**UNIT 1**

Binary attribute are

A) This takes these values will be 0 and 1 and they can be coded as one bit

B) The natural environment of a certain species

C) Systems that can be used without knowledge of internal operations

D) None of these

ANSWER: A

A definition of a concept is-----if it recognizes all the instances of that concept

A) Complete

B) Consistent

C) Constant

D) None of these

ANSWER: A

Data mining is

A) The actual discovery phase of a knowledge discovery process

B) The stage of selecting the right data for a KDD process

C) A subject-oriented integrated time variant non-volatile collection of data in support of management

D) None of these

ANSWER: A

Data selection is

A) The actual discovery phase of a knowledge discovery process

B) The stage of selecting the right data for a KDD process

C) A subject-oriented integrated time variant non-volatile collection of data in support of management

D) None of these

ANSWER: B

Discovery is

A) It is hidden within a database and can only be recovered if one is given certain clues

B) The process of executing implicit previously unknown and potentially useful information from data

C) An extremely complex molecule that occurs in human chromosomes and that carries genetic information in the form of genes.

D) None of these

ANSWER: B

Euclidean distance measure is

A) stage of the KDD process in which new data is added to the existing selection.

B) The process of finding a solution for a problem simply by enumerating all possible solutions according to some pre-defined order and then testing them

C) The distance between two points as calculated using the Pythagoras theorem

D) None of these

ANSWER: C

Hidden knowledge referred to

A) set of databases from different vendors, possibly using different database paradigms

B) An approach to a problem that is not guaranteed to work but performs well in most cases

C) Information that is hidden in a database and that cannot be recovered by a simple SQL query.

D) None of these

ANSWER: C

Heterogeneous databases referred to

A) set of databases from different b vendors, possibly using different database paradigms

B) An approach to a problem that is not guaranteed to work but performs well in most cases.

C) Information that is hidden in a database and that cannot be recovered by a simple SQL query.

D) None of these

ANSWER: A

Enumeration is referred to

A) stage of the KDD process in which new data is added to the existing selection.

B) The process of finding a solution for a problem simply by enumerating all possible solutions according to some pre-defined order and then testing them

C) The distance between two points as calculated using the Pythagoras theorem.

D) None of these

ANSWER: B

Heuristic is

A) set of databases from different vendors, possibly using different database paradigms

B) An approach to a problem that is not guaranteed to work but performs well in most cases

C) Information that is hidden in a database and that cannot be recovered by a simple SQL query.

D) None of these

ANSWER: B

Hybrid learning is

A) Machine-learning involving different techniques

B) The learning algorithmic analyzes the examples on a systematic basis 2nd makes incremental adjustments to the theory that is learned

C) Learning by generalizing from examples

D) None of these

ANSWER: A

Knowledge engineering is

A) The process of finding the right formal representation of a certain body of knowledge in order to represent it in a knowledge-based system

B) It automatically maps an external signal space into a system's internal representational space. They are useful in the performance of classification tasks.

C) A process where an individual learns how to carry out a certain task when making a transition from a situation in which the task cannot be carried out to a situation in which the same task under the same circumstances can be carried out.

D) None of these

ANSWER: A

Information content is

A) The amount of information with in data as opposed to the amount of redundancy or noise

B) One of the defining aspects of a data warehouse

C) Restriction that requires data in one column of a database table to the a subset of another-column.

D) None of these

ANSWER: A

KDD (Knowledge Discovery in Databases) is referred to

A) Non-trivial extraction of implicit previously unknown and potentially useful information from data

B) Set of columns in a database table that can be used to identify each record within this table uniquely.

C) collection of interesting and useful patterns in a database

D) none of these

ANSWER: A

Machine learning is

A) An algorithm that can learn

B) A sub-discipline of computer science that deals with the design and implementation of learning algorithms

C) An approach that abstracts from the actual strategy of an individual algorithm and can therefore be applied to any other form of machine learning.

D) None of these

ANSWER: B

Multi-dimensional knowledge is

A) class of learning algorithms that try to derive a Prolog program from examples

B) A table with n independent attributes can be seen as an n-dimensional space

C) A prediction made using an extremely simple method, such as always predicting the same output.

D) None of these

ANSWER: B

Prediction is

A) The result of the application of a theory or a rule in a specific case

B) One of several possible enters within a database table that is chosen by the designer as the primary means of accessing the data in the table.

C) Discipline in statistics that studies ways to find the most interesting projections of multi-dimensional spaces.

D) None of these

ANSWER: A

Operational database is

A) A measure of the desired maximal complexity of data mining algorithms

B) A database containing volatile data used for the daily operation of an organization

C) Relational database management system

D) None of these

ANSWER: B

Noise is

A) A component of a network

B) In the context of KDD and data mining, this refers to random errors in a database table.

C) One of the defining aspects of a data warehouse

D) None of these

ANSWER: B

Computers are best at learning

A) facts.

B) concepts.

C) procedures.

D) principles.

ANSWER: B

Like the probabilistic view, the \_\_\_\_\_\_\_\_ view allows us to associate a probability of membership with each classification.

A) exemplar

B) deductive

C) classical

D) inductive

ANSWER: A

Data used to build a data mining model.

A) validation data

B) training data

C) test data

D) hidden data

ANSWER: B

Supervised learning and unsupervised clustering both require at least one

A) hidden attribute.

B) output attribute.

C) input attribute.

D) categorical attribute.

ANSWER: C

Database query is used to uncover this type of knowledge.

A) deep

B) hidden

C) shallow

D) multidimensional

ANSWER: C

A statement to be tested.

A) theory

B) procedure

C) principle

D) hypothesis

ANSWER: D

A person trained to interact with a human expert in order to capture their knowledge.

A) knowledge programmer

B) knowledge developer

C) knowledge engineer

D) knowledge extractor

ANSWER: C

Which of the following is not a characteristic of a data warehouse?

A) contains historical data

B) designed for decision support

C) stores data in normalized tables

D) Promotes data redundancy

ANSWER: C

A structure designed to store data for decision support.

A) operational database

B) flat file

C) decision tree

D) data warehouse

ANSWER: D

If a customer is spending more than expected, the customer�s intrinsic value is \_\_\_\_\_\_\_\_ their actual value.

