Test Bank Questions ([# of questions])

Data Mining and Data Warehousing - IT 446

Developed by Bradley C. Watson

* Reference: Tan, Steinbach and Kumar (2006)
* Reference: [IBM, “Descriptive, predictive, prescriptive: Transforming asset and facilities management with analytics” (2013)](https://static.ibmserviceengage.com/TIW14162USEN.PDF)

**Week 2**

1. Which of the following is a motivating challenge for developing data mining:.
2. No one cares about data mining
3. Scalability
4. Algorithms are very complex
5. Boredom
6. Data mining, as a research discipline, does not draw ideas from this research discipline:
7. Statistics
8. Linguistics
9. Artificial Intelligence
10. Information Theory
11. Database technology does not support data mining in terms of
12. Theoretical basis of data mining research
13. Efficient storage
14. Query processing
15. Indexing
16. According to the Tan, Steinbach and Kumar text, which of these is not considered a major task of data mining:
17. Predictive tasks
18. Prescriptive tasks
19. Descriptive tasks
20. None of the above
21. According to IBM, which these is not considered a major task of data mining:
22. Predictive analytics
23. Prescriptive analytics
24. Descriptive analytics
25. None of the above
26. Which of the following is an important data-related success factor in data mining efforts:
27. Data type(s) being mined.
28. Processing power available
29. Algorithm design
30. Descriptive statistics-based tools
31. Graph-Based data (choose the false answer)
32. Is a rare form of data that is only used in data mining if absolutely necessary given the difficulty of obtaining it and processing it
33. Can consist of data objects whose relationship to each other is representable by placement as nodes in a graph
34. Can be data objects that are themselves graphs
35. Can, in some instances, be mined in terms of substructures
36. Ordered data (choose the true answer)
37. Is never ordered based on one or more time attributes (stamps)
38. Can be sequentially ordered in terms of spatial (positional)- or time-based attributes
39. Is of no use in data mining
40. Is not a method for modeling genetic information, such as genes in human DNA
41. Which of these attributes of data is of interest when performing mining operations:
42. Dissimilarity in terms of the Euclidean definition of distance between points
43. Dissimilarity in terms of the Hamming distance of the bits in two data objects
44. Dissimilarity in terms of the Supremum distance between any given attribute of data objects
45. All of the above
46. To select the “right” proximity measure, which of these is a useful heuristic:
47. Choose a proximity measure that fits the data.
48. Choose a measure that ignores attributes that neither object has (0-0 matches) for sparse, asymmetric data where similarity is important, c
49. Choose a proximity measure based on differences if the data consists of attributes that are continuous
50. All of the above
51. The terms “Analytics” and “Big Data” have become essentially synonymous with the term “data mining” in recent years.

Answer true

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

Answer false

1. When using feature subset selection to reduce dimensionality, \_redundant\_\_ and/or \_irrelevant\_\_ features must be present in order to achieve a feature subset that truly represents the initial data set, but is significantly smaller in terms of number of features per data object.
2. Principle Components Analysis is a linearalgebra technique that finds new attributes that are, among other things, linear combinations of the original attributes.
3. Singular Value Decomposition is a linear algebra technique often used for dimensionalityreduction that depends on the concepts of eigenvalues and eigenvectorswhere for Matrix **A**, of dimensions m by n, the eigenvalue is the scalar value **λ** and eigenvectors are the vectors **u**, such that they are solutions to this equation:

Au = λu

Answer true

1. The term “noise” has a technical meaning in data mining referring to the distortion of data from their true value and/or the addition of spurious objects.

Answer true

1. Effectively handling “noisy” data is the object of much data mining research, as data used in the real world often contains significant amounts of noise, thus potentially contaminating data mining results to the point where those results are useless.

Answer true

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

Answer false

1. Cluster analysis is a means for discovering and grouping together (clustering) sets of observations that are closely related

Answer true

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

Answer false

**Week 3**

1. Which of the following is not a factor in data quality:
2. Accuracy
3. Completeness
4. Relevance
5. Timeliness
6. Which of the following data attributes is not one of the most common found in real-world databases and data warehouses
   1. Inaccuracy
   2. Interpretability
   3. Incompleteness
   4. Timeliness
7. Data preprocessing does not include which of these tasks:
8. Data classification
9. Data integration
10. Data reduction
11. Data cleaning
12. Dirty data can cause which of the following problems regarding data mining results
13. Distrust of the results by those who must rely on them to make important decisions
14. Inaccurate results
15. Incomplete results
16. All of the above
17. Data integration often involves:
18. Reducing the dimensionality of the data set to be mined
19. Removing outliers
20. Determining which data objects in the various data sources match to each other
21. Transforming the data from one form to another
22. Correlation analysis applied in the data integration stage is an effective means to apply
23. When attempting to eliminate redundancy in a data set.
24. Identify matching objects across different data sources
25. Removing outlier data values
26. Detecting non-interpretable and/or non-believable data values
27. This equation, , is used to compute the Pearson Statistic (correlation test) for
28. Unbelievable data
29. Impossible to interpret data
30. Nominal data
31. None of the above
32. Pearson’s product moment coefficient is used to evaluate correlation for which type of these attributes:
33. Nominal
34. Numeric
35. Alphabetic
36. Is not used to correlate any type of attribute
37. Covariance and correlation measures can be used to compute:
38. How often two numeric attributes change together across a given data set
39. How often two alphabetic attributes change together across a given data set
40. Whether two attributes are believable when they change together more often than they change independently of each other
41. None of the above
42. Data conflicts are resolved during data
43. Cleaning
44. Reduction
45. Transformation
46. Integration
47. Multidimensional Scaling is a technique for projecting data from a higher dimensionality to a lower dimensionality, often referred to as a *p*-dimensional space, where the goal is to assign data objects from the original data set to a point in the *p*-dimensional space where the quantity **stress** for each such object is minimized.

Answer true

1. Range and variance are measures of location.

Answer false

1. Median is a summary statistic, where the function**median(x)** returns the middle value in a data set with an odd number of values and the average of the two middle values if the number of values is even.

Answer true

1. Visualization techniques are often specialized to the type of data being analyzed.

Answer true

1. A \_Contour\_\_Plot\_ is a useful visualization format for three-dimensional data where two attributes can define a point in a plane and the third represents a continuous function.
2. A multidimensional representation of data with all possible totals (aggregates) is known as a \_data\_ \_cube\_.
3. If we aggregate over all the dimensions of a data set except for two, we are creating a two dimensional table using a data reduction approach through aggregation known as \_pivoting\_.
4. A \_MOLAP\_ is a database system that might use as its base data model a data cube representation.
5. If we analyze monthly data in terms of the days of each month we are \_drilling\_\_down\_.
6. Two steps are necessary to define data in a multidimensional array representation: \_identification of dimensions\_ and an\_attribute\_ that is the analysis focus.

**Week 4**

1. Which of the following is/are unordered categorical variable(s):.
2. Gender(male, female)
3. Economic status (low, middle, high)
4. Set of set of even integers from n=2 to n=30
5. Hair color (brown, brown, red)
6. All of the above
7. The set of odd integers from n=5 to n=41 is which of these types of variables:
8. Categorical
9. Interval
10. Independent
11. Ordinal
12. Dependent
13. Independent variables in an experiment
14. Have no impact on each other
15. Are varied in value to determine their influence on dependent variables
16. Satisfy this equation, P(X,Y) = P(X) \* P(Y), where X and Y are the variables and P is the probability function.
17. All of the above
18. According to the Tan, Steinbach and Kumar text, which of these is not considered a technique for visualizing small numbers of attributes:
19. Pivot tables
20. Stem and Leaf Plots
21. Histograms
22. Box Plots
23. Contour Plots and Surface Plots are:
24. Used for visualizing data with high dimensionality
25. Only useful for data describing geographically spatially related entities, such as average sea temperatures around the earth.
26. Are useful if the data can be spatially related, in reality or virtually.
27. None of the above
28. Which of the following is not part of the ACCENT Principles
29. Complexity -- ideas are to be communicated with clarity, precision and efficiency
30. Apprehension – graphs should maximize this attribute for their audience
31. Clarity – the most important attributes should be clearly the most visually prominent
32. Truthfulness – graph elements must be scaled and positioned accurately
33. None of the above
34. Graph-Based data (choose the false answer)
35. Is a rare form of data that is only used in data mining if absolutely necessary given the difficulty of obtaining it and processing it
36. Can consist of data objects whose relationship to each other is representable by placement as nodes in a graph
37. Can be data objects that are themselves graphs
38. Can, in some instances, be mined in terms of substructures
39. Spatially related data
40. Is always best understood through visualization techniques based on those used by map makers
41. Often is best analyzed through visualization techniques
42. Cannot be adequately mined using current data mining approaches
43. Includes genetic information, such as genes in human DNA.
44. Measures of location for multivariate data can often be obtained by computing
45. The mean and/or the median for each attribute separately
46. The range and variance of the most important attribute and dividing the range by the variance
47. A roll-up value for each attribute and choosing the attribute with the lowest value
48. The probability that each variable is an independent variable and reporting the attribute with the highest probability
49. A person who scores at the 50th percentile on a standardized test
50. Did as well or better than ½ of those who took the test
51. Missed exactly 50% of the questions on the test
52. Clearly is not knowledgeable about the field in which they took the test
53. Is an outlier from the norm amongst people taking the test
54. According to Han, Kamber and Pei, data warehouses are a multidimensional space for storing data

Answer true

1. The four key words in William H. Inmon’s definition of a data warehouse that separate them from other types of data storage structures are: subject-oriented, integrated, time-variant and nonvolatile.

