***Business Analytics, 2e, GE* (Evans)**

**Chapter 10 Introduction to Data Mining**

1) Which of the following is included in the data mining approach of data exploration and reduction?

A) analyzing data to predict how to classify a new data element

B) identifying groups in which the elements of the groups are in some way similar

C) creating rules for target marketing based on association of variables

D) developing analytic models to describe the relationship between metrics

Answer: B

Diff: 1

Blooms: Remember

Topic: The Scope of Data Mining

LO1: Define data mining and some common approaches used in data mining.

LO2: Identify different business uses for statistics and the major statistical tools businesses use

2) Spam filtering for e-mails can be seen as an example of which of the following types of approaches of data mining?

A) reduction

B) association

C) cause-and-effect modeling

**D) classification**

Answer: D

Diff: 1

Blooms: Remember

Topic: The Scope of Data Mining

LO1: Define data mining and some common approaches used in data mining.

LO2: Identify different business uses for statistics and the major statistical tools businesses use

3) U-botit.com is an electronic commerce company that sells music online. It keeps a tab of what genre of music their registered customers buy. If u-botit.com were to use the data mining approach of association, which of the following actions would it take?

**A) send recommendations to customers based on their buying habits**

B) conduct surveys to customers to gauge customer satisfaction

C) classify the customers based on the genre of music they listen to

D) provide discounts to certain genre buyers where sales are less

Answer: A

Diff: 1

Blooms: Apply

AACSB: Analytic Skills

Topic: The Scope of Data Mining

LO1: Describe association rule mining and its use in market basket analysis.

LO2: Identify different business uses for statistics and the major statistical tools businesses use

4) The data mining approach called \_\_\_\_\_\_\_\_ involves the developing of analytic models to describe the relationship between metrics that drive business performance like profitability, customer satisfaction, or employee satisfaction.

A) association

B) reduction

**C) cause-and-effect modeling**

D) classification

Answer: C

Diff: 1

Blooms: Remember

AACSB: Analytic Skills

Topic: The Scope of Data Mining

LO1: Define data mining and some common approaches used in data mining.

LO2: Identify different business uses for statistics and the major statistical tools businesses use

5) Sendstars is a package delivering company that recently made a study on its customer retention and service renewal metrics. They found that most customers defected from using Sendstars' services due to customer dissatisfaction stemming from delivery personnel being rude or ill-mannered. To curb this issue, Sendstars gave special training to its employees in customer service. Which of the following data mining approaches did Sendstars employ when they decided to train their employees in customer care based on the study?

A) association

B) cause-and-effect modeling

C) classification

D) reduction

Answer: B

Diff: 1

Blooms: Apply

AACSB: Analytic Skills

Topic: The Scope of Data Mining

LO1: Define data mining and some common approaches used in data mining.

LO2: Discuss ways in which statistics are used for quality management

6) \_\_\_\_\_\_\_\_ is a collection of techniques that seek to group or segment a collection of objects or observations into subsets, such that those within each subset are more closely related to one another than objects assigned to different subsets.

A) Association rule mining

B) Discriminant analysis

**C) Cluster analysis**

D) Logistic regression

Answer: C

Diff: 1

Blooms: Apply

AACSB: Analytic Skills

Topic: Data Exploration and Reduction

LO1: Explain how cluster analysis is used to explore and reduce data.

LO2: Identify different business uses for statistics and the major statistical tools businesses use

7) Which of the following is true of cluster analysis?

A) It is a cause-and-modeling type of analytic model.

B) It does not provide a definitive answer from analyzing the data.

C) It is primarily a prescriptive analytical method.

D) It uses clustered data that are not affected by the specific method used.

Answer: B

Diff: 2

Blooms: Understand

Topic: Data Exploration and Reduction

LO1: Explain how cluster analysis is used to explore and reduce data.

LO2: Identify different business uses for statistics and the major statistical tools businesses use

8) Which of the following is true of hierarchical clustering?

A) All clusters must have the same number of data.

B) No single cluster can have all objects

**C) The data are not partitioned into a particular cluster in a single step.**

D) All clusters must have more than one object in it.