A) greater than

B) less than

C) less than or equal to

D) equal to

ANSWER: B

The most laborious and time consuming step in the overall Knowledge discovery process is

A) Data preprocessing

B) Data postprocessing

C) Data mining

D) All are correct

ANSWER: A

An example of post processing is

A) Feature selection

B) Feature Extraction

C) **Visualization**

D) None of the above

ANSWER: C

To remove noise, inconsistent data and irrelevant data called

A) Data integration

B) **Data Cleaning**

C) Data mining

D) Data selection

ANSWER: B

Traditional Techniques may be unsuitable due to

A) Homogeneous data

B) Low dimensionality

C) **Enormity of data**

D) All the above

ANSWER: C

Motivating challenges in data mining are

A) Scalability

B) High Dimensionality

C) Data Ownership and Distribution

D) **All the above**

ANSWER: D

Data mining has also been quick to adopt ideas from other areas including

A) Optimization

B) Evolutionary computing

C) Information theory

D) **All the above**

ANSWER: D

An ............. is a property or characteristic of an object

A) **Attribute**

B) Object

C) Record

D) Pattern

ANSWER: A

A data set is collection of .......

A) Tuples

B) Instance

C) Objects

D) **All the Options**

ANSWER: D

Measurement scale is a function that associates a ..............

A) **Numerical or symbolic value**

B) Symbolic

C) Numerical

D) Floating values

ANSWER: A

Object is also known as

A) Dimension

B) **Observation**

C) Characteristic

D) Field

ANSWER: B

The task of identifying observations whose characteristics are significantly different from the rest of the data. Such observations are known as .............

A) **Anomalies or Outliers**

B) Outliers

C) Anomalies

D) None

ANSWER: A

Real numbers as attribute values called.......

A) Categorical Attribute

B) Dicrete Attribute

C) Binary Attribute

D) **Continuous Attribute**

ANSWER: D

The difficulties are associated with analyzing high dimensional data are referred as.........

A) Sampling

B) Association

C) **Curse of Dimensionality**

D) Clustering

ANSWER: C

Sequential data is also called as..............

A) **Temporal Data**

B) Time series data

C) Spatial data

D) Genetic sequence data

ANSWER: A

Numerical difference of the measured and true value is called...

A) Dimension

B) **Error**

C) Field

D) Observations

ANSWER: B

A small crack in the lens of a camera produces images with distortion or a mark at that place called as.............

A) **Artifacts**

B) Noise

C) Error

D) Measurement

ANSWER: A

The closeness of measurements to the true value of the quantity being measured called as.......

A) Sampling

B) Precision

C) Bias

D) **Accuracy**

ANSWER: D

Issues related to applications are:

A) Timeliness

B) Relevance

C) Knowledge about the data

D) **All**

ANSWER: D

Objects are not removed from the population as they are selected for the sample is called...

A) Sampling without replacement

B) **Sampling with replacement**

C) Sampling

D) None

ANSWER: B

Process of transforming a continuous attribute into a categorical attribute called as.............

A) **Discretization**

B) Binarization and Discretization

C) Binarization

D) None

ANSWER: A

Numerical measure of the degree to which the 2 objects are alike called as.......

A) Dissimilarity

B) Dicrete

C) **Similarity**

D) All Options

ANSWER: C

SMC stands for:

A) Sample Matching Coefficient

B) Simple Matching Coordinate

C) **Simple Matching Coefficient**

D) Simple Match Coefficient

ANSWER: C

Visualization Techniques are:

A) Scatter plots

B) Contour plots

C) Histograms

D) **All the options**

ANSWER: D

Process of transforming a continuous attribute into a categorical attribute called as.............

A) **Discretization**

B) Binarization and Discretization

C) Binarization

D) None

ANSWER: A

Numerical measure of the degree to which the 2 objects are alike called as.......

A) Dissimilarity

B) Dicrete

C) **Similarity**

D) All Options

ANSWER: C

The .......... is defined as one of the most common measure of document similarity.

A) **Cosine similarity**

B) Correlation

C) SMC

D) Bregman Divergence

ANSWER: A

Visualization Techniques are:

A) Scatter plots

B) Contour plots

C) Histograms

D) **All the options**

ANSWER: D

data mining algorithm is unstable if

A) test set accuracy depends on the ordering of test set instances.

B) the algorithm builds models unable to classify outliers.

C) **the algorithm is highly sensitive to small changes** **in the training data.**

D) test set accuracy depends on the choice of input attributes.

ANSWER: C

The correlation coefficient for two real-valued attributes is �0.85. What does this value tell you

A) The attributes are not linearly related

B) **As the value of one attribute decreases the value of the second attribute increases**

C) As the value of one attribute increases the value of the second attribute also increases

D) The attributes show a linear relationship

ANSWER: B

Which of the following is not a characteristic of a data warehouse?

A) contains historical data

B) designed for decision support

C) **stores data in normalized tables**

D) promotes data redundancy

ANSWER: C

KDD has been described as the application of \_\_\_ to data mining.

A) the waterfall model

B) object-oriented programming

C) **the scientific method**

D) procedural intuition

ANSWER: C

.......................is useful for discovering interesting relationships hidden in large data set

A) **Association Analysis**

B) Cluster Analysis

C) Classification

D) None

ANSWER: A

The uncovered relationships can be represented in the form of........................

A) Sets of frequent items

B) Collection of frequent items

C) Association Rules

D) **All Options**

ANSWER: D

The number of transactions that contain a particular Itemset

A) **Support Count**

B) Confidence

C) Support

D) None

ANSWER: A

Apriori is the first association rule mining algorithm that pioneered the use of...........

A) None

B) Support and Confidence based pruning

C) Cnfidence based pruning

D) **Support based pruning**

ANSWER: D

Which of the following issue is considered before investing in Data Mining?

A) Functionality

B) Vendor consideration

C) Compatibility

D) **All of the mentioned**

ANSWER: D

Apriori Principle uses .............. property

A) monotype property

B) Anti-monotype property

C) Monotone property

D) **Anti-monotone property**

ANSWER: D

Apriori generation function uses .............. candidate generation procedure

A) Brute force method

B) Fk-1 \* F1 method

C) **Fk-1 \* Fk-1 method**

D) All the option

ANSWER: C

Computational Complexity of Apriori Algorithm can be affected by the following factors...

A) Support Threshold

B) Dimensionality

C) Average Transaction Width

D) **All the option**

ANSWER: D

A collection of zero or more items is termed as an ....

A) N-Itemset

B) Set

C) Item

D) **Itemset**

ANSWER: D

Each frequent k-Itemset, Y, can produce up to ..............association rules

A) 2k +2

B) 2k

C) 2k -1

D) **2k -2**

ANSWER: D

Unlike the support measure, confidence does not have .............. property

A) monotype property

B) Anti-monotype property

C) Monotonic property

D) **Monotone property**

ANSWER: D

The number of iterations in apriori \_\_\_\_\_\_\_\_\_\_\_ Select one:

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

.....................are useful for removing some of the redundant association rules.