Answer true

1. A library maintained by a business does not fit William H. Inmon’s definition of a data warehouse.

Answer false

1. OLTP is an acronym for online\_ \_transaction\_\_processing.
2. Data warehouses are not a form of online analytical processing (OLAP) systems

Answer false

1. The term “data cube” refers to a data structure that is defined by dimensions and facts.

Answer true

1. Data cubes consist of “n” dimensions, where n ≥ 2.

Answer true

1. If one dimension of a data cube is “location”, described by the attributes *number, street, city, province\_or\_state, zip-code and country,* then there is implicit in the location dimension a concept hierarchy defined in the schema for that database.

Answer true

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

Answer false

1. The slice and dice operations both form subcubes, but the slice is done on one dimension, while the dice is done on multiple dimensions.

Answer true

**Week 5**

1. A key motivation for using a multidimensional approach to the data to be analyzed is that aggregating data in multiple ways is:
2. Data cubes are an easy data structure for high-level managers to grasp
3. The computational efficiency of algorithms that work with more than two dimensions is very high
4. Important
5. Non-existent, there is no key motivation to do so
6. Data cubes:
   1. Do not need to consist of dimensions (D) of equal size (number of attributes per Di)
   2. Are not necessarily three dimensional, despite the name, data *cube*.
   3. Are a generalization of what is meant by the term “cross-tabulation”.
   4. Are defined as being multidimensional representations of data and all possible aggregations (totals) of that data.
   5. None of the above.
7. The least squares method is often used to fit data to which of these models:
8. *f*(x) = ω1x + ω0
9. *y* = *f*(x) + [*y* – *f*(x)]
10. *y* = *f*(x, Ω) + ε
11. None of the above
12. Regression
13. Is used to model data via a mathematical function that minimizes the error term.
14. Involves determining a function, *f*(x), which describes (fits) the data without error.
15. Is a difficult task that involves guessing what values of the regression coefficients, ω1, and ω0, will be most efficiently used in linear regression computations.
16. Is not used in solving data mining problems.
17. The error function in regression analysis can be determined by
18. Removing outliers
19. Summing either the Absolute or the Squared Error, as defined by these equations:

(Absolute Error) and (Squared Error)

1. Solving this equation: *f*(x) = *f*(x, Ω) + ε.
2. No known method – it is undeterminable, as we can understand it.
3. Regression analysis
4. Is useless with data sets that can be represented multidimensionally.
5. Is based on resolving this equation: D = { (x*i*, y*i*) | *i* = 1, 2, 3, …, *N*}
6. Involves the concept of explanatory attributes that are either discrete or continuous.
7. All of the above
8. None of the above
9. The response variable, *y*, in regression analysis
10. Is always deterministic – for a given attribute set, *y* always resolves to the same value
11. Like the response variable *x*, can be either discrete or continuous.
12. Can be non-deterministic, which can be modeled by probabilistic approaches.
13. Represents the random noise in any given data set being analyzed
14. Computing aggregate totals in data cubes:
15. Is always computationally efficient.
16. Involves fixing specific values for some set of attributes that define the dimensions of the cube and then summing over all the possible values for the remaining attributes that make of the cube’s dimensions.
17. Is not part of the analysis effort in data mining that uses data cubes
18. None of the above
19. Drill down operations are needed
20. When the summary data at a given level of abstraction is insufficient to reveal important patterns in the data, such as sales of milk and honey on specific days of the week when the data you have is only at the monthly level of abstraction.
21. Involve summarizing to a higher level of abstraction, such as summarizing daily sales information into weekly or monthly.
22. Are very expensive computationally, and so are to avoided at all costs.
23. All of the above
24. A two-dimensional table
25. Is too simple a data structure to reveal significant patterns in data sets.
26. Is a form of the data structure known as a “data cube”.
27. Is a complex data structure that is computationally difficult to analyze
28. While basic to the definition of relational databases is never found in serious enterprise level data warehouses.
29. Data cubes can store precomputed measures, such as aggregated values of dimensions of an attribute such as daily sales totals.

Answer true

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

Answer false

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

Answer false

1. Preprocessing, or computing, the values to be stored in a given data cube allows for more efficient real-time querying in a multidimensional online analysis processing environment.

Answer true

1. A data cube of n dimensions contains2n\_cuboids\_.
2. A data cube can be usefully viewed as a lattice of cuboids.

Answer true

1. If we aggregate only those cells where the number of items bought by a particular customer on a given day is greater than x, then the resulting partially materialized data cube is an\_Iceberg\_ data cube.
2. The following SQL code precomputes cells in a data cube that contain a count value that is equal to or above a specified minimum support value:

compute cube sales iceberg as select month, city, customer group, count(\*) from salesInfo cube by month, city, customer group having count(\*) >=min\_sup

Answer true

1. Drill down operations on a prediction cube is a computational challenge, given the need to materialize cell values at many different granularity levels.

Answer true

1. A prediction cube differs from a data cube in that the cell values in the latter are aggregates computed over the data subset of the cell, while in a prediction cube is computed by evaluating a \_predictive model\_ built on the data subset of the cell.

**Week 6**

1. Which of the following is / aretrue of arule extracted via association analysis (AA):
2. Always describes an association between / among two or more items that is non-random, not based in chance
3. Can be, but is not always, reliably useful to predict future behaviors of the population described by the data from which the rule is derived
4. Is often computationally expensive to discover via AA
5. Must always be assumed to be true in 100% of all future behaviors of the population that is described by the data from which the rule was extracted
6. If a binary representation is chosen for market basket data, then
7. It is true that the presence of an item is often less important than its absence
8. The item variable is a binary type referred to as asymmetric, because the value one is often more important to the final results than a value zero for any given item
9. The only two possible values for an item variable are ‘-1’ and ‘1’
10. Calculations on item variables become very computationally inefficient because computers cannot easily handle binary valued variables
11. If the equation, , represents the value known as the support count in associative analysis, then
    1. It must be true that *ti* represents a member of the itemset X
    2. Then the higher the value, is determined to be, the less likely that the itemset is meaningful in the final analysis results
    3. cannot be the support value, because the set T is clearly always going to be equal to the null set
    4. X is the itemset for which we are trying to determine the number of transactions, , in the data set of transactions, T, that contain X
12. The theorem that states “If an itemset is frequent, then all of its subsets must also be frequent” is often referred to as the
13. Theorem that can never be proved
14. Apriori Principle
15. The key to understanding market basket analysis
16. All of the above
17. In associative analysis, “frequent” or, put differently, the minimum support value, that an itemset must have to considered in the final results
18. Is specific to the data being analyzed
19. Is a constant for a given organization’s data, but can be different across organizations
20. Is often found to be non-determinable
21. Is often just a random value chosen by the analyst and so cannot be tested as to its validity
22. If there exists these two confidence rules, , , where for X, Y, , both of these are true, , which of the following is true:
23. and
24. No such confidence rules can co-exist
25. The value of , can be higher, equal to, lower than the value of
26. The value , can only be equal to the value of ,
27. Maximal and closed itemsets are
28. Rare forms of itemsets seldom used in associative analysis (market basket analysis) that is only used in data mining if absolutely necessary given the difficulty of obtaining it and processing it
29. Useful as a compaction of large collections of itemsets where the originating collection is too large to compute associative rules cost effectively
30. Non-existent
31. None of the above
32. A closed itemset (choose the true answer)
33. Has one or more immediate supersets where the support count is identical to its support count
34. Has no immediate supersets where the confidence value is identical to its own
35. Is of no use in data mining
36. Has no immediate supersets where the support count is identical to its own
37. Which of these possible attributes of a maximal itemset actually is the defining one:
38. None of the immediate subsets of the given itemset are frequent, but it is frequent
39. All of the immediate supersets of the given itemset are frequent, but it is frequent
40. None of the immediate supersets of the given itemset are frequent, but it is frequent
41. One or more of the immediate supersets of the given itemset are frequent, but it is not frequent
42. Which of these algorithms is often useful in finding maximal itemsets in a collection:
43. Shallow first
44. Binary sort
45. Backtrack
46. Depth-first
47. Because the FP-Growth algorithm abandons the *generate and test* approach of the Apriori algorithm in favor a significantly more direct paradigm of storing into a compact data structure and directly selecting the frequent itemsets from the structure, one can safely say the FP-Growth algorithm is a radical departure from Apriori.