Answer: C

Diff: 1

Blooms: Remember

Topic: Data Exploration and Reduction

LO1: Explain how cluster analysis is used to explore and reduce data.

LO2: Identify different business uses for statistics and the major statistical tools businesses use

9) Which of the following is the first stage of joining clusters in agglomerative hierarchical clustering?

A) by separating clusters into finer groups

B) by joining two clusters farthest away from each other

C) by joining two clusters that are not at a Euclidean distance

**D) by joining two clusters that are closest to each other**

Answer: D

Diff: 1

Blooms: Remember

Topic: Data Exploration and Reduction

LO1: Explain how cluster analysis is used to explore and reduce data.

LO2: Identify different business uses for statistics and the major statistical tools businesses use

10) Divisive clustering method is different from agglomerative clustering methods in that divisive clustering methods \_\_\_\_\_\_\_\_.

A) can only have a pair of subjects in each cluster

B) separate objects into a particular cluster in one step

**C) separate*n* objects successively into finer groupings**

D) can only have a single subject in each cluster

Answer: C

Diff: 1

Blooms: Remember

Topic: Data Exploration and Reduction

LO1: Explain how cluster analysis is used to explore and reduce data.

LO2: Identify different business uses for statistics and the major statistical tools businesses use

11) If the Euclidean distance were to be represented in a right triangle, which of the following would be considered the distance between two objects of a cluster?

**A) the hypotenuse**

B) the small leg

C) the long leg

D) the average of the sum of both the legs

Answer: A

Diff: 1

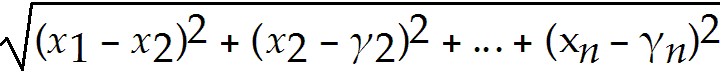
Blooms: Remember

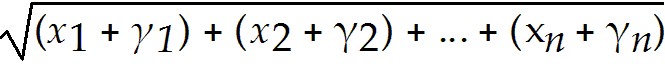
Topic: Data Exploration and Reduction

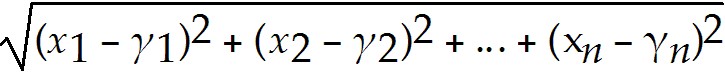
LO1: Explain how cluster analysis is used to explore and reduce data.

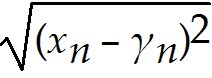
LO2: Identify different business uses for statistics and the major statistical tools businesses use

12) Which of the following formulas calculates the Euclidean distance between X and Y?

A) 

B) 

C) 

D) 

Answer: C

Diff: 1

Blooms: Remember

Topic: Data Exploration and Reduction

LO1: Explain how cluster analysis is used to explore and reduce data.

LO2: Identify different business uses for statistics and the major statistical tools businesses use

13) In the \_\_\_\_\_\_\_\_ method, the distance between groups is defined as the distance between the closest pair of objects, where only pairs consisting of one object from each group are considered.

A) Ward's linkage clustering

**B) single linkage clustering**

C) divisive clustering

D) average group linkage clustering

Answer: B

Diff: 1

Blooms: Remember

Topic: Data Exploration and Reduction

LO1: Explain how cluster analysis is used to explore and reduce data.

LO2: Identify different business uses for statistics and the major statistical tools businesses use

14) Which of the following is a definition of distance between two clusters in a complete linkage clustering?

A) the average of distances between all pairs of objects, where each pair is made up of one object from each group

**B) the distance between the most distant pair of objects, one from each group**

C) the sum of squares of the distance between clusters

D) the distance between the value of the shortest link between the clusters

Answer: B

Diff: 2

Blooms: Remember

Topic: Data Exploration and Reduction

LO1: Explain how cluster analysis is used to explore and reduce data.

LO2: Identify different business uses for statistics and the major statistical tools businesses use

15) Which of the following uses the sum of squares between the objects in the cluster when measuring their distances?

A) divisive clustering

B) average group linkage clustering

C**) Ward's hierarchical clustering**

D) single linkage clustering

Answer: C

Diff: 1

Blooms: Remember

Topic: Data Exploration and Reduction

LO1: Explain how cluster analysis is used to explore and reduce data.

