d) all of the mentioned

Answer: b

Explanation: Hierarchical clustering is an agglomerative approach.

3. Which of the following clustering requires merging approach?

a) Partitional

b) Hierarchical

c) Naive Bayes

d) None of the mentioned

Answer: b

Explanation: Hierarchical clustering requires a defined distance as well.

4.[True or False] It is not necessary to have a target variable for applying dimensionality reduction algorithms.

A. TRUE

B. FALSE

Solution: (A)

Explaination : LDA is an example of supervised dimensionality reduction algorithm

5. Which of the following options are correct, when you are applying PCA on a image dataset?

It can be used to effectively detect deformable objects.

It is invariant to affine transforms.

It can be used for lossy image compression.

It is not invariant to shadows.

A. 1 and 2

B. 2 and 3

C. 3 and 4

D. 1 and 4

Solution: (C)

Option C is correct

6.Hierarchical clustering should be primarily used for exploration.

a) True

b) False

Answer: a

Explanation: Hierarchical clustering is deterministic.

By- Achal Dudhbhate (branch - ENTC)

Q.1 _____ helps in designing effective tables and charts for data visualization.

A) Data-ink ratio

B) Crosstabulation

C) PivotTable

D) Scatter charts

Ans. A

Explanation: One of the most helpful ideas for creating effective tables and charts for data visualization is the idea of the data-ink ratio.

Q.2 Hierarchical clustering should be primarily used for exploration.

a) True

b) False

Answer: a

Explanation: Hierarchical clustering is deterministic.

Q.3 Which of the following is required by K-means clustering?

a) defined distance metric

b) number of clusters

c) initial guess as to cluster centroids

d) all of the mentioned

Answer: d

Explanation: K-means clustering follows partitioning approach

Q.4 Point out the wrong statement.

a) k-means clustering is a method of vector quantization

b) k-means clustering aims to partition n observations into k clusters

c) k-nearest neighbor is same as k-means

d) none of the mentioned

Answer: c

Explanation: k-nearest neighbor has nothing to do with k-means.

Q.5 Which of the following clustering requires merging approach?

a) Partitional

b) Hierarchical

c) Naive Bayes

d) None of the mentioned

Answer: b

Explanation: Hierarchical clustering requires a defined distance as well.

Q.6 A data visualization tool that updates in real time and gives multiple outputs is called

A)a data table.

B)a metrics table.

C)the GIS.

D)a data dashboard.

Ans. D) A data dashboard

Explaination : Is a data-visualization tool that illustrates multiple metrics and automatically updates these metrics as new data become available