Answer true

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

Answer false

1. Correlationanalysis can be used to supplement support-confidence frameworks to discover interesting patterns, especially when low support thresholds are being used.
2. The end user of an analysis is the only one who can ultimately judge if a given rule is interesting in terms of the results it produces.

Answer true

1. A good interestingness measure will not be impacted by transactions that do not contain itemsets of interest, because measures that are so impacted generate unstable results
2. If the function Lift(X, Y) returns a value less than 1, then the presence of event X in a given set of events most likely means event Y is absent, and the reverse is also true.

Answer true

1. If the χ2 value is greater than 1 and the observed value of a slot (X, Y) is less than the expected value, then there is a negativecorrelation between members X and Y of the slot
2. To say that the measures for interestingness, *lift* and χ2 are not null-invariant, is to assert that their values are not independent of the number of null transactions in the data set being analyzed.

Answer true

1. Han, Kamber and Pei recommend the use of the *Kulc* null-invariant measure in conjunction with the imbalance ratio in determining interestedness.

Answer true

1. All strong association rules are interesting

Answer false

1. \_\_\_\_\_\_\_\_\_\_ is a methodology useful for discovering interesting relationships within large sets of data.
2. Big Data
3. Association analysis
4. Data Mining
5. Algorithm
6. Market basket transactions show:
7. eggs, milk, and bread
8. monthly customer purchases
9. data relationships
10. daily customer purchase data.
11. Sets of frequent items hidden in large data sets are called \_\_\_\_\_\_\_\_\_.
12. associationrules.
13. big data.
14. binary representation.
15. analysis.
16. Two key issues that must be addressed when applying association analysis are:
17. Overpopulation
18. Computational Expense
19. Discovering spurious patterns
20. Data density
21. The strength of an association rule can be measured in terms of its\_\_\_\_ and \_\_\_\_\_.
22. support
23. finances
24. confidence
25. size
26. The \_\_\_\_\_\_ approach for mining association rules is to compute the support and confidence for every rule.
27. Hard-nosed
28. Brute force
29. Comprehensive
30. Association Rule
31. The objective of the \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ strategy is to find all the items that satisfy the minsup threshold.
32. Association Rule Discovery
33. Correct Answer
34. Frequent Itemset Generation
35. Incorrect Answer
36. The objective of the \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ strategy is to extract all the high-confidence rules from the frequent itemsets found in the Frequent Itemset Generation.
37. Association Rule
38. Rule Generation
39. Association Analysis
40. Association Generation
41. Trimming the exponential search space based on the support measure is known as:
42. Exponential Pruning
43. Apriori Algorithm
44. Trimming
45. Support-Based Pruning
46. \_\_\_\_\_\_\_\_\_\_ are visual structures that use branches and leaf nodes to search an item or itemset.
47. Functions
48. Hash Trees
49. Root Node
50. Algorithms
51. Market-based analysis studies customers’ buying habits by searching for itemsets that are frequently purchased together (or in sequence).

TRUE

1. The process of flipping a bit vector is called inversion.

TRUE

1. Simpson’s paradox is a phenomenon where hidden variables may cause the observed relationship to multiply.

FALSE

1. Cross support patters are likely to be spurious because their correlations tend to be weak.

TRUE

1. Objective measures rank patterns according to user’s interpretations.

FALSE

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

FALSE

1. The Apriori Principle says that if an itemset is frequent then all of its subsets must also be frequent.

TRUE

1. In general, confidence has an anti-monotone property.

FALSE

1. Highly correlated and strongly associated patterns are called hyperclique patterns.

TRUE

1. Frequent pattern mining has reached far beyond basics due to substantial research, numerous extensions of the problem scope, and broad application studies.

TRUE

**Week 7**

1. According to Han, Kamber and Pei, which of the following is not a high-level pattern mining research field:
2. Mining Methods
3. Kinds of Patterns and Rules
4. Extended Data Types
5. Extensions and Applications
6. According to Han, Kamber and Pei, which of the following is not an extended data type that researchers in the pattern mining field are studying
   1. Spatial (such as colocation)
   2. Rational and Integer number patterns
   3. Network patterns
   4. Temporal patterns
7. The pattern of buying by customers of specific items such as (Tablet Phone Headset) is known as a
8. Temporal Pattern
9. Sequential Pattern
10. Structural Pattern
11. Indeterminate Pattern
12. None of the above
13. Using reduced minimum support at lower levels is a
14. Technique applicable to mining data where multiple levels of abstractions are used to describe the data
15. Technique in data mining that can be very rewarding, but computation costs are high
16. Incomplete results
17. All of the above
18. When mining under reduced or group-based support, the Apriori property
19. Is never impacted
20. Cannot be relied on to hold uniformly across all of the items being analyzed
21. Is very reliable across all the items being analyzed
22. None of the above
23. Pattern-Fusion mining is useful for which of these data set situations:
24. Large number of dimensions (features / items) to analyze but a small sample size
25. Few dimensions, large sample size
26. Large sample size, large number of dimensions
27. All of the above
28. None of the above
29. The Pattern-Fusion data mining strategy uses which of these tree traversal approaches
30. Depth-first
31. Bounded-breadth
32. Breadth-first
33. None of the above
34. Mining for colossal patterns is computationally beyond the capability of
35. Methods that incrementally increase patterns one item at a time
36. Depth-first methods that generate too many sub-trees, such as FP-Growth
37. Modern computers, except the absolutely latest entries in the supercomputer field.
38. None of the above
39. The output structures of one approach to image analysis and recognition is referred to as
40. Visual bytes
41. Sound bits
42. Visual words
43. None of the above
44. Mining using multilevel associations
45. Involves data which is strictly numeric
46. Is always far too computationally costly to actually do – thus it is strictly a theoretical model for data mining
47. Means that there exist multiple levels of abstraction to describe the data
48. All of the above
49. None of the above
50. Optimization in the world of statistics means to discover the minimum or maximum values of a given function.

Answer true

1. The K-means clustering algorithm seeks to determine a set of clusters where the sum of the squared error is optimized in the sense of a minimal value is sought

Answer true

1. When analytical methods for determining optimal values of a given function fail, we often turn to numerical methods for approximating those values

Answer true

1. Newton’s method determines exact / approximate optimal values
2. Two types of constraints must be dealt with in some forms of optimization: Equality and Inequality
3. The following equation holds true for negatively correlated association rules,

ϱ(Χ ∪ Υ)< σ(Χ)ϱ(Υ), where Χ and Υ are disjoint itemsets,

Χ ∪ Υ =∅

1. The graph isomorphism problem comes into play when pruning candidate subgraphs during subgraph mining operations
2. A candidate *k-*sequence is pruned if at least one of its (*k* – 1)-sequences is infrequent.
3. A sequence *t* is a subsequence of another sequence, *s*, if each ordered element in *t* is a subset of an ordered element in *s*.
4. A concept hierarchy is a set of concepts arranged hierarchically and drawn from a specific knowledge domain.

Answer: true

1. If a rule describes associations between quantitative items or attributes, then it is a:
2. Qualitative Association Rule
3. Quantitative Association Rule
4. Association Rule
5. Boolean Association Rule
6. A pattern that satisfies a minimum support threshold is a(n):
7. Infrequent Pattern
8. Closed pattern
9. Max-Pattern
10. Frequent Pattern
11. If a rule involves associations between the presence or absence of items, it is called a:
12. Boolean Association Rule
13. Multidimensional Association Rule
14. Quantitative Association Rule
15. Approximation Rule
16. A concept hierarchy can be represented using (a):
17. Hierarchy Tree
18. Taxonomy
19. Nodes
20. Directed Acyclic Graph
21. Event based data have an inherent \_\_\_\_\_\_\_\_\_\_\_\_\_ nature to them.
22. Mysterious
23. Misunderstood
24. Sequential
25. Unorganized
26. \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ recommend(s) information items that are likely to be of interest to the user based on similar users’ patterns.
27. Recommender Systems
28. Tendency Reports
29. E-Commerce Data
30. Analysis Tasks
31. A \_\_\_\_\_\_- sized threshold can be defined to specify the maximum allowed time difference between the latest and earliest occurrences of events in any element of a sequential pattern.
32. Giant
33. Window
34. Pattern
35. Limited
36. Quantitative attributes are \_\_\_\_\_\_\_\_\_\_\_.
37. Alphabetical
38. Random
39. Numeric
40. Nominal
41. A(n) \_\_\_\_\_\_\_\_\_\_ pattern has a frequency support that is below a user specified minimum support threshold.
42. Infrequent/Rare
43. Frequent/Regular
44. Large/Significant
45. Small/Minimal
46. The lower-level core patterns of a colossal pattern are called \_\_\_\_\_\_\_\_\_\_.
47. Basic patterns
48. Central descendants
49. Basic descendants
50. Core descendants
51. Mutual information is widely used in information theory to measure the mutual independency of two random variables.