LO2: Identify different business uses for statistics and the major statistical tools businesses use

16) In classification, which of the following would be considered as a categorical variable of interest for a credit approval decision for a requester?

A) age of the requester

B) income of the requester

C) revolving balance of the requester

**D) reject or accept credit approval**

Answer: D

Diff: 2

Blooms: Remember

Topic: Classification

LO1: Explain the purpose of classification methods, how to measure classification performance, and the use of training and validation data.

LO2: Identify different business uses for statistics and the major statistical tools businesses use

17) Which of the following features of classification, used in Excel, for a particular database will necessarily be coded to a certain value?

A) categorical variables

**B) non-numerical variables**

C) predictor variables

D) numerical variables

Answer: B

Diff: 1

Blooms: Remember

Topic: Classification

LO1: Explain the purpose of classification methods, how to measure classification performance, and the use of training and validation data.

LO2: Identify different business uses for statistics and the major statistical tools businesses use

18) The effectiveness of a classification rule can be judged making a probability of misclassification errors and summarizing the results in a \_\_\_\_\_\_\_\_.

**A) classification matrix**

B) classification hierarchy

C) dendogram

D) classification algorithm

Answer: A

Diff: 1

Blooms: Remember

Topic: Classification

LO1: Explain the purpose of classification methods, how to measure classification performance, and the use of training and validation data.

LO2: Identify different business uses for statistics and the major statistical tools businesses use

19) Which of the following data sets provides the most realistic estimate of the performance of a model on completely unseen data?

A) validation data set

**B) test data set**

C) training data set

D) linear regression data set

Answer: B

Diff: 1

Blooms: Remember

Topic: Classification

LO1: Explain the purpose of classification methods, how to measure classification performance, and the use of training and validation data.

LO2: Identify different business uses for statistics and the major statistical tools businesses use

20) Which of the following is true of a training data set?

A) They are primarily used to fine-tune models.

B) They provide the most realistic estimate for a model's performance.

C) They are used to build models where the data is unknown.

D) They have known outcomes.

Answer: D

Diff: 1

Blooms: Remember

Topic: Classification

LO1: Explain the purpose of classification methods, how to measure classification performance, and the use of training and validation data.

LO2: Identify different business uses for statistics and the major statistical tools businesses use

21) Validation data sets differ from training data sets in that validation data sets \_\_\_\_\_\_\_\_.

A) have known outcomes

B) test a model with unseen data

C) are used to teach data-mining algorithms

D) provide the most realistic test for models with known data

Answer: B

Diff: 1

Blooms: Remember

Topic: Classification

LO1: Explain the purpose of classification methods, how to measure classification performance, and the use of training and validation data.

LO2: Identify different business uses for statistics and the major statistical tools businesses use

22) Which of the following is true of random partitioning?

A) It cannot be used for creating training data sets.

**B) It can contain any non-negative value from the observations.**

C) It can include negative-value observations in its data.

D) It selects its data based on observations that have similar properties.

Answer: B

Diff: 2

Blooms: Understand

Topic: Classification

LO1: Explain the purpose of classification methods, how to measure classification performance, and the use of training and validation data.

LO2: Identify different business uses for statistics and the major statistical tools businesses use

23) The \_\_\_\_\_\_\_\_ algorithm is a classification scheme that attempts to find records in a database that are similar to one we wish to classify.

A) linear regression

B) logistic regression

**C) *k*-Nearest Neighbors**

D) discriminant analysis

Answer: C

Diff: 1

Blooms: Understand

Topic: Classification Techniques

LO1: Explain the purpose of classification methods, how to measure classification performance, and the use of training and validation data.

LO2: Identify different business uses for statistics and the major statistical tools businesses use

24) In a *k*-Nearest Neighbors algorithm, similarity of records is based on the \_\_\_\_\_\_\_\_.

A) closeness of a record to numerical predictors in the other records

B) sum of the squares of the distance between the numerical predictors

C) set of linear functions of predictors called discriminant functions

D) nearness of a record to its own observations

Answer: A

Diff: 1

Blooms: Understand

Topic: Classification Techniques

LO1: Explain the purpose of classification methods, how to measure classification performance, and the use of training and validation data.