A) Minimum frequent itemsets

B) Maximal frequent itemsets

C) **Closed frequent itemsets**

D) All the option

ANSWER: C

The Apriori algorithm uses a ............ search strategy

A) Specific-to-specific

B) General-to-general

C) Specific-to-General

D) **General-to-specific**

ANSWER: D

In Apriori algorithm, if 1 item-sets are 100, then the number of candidate 2 item-sets are Select one:

A) 100

B) **4950**

C) 200

D) 5000

ANSWER: B

The most general form of distance is Select one:

A) Manhattan

B) Eucledian

C) Mean

D) **Minkowski**

ANSWER: D

The................ strategy is useful to discover maximal frequent itemsets

A) **Specific-to-general**

B) General-to-general

C) Specific-to-specific

D) General-to-general

ANSWER: A

.................can also be defined according to the prefix or suffix labels of an itemset.

A) **Equivalence classes**

B) Inequivalence classes

C) Frequent Classes

D) All options

ANSWER: A

The Apriori algorithm traverses the lattice in a ........

A) Both Breadth and Depth First

B) None

C) Depth-first manner

D) **Breadth-first manner**

ANSWER: D

Data Visualization in mining cannot be done using Select one:

A) **Photos**

B) Graphs

C) Charts

D) Information Graphics

ANSWER: A

Dimensionality reduction reduces the data set size by removing \_\_\_\_\_\_\_\_\_ Select one:

A) composite attributes

B) derived attributes

C) relevant attributes

D) **irrelevant attributes**

ANSWER: D

Incorrect or invalid data is known as \_\_\_\_\_\_\_\_\_ Select one:

A) Missing data

B) Outlier

C) Changing data

D) **Noisy data**

ANSWER: D

....................algorithm is a tree based Algorithm.

A) None

B) Both FP-growth and Apriori

C) **FP-Growth**

D) Apriori

ANSWER: C

Objective interestingness measure include...............

A) Support

B) Confidence

C) Correlation

D) **All options**

ANSWER: D

Unlike traditional production rules, association rules

A) **allow the same variable to be an input attribute in one rule and an output attribute in another rule.**

B) allow more than one input attribute in a single rule.

C) require input attributes to take on numeric values.

D) require each rule to have exactly one categorical output attribute.

ANSWER: A

Association rule support is defined as

A) the percentage of instances that contain the antecendent conditional items listed in the association rule.

B) the percentage of instances that contain the consequent conditions listed in the association rule.

C) **the percentage of instances that contain all items listed in the association rule.**

D) the percentage of instances in the database that contain at least one of the antecendent conditional items listed in the association rule.

ANSWER: C

Confidence is the conditional probability that a randomly selected Transaction will include all the items in the transaction includes all the items in the antecedent

A) **True.**

B) false

ANSWER: A

Which data mining method is used to analyse transaction data?

A) clustering

B) classification

C) **Association analysis**

D) visualization

ANSWER: C

The Apriori algorithm traverses the lattice in a ........

A) Both Breadth and Depth First

B) None

C) Depth-first manner

D) **Breadth-first manner**

ANSWER: D

Unlike the support measure, confidence does not have .............. property

A) monotype property

B) Anti-monotype property

C) Monotonic property

D) **Monotone property**

ANSWER: D

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

ANSWER: A

2.Decision Tree is a display of an algorithm.

A) **True**

B) False

ANSWER: A

3.Decision Trees can be used for Classification Tasks.

A) **True**

B) False

ANSWER: A

4. 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) **Flow-Chart & Structure in which internal node represents test on an attribute, each branch represents outcome of test and each leaf node represents class label**

D) None of the mentioned

ANSWER: C

5. Decision Nodes are represented by \_\_\_\_\_\_\_\_\_\_\_\_

A) Disks

B) **Squares**

C) Circles

D) Triangles

ANSWER: B

6. 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**

ANSWER: D

7. Which of the following is a widely used and effective machine learning algorithm based on the idea of bagging?

A) Decision Tree

B) Regression

C) Classification

D) **Random Forest**

ANSWER: D

8. The most widely used metrics and tools to assess a classification model are:

A) Confusion matrix

B) Cost-sensitive accuracy

C) Area under the ROC curve

D) **All of the above**

ANSWER: D

9. Which of the following is a disadvantage of decision trees?

A) Factor analysis

B) Decision trees are robust to outliers

C) **Decision trees are prone to be overfit**

D) None of the above

ANSWER: C

10. Which of the following is true about Naive Bayes ?

A) Assumes that all the features in a dataset are equally important

B) Assumes that all the features in a dataset are independent

C) **Both A and B**

D) None of the above options

ANSWER: C

11. \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ is used to find interesting relationships in large data sets

A) Classification

B) Regression

C) Clustering

D) **Association analysis**

ANSWER: D

12. A \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ is a decision support tool that uses a tree like graph or model of decisions and their possible consequences, including chance event outcomes

A) **Decision tree**

B) Graphs

C) Trees

D) Neural Network

ANSWER: A

13. Decision Trees can be used for Classification Tasks.

A) **True**

B) False

ANSWER: A

14. Impurity measures used by decision tress

A) Entropy

B) Gini Index

C) classification error

D) **All of the above**

ANSWER: D

15. A pure data set has entropy value of

A) infinity

B) **zero**

C) one

D) less than one

ANSWER: B

16. An impure set has entropy value of

A) less than zero

B) zero

C) **greater than zero but less than or equal to one**

D) none of the above

ANSWER: C

17. A pure data set has Gini Index value of

A) infinity

B) **zero**

C) one

D) less than one

ANSWER: B

18. An impure set has Gini Index value of

A) less than zero

B) zero

C) **greater than zero but less than or equal to one**

D) none of the above

ANSWER: C

19. For a binary classifier, if all the data sets belong to the same class, then entropy value is

A) less than zero

B) **zero**

C) greater than zero but less than or equal to one

D) none of the above

ANSWER: B

20. Hunts algorithm is used for building decision trees

A) **True**

B) False

ANSWER: A

21. Which algorithm can be used for building decision tree

A) ID3

B) CART

C) C4.5

D) **All of the above**

ANSWER: D

22. These nodes make up decision trees

A)Root node, internal node

B) leaf node

C) **Root node, internal node, leaf nodes**

D) None of the above

ANSWER: C

23. For a binary attribute the number of splits(branches) in decision tree would be

A) one

B) **two**

C) more than one

D) more than two

ANSWER: B

24. For ordinal attribute, splitting should maintain order

A) False

B) **True**

ANSWER: B

25. For a binary classifier, involving 8 datasets, with 4 datasets having label true and 4 datasets having label false, the entropy would be