TRUE

1. Which of the following is not a valid timing constraint of a pattern?
2. Mingap
3. Maxgap
4. Maxspan
5. Minmax
6. The objective measure ranks patterns based on hypothetical predications.

FALSE

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

FALSE

1. The approach of iteratively expanding a subgraph by adding an extra vertex is known as vertex growing.

TRUE

1. Determining whether two graphs are topologically equivalent is known as graph isomorphism.

TRUE

1. Object measures alone may be sufficient to eliminate uninteresting infrequent patterns.

FALSE

1. Interesting, infrequent patterns are also known indirect association patterns.

TRUE

1. Negative itemsets and negative association rules are collectively known as positive patterns.

FALSE

1. Negative-correlated patterns are useful for identifying competing items, or items that can be substituted for one another.

TRUE

**Week 9**

1. Choose all of the following that are a prediction problem that comes under the task of classification:
2. Please tell me which of my current mortgage applicants I can safely approve for a mortgage and which are unsafe at this time to approve for any kind of mortgage.
3. Please tell me the likelihood of each of my current mortgage clients to default on their mortgage in the next 12 months.
4. Please tell me the maximum monthly payment Client X can safely make, that is avoid becoming a risk for foreclosure
5. Please tell me which of my current mortgage clients are safe candidates for a low fixed interest rate mortgage, which for an initial low-rate, variable rate mortgage, and which for a high,fixed interest rate mortgage.
6. Which of these terms describes the first major task in the data classification process:
7. Classify
8. Learning
9. Choose training data
10. Analyze the training data for possible “noise” in the database
11. None of the above
12. If the classes of the training data are an unknown, we can apply which of these computational algorithms to attempt to find useable classes
13. Binary Sort
14. Bayesian Analysis
15. Pruning Analysis
16. Clustering
17. Which of these are typically used to represent the sought-after mapping function in the first step of the classification process:
18. Decision Trees
19. Mathematical formula
20. Classification rules
21. All of the above
22. None of the above
23. Mapping function accuracy:
24. Is measured in terms of the percentage of test tuples/samples that the function correctly classified/labeled
25. Must be 98% to be useful
26. Are useful if the data can be spatially related, in reality or virtually.
27. None of the above
28. Bayesian classifiers
29. Are the least accurate of the three kinds of classifiers
30. Require the maximization of the product of the posterior probability that our evidence, X, is in a given class when we know our evidence, X, satisfies our hypothesis and the prior probability that our hypothesis holds for any given evidence, X
31. Are computationally nearly impossible to use in the real world of data mining, but provide a good theoretical basis for other classifiers
32. All of the above
33. The zero-probability issue can be handled by which of these approaches:
    1. Applying the Laplacian correction technique
    2. Applying the Laplacian estimator technique
    3. Applying the add 1 to the number of tuples in each class in our data set technique
    4. All of the above
    5. None of the above
34. We often perform rule extraction from decisions trees because
    1. Humans understand tree structures much better than they do IF-THEN rules
    2. Humans are rule-based thinkers.
    3. Decision trees can become so large that interpretation can become difficult, if not impossible for the human mind, but rules are easier to interpret for many humans.
    4. None of the above
35. Which of these attribute lists is most true for the rule set extracted from a decision tree:
    1. Such rules are mutually exclusive, exhaustive and unordered
    2. Such rules are non-exclusive, exhaustive, ordered
    3. Logical OR exists between such rules, they are unordered.
    4. None of the above
36. Ensemble methods for improving accuracy most resemble which of these situations:
    1. A person gathers the opinions of 10 doctors for how best to cure that person’s illness, the person chooses the cure approach most often recommended by the 10 doctors
    2. A person gathers the opinions of 100 randomly chosen doctors for how best to cure that person’s illness, the person chooses the cure approach recommended the most by doctors with incomes greater than the median for the group
    3. A person gathers the opinions of the three top-rated doctors in the world in the field of medicine that his illness falls within on how best to cure the person’s illness and the opinions of 5 other randomly chosen doctors within the same field, and chooses the opinion of the top doctor who is matched by the most doctors in the random chosen group.
    4. None of the above
37. According to Tan, Steinbach and Kumar, classification, the act of assigning objects to one of a set of predefined categories, is a problem area for many data mining applications

Answer true

1. Tan, Steinbach and Kumar assert that the input data for a classification task is a collection of records. Han, Kamber and Pei prefer the term tupleinstead of record.
2. Tan, Steinbach and Kumar assert that classification techniques work much better for data sets that are classifiable by binary or nominal categories than those that are classified by ordinal categories.
3. Hunt’s algorithm depends on an attribute test conditionto determine how to split a set of records that can be labeled by more than one category or class.
4. To say that a given node in classification tree is pure is to say that all the records in the node belong to one and only one of the predetermined classes

Answer true

1. Entropy, in the case of determining the best split, is a measure of the skewedness of the class distribution, or put differently, the degree of impurity in the child nodes.
2. To inductively grow a decision tree, one needs to be able to create nodes, find the best split of records in a node into two or more nodes, classify the leaf nodes, and determine when to stop trying to grow the tree.

Answer true

1. If a training set lacks sufficient example records of given class, then misclassification or overfitting occurs.
2. Einstein’s assertion that we should make everything as simple as possible, but no simpler is another way of stating the test for selecting the best model from several possible models for data mining that is both simple in its structure and competent, i.e., not simpler than what it is possible.

Answer true

1. Prepruning requires a sufficiently restrictive stopping condition that prevents the decision tree induction algorithm from growing a tree that is highly prone to overfitting the data.
2. \_\_\_\_\_\_\_\_\_\_ is a form of data analysis that extracts models describing important data classes.

A) Classification

B) Model

C) Attribute set

D) Class model

1. Data classification is a \_\_\_\_\_\_\_ step process.

A) three

B) one

C) two

D) four

1. A(n) \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ is a flowchart-like structure used as a classification technique.

A) algorithm

B) node

C) branch

D) decision tree

1. Each classification technique employs a \_\_\_\_\_\_\_\_\_\_\_\_\_\_ to identify a model that best fits the relationship between the attribute set and class level of the input data.

A) learning algorithm

B) binary split

C) test set

D) internal node

1. A(n) \_\_\_\_\_\_\_\_\_\_\_\_\_ shows counts of test records correctly and incorrectly predicted by a classification model.

A) decision tree

B) attribute class

C) confusion matrix

D) learning model

1. Attribute selection measures are also known as \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_.

A) splitting rules

B) majority voting

C) backpropagation

D) classifiers

1. The basis of existing decision tree algorithms ID3, C4.5, and CART is called \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_.

A) ID4

B) Gini index

C) Hunt’s Algorithm

D) Information gain

1. Gain ratio tends to prefer \_\_\_\_\_\_\_\_\_\_\_\_ splits.

A) balanced

B) multiple

C) no

D) unbalanced

1. A statistical technique that creates several smaller samples (subsets) is called:

A) scaling

B) division

C) bootstrapping

D) subsetting

1. \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ assumes that the effect of an attribute value on a given class is independent of the values of the other attributes.

A) RainForest

B) Tree Pruning

C) Oversampling

D) Naïve Bayesian classificant

1. There cannot be more than one decision tree that fits the same data.

FALSE

1. A classification model can also be used to predict the class label of unknown records.

TRUE

1. “Overfitting” is when a tree becomes too large and its test error rate begins to increase, even though its training error rate continues to decrease.

