1) Which of these attributes of data is of interest when performing mining operations:
   A. Dissimilarity in terms of the Euclidean definition of distance between points
   B. Dissimilarity in terms of the Hamming distance of the bits in two data objects
   C. Dissimilarity in terms of the Supremum distance between any given attribute of data objects
   D. All of the above

2) To select the "right" proximity measure, which of these is a useful heuristic:
   A. Choose a proximity measure that fits the data.
   B. Choose a measure that ignores attributes that neither object has (0-0 matches) for sparse, asymmetric data where similarity is important, c
   C. Choose a proximity measure based on differences if the data consists of attributes that are continuous
   D. All of the above

3) Dirty data can cause which of the following problems regarding data mining results
   A. Distrust of the results by those who must rely on them to make important decisions
   B. Inaccurate results
   C. Incomplete results
   D. All of the above

4) Independent variables in an experiment
   A. Have no impact on each other
   B. Are varied in value to determine their influence on dependent variables
   C. Satisfy this equation, P(X,Y) = P(X) * P(Y), where X and Y are the variables and P is the probability function.
   D. All of the above

5) Which of these are typically used to represent the sought-after mapping function in the first step of the classification process:
   A. Decision Trees
   B. Mathematical formula
   C. Classification rules
   D. All of the above
   E. None of the above

6) The zero-probability issue can be handled by which of these approaches:
   A. Applying the Laplacian correction technique
   B. Applying the Laplacian estimator technique
   C. Applying the add 1 to the number of tuples in each class in our data set technique
   D. All of the above
   E. None of the above

7) Classification rules can be represented by $r_i : (Condition_i) \longrightarrow y_i$, where
   A. r_i is the i_{th} rule in a set of rules, representing a rule precondition or antecedent
   B. $(Condition_i)$ represents a conjunction of attribute tests
   C. y_i is the predicted class, or rule consequent
   D. All of the above
   E. None of the above

8) K-means is a
   A. Method for creating clusters based on similarities amongst members of a cluster
   B. Method for clustering that requires an initial value, k, that predetermines the number of clusters to create from a given data set
   C. Method that avoids the fact that optimizing the within-cluster variation is NP-hard for k clusters in 2-D Euclidean space
   D. All of the above

9) Density-based methods
   A. Resolve the issue of finding clusters of arbitrary shapes
   B. Includes DBSCAN, which starts with a parameter, $\varepsilon > 0$, specifying the radius of an object's neighborhood, and a second parameter, *MinPts*, which specifies the minimum number of objects in a neighborhood of an object in order for that object to be considered a core object
   C. Include the concept of density-connectedness to form clusters of small of dense regions
   D. All of the above

10) CLIQUE (Clustering In QUEst)
   A. Relies on the monotonicity property of density-based clusters
   B. Finds subspace clusters despite the fact that the number of subspaces involved is exponential to the number of dimensions
   C. Uses an algorithm that is conceptually similar to the *Apriori* algorithm for finding frequent itemsets
   D. All of the above

11) Which of these does not make the use of clustering problematic in detecting outliers:
   A. Outliers impact the clustering process indeterminately, thus it is problematic as to whether the clusters are real clusters, and thus which members are legitimate outliers
   B. Clustering techniques that do not automatically determine the number of clusters can give different results as to which objects are true outliers based on the number of predetermined clusters with which the algorithm is seeded
   C. Assessing accurately the extent to which a given object belongs to a particular cluster is not a well-developed, well understood process at this time
   D. All of the above
   E. None of the above

==========

12) According to IBM, which these is not considered a major task of data mining:
   A. Predictive analytics
   B. Prescriptive analytics
   C. Descriptive analytics
   D. None of the above

13) Contour Plots and Surface Plots are:
   A. Used for visualizing data with high dimensionality
   B. Only useful for data describing geographically spatially related entities, such as average sea temperatures around the earth.
   C. Are useful if the data can be spatially related, in reality or virtually.
   D. None of the above

14) Mapping function accuracy:

A. Is measured in terms of the percentage of test tuples/samples that the function correctly classified/labeled
B. Must be 98% to be useful
C. Are useful if the data can be spatially related, in reality or virtually.
D. None of the above

15) Processing order of the data
A. Never impacts the accuracy of any know clustering algorithm
B. While important to how accurately certain clustering algorithms perform, is in some cases an insufficient reason for not using a given algorithm, depending on other worthy characteristics of the algorithm and the amount of inaccuracy introduced
C. While important to how accurately certain clustering algorithms perform, the inaccuracies are determinant, and thus can be accounted for through normalization techniques
D. None of the above

16) Which of these situations is never true about an anomalous object
A. The value of a single attribute is anomalous, thus the object is anomalous
B. The values of some attributes are somewhat anomalous, but most are normal, thus the object is anomalous
C. No single attribute has an anomalous value, but when analyzed holistically, the particular set of values for a given object makes it an anomalous object
D. All of the above
E. None of the above

==========

1) Search algorithms are the only AI techniques of interest to data mining researchers

Answer false

2) Cluster analysis is a means for discovering patterns in the data based on highly associated features of the data

Answer false

3) Association analysis is a means for discovering and grouping together (clustering) sets of observations that are closely related

Answer false

4) Range and variance are measures of location.