A) can't say

B) greater than zero

C) zero

D) **one**

ANSWER: D

26. Decision tree requires prior information about the classes

A) True

B) **False**

ANSWER: B

27. These are classifiers

A) Decision trees

B) Naive bayes

C) **Both Decision trees and Naive bayes**

D) market basket analysis

ANSWER: C

28. In k-NN, k represents

A) number of classes

B) **number of neighbours**

C) number of training records

D) number of test records

ANSWER: B

29. What is false about k-fold cross validation

A) measuring the performance of a machine learning model

B) the value of k represents the number of groups into which data is split

C) k groups include both training and testing data

D) **k groups include only training set**

ANSWER: D

30. Which of the following will be Euclidean Distance between the two data point A(1,3) and B(2,3)?

A) **1**

B) 2

C) 4

D) 8

ANSWER: A

31. Which of the following option is true about k-NN algorithm?

A) It can be used for classification

B) It can be used for regression

C) **It can be used in both classification and regression**

D) It can't be used in both classification and regression

ANSWER: C

32. The quality of a classification rule can be evaluated using measures

A) Coverage

B) **Accuracy**

C) Both coverage and accuracy

D) neither coverage and accuracy

ANSWER: B

33. Which of the following is true about Naive Bayes ?

A. Assumes that all the features in a dataset are equally important

B. Assumes that all the features in a dataset are independent

C. **Both A and B**

D. None of the above options

ANSWER: C

34. Naive Bayes classifier works on principle of

A) Conditional probability

B) **unconditional probability**

C) neither conditional or unconditional probability

D) both conditional or unconditional probability

ANSWER: B

35. Strategies for rule growing

A) general to specific

B) specific to general

C) **both specific to general and general to specific**

D) neither of these

ANSWER: C

36. what is false about RIPPER

A) is a rule induction algorithm

B) **works well with noisy data sets**

C) employs general to specific strategy

D) none of the above

ANSWER: B

37. k- nearest algorithm is

A) **lazy learner**

B) eager learner

C) neither lazy nor eager learner

D) both lazy and eager learner

ANSWER: A

38. Binary attribute are

A. **This takes only two values. In general, these values will be 0 and 1 and they can be coded as one bit**

B. The natural environment of a certain species

C. Systems that can be used without knowledge of internal operations

D. None of these

Answer: A

39. Classification task referred to

A. A subdivision of a set of examples into a number of classes

B. A measure of the accuracy, of the classification of a concept that is given by a certain theory

C. **The task of assigning a classification label to a set of examples**

D. None of these

Answer: C

40. Euclidean distance measure is

A. A stage of the KDD process in which new data is added to the existing selection.

B. The process of finding a solution for a problem simply by enumerating all possible solutions according to some pre-defined order and then testing them

C. **The distance between two points as calculated using the Pythagoras theorem**

D. None of these

ANSWER: C

41. In a Rule based classifier, If there is a rule for each combination of attribute values, what do you called that rule set R

A. **Exhaustive**

B. Inclusive

C. Comprehensive

D. Mutually exclusive

ANSWER: A

42. Which of the following algorithm comes under the classification

A. Apriori

B. Brute force

C. DBSCAN

D. **K-nearest neighbor**

ANSWER: D

43. Classification rules are extracted from \_\_\_\_\_\_\_\_\_\_\_\_\_

A. **decision tree**

B. root node

C. branches

D. siblings

ANSWER: A

44. Classification is

A. **A subdivision of a set of examples into a number of classifiers**

B. A measure of the accuracy, of the classification concept that is given by a certain theory

C. The task of assigning a classification to a set of examples

D. None

ANSWER: A

45.Discrete variables and continuous variables are two types of

A. open end classification

B. time series classification

C. qualitative classification

D. **quantitative classification**

ANSWER: D

46. Classification method in which the upper limit of interval is same as of lower limit class interval is called

A. **exclusive method**

B. inclusive method

C. mid-point method

D. ratio method

ANSWER: A

47. Largest value is 60 and the smallest value is 40 and the number of classes desired is 5 then the class interval is

A. 20

B. **4**

C. 25

D. 15

ANSWER: B

48. Summary and presentation of data in tabular form with several non-overlapping classes is referred as

A. nominal distribution

B. ordinal distribution

C. chronological distribution

D. **frequency distribution**

ANSWER: D

49. Classification method in which upper and lower limits of interval is also in class interval itself is called

A. exclusive method

B. **inclusive method**

C. mid-point method

D. ratio method

ANSWER: B

50. Bayesian classifiers is

A. **class of learning algorithm that tries to find an optimum classification of a set of examples using the probabilistic theory.**

B. Any mechanism employed by a learning system to constrain the search space of a hypothesis

C. An approach to the design of learning algorithms that is inspired by the fact that when people encounter new situations, they often explain them by reference to familiar experiences, adapting the explanations to fit the new situation

D. None of these

ANSWER: A

51. The most widely used metrics and tools to assess a classification model are

A. Confusion matrix

B. Cost-sensitive accuracy

C. Area under the ROC curve

D. **All of the above**

ANSWER: D

52. Which of the following is a disadvantage of decision trees

A. Factor analysis

B. Decision trees are robust to outliers

C. **Decision trees are prone to be overfit**

D. None of the above

ANSWER: C

53. 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. None of the above

ANSWER: A

54. Bayesian classifiers is

A) **A class of learning algorithm that tries to find an optimum classification of a set of examples using the probabilistic theory**

B) Any mechanism employed by a learning system to constrain the search space of a hypothesis

C) An approach to the design of learning algorithms that is inspired by the fact that when people encounter new situations, they often explain them by reference to familiar experiences, adapting the explanations to fit the new situation

D) None of these

ANSWER: A

55. Naive Bayes uses a similar method to predict the probability of different class based on various attributes. This algorithm is mostly used in text classification and with problems having multiple classes.

A) **TRUE**

B) FALSE

ANSWER: A

56. Why is naive Bayes so naive

A) **naive Bayes is so naive because it assumes that all of the features in a data set are equally important and independent. As we know, these assumption are rarely true in real world scenario.**

B) naive Bayes is so naive because it assumes that all of the features in a data set are equally unimportant and dependent. As we know, these assumption are rarely true in real world scenario.

ANSWER : A

57. Advantages of Naive Bayes Classifier, it handles

A) continuous and discrete data

B) Scalable

C) Fast

D) **All of the above**

ANSWER: D

58. Bayesian classifiers is

A) **A class of learning algorithm that tries to find an optimum classification of a set of examples using the probabilistic theory.**

B) Any mechanism employed by a learning system to constrain the search space of a hypothesis

C) An approach to the design of learning algorithms that is inspired by the fact that when people encounter new situations, they often explain them by reference to familiar experiences, adapting the explanations to fit the new situation.

D)None of these

ANSWER: A

59. Naive prediction is

A) A class of learning algorithms that try to derive a Prolog program from examples

B) A table with n independent attributes can be seen as an n- dimensional space.

