1. **The process of forming general concept definitions from examples of concepts to be learned.**

A. Deduction
B. abduction
C. induction
D. conjunction

2. **Computers are best at learning**

A. facts.
B. concepts.
C. procedures.
D. principles.

3. **Data used to build a data mining model.**

A. validation data
B. training data
C. test data
D. hidden data

4. **Supervised learning and unsupervised clustering both require at least one**

A. hidden attribute.
B. output attribute.
C. input attribute.
D. categorical attribute.

5. **Supervised learning differs from unsupervised clustering in that supervised learning requires**

A. at least one input attribute.
B. input attributes to be categorical.
C. at least one output attribute.
D. ouput attriubutes to be categorical.

6. **A regression model in which more than one independent variable is used to predict the dependent variable is called**

A. a simple linear regression model
B. a multiple regression models
C. an independent model
D. none of the above

7. A term used to describe the case when the independent variables in a multiple regression model are correlated is

A. regression
B. correlation
C. multicollinearity
D. none of the above

8. A multiple regression model has the form: y = 2 + 3x1 + 4x2. As x1 increases by 1 unit (holding x2 constant), y will

A. increase by 3 units
B. decrease by 3 units
C. increase by 4 units
D. decrease by 4 units

9. A multiple regression model has

A. only one independent variable
B. more than one dependent variable
C. more than one independent variable
D. none of the above

10. A measure of goodness of fit for the estimated regression equation is the

A. multiple coefficient of determination
B. mean square due to error
C. mean square due to regression
D. none of the above

11. The adjusted multiple coefficient of determination accounts for

A. the number of dependent variables in the model
B. the number of independent variables in the model
C. unusually large predictors
D. none of the above

12. The multiple coefficient of determination is computed by

A. dividing SSR by SST
B. dividing SST by SSR
C. dividing SST by SSE
D. none of the above

13. For a multiple regression model, SST = 200 and SSE = 50. The multiple coefficient of determination is

A. 0.25
B. 4.00
C. 0.75
D. none of the above

14. A nearest neighbor approach is best used

A. with large-sized datasets.
B. when irrelevant attributes have been removed from the data.
C. when a generalized model of the data is desireable.
D. when an explanation of what has been found is of primary importance.

15. Determine which is the best approach for each problem.

A. *supervised learning*
B. *unsupervised clustering*
C. *data query*

1. What is the average weekly salary of all female employees under forty years of age? **(C)**
2. Develop a profile for credit card customers likely to carry an average monthly balance of more than $1000.00. **(A)**
3. Determine the characteristics of a successful used car salesperson. **(A)**
4. What attribute similarities group customers holding one or several insurance policies? **(A)**
5. Do meaningful attribute relationships exist in a database containing information about credit card customers? **(B)**
6. Do single men play more golf than married men? **(C)**
7. Determine whether a credit card transaction is valid or fraudulent **(A)**

16. Another name for an output attribute.

A. predictive variable
B. independent variable
C. estimated variable
D. dependent variable

17. Classification problems are distinguished from estimation problems in that

A. classification problems require the output attribute to be numeric.
B. classification problems require the output attribute to be categorical.
C. classification problems do not allow an output attribute.
D. classification problems are designed to predict future outcome.

18. Which statement is true about prediction problems?

A. The output attribute must be categorical.
B. The output attribute must be numeric.
C. The resultant model is designed to determine future outcomes.
D. The resultant model is designed to classify current behavior.

19. Which statement about outliers is true?

A. Outliers should be identified and removed from a dataset.
B. Outliers should be part of the training dataset but should not be present in the test data.
C. Outliers should be part of the test dataset but should not be present in the training data.
D. The nature of the problem determines how outliers are used.
E. More than one of a,b,c or d is true.

20. Which statement is true about neural network and linear regression models?

A. Both models require input attributes to be numeric.
B. Both models require numeric attributes to range between 0 and 1.
C. The output of both models is a categorical attribute value.
D. Both techniques build models whose output is determined by a linear sum of weighted input attribute values.
E. More than one of a,b,c or d is true.

21. Which of the following is a common use of unsupervised clustering?

A. detect outliers
B. determine a best set of input attributes for supervised learning
C. evaluate the likely performance of a supervised learner model
D. determine if meaningful relationships can be found in a dataset
E. All of a,b,c, and d are common uses of unsupervised clustering.

22. The average positive difference between computed and desired outcome values.

A. root mean squared error
B. mean squared error
C. mean absolute error
D. mean positive error

23. Selecting data so as to assure that each class is properly represented in both the training and test set.

A. cross validation
B. stratification
C. verification
D. bootstrapping

24. The standard error is defined as the square root of this computation.

A. The sample variance divided by the total number of sample instances.
B. The population variance divided by the total number of sample instances.
C. The sample variance divided by the sample mean.
D. The population variance divided by the sample mean.

25. Data used to optimize the parameter settings of a supervised learner model.

A. training
B. test
C. verification
D. validation

26. Bootstrapping allows us to

A. choose the same training instance several times.
B. choose the same test set instance several times.
C. build models with alternative subsets of the training data several times.
D. test a model with alternative subsets of the test data several times.

27. The correlation between the number of years an employee has worked for a company and the salary of the employee is 0.75. What can be said about employee salary and years worked?

A. There is no relationship between salary and years worked.
B. Individuals that have worked for the company the longest have higher salaries.
C. Individuals that have worked for the company the longest have lower salaries.
D. The majority of employees have been with the company a long time.
E. The majority of employees have been with the company a short period of time.

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

A. The attributes are not linearly related.
B. As the value of one attribute increases the value of the second attribute also increases.
C. As the value of one attribute decreases the value of the second attribute increases.
D. The attributes show a curvilinear relationship.

29. The average squared difference between classifier predicted output and actual output.

A. mean squared error
B. root mean squared error
C. mean absolute error
D. mean relative error

30. Simple regression assumes a _____ relationship between the input attribute and output attribute.

A. linear
B. quadratic
C. reciprocal
D. inverse

31. Regression trees are often used to model _____ data.

A. linear
B. nonlinear
C. categorical
D. symmetrical

32. The leaf nodes of a model tree are

A. averages of numeric output attribute values.
B. nonlinear regression equations.
C. linear regression equations.
D. sums of numeric output attribute values.

33. Logistic regression is a _____ regression technique that is used to model data having a _____outcome.

A. linear, numeric
B. linear, binary
C. nonlinear, numeric
D. nonlinear, binary

34. This technique associates a conditional probability value with each data instance.

A. linear regression
B. logistic regression
C. simple regression
D. multiple linear regression

35. This supervised learning technique can process both numeric and categorical input attributes.

A. linear regression
B. Bayes classifier
C. logistic regression
D. backpropagation learning

36. With Bayes classifier, missing data items are

A. treated as equal compares.
B. treated as unequal compares.
C. replaced with a default value.
D. ignored.

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

A. agglomerative clustering
B. expectation maximization
C. conceptual clustering
D. K-Means clustering

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

A. agglomerative clustering
B. conceptual clustering
C. K-Means clustering
D. expectation maximization

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

A. agglomerative clustering
B. conceptual clustering
C. K-Means clustering
D. expectation maximization

40. Machine learning techniques differ from statistical techniques in that machine learning methods

A. typically assume an underlying distribution for the data.
B. are better able to deal with missing and noisy data.
C. are not able to explain their behavior.
D. have trouble with large-sized datasets.