Answer false

5) A library maintained by a business does not fit William H. Inmon's definition of a data warehouse.

Answer false

6) Data warehouses are not a form of online analytical processing (OLAP) systems

Answer false

7) Concept hierarchies in a data warehouse represent easily materialized views of the data using several non-interactive data cube operations.

Answer false

8) Explanatory multidimensional data mining is never an interactive process, given the intense computations that necessarily involved.

Answer false

9) Computation costs are not a factor in performing knowledge discovery in a multidimensional online analysis environment.

Answer false

10) Overlapping paths in a FP-tree are indicative of corrupted input data.

Answer false

11) All strong association rules are interesting

Answer false

12) Simpson's paradox is a phenomenon where hidden variables may cause the observed relationship to multiply.
FALSE

13) Objective measures rank patterns according to user's interpretations.
FALSE

14) None of the association rule mining algorithms use support measure to prune rules and itemsets.
FALSE

15) In general, confidence has an anti-monotone property.
FALSE

16) The objective measure ranks patterns based on hypothetical predications.
FALSE

17) The approach of iteratively expanding a subgraph by adding an extra vertex is known as edge growing.
FALSE

18) Object measures alone may be sufficient to eliminate uninteresting infrequent patterns.
FALSE

19) Negative itemsets and negative association rules are collectively known as positive patterns.
FALSE

20) There cannot be more than one decision tree that fits the same data.
FALSE

21) The “Scalability” of a database is defined as classifying data sets with millions of examples and hundreds of attributes inefficiently, over a very long period of time.
FALSE

22) A partition is “pure” if all the tuples in it belong to different classes.
FALSE

23) Bootstrapping does not work well with small sets.
FALSE

24) Significance tests and ROC curves are useless for model selection.
FALSE

25) Rule-ordering classification approaches can only be done rule-by-rule. It is not possible to order classes.

Answer false

26) When generating rules using the sequential covering algorithm, a rule is considered to be acceptable if it covers all positive examples in the training set. The number of negative examples it covers does not come into play.

Answer false

27) Decision tree and rule-based classifiers are examples of lazy learners.
FALSE

28) An example of an eager learner would be Rote classifier.
FALSE

29) A naïve Bayesian classifier estimates the class-conditional probability by assuming that the attributes are conditionally dependent.
FALSE

30) The kernel trick is a method that solves issues with irrelevant attributes.
FALSE

31) The higher the total sum of the squares (SSB) for a group of clusters, the lower the separation of the clusters.

Answer: false

32) A density-based method clusters objects based on the notion of clique.
FALSE

33) In a fuzzy clustering, every object belongs to every cluster with a membership weight that ranges between 0 and 5.

FALSE

34) Clustering is a relatively simple field.
FALSE

35) CLIQUE is a simple grid-based method for finding density-based cluster in subspaces.
TRUE

36) If ground truth is available, it can be used by intrinsic methods.
FALSE

37) If the ground truth is available, it can be used by extrinsic methods.
FALSE

38) High dimensional data typically contains larger clusters then does low-dimensional data

Answer false

39) Traditional clustering methods require each object to belong to many clusters.
FALSE

40) A sparsest cut will not likely lead to good clustering.
FALSE

41) The optimization-based method is not a type of biclustering method.
FALSE

42) High dimensional data rarely pose a problem cluster analysis.
FALSE

43) If someone steals your credit card, the fact that they use it to buy breakfast at a different place in your town then you normally go is a clear signal of an outlier transaction

Answer false

44) Noise in the data never interferes with any known outlier detection methods, so preprocessing of data sets to remove noise data points is not necessary

Answer false

45) Many discordancy tests are common knowledge and do not require statistical knowledge.
FALSE

46) Cluster techniques like K-means automatically determine the number of clusters.
FALSE

47) When outliers are detected, there is no question of whether the results are valid.

FALSE

48) A mixture model approach for anomaly detection assume that data comes from just a single probability distributions.
FALSE

49) According to Han, Kamber and Peidata reduction is a theory concerning the basis of data mining that is the only useful theory concerning data mining

Answer: false

50) According to Han, Kamber and Pei, the ideal theoretical framework for data mining should model typical data mining tasks, have a probabilistic nature, handle multiple forms of data, but can safely ignore the iterative and interactive essence of data mining.

Answer: false

51) According to Han, Kamber and Pei, the high dimensionality of retail data on sales, customers, products, time, and region makes such applications determining effective sales campaigns and NP-hard problem, that is, impossible.

Answer: false

52) Research on the theoretical frameworks of data mining have matured significantly.
FALSE

53) The microeconomic view is a theory of data mining that considers finding patterns is not concerned with the utility of patterns.
FALSE

54) Text mining is by its very nature, *not* interdisciplinary.
FALSE

55) Few classification methods perform model construction based on feature vectors.
FALSE

56) Theories of data mining are mutually exclusive.
FALSE