C) **A prediction made using an extremely simple method, such as always predicting the same output.**

D) None of these

ANSWER:C

60. A definition or a concept is if it classifies any examples as coming within the concept

A) **Complete**

B) Consistent

C) Constant

D) None of these

ANSWER: A

61. Hunts algorithm grows a decision tree in a recursive fashion by partitioning the trainig records into successively purer subsets

A)**TRUE**

B)FALSE

ANSWER: A

62. Method in which the previously calculated probabilities are revised with new probabilities is classified as

A) updating theorem

B) revised theorem

C) **Bayes theorem**

D) dependency theorem

ANSWER: C

63. Previous probabilities in Bayes Theorem that are changed with the help of new available information are classified as

A) independent probabilities

B) **posterior probabilities**

C) interior probabilities

D) dependent probabilities

ANSWER: B

64. A .................set is used to determine the accuracy of the model

A) Training

B) Both Training and Test

C) **Test**

D) None of these

ANSWER: C

65. Classification is called ....................

A) Unsupervised learning

B) **Supervised learning**

C) Both Supervised and Unsupervised learning

D) None of these

ANSWER: B

66. Classification techniques are less effective for ....................

A) **Ordinal categories**

B) Binary categories

C) Nominal categories

D) All options

ANSWER: A

67. Classification Techniques are..............

A) Support Vector Machines

B) Rule-based classifiers

C) Neural Networks

D) **All options**

ANSWER: D

68. ................is the technique of accomplishing a task by providing training, input and output patterns to the systems

A) Unsupervised learning

B) **Supervised learning**

C) Both Unsupervised and supervised

D) None of these

ANSWER: B

69. A confusion matrix contains information about .....................................classifications done by a classification system

A) Actual

B) **Actual and Predicted**

C) Predicted

D) None of the Mentioned

ANSWER: B

70. Which are the performance matric to compare the performance of different models?

A) Accuracy

B) Error Rate

C) **Accuracy and Error Rate**

D) None of these

ANSWER: C

71. Performance matric accuracy is defined as...............

A) **Number of correct predictions/Total number of predictions**

B) Number of wrong predictions/Total number of predictions

C) Number of correct predictions/Number of wrong predictions

D) all of the Mentioned

ANSWER: A

72. Performance matric error rate is defined as...............

A) Number of correct predictions/Total number of predictions

B) **Number of wrong predictions/Total number of predictions**

C) Number of correct predictions/Number of wrong predictions

D) all of the Mentioned

ANSWER: B

73. Classification techniques are most suited for predicting or describing data sets with .............

A) **Binary or nominal categories**

B) Only Binary categories

C) Only Nominal categories

D) Only ordinal Categories

ANSWER: A

74. The problem of finding hidden structure in unlabeled data is called

A) Supervised learning

B) **Unsupervised learning**

C) Reinforcement learning

D) None of these

ANSWER: B

75. Some telecommunication company wants to segment their customers into distinct groups in order to send appropriate subscription offers, this is an example of

A) Supervised learning

B) Data extraction

C) Serration

D) **Unsupervised learning**

ANSWER: D

76. Self-organizing maps are an example of

A) **Unsupervised learning**

B) Supervised learning

C) Reinforcement learning

D) Missing data imputation

ANSER: A

77. Which of the following issue is considered before investing in Data Mining?

A) Functionality

B) Vendor consideration

C) Compatibility

D) **All of the above**

ANSWER: D

78. Bayesian classifiers is

A) **A class of learning algorithm that tries to find an optimum classification of a set of examples using the probabilistic theory**

B) Any mechanism employed by a learning system to constrain the search space of a hypothesis

C) An approach to the design of learning algorithms that is inspired by the fact that when people encounter new situations, they often explain them by reference to familiar experiences, adapting the explanations to fit the new situation

D) None of these

ANSWER: A

79. The number of iterations in apriori \_\_\_\_\_\_\_\_\_\_\_ Select one:

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

80. Classification and regression are the properties of

A) data analysis

B) data manipulation

C) **data mining**

D) none of these

ANSWER: C

81. A class of learning algorithm that tries to find an optimum classification of a set of examples using the probabilistic theory is named as

A) **Bayesian classifiers**

B) Dijkstra classifiers

C) doppler classifiers

D) all of these

ANSWER: A

82. We define a ..... as a subdivision of a set of examples into a number of classes

A) kingdom

B) tree

C) **classification**

D) array

ANSWER: C

83. Group of similar objects that differ significantly from other objects is named as

A) classification

B) **cluster**

C) community

D) none of these

ANSWER: B

84. What is the name of database having a set of databases from different vendors, possibly using different database paradigms

A) homogeneous database

B) **heterogeneous database**

C) hybrid database

D) none of these

ANSWER: B

85. What is inductive learning

A) learning by hypothesis

B) learning by analyzing

C) **learning by generalizing**

D) none of these

ANSWER: C

86. Which of the following is true for Classification

A) **A subdivision of a set**

B) A measure of the accuracy

C) The task of assigning a classification

D) All of these

ANSWER: A

87. ....... is not a data mining functionality

A) Clustering and Analysis

B) **Selection and interpretation**

C) Classification and regression

D) Characterization and Discrimination

ANSWER: B

88. The problem of finding hidden structure in unlabeled data is called

A) Supervised learning

B) **Unsupervised learning**

C) Reinforcement learning

ANSWER: B

89. Task of inferring a model from labeled training data is called

A) Unsupervised learning

B) **Supervised learning**

C) Reinforcement learning

ANSWER: B

90. Some telecommunication company wants to segment their customers into distinct groups in order to send appropriate subscription offers this is an example of

A) Supervised learning

B) Data extraction

C) Serration

D) **Unsupervised learning**

ANSWER: D

91. Self organizing maps are an example of

A) **Unsupervised learning**

B) Supervised learning

C) Reinforcement learning

D) Missing data imputation

ANSWER: A

92. You are given data about seismic activity in Japan, and you want to predict a magnitude of the next earthquake, this is in an example of

A) **Supervised learning**

B) Unsupervised learning

C) Serration

D) Dimensionality reduction

ANSWER: A

93. Assume you want to perform supervised learning and to predict number of newborns according to size of storks population it is an example of

A) Classification

B) **Regression**

C) Clustering

D) Structural equation modeling

ANSWER: B

94. Discriminating between spam and ham e-mails is a classification task true or false

A) **True**

B) False

ANSWER: A

95. In the example of predicting number of babies based on storks population size number of babies is

A) **outcome**

B) feature

C) attribute

D) observation

ANSWER: A

96. It may be better to avoid the metric of ROC curve as it can suffer from accuracy paradox

A) True

B) **False**

ANSWER: B

97. Which of the following is not involve in data mining

A) Knowledge extraction

B) Data archaeology

C) Data exploration

D) **Data transformation**

ANSWER: D

98. Discrete variables and continuous variables are two types of

A) open end classification

B) time series classification

C) qualitative classification

D) **quantitative classification**

ANSWER: D

99. Classification method in which the upper limit of interval is same as of lower limit class interval is called

A) **exclusive methoD**

B) inclusive method

C) mid point method

D) ratio method

ANSWER: A

100. Largest value is 60 and the smallest value is 40 and the number of classes desired is 5 then the class interval is

A) 20

B) **4**

C) 25

D) 15

ANSWER: B

Summary and presentation of data in tabular form with several non-overlapping classes is referred as