TRUE

1. 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

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

FALSE

1. A precise measure will label a high % of tuples correctly.

TRUE

1. Bootstrapping does not work well with small sets.

FALSE

1. A “nulled” hypothesis is when two variables, such as M1 and M2 are the same.

TRUE

1. Significance tests and ROC curves are useless for model selection.

FALSE

1. No method has been found superior for all data sets.

TRUE

**Week 10**

1. Bayesian belief networks (BBN) differ from naïve Bayesian classifiers (BC) in that:
2. BCs assume the values of the attributes in a tuple are conditionally independent, while BBNs do not make that assumption
3. BBNs are always more computationally more efficient than are BCs
4. BCs are always more accurate, even when the data tuples have attribute values that are not necessarily conditionally independent
5. No significant differences exist
6. Which of these best describe the components of a Bayesian Belief Network:
   1. Directed acyclic graph where each node represents a non-random attribute of the tuples/records in the data set and a set of conditional probability tables
   2. Directed acyclic graph and a set of conditional probability tables
   3. Acyclic graph and a related set of probability tables
   4. Directed acyclic graph where each node represents a random variable corresponding to actual of the tuples/records in the data set or to ‘hidden variables’ and a set of conditional probability tables
7. Classification rules can be represented by , where
8. is the *ith* rule in a set of rules, representing a rule precondition or antecedent
9. represents a conjunction of attribute tests
10. is the predicted class, or rule consequent
11. All of the above
12. None of the above
13. In BBNs each arc in the acyclic graph represents
14. The probabilistic dependency of the two connected nodes
15. A non-dependent relationship of the two connected nodes
16. An educated guess as to a causal relationship between the two connected nodes
17. None of the above
18. Use of ordered rule sets
19. Guarantees computability
20. Avoids the problem of conflicting assignment of class labels, since the highest rated rule always wins
21. Is often non-desirable because of the high computational costs involved
22. Is not recommended by experts in the field, because it restricts the possibility of finding all interesting/useful categories for a given record/tuple
23. Bayesian belief networks
24. Are best built using completely automated means, with no human intervention
25. Are always above 90% in their accuracy, assuming they are correctly built
26. Can be built at a faster learning rate if human experts can provide prior knowledge as to the possible best network topology and/or conditional probability values
27. All of the above
28. Backpropagation
29. Is a training method used for Bayesian Belief Networks
30. Involves updating the weights in a neural network during its training by computing the error margin between a given class prediction and the actual class and feeding that information back to the beginning hidden layer, re-computing the weight at that point, and moving forward again through to the output layer until weight(s) are optimal, or a threshold of optimality is reached
31. Is a determinate approach to iteratively reaching an optimal set of weights for a neural network
32. Is an indeterminate approach to discovering and communicating to the whole neural network the best possible weights
33. Support vector machines involve:
34. Projecting the training data into a lower dimension
35. Discovering support vectors (data tuples) in different classes that are furthest from the hyperplane separating classes
36. Discovering hyperplanes that separate classes from each other with no overlap
37. None of the above
38. Classification using frequent patterns
39. Involves classifying data by the generating rules of association of a given frequent pattern and a given class, with the goal of seeking patterns that are the least associated with a given class
40. Is non-existent.
41. Isquite inexpensive, and so worth trying even if the data isn’t all that wonderful
42. Involves classifying data by discovering and applying rules of pattern association based on frequent patterns that exist in the data set
43. Case-based reasoning is
44. A very complex method of classification that involves attaining the highest form of artificial intelligence yet constructed
45. A form of classification that involves the matching of a given case or situation with the data stored in a database concerning past solutions to problems. If a match or near match is found, then the solution is given.
46. Is theoretically achievable, but no one has actually succeeded with it since it involves using artificial intelligence techniques to read and understand textual information
47. None of the above
48. Rule-ordering classification approaches can only be done rule-by-rule. It is not possible to order classes.

Answer false

1. 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

1. The Learn-One-Rule function handles the computational complexity issue, where search costs can grow exponentially, by growing the rule in a greedy fashion, stopping only when a preset stopping condition is met
2. Which of these is a Rule-Growing Strategy (select all that meet this criterion)?
   1. General to specific
   2. Start-in-middle
   3. Start at end
   4. Start at beginning
   5. Specific to general
3. Given the true probability distribution governing P(X|Y), the Bayesian classification method allows for the determination of the decision boundary that is ideal | most problematic. (circle the correct term)
4. What is the single equation that the parameters w and *b*, defining the decision boundary for a support vector machine being trained, must meet? Assume *yi* can take on one of two values, 1 and -1

*y1* (w \* w*i* + *b*) ≥ 1, *i* = 1, 2, ..., N.

1. For support vector machines, the decision boundary the w and *b* parameters define must be maximal.
2. The base classifiers used in an ensemble classification approach must meet two necessary conditions for the ensemble classifier to perform better than a single classifier.
3. When an ensemble classifier is built by manipulating the training set, there are how many classifiers?

One per training set

1. A false positive result when classifying a particular record / tuple can be very significant when the goal is detect records that fall into a rare class in the target population.
2. This belief network, used for classification via graphic models, allows the representation of dependencies among subsets of attributes.
   1. Atkinson Index
   2. Bayesian belief network
   3. Theil Index
   4. Simpson Index
3. \_\_\_\_\_\_\_\_\_\_\_\_\_\_ learns by iteratively processing a data set of training tuples, comparing the network’s prediction for each tuple with the actual known target value.

A) split

B) set

C) Backpropagation

D) classifier

1. This theory can be used to approximately define classes that are not distinguishable based on the available attributes.
   1. Backpropagation
   2. Genetic algorithms
   3. Fuzzy set
   4. Rough set theory
2. A form of supervised learning that is also suitable for situations where data is abundant, yet the class labels are scarce or expensive.
   1. Passive learning
   2. Active teaching
   3. Active learning
   4. Passive teaching
3. This type of classification builds a classifier using both labeled and unlabeled data.
   1. Multi-class classification
   2. Semi-supervised classification
   3. Support vectors
   4. Instance based
4. The right side of the rule is called the \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_, which contains the predicted class.
   1. rule consequent
   2. precondition
   3. rule antecedent
   4. data set
5. An ordered rule set is also known as a \_\_\_\_\_\_\_.
   1. Shopping list
   2. Decision list
   3. Ordering scheme
   4. Exhaustive rule
6. \_\_\_\_\_\_\_\_\_\_\_\_\_ aims to extract the knowledge from one or more source tasks and applt the knowledge to a target task.
   1. Instance-based approach
   2. TrAdaBoost
   3. Transfer learning
   4. Negative transfer
7. In order to improve generalization errors in rules generate by the Learn-One-Rule function, one can use \_\_\_\_\_\_\_\_\_\_\_\_\_\_.
   1. Sequential covering
   2. Alternative techniques
   3. Rule growing
   4. Rule pruning
8. \_\_\_\_\_\_\_\_ chooses the majority class as its default class and learns the rules for detecting the minority class.
   1. FOIL’s information gain
   2. RIPPER
   3. Rote classifier
   4. Support vector machine
9. A support vector machine is an algorithm for the classification of both linear and nonlinear data.

TRUE

1. Decision tree and rule-based classifiers are examples of lazy learners.

FALSE

1. An example of an eager learner would be Rote classifier.

FALSE

1. Training examples that are relatively similar to the attributes of the test example are called nearest neighbors.

TRUE

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

FALSE

1. Naïve Bayes classifiers are robust to irrelevant attributes.

TRUE

1. Which of the following are two key elements of a Bayesian network?
   1. a DAG
   2. a decision tree
   3. an error rate
   4. a probability table
2. An illustration of a simple neural network architecture is called a Perceptron.

TRUE

1. ANN stands for:
   1. Artistic Neutral Neurons
   2. Artificial Neural Network
   3. Algorithm Network Neuron
   4. Attributes Neural Naiveté
2. The kernel trick is a method that solves issues with irrelevant attributes.

FALSE

**Week 11**

1. According to Han, Kamber and Pei, which of the following is the best tool for dealing with data that needs to be labeled, but there are no predetermined labels:
2. Pattern Matchers
3. Classifiers
4. Clusterers
5. Recognizers
6. According to Han, Kamber and Pei, which of the following is an application where data segmentation is useful
   1. Market basket analysis
   2. Outlier detection
   3. Trend prediction
   4. Temporal pattern discovery
7. Which of these is not considered a basic clustering technique by Han, Kamber and Pei
8. Partitioning methods
9. Sequential methods
10. Hierarchical methods
11. Density-based methods
12. Grid-based methods
13. Clustering is an approach that uses which learning technique:
14. By observation
15. By examples
16. By listening
17. All of the above
18. Which of these is an example of clustering under constraints
19. Grouping a set of randomly chosen web sites by primary subject categories that are not predetermined
20. Discovering groups of people in a city with similar buying habits regarding purchasing of automobiles living within a 20 minute walk of each other
21. Discovering the best places in a city to place parks, taking into account physical barriers for people to get there by walking, such as rivers, highways, and distance, so that all people in the city can get to at least one park by foot in 15 minutes or less
22. None of the above
23. K-means is a
24. Method for creating clusters based on similarities amongst members of a cluster
25. Method for clustering that requires an initial value, k, that predetermines the number of clusters to create from a given data set
26. Method that avoids the fact that optimizing the within-cluster variation is NP-hard for k clusters in 2-D Euclidean space
27. All of the above
28. Hierarchical clustering methods
29. Are always top-down in their approach to forming hierarchies
30. Are impervious to the choice of split / merge points
31. Start either with individual objects as clusters or with the entire data set as a single cluster
32. All of the above
33. None of the above
34. Density-based methods
35. Resolve the issue of finding clusters of arbitrary shapes
36. Includes DBSCAN, which starts with a parameter, ε > 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
37. Include the concept of density-connectedness to form clusters of small of dense regions
38. All of the above
39. Grid-based clustering methods
40. Enjoy fast processing times because the time is relationally dependent upon the number of data objects
41. Include STING, which can also be considered a density-based clustering method
42. Are computationally expensive
43. None of the above
44. According to Han, Kamber and Pei, which of these is not a measure used to evaluate clustering methods
45. Clustering quality
46. Computational costs
47. Clustering tendency
48. Determination of number of clusters in a data set
49. If the sum of the squared error of a set of clusters is near 0, then we can be reasonably certain that the clusters are valid.