LO2: Identify different business uses for statistics and the major statistical tools businesses use

25) Which of the following is true of the value of *k* in the *k*-Nearest Neighbors algorithm?

A) The value of *k* is always taken as a constant and is equal to 1.

B) If the value of *k* is large, it drastically increases variability.

C) If the value *k* is very large, it introduces biases into the classification decisions.

D) If the value *k* is large, the classification of a record is very sensitive to the classification of the single record to which it is closest.

Answer: C

Diff: 2

Blooms: Understand

Topic: Classification Techniques

LO1: Explain the purpose of classification methods, how to measure classification performance, and the use of training and validation data.

LO2: Identify different business uses for statistics and the major statistical tools businesses use

26) \_\_\_\_\_\_\_\_ is a data-mining technique used for classifying a set of observations into predefined classes.

A) Logistic regression

B) *k*-Nearest Neighbors algorithm

C) Discriminant analysis

D) Linear regression

Answer: C

Diff: 1

Blooms: Remember

Topic: Classification Techniques

LO1: Explain the purpose of classification methods, how to measure classification performance, and the use of training and validation data.

LO2: Identify different business uses for statistics and the major statistical tools businesses use

27) The weights for determining the discriminant functions are determined by \_\_\_\_\_\_\_\_.

A) assessing the number of outliers that are present in each group

B) calculating the distance between the two closest observations in each group

C) measuring the closeness between predictor values of each set

D) maximizing the between-group variance relative to the within-group variance

Answer: D

Diff: 2

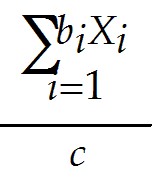
Blooms: Understand

Topic: Classification Techniques

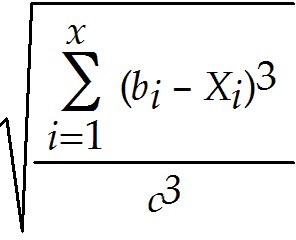
LO1: Explain the purpose of classification methods, how to measure classification performance, and the use of training and validation data.

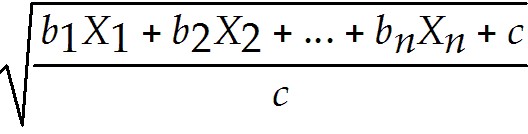
LO2: Identify different business uses for statistics and the major statistical tools businesses use

28) If *b*s are weights, *X*s are input values, and *c* is a constant or intercept, provide the equation for discriminant functions, *L*.

A) *L* = 

B) *L* = *b*1*X*1 + *b*2*X2*+ ... + *bnXn* + *c*

C) *L =* 

D) *L* = 

Answer: B

Diff: 1

Blooms: Understand

Topic: Classification Techniques

LO1: Explain the purpose of classification methods, how to measure classification performance, and the use of training and validation data.

LO2: Identify different business uses for statistics and the major statistical tools businesses use

29) Logistic regression is different from discriminant analysis in that logistic regression \_\_\_\_\_\_\_\_.

A) does not predict the weights

B) sets observation into predefined classes

C) does not depend on assumptions

D) depends on assumptions such as normalization of independent variables

Answer: C

Diff: 1

Blooms: Understand

Topic: Classification Techniques

LO1: Explain the purpose of classification methods, how to measure classification performance, and the use of training and validation data.

LO2: Identify different business uses for statistics and the major statistical tools businesses use

30) Which of the following is true of logistic regression as a classifying method?

A) Its dependent variable is always categorical.

B) Its independent variable is always continuous or numerical.

C) It predicts the probability of output variables based on dependent variables.

D) It cannot be used when the dependent variable is binary.

Answer: A

Diff: 1

Blooms: Remember

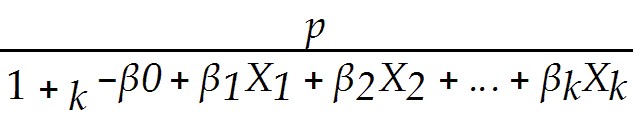
Topic: Classification Techniques

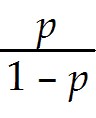
LO1: Explain the purpose of classification methods, how to measure classification performance, and the use of training and validation data.