A) nominal distribution

B) ordinal distribution

C) chronological distribution

D) **frequency distribution**

ANSWER: D

1. Cluster is

A) **Group of similar objects that differ significantly from objects belonging to different group**

B) Operations on a database to transform or simplify data in order to prepare it for a machine-learning algorithm

C) Symbolic representation of facts or ideas from which information can potentially be extracted

D) None of these

ANSWER: A

2. Sentiment Analysis is not the example of:

A) Regression

B) Classification

C) **Clustering**

D) Reinforcement Learning

ANSWER: C

3. Movie Recommendation systems are an example of:

A) **Clustering**

B) Classification

C) Both A And B

D) None

ANSWER: A

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

A) 0

B) **1**

C) 2

D) 3

ANSWER: B

5. Clustering is: ------------

A) Predictive and unsupervised

B) Predictive and supervised

C) Descriptive and supervised

D) **Descriptive and unsupervised**

ANSWER: D

6. Which type of clustering groups data over a variety of scales by creating a cluster tree or dendrogram.

A) Partitional

B) **Hierarchical**

C) Naive Bayes

D) None of the Mentioned

ANSWER: B

7. 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) In general, the merges and splits are determined in a greedy manner

D) **All of the Mentioned**

ANSWER: D

8. 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

9. 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

10. 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

11. 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

12. Which of the following clustering requires merging approach ?

A) Partitional

B) **Hierarchical**

C) Naive Bayes

D) None of the Mentioned

ANSWER: B

13. This clustering algorithm initially assumes that each data instance represents a single cluster.

A) **agglomerative clustering**

B) conceptual clustering

C) K-Means clustering

D) expectation maximization

ANSWER: A

14. This clustering algorithm merges and splits nodes to help modify nonoptimal partitions.

A) agglomerative clustering

B) expectation maximization

C) **conceptual clustering**

D) K-Means clustering

ANSWER: C

15. This unsupervised clustering algorithm terminates when mean values computed for the current iteration of the algorithm are identical to the computed mean values for the previous iteration.

A) agglomerative clustering

B) conceptual clustering

C) **K-Means clustering**

D) expectation maximization

ANSWER: C

16. Which of the following requirements are not necessary for cluster analysis?

A) We need to consider how to incorporate user preference for cluster size and shape into the clustering algorithm.

B) In order to perform cluster analysis, we need to have a similarity measure between data objects.

C) We need to be able to handle a mixture of different types of attributes (e.g., numerical, categorical).

D) **We must know the number of output clusters a priori for all clustering algorithms.**

ANSWER: D

17. Which of the following statements are true?

A) Graphs, time-series data, text, and multimedia data are all examples of data types on which cluster analysis can be performed.

B) Agglomerative clustering is an example of a distance-based clustering method.

C) When dealing with high-dimensional data, we sometimes consider only a subset of the dimensions when performing cluster analysis.

D) **All the Above**

ANSWER: D

18. Can decision trees be used for performing clustering?

A) **True**

B) False

ANSWER: A

19. Which points are eliminated by the DBSCAN algorithm

A) Core points

B) Border points

C) **Noise points**

D) None of the Mentioned

ANSWER: C

20. How is the density of point p at the density based clustering defined?

A) MinPts minus number of data points in an epsilon-neighbourhood

B) **Number of data points in an epsilon-neighbourhood of p**

C) Reciprocal value of the distance from p to the nearest neighbour

D) None of the Mentioned

ANSWER: B

21. DBSCAN can handle clusters of different sizes and shapes and is not strongly affected by noise or outliers.

A) **Yes**

B) No

C) Can�t say

D) None of these

ANSWER: A

22. DBSCAN Uses a --------------notion of a cluster

A) **Density based**

B) Prototype based

C) Centroid based

D) Mediod based

ANSWER: A

23. Which of the following is/are not true about DBSCAN clustering algorithm

1.For data points to be in a cluster, they must be in a distance threshold to a core point

2.It has strong assumptions for the distribution of data points in dataspace

3.It has substantially high time complexity of order O(n3)

4.It does not require prior knowledge of the no. of desired clusters

A) 1 only

B) 2 only

C) 4 only

D) **2 and 3**

ANSWER:D

24. What is the full form of DBSCAN

A)**Density based spatial clustering of applications with noise**

B)Database spatial clustering of applications with noise

C)Density based spatial clustering approach with noise

D) None of the above

ANSWER:A

25. DBSCAN algorithm is proposed for----------- databases.

A)Multimedia databases

B)Temporal databases

C)Geographic databases

D)**Spatial databases**

ANSWER:D

26. Which of the following is a clustering technique.

A)**DBSCAN**

B)FP Growth

C)Naive Bayes

D)Apriori

ANSWER: A

27. Noise point is defined as

A)**Any point that is neither a core point nor a border point**

B)Any point that is not a core point and falls within the neighborhood

C)The point which are in the interior of a density based cluster

D)None of the above

ANSWER:A

28. In DBSCAN,the worst case time complexity is given by

A)**O(M�)**

B)O(mlogm)

C)O(M)

D)O(logn)

ANSWER:A

29. 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

30. 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

31. K-means is not deterministic and it also consist of number of iterations.

A) **True**

B) False

ANSWER: A

32. 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

33. \_\_\_\_\_ is frequently referred to as k-means clustering.

A) Non-hierarchical clustering

B) partitioning

C) **Divisive clustering**

D) Agglomerative clustering

ANSWER : D

34. Is it possible that Assignment of observations to clusters does not change between successive iterations in K-Means

A) **Yes**

B) No

C) Can�t say

D) None of these

ANSWER : A

35. 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.

A) 1, 3 and 4

B) 1, 2 and 3

C) 1, 2 and 4

D) **All of the above**

ANSWER : D

36. 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

37. 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**

E) 1, 2, 3 and 4

ANSWER : D

38. Assume, you want to cluster 7 observations into 3 clusters using K-Means clustering algorithm. After first iteration clusters, C1, C2, C3 has following observations: C1: {(2,2), (4,4), (6,6)} C2: {(0,4), (4,0)} C3: {(5,5), (9,9)} What will be the cluster centroids if you want to proceed for second iteration?