Answer true

1. Even if the difference between two clusters is statistically significant (repeatable), that does not necessarily mean that the magnitude of that difference is significant for any given application. That is, a difference of 0.1% for application A might be significant, while for application B, such a low difference is not important.

Answer true

1. Many clustering approaches assume the following:
   * 1. If two objects were created at essentially the same time, then they must belong to the same cluster, and thus have the same class label
     2. If two objects have the same class label, then they must belong to the same cluster, but not the reverse.
     3. If two objects are in the same cluster, then they must have the same class label, but not the reverse.
     4. If two objects are in the same cluster, then they must have the same class label, and the reverse is also true
     5. None of the above
2. If we maximize the recall competency of a clustering algorithm we are most likely lowering the precision competency.
3. Maximizing the recall competency increases the likelihood of false positives
4. By minimizing separation of clusters, we are maximizing the cohesion of the individual clusters.
5. In graph-based clusters, the measure of the cohesion of a given cluster is found by summing the weights associated with links between members of the cluster
6. In graph-based clusters, the measure of the separation between two given clusters is found by summing the weights associated with links between the points in one cluster and points in the second cluster.

Answer: true

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

Answer: false

1. If neither one of two given clusters are clearly cohesive, but they are in close proximity to each other, merging the two clusters is often a good option.
2. A(n) \_\_\_\_\_\_\_ is a collection of data objects that are similar to one another within the same cluster and are dissimilar to the objects in other groups.
3. application
4. cluster
5. data group
6. field
7. Which of the following is *not* a basis for grouping cluster.
8. Prototype-based
9. Graph-based
10. Density-based
11. Quality-based
12. This is a prototype-based, partitional clustering technique that attempts to find a user-specified number of clusters, which are represented by their centroids.
    * 1. K-means
      2. Agglomerative Hierarchical Clustering
      3. DBSCAN
13. This is a density-based clustering algorithm that produces a partitional clustering in which the number of clusters is automatically determined by the algorithm.
14. K-means
15. Agglomerative Hierarchical Clustering
16. DBSCAN
17. A clustering technique that starts with each point as a singleton cluster and then repeatedly merging the two closest clusters until a single, all-encompassing cluster remains.
18. K-means
19. Agglomerative Hierarchical Clustering
20. DBSCAN
21. \_\_\_\_\_\_\_\_\_\_\_\_\_ assesses the feasibility of clustering analysis on a data set and the quality of the results generated by a clustering method.
22. Grid-based method
23. Analytics
24. Clustering evaluations
25. Hierarchical method
26. We commonly refer to prototype based clusters as:
27. Prototype clusters
28. Center-based clusters
29. Clique
30. Contiguity-based
31. This technique is based on picking an initial solution and then repeating two steps: compute the change and update the solution.
32. Gradient descent
33. K-means
34. Centroids
35. Hierarchical methods
36. Which of the following are two basic approaches for generational hierarchical clustering?
37. Agglomerative
38. Divisive
39. Inclusive
40. Conglomerate
41. \_\_\_\_\_\_\_\_\_ aims to overcome some of the disadvantages of a cluster hierarchy by using probabilistic models to measure distances between clusters.
42. Multiphase hierarchical clustering
43. Divisive hierarchical clustering
44. Density-based clustering
45. Probabilistic hierarchical clustering
46. A tree like diagram that graphically displays hierarchical clustering:
47. Centroid
48. Graph
49. Clique
50. Dendrogram
51. A density-based method clusters objects based on the notion of clique.

FALSE

1. Grid based clustering approach quantizes the object space into a finite number of cells.

TRUE

1. The process of grouping a set of physical or abstract objects into classes of similar objects is called clustering.

TRUE

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

FALSE

1. Clustering is a relatively simple field.

FALSE

1. CLIQUE is a simple grid-based method for finding density-based cluster in subspaces.

TRUE

1. If ground truth is available, it can be used by intrinsic methods.

FALSE

1. If the ground truth is available, it can be used by extrinsic methods.

FALSE

1. Biology is one of many applications of cluster analysis.

TRUE

**Week 12**

1. Which of the these is not an accurate comparison between K-means and DBSCAN:
2. K-means can be applied to sparse, high-dimensional data, such as document data. DBSCAN typically performs poorly for such data because the traditional Euclidean definition of density does not work well for high dimensional data
3. DBSCAN produces the same set of clusters from one run to another, while K-means does not, as it is typically used with random initialization of centroids
4. K-means uses a prototype-based notion of a cluster, while DBSCAN uses a density-based
5. K-means uses a density-based notion of a cluster, while DBSCAN uses a prototype-based concept
6. The example of holding the number of data points constant, while increasing the number of dimensions, and thus the volume occupied by those points
7. Illustrates the flexibility of K-means to handle high dimensionality, low number of data points
8. Illustrates why DBSCAN performs poorly on high dimensional data sets, as the Euclidean definition of density as being the number of data points per unit of volume becomes meaningless as the volume increases with the adding of dimensions
9. Illustrates how the higher the number of dimensions forces the number of clusters higher, until at some point every cluster will consist of one data point
10. All of the above
11. None of the above
12. If the classes of the training data are an unknown, we can apply which of these computational algorithms to attempt to find useable classes
13. Binary Sort
14. Bayesian Analysis
15. Pruning Analysis
16. Clustering
17. The example of scale as an issue in clustering where the attributes used to do the clustering are height and weight of a group of people, and height is measured in meters and weight is measured in kilograms, thus favoring weight as the main attribute determining cluster membership is true because:
18. The numeric value of meters will always be a much smaller number than the numeric value of kilograms associated with any given individual (e.g. 2 meters, 100 kg), thus favoring weight in determining similarity or dissimilarity between two people.
19. The two attributes are simple too different to be useable by any numeric-based clustering algorithm
20. Height is not a serious differential attribute when comparing people, while weight is very much more a determinant of the similarity of two or more people
21. None of the above
22. Processing order of the data
23. Never impacts the accuracy of any know clustering algorithm
24. 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
25. While important to how accurately certain clustering algorithms perform, the inaccuracies are determinant, and thus can be accounted for through normalization techniques
26. None of the above
27. K-means clustering algorithms are indeterminate as to results from run to run on the same data because
28. The clusters that are presented at the end are chosen from a group of clusters produced during processing using a randomly initialized selecting function
29. The initialization step for K-means algorithms includes a random selection of the beginning population of centroid objects
30. As clusters are generated and regenerated, the algorithm calls a randomization function to choose a different centroid for each of the current population of clusters
31. None of the above
32. CLIQUE (Clustering In QUEst)
    1. Relies on the monotonicity property of density-based clusters
    2. Finds subspace clusters despite the fact that the number of subspaces involved is exponential to the number of dimensions
    3. Uses an algorithm that is conceptually similar to the *Apriori* algorithm for finding frequent itemsets
    4. All of the above
33. Since graphing the likelihood of the data for different values of the parameters used in a clustering algorithm is not generally feasible, we
    1. Choose the parameter values randomly initially, and then keep iterating with the algorithm with different values until we get values that are accurate for the data.
    2. Finesse the issue by normalizing the values based on certain known characteristics of the data, such as the number of data objects to be clustered and number of attributes being used to determine cluster membership
    3. Rely on the fact that for a Gaussian distribution the mean and standard deviation of the sample data points are the maximum likelihood estimates for the corresponding parameters for the underlying distribution
    4. None of the above
34. The most useful feature of CLIQUE is
    1. That it finds clusters in subspaces efficiently
    2. That it is based on an already familiar algorithm, so its properties are well understood
    3. That it can summarize the list of cells that comprise a cluster with a small set of inequalities
    4. None of the above
35. Assume that you want to analyze legal issues involved in the lives of people with certain medical conditions, which of these clustering algorithms should be used to discover the appropriate documents for helpful information:
    1. An algorithm for determining clusters of documents based on the most prominent thematic element in each document
    2. An algorithm that determines clusters based on the two most prominent thematic elements in each document
    3. An algorithm that assigns a probability of cluster membership for each document across all of the clusters using all the identified thematic elements of each document
    4. None of the above
36. If a particular document belongs to the cluster LEGAL THEME with a degree of participation above 0.5 and to the cluster MEDICAL THEME with a degree above 0.7, then we can say that the likelihood of that document being of interest to those studying the connections between medicine and the law is strong.