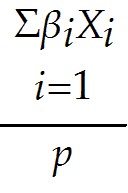
LO2: Identify different business uses for statistics and the major statistical tools businesses use

31) When using logistic regression, where *p* being the probability that the dependent variable *Y* = 1, *X*1, *X*2 ...,*Xk* are the independent variables, and*β*0, *β*1, *β*2 ..., *βk* are unknown regression coefficients, \_\_\_\_\_\_\_\_ is called the odds of belonging to category 1(*Y* = 1).

A) *p*(β0 + β1*X*1 + β2*X*2 + ...+ β*kXk*)

B) 

C) 

D) 

Answer: C

Diff: 2

Blooms: Remember

Topic: Classification Techniques

LO1: Explain the purpose of classification methods, how to measure classification performance, and the use of training and validation data.

LO2: Identify different business uses for statistics and the major statistical tools businesses use

32) Which of the following is true of association rule mining?

A) It develops analytic models to describe the relationship between metrics that drive business performance.

B) It identifies attributes that occur frequently together in a given data set.

C) It seeks to classify a categorical outcome into one of two or more categories.

D) It is a data reduction technique that reduces large information into smaller heterogeneous groups.

Answer: B

Diff: 1

Blooms: Remember

Topic: Association Rule Mining

LO1: Describe association rule mining and its use in market basket analysis.

LO2: Identify different business uses for statistics and the major statistical tools businesses use

33) Which of the following types of data-mining methods provides probabilistic if-then statements?

A) association rule

B) logistic regression

C) reduction

D) cause-and effect modeling

Answer: A

Diff: 1

Blooms: Remember

Topic: Association Rule Mining

LO1: Describe association rule mining and its use in market basket analysis.

LO2: Identify different business uses for statistics and the major statistical tools businesses use

34) Which of the following typically describes the support for the association rule?

A) the number of transactions that include all items that are in the antecedent parts of the rule

B) the number of transactions that include all items in the antecedent and consequent parts of the rule

C) the number of transactions that include half from the antecedent and half from the consequent

D) the number of transactions that include all items that are in the consequent parts of the rule

Answer: B

Diff: 1

Blooms: Understand

Topic: Association Rule Mining

LO1: Describe association rule mining and its use in market basket analysis.

LO2: Identify different business uses for statistics and the major statistical tools businesses use

35) \_\_\_\_\_\_\_\_ is the ratio of the number of transactions that include all items in the consequent as well as the antecedent to the number of transactions that include all items in the antecedent.

A) Lift

B) Logit

C) Support for the association rule

D) Confidence of the association rule

Answer: D

Diff: 1

Blooms: Remember

Topic: Association Rule Mining

LO1: Describe association rule mining and its use in market basket analysis.

LO2: Identify different business uses for statistics and the major statistical tools businesses use

36) The strength of the association rule, known as lift, is calculated as the ratio of the \_\_\_\_\_\_\_\_.

A) sum of the antecedents and the consequents to the antecedents

B) antecedents to the consequents

C) support to the confidence level

D) confidence to expected confidence

Answer: D

Diff: 1

Blooms: Remember

Topic: Association Rule Mining

LO1: Describe association rule mining and its use in market basket analysis.

LO2: Identify different business uses for statistics and the major statistical tools businesses use

37) Which of the following is true of the lift value in association rule?

A) Its strength is directly proportional to the strength of the association rule.

B) It cannot have a value that is more than 1 to be considered a good minimum.

C) It provides information on the increase of the antecedent given the consequent.

D) It does not require the expected confidence to be calculated.

Answer: A

Diff: 1

Blooms: Understand

Topic: Association Rule Mining

LO1: Describe association rule mining and its use in market basket analysis.

LO2: Identify different business uses for statistics and the major statistical tools businesses use

38) In the cause-and-effect modeling, internal metrics, such as employee satisfaction, productivity, and turnover are considered to be \_\_\_\_\_\_\_\_ measures.

A) logit

B) life

C) leading

D) lagging

Answer: C

Diff: 1

Blooms: Remember