A) **C1: (4,4), C2: (2,2), C3: (7,7)**

B) C1: (6,6), C2: (4,4), C3: (9,9)

C) C1: (2,2), C2: (0,0), C3: (5,5)

D) None of these

ANSWER : A

39. Feature scaling is an important step before applying K-Mean algorithm. What is reason behind this?

A. **In distance calculation it will give the same weights for all features**

B. You always get the same clusters. If you use or don�t use feature scaling

C. In Manhattan distance it is an important step but in Euclidian it is not

D. None of these

ANSWER : A

40. In \_\_\_\_\_\_\_\_ algorithm each cluster is represented by the center of gravity of the cluster.

A) k-medoid

B) **k-means**

C) STIRR

D) ROCK

ANSWER : B

41. Considering the K-means algorithm, if points (-1, 3), (-3, 1), and (-2, -1) are the only points which are assigned to the first cluster now, what is the new centroid for this cluster?

A) **(-2, 1)**

B) (1, 1)

C) (-1, 2)

D) (2, 2)

ANSWER : A

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

A) **The K-means algorithm is 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 be any observed data points.

D) The K-means algorithm can detect non-convex clusters

ANSWER : A

43. Tree graph is a graphical device for displaying clustering results. Vertical lines represent clusters that are joined together. The position of the line on the scale indicates the distances at which clusters were joined. Identify the tree graph from below.

A) **Dendrogram**

B) Scattergram

C) Screen plot

D) icicle diagram

ANSWER: A

44. Identify the clustering procedure characterized by the development of a tree-like structure.

A) Non-hierarchical clustering

B) **Hierarchical clustering**

C) Density based clustering

D) K-Means clustering

ANSWER: B

45. Name the clustering procedure where all objects start out in one giant cluster. Clusters are formed by dividing this cluster into smaller and smaller clusters.

A) Non-hierarchical clustering

B) Density based clustering

C) **Divisive clustering**

D) Agglomerative clustering

ANSWER: C

46. K-means clustering can be categorized under which category.

A) **Non-hierarchical clustering**

B) Optimizing partitioning

C) Divisive clustering

D) Agglomerative clustering

ANSWER: A

47. Which method uses information on all pairs of distances, not merely the minimum or maximum distances.

A) single linkage

B) medium linkage

C) complete linkage

D) **average linkage**

ANSWER: D

48. The single linkage method of hierarchical clustering is uses

A) **Min distance to update proximity**

B) Max distance to update proximity

C) Average distance to update proximity

D) None of the above

ANSWER: A

49. The complete linkage method of hierarchical clustering is uses

A) Min distance to update proximity

B) **Max distance to update proximity**

C) Average distance to update proximity

D) None of the above

ANSWER: B

50. Pick out hierarchical clustering method.�

A) Agglomerative clustering�

B) Divisive clustering

C) **A and B**

D) K-Means clustering

ANSWER: C

51. 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

52. Hierarchical clustering should be primarily used for exploration.

A) **True**

B) False

Answer: A

53. Which of the following metrics, do we have for finding dissimilarity between two

clusters in hierarchical clustering?

A)only single link

B)both single and complete link

C)both complete and average link

D)**All the above**

ANSWER: D

54. Identify the clustering technique start with as many clusters as there are records, with each

clusterhaving only one record.

A) **Agglomerative**

B) divisive

C) Partition

D) Numeric

ANSWER: A

55. Clustering is

A) Supervised

B) **Unsupervised**

C) both A and B

D) None of the above

ANSWER: B

56. In exclusive clustering technique, Each data point

A) **belongs to one cluster only**

B) can belong to any number of clusters

C) Both A & B

D) None of the above

ANSWER: A

57. In partial clustering

A) All data points belong to atleast one cluster

B) **Some data points do not belong any cluster**

C) Both A & B

D) None of the above

ANSWER: B

58. In K-Means clustering, K denotes

A) distance between clusters

B) distance between data points

C) **number of clusters**

D) centoid or mean value

ANSWER: C

59. The problem of finding hidden structure in unlabeled data is called

A) Supervised learning

B) **Unsupervised learning**

C) Reinforcement learning

D) None of these

ANSWER: B

60. Some telecommunication company wants to segment their customers into distinct groups in order to send appropriate subscription offers, this is an example of

A) Supervised learning

B) Data extraction

C) Serration

D) **Unsupervised learning**

ANSWER: D

61. Self-organizing maps are an example of

A) **Unsupervised learning**

B) Supervised learning

C) Reinforcement learning

D) Missing data imputation

ANSER: A

62. Which of the following issue is considered before investing in Data Mining?

A) Functionality

B) Vendor consideration

C) Compatibility

D) **All of the above**

ANSWER: D

63. Cluster is

A) **Group of similar objects that differ significantly from other objects**

B) Operations on a database to transform or simplify data in order to prepare it for a machine-learning algorithm

C) Symbolic representation of facts or ideas from which information can potentially be extracted

D) None of these

ANSWER: A

64. Bayesian classifiers is

A) **A class of learning algorithm that tries to find an optimum classification of a set of examples using the probabilistic theory**

B) Any mechanism employed by a learning system to constrain the search space of a hypothesis

C) An approach to the design of learning algorithms that is inspired by the fact that when people encounter new situations, they often explain them by reference to familiar experiences, adapting the explanations to fit the new situation

D) None of these

ANSWER: A

65. This clustering algorithm terminates when mean values computed for the current iteration of the algorithm are identical to the computed mean values for the previous iteration Select one:

A) **K-Means clustering**

B) conceptual clustering

C) expectation maximization

D) agglomerative clustering

ANSWER: A

66. The correlation coefficient for two real-valued attributes is �0.85. What does this value tell you? Select one:

A) The attributes are not linearly related.

B) **As the value of one attribute decreases the value of the second attribute increases.**

C) As the value of one attribute increases the value of the second attribute also increases.

D) The attributes show a linear relationship

ANSWER: B

67. Time Complexity of k-means is given by Select one:

A) O(mn)

B) **O(tkn)**

C) O(kn)

D) O(t2kn)

ANSWER: B

68. \_\_\_\_\_\_\_\_\_ clusterings, points may belong to multiple clusters Select one:

A) Non exclusivce

B) Partial

C) **Fuzzy**

D) Exclusive

ANSWER: C

69. Which statement is true about the K-Means algorithm? Select one:

A) The output attribute must be cateogrical.

B) All attribute values must be categorical.