Answer true

1. If a particular document belongs to the cluster LEGAL THEME with a degree of participation of 1 and to the cluster MEDICAL THEME with a degree of 0, then we can say that the likelihood of that document being of interest to those studying the connections between medicine and the law is low, but not necessarily non-existent.

Answer true

1. If the objects to be analyzed have ten or fewer attributes, according to Han, Kamber and Pei, the data are of low dimensionality
2. When analyzing high dimensional data, the attributes of the objects in a cluster that specifically determined those objects to be placed in that cluster are needed for the final analysis to be accurately informed, unlike with low dimensional data, where only the cluster objects need to be given over for analysis

Answer true

1. High dimensional data typically contains larger clusters then does low-dimensional data

Answer false

1. Computationally, highdimensional data sets often are very costly because they contain an exponential number of subspaces
2. Top-down approaches to clustering high dimensional data only work effectively when the subspace of a given cluster can be relied upon to be determinable by the local neighborhood (locality assumption)
3. Bi-clustering involves clustering both by the objects and the attributes, simultaneously
4. Which of these clustering algorithms allows an object to participate in multiple clusters, or even to not participate in any cluster
5. Fuzzy Clustering
6. Probabilistic Model-Based Clustering
7. Graph Clustering
8. Bi-Clustering
9. A constraint that an analyst is willing to change in order to obtain a more realistic clustering of the data is called a soft constraint
10. A \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ records the membership degree of objects.
11. Decision tree
12. Partition matrix
13. Mixture model
14. Constraint
15. These two methods of cluster analysis assign an object to one or more clusters.
16. Partition matrix
17. Fuzzy clustering
18. K-means clustering
19. Probabilistic model-based clustering
20. A \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ assumes that a set of observed objects is a micture of isntances from multiple probabilistic clusters.
21. Matrix model
22. Decision tree
23. Partition matrix
24. Expectation-maximization algorithm
25. This assumes that a cluster is a parameterized distribution.
26. Probabilistic model-based clustering
27. Partition matrix
28. Fuzzy clustering
29. High-dimensional data
30. A \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ records the membership degree of objects belonging to clusters.
31. High-dimensional data
32. Expectation-maximization algorithm
33. Subspace clustering method
34. Partition matrix
35. The number of edges between two vertices on a graph:
36. Spectral clustering
37. Modularity
38. Geodesic distance
39. Constraints
40. \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ cluster objects and attributes simultaneously.
41. Biclustering methods
42. Constant values
43. Optimization-based methods
44. Enumeration methods
45. Graph clustering can be modeled as computing \_\_\_\_\_\_\_\_\_\_\_\_\_.
46. Clusters
47. Graph cuts
48. Sparsest cut
49. Modularity
50. By strictly respective the constraints in the cluster assignment process, \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ can be enforced.
51. Soft constraints for clustering
52. SCAN
53. Hard constraints for clustering
54. Cannot-link constraints
55. \_\_\_\_\_\_\_\_\_\_\_\_\_ may be used to speed up constrained clustering.
56. Similarity measurement
57. Heuristics
58. SimRank
59. SCAN
60. Traditional clustering methods require each object to belong to many clusters.

FALSE

1. A sparsest cut will not likely lead to good clustering.

FALSE

1. Clustering with soft constraints can be considered an optimization problem.

TRUE

1. Biclustering methods cluster objects and attributes simultaneously.

TRUE

1. The optimization-based method is not a type of biclustering method.

FALSE

1. High dimensional data rarely pose a problem cluster analysis.

FALSE

1. Spectral clustering is a dimensionality reduction method.

TRUE

1. Constraints can be used to express application-specific requirements or background knowledge for cluster analysis.

TRUE

1. Name three categories of constraints for clustering:

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

1. Compare the differences between SCAN algorithm and DBSCAN. What are their similarities and differences?

**Week 13**

1. Which of these is not typically thought to be an application area where outlier or anomaly detection is paramount
2. Fraud detection
3. Intrusion detection
4. Ecosystem disturbances
5. Detection of books of interest to a library client
6. Which is not a known cause of anomalies
   1. Data object in a group of data objects that does not belong to the same class as all the other objects
   2. Data objects with one or more given attributes whose values differ significantly from the norm for that class of data objects
   3. Data errors
   4. Infrequent, random changes in the data caused by natural phenomena such as gamma rays, meteor strikes, etc.
7. Model-based techniques for anomaly / outlier detection
8. Do not work when no model currently exists
9. Are known to be brittle – changes in the data not seen before break the model
10. Should not be used when building the model is more difficult than using other well understood techniques that will do the job exist
11. All of the above
12. None of the above
13. Which of these are not normally found to be useful in outlier / anomaly detection
14. Fuzzy-based techniques
15. Model-based techniques
16. Proximity-based techniques
17. Density-based techniques
18. Distance measures are often preferred to objective function measures when assessing the extent a given object belongs to as cluster because
19. Distance measures are always more accurate then measures of objective function improvement
20. Determining objective function improvement when the given object is eliminated from the cluster is quite often computationally intensive, thus costly
21. Very few clustering techniques exist that use objective functions, and even when they do, distance measures still can be applied, and we understand distance measure more deeply than we do objective functions
22. Is not recommended by experts in the field, as our theoretical knowledge of objective functions is very sparse – more research is needed
23. Which of these does not make the use of clustering problematic in detecting outliers:
24. 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
25. 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
26. 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
27. All of the above
28. None of the above
29. Which of these situations would be considered anomaly detection in a semi-supervised mode:
30. Training set exists with full information needed about both normal and anomalous objects
31. Training set exists with information needed for normal objects only
32. No training set exists with even class labels for normal or anomalous objects
33. None of the above
34. Which of these situations is never true about an anomalous object
35. The value of a single attribute is anomalous, thus the object is anomalous
36. The values of some attributes are somewhat anomalous, but most are normal, thus the object is anomalous
37. 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
38. All of the above
39. None of the above
40. Which of these situations is referred to as a masking problem when attempting to detect outliers
41. Anomalies in the data set distort the model being applied to detect outliers, making outlier objects appear as normal objects
42. There is no such situation, no such problem as masking in outlier detection
43. Using one at a time anomaly detection fails to detect objects where the presence of two or more anomalies hides the presence of all the anomalies for that object
44. None of the above
45. When detecting outliers in a univariate normal distribution, which of these situations could negatively impact the accuracy of the detection:
46. Mean and standard deviation values are known, the number of observations is large
47. Mean and standard deviation are not known, and so must be estimated and the number of observationsis small
48. Mean and standard deviation are not known, and so must be estimated and the number of observations is very large
49. Mean and standard deviation values are known, the number of observations is small
50. None of the above
51. Outlier detection is the inverse of cluster analysis in that the former process seeks to find objects that do not fit the majority patterns, while cluster analysis seeks to find the majority patterns in a data set and organize the data using those patterns.

Answer true

1. 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

1. 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

1. An object is a contextual or conditional outlier if it deviates significantly with respect to a specific context of the object
2. Collective outliers are a group of objects that, as individual objects might not be outliers, but as a group, they are

Answer true

1. As the dimensionality of the data being analyzed increases, the distance between objects may become a function of the noise in the data, thus the distance and similarity between two points in a high-dimension data space may not accurately reflect the true relationship between those two points.
2. The HilOut algorithm detects outliers in the fullspace of high dimensioned data sets, not resorting to examining sub-spaces with lower dimensionality
3. Using algorithms that determine outliers in high dimensioned data sets in its sub-spaces is uniquely useful because there is critical information exists there that support interpreting why and to what extent an outlier is deemed an outlier
4. Using classification-based approaches for outlier detection requires the existence of a training set with data that is labeled in a way that explicitly separates normal data from outlier data.
5. If an object falls outside the decisionboundary of the normal class when applying a one-class model, the object is an outlier, even if the training set did not include an example of that kind of outlier