C) **All attributes must be numeric**

D) Attribute values may be either categorical or numeric

ANSWER: C

70. The number of iterations in apriori \_\_\_\_\_\_\_\_\_\_\_ Select one:

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

71. A good clustering method will produce high quality clusters with Select one:

A) high inter class similarity

B) low intra class similarity

C) **high intra class similarity**

D) no inter class similarity

72. Which two parameters are needed for DBSCAN Select one:

A) Min threshold

B) **Min points and** eps

C) Min sup and min confidence

D) Number of centroids

ANSWER: B

73. In Apriori algorithm, if 1 item-sets are 100, then the number of candidate 2 item-sets are Select one:

A) 100

B) **4950**

C) 200

D) 5000

ANSWER: B

74. The most general form of distance is Select one:

A) Manhattan

B) Eucledian

C) Mean

D) **Minkowski**

ANSWER: D

75. Simple regression assumes a \_\_\_\_\_\_\_\_\_\_ relationship between the input attribute and output attribute. Select one:

A) quadratic

B) inverse

C) **linear**

D) reciprocal

ANSWER: C

76. Hierarchical agglomerative clustering is typically visualized as? Select one:

A) **Dendrogram**

B) Binary trees

C) Block diagram

D) Graph

ANSWER: A

77. This clustering algorithm terminates when mean values computed for the current iteration of the algorithm are identical to the computed mean values for the previous iterationSelect one:

A) conceptual clustering

B) **K-Means clustering**

C) expectation maximization

D) agglomerative clustering

ANSWER: B

78. Data Visualization in mining cannot be done using Select one:

A) **Photos**

B) Graphs

C) Charts

D) Information Graphics

ANSWER: A

79. Dimensionality reduction reduces the data set size by removing \_\_\_\_\_\_\_\_\_ Select one:

A) composite attributes

B) derived attributes

C) relevant attributes

D) **irrelevant attributes**

ANSWER: D

80. Incorrect or invalid data is known as \_\_\_\_\_\_\_\_\_ Select one:

A) Missing data

B) Outlier

C) Changing data

D) **Noisy data**

ANSWER: D

81. Movie Recommendation systems are an example of

1. Classification

2. Clustering

3. Reinforcement Learning

4. Regression

A. 1 only

B. 2 only

C. **2 and 3**

D. None

ANSWER: C

82.Sentiment Analysis is an example of

1. Regression

2. Classification

3. Clustering

4. Reinforcement Learning

A. 1 Only

B. 1 and 2

C. 1, 2 and 3

D. **1, 2 and 4**

ANSWER: D

83. 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

84. 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. 1 and 2

C. 2 and 3

D. **1, 2 and 3**

ANSWER: D

85. This clustering algorithm terminates when mean values computed for the current iteration of the algorithm are identical to the computed mean values for the previous iteration

A. **K-Means clustering**

B. conceptual clustering

C. expectation maximization

D. agglomerative clustering

ANSWER: A

86. This clustering approach initially assumes that each data instance represents a single cluster

A. expectation maximization

B. K-Means clustering

C. agglomerative clustering

D. **NONE**

ANSWER: D

87. The correlation coefficient for two real-valued attributes is �0.85. What does this value tell you

A. The attributes are not linearly related

B. **As the value of one attribute decreases the value of the second attribute increases**

C. As the value of one attribute increases the value of the second attribute also increases

D. The attributes show a linear relationship

ANSWER: B

88. In \_\_\_\_\_\_\_\_\_ clusterings, points may belong to multiple clusters

A. Non exclusivce

B. Partial

C. **Fuzzy**

D. Exclusive

ANSWER: C

89. Which statement is true about the K-Means algorithm

A. The output attribute must be cateogrical

B. All attribute values must be categorical

C. **All attributes must be numeric**

D. Attribute values may be either categorical or numeric

ANSWER: C

90. This data transformation technique works well when minimum and maximum values for a real-valued attribute are known

A. z-score normalization

B. **min-max normalization**

C. logarithmic normalization

D. decimal scaling

ANSWER: B

91. Find odd man out

A. K medoid

B. K mean

C. **DBSCAN**

D. PAM

ANSWER: C

92. Which of the following is cluster analysis

A. Simple segmentation

B. **Grouping similar objects**

C. Labeled classification

D. Query results grouping

ANSWER: B

93. A good clustering method will produce high quality clusters with

A. high inter class similarity

B. low intra class similarity

C. **high intra class similarity**

D. no inter class similarity

ANSWER: C

94. Which two parameters are needed for DBSCAN

A. Min threshold

B. **Min points and eps**

C. Min sup and min confidence

D. Number of centroids

ANSWER: B

95. The concept of core, border and noise points fall into this ategory

A. DENCLUE

B. Subspace clustering

C. Grid based

D. **DBSCAN**

ANSWER: D

96. Arbitrary shaped clusters can be found by using

A. **Density methods**

B. Partitional methods

C. Hierarchical methods

D. Agglomerative

ANSWER: A

97. Clustering is \_\_\_\_\_\_\_\_\_\_\_ and is example of \_\_\_\_\_\_\_\_\_\_\_\_learning

A. Predictive and supervised

B. Predictive and unsupervised

C. Descriptive and supervised

D. **Descriptive and unsupervised**

ANSWER: D

98. Hierarchical agglomerative clustering is typically visualized as

A. **Dendrogram**

B. Binary trees

C. Block diagram

D. Graph

ANSWER: A

99. What is the final resultant cluster size in Divisive algorithm, which is one of the hierarchical clustering approaches

A. Zero

B. Three

C. **singleton**

D. Two

ANSWER: C

100. This clustering algorithm terminates when mean values computed for the current iteration of the algorithm are identical to the computed mean values for the previous iteration

A. conceptual clustering

B. **K-Means clustering**

C. expectation maximization

D. agglomerative clustering

ANSWER: B

101. What does K refers in the K-Means algorithm which is a non-hierarchical clustering approach

A. Complexity

B. Fixed value

C. No of iterations

D. **number of clusters**

ANSWER: D

102. Cluster is

A. **Group of similar objects that differ significantly from other objects**

B. Operations on a database to transform or simplify data in order to prepare it for Machine Learning algorithm

C. Symbolic representation of facts or ideas from which information can potentially be extracted

D.None

ANSWER: A

103. A Clustering approach initially assumes that each data instance represents a single cluster

A. Expectation maximization

B. K-means clustering

C. **Agglomerative clustering**

D. Conceptual Clustering

ANSWER: C

104. Which statement is true about the K-Means algorithm?

A. The output attribute must be categorical

B. All attribute values must be categorical

C. **All attributes must be numeric**

D. Attribute values may be either categorical or numeric

ANWER: C

105. Can decision trees be used for performing clustering

A. **True**

B. False

ANSWER: A

106. For two runs of K-Mean clustering is it expected to get same clustering results

A. Yes

B. **No**

ANSWER: B

107. Is it possible that Assignment of observations to clusters does not change between successive iterations in K-Means

A. **Yes**

B. No

C. Can�t say

D. None of these

ANSWER: A

108. 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