Answer: true

1. A (n) \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ is a data object that deviates significantly from the rest of the objects, as if it were generated by a different mechanism.
2. deviant
3. exception
4. outlier
5. preliminary
6. Which two of the following are types of outliers?
7. local
8. global
9. supervised
10. contextual
11. These outliers are the simplest form and the easiest to detect.
12. Global
13. Contextual
14. Collective
15. Local
16. Outliers are most commonly also be called \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_.
17. Deviants
18. Exceptions
19. Preliminaries
20. Anomalies
21. \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ assume that the normal data objects follow a statistical model.
22. Theoretical outlier detection methods
23. Proximity based outlier detection methods
24. Statistical outlier detection methods
25. Clustering based outlier detection based methods
26. A subset of data objects forms a \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ if the objects as a whole deviate significantly from the entire data set.
27. Cluster
28. Collective outlier
29. Contextual outlier
30. Global outlier
31. The assessment of the degree to which something is an outlier is called \_\_\_\_\_\_\_\_\_\_\_\_.
32. Outlier score
33. Evaluation score
34. Masking score
35. Outlier degree
36. When the presence of several anomalies masks the presence of all, it is called:
37. Swamping
38. Efficiency
39. Anomaly detection
40. Masking
41. Statistical distributions with values far from the mean are common in practice and are known as \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_.
42. Heavy-tailed distributions
43. Discordant observation
44. Gaussian distribution
45. Binomial distribution
46. \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ assume that the normal data objects belong to large and dense clusters, whereas outliers belong to small or sparse clusters, or not not belong to any clusters.
47. Classification-based outlier detection method
48. Clustering-based outlier detection method
49. Contextual-based outlier detection method
50. Proximity-based outlier detection method
51. An object may be more than one type of outlier.

TRUE

1. Many discordancy tests are common knowledge and do not require statistical knowledge.

FALSE

1. When there is sufficient knowledge of the data and the type of test that should be applied, anomaly detection tests can be very effective.

TRUE

1. Because clusters may form a hierarchy, outliers may belong to different granularity levels.

TRUE

1. Cluster techniques like K-means automatically determine the number of clusters.

FALSE

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

FALSE

1. Proximity based outlier detections methods generally take more time that distance-based methods.

TRUE

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

FALSE

1. Name three main approaches for outlier detection methods of high dimensional data.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

1. Using an equal-depth histogram, design a way to assign an object with an outlier score.

**Week 14**

1. If you were to attempt to mine a set of digitized Hollywood-produced films for major thematic references, you would be working with what kind of data from a data mining perspective
2. Mixed format
3. Problematic
4. Complex
5. Visual
6. Mining time-series data often involves
   1. Analyzing random sub-sequences from the series for known patterns
   2. Exactly matching sub-sequences of the time-series against a preselected subsequence of similar time-series data
   3. Searching for data sequences in the series that are similar to a preselected “query” sequence
   4. None of the above
7. If we find the same four five-second sub-sequences occur, in order, every 45 seconds in a radio signal, this is an example of a:
8. Random movement over time
9. Seasonal variation movement over time
10. Trend movement over time
11. Cyclic movement over time
12. None of the above
13. When pattern matching in sequence data where gaps between sub-sequences in a derived pattern need to be very close together, we apply
14. A constraint during processing that skips any two subsequences that are too far apart.
15. An algorithm that finds all possible patterns of interest, then resift these patterns, throwing out those where the sub-sequences are too far apart
16. A gap constraint in our algorithm that does not allow a pattern to be derived if the sub-sequences are too close together
17. B & C
18. None of the above
19. In biological sequence mining activities,
20. The goal is to discover sub-sequences in one sequence that can also be found in a second sequence
21. The goal is usually to align the sub-sequences in one sequence with those in one or more other sequences in order to determine the degree of similarity of the input sequences, either on a local or global level
22. The fact that optimal alignment is an NP-hard problem keeps us from doing anything useful – all such activities at this point are purely theoretical
23. None of the above
24. If we find in a network of nodes representing the occurrence of specific symbolic references in order through a given text there exists a pattern of “Fool” and “Violence” references always co-occurring in the text, then
25. We can dismiss this phenomena as random, accidental, thus meaningless
26. We must then ask the author if this pattern was intentional, because if the author did not intend the pattern, then it is as if the phenomena an artifact of random, accidental insertions into the text
27. We can conclude that the author of the text, consciously or unconsciously, views violence and fools to be closely aligned with each other
28. None of the above
29. If we want to discover what set of games a given Facebook user is likely to choose to play
30. We need to talk to the user
31. We need to perform a link prediction analysis looking at current and recent games the user is or did play in relation to the current and recent games that the Facebook friends of the user are or have played
32. We are embarking on an impossible task
33. We need to perform a cluster analysis of the games played by the user and the user’s friends, looking for outliers
34. If we want to have a system to warn drivers in a city of impending massive traffic congestions, we will need to do what kind of mining:
35. Multimedia data mining
36. Text mining
37. Cyber-Physical System Data mining
38. None of the above
39. According to Han, Kamber and Pei, text mining does not involve the input from which of these research disciplines
40. Statistics
41. Computational linguistics
42. Information Retrieval
43. Library Science
44. Mining Web content
45. Includes text, multimedia and structured data mining
46. Is an NP-hard problem, thus no one is doing research in this area
47. Can only be done on the surface Web – the content on the deep Web is never accessible for analysis
48. None of the above
49. Webstructuremininginvolves the use of network and graph mining theory and methods to analyze the nodes and connections on the Web
50. Mining data streams is best done at the current time using single-scan or a-very-few scan algorithms, as they are often large in volume, change dynamically, are possibly infinite and are multidimensional in their features.

Answer true

1. Regression is a statistical approach to data mining involves predicting the value of a response variable based on the values of one or more predictor variables, all of which must be numeric

Answer: true

1. 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

1. According to Han, Kamber and Pei, the various theories for a basis of data mining are not mutuallyexclusive
2. 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

1. Data visualization approaches can be combined with data mining in a several ways. One such way is to provide an interactive visual display to allow users to manipulate in real time a given data mining process

Answer: true

1. One method for improving audio data mining is to transform the patterns in the signals into sound, perhaps with a musical quality, eliminating a large part of the boredom issue involved in watching graphical display versions of audio data

Answer: true

1. 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

1. According to Han, Kamber and Pei, since scientific data can now be amassed at high speed and large volumes, scientists have turned to a new paradigm of research. Instead of the researcher-generated hypothesis and test approach, they now collect, store and data mine for the hypotheses, which are then confirmed with data and/or experimentation

Answer: true

1. \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ integrates data mining and data visualization to discover implicit and useful knowledge from large data sets.
2. Audio data mining
3. Theoretical data mining
4. Invisible data mining
5. Visual data mining
6. \_\_\_\_\_\_\_\_\_\_ uses audio signals to indicate data patterns or features of data mining results.
7. Audio data mining
8. Theoretical data mining
9. Invisible data mining
10. Visual data mining
11. The constant presence of data mining in many aspects of our daily lives is called:
12. Invisible data mining
13. Audio data mining
14. Ubiquitous data mining
15. Visual data mining
16. This is a major social concern of data mining.
17. Privacy and data security
18. Data warehousing
19. Cyber-physical system data
20. Improved scalability
21. Data mining tools for particular industries, such as finance, retail, and telecommunications, are called:
22. Statistical methods
23. Theoretical data mining
24. Audio data mining
25. Domain-specific applications
26. \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ discovers patterns and knowledge from spatial data.
27. Moving-object data mining
28. Spatial Data mining
29. Spatiotemporal data mining
30. Multimedial data mining
31. Link prediction can also be referred to as \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_.
32. Link mining
33. Data mining
34. Node prediction
35. Data prediction
36. Data mining that relates to both space and time is called:
37. Moving-object data mining
38. Spatial Data mining
39. Spatiotemporal data mining
40. Multimedial data mining
41. The discovery of relationships among multiple moving objects such as moving clusters, leaders, and followers:
42. Moving-object data mining
43. Spatial Data mining
44. Spatiotemporal data mining
45. Multimedial data mining
46. \_\_\_\_\_\_\_\_\_\_\_\_\_\_ is the mining of frequent subgraphs, or subgraph patterns.
47. Biological sequence analysis
48. Graph pattern mining
49. Substitution matrices
50. Sequence alignment
51. An abundance of personal or confidential information available in electronic forms poses a threat to data privacy and security.

TRUE

1. Research on the theoretical frameworks of data mining have matured significantly.

FALSE

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

FALSE

1. According to the data mining theory of data compression, the basis of data mining is to compress given data by encoding in terms of bits, association rules, decision trees, and clusters.

TRUE

1. Web usage mining is the process of extracting useful information from server logs.

TRUE

1. Text mining is by its very nature, *not* interdisciplinary.

FALSE

1. Web structure mining utilizes graph and network theories to analyze the connection structures of the web.

TRUE

1. Few classification methods perform model construction based on feature vectors.

FALSE

1. Theories of data mining are mutually exclusive.

FALSE

1. Data mining trends include further efforts toward the exploration of new application areas.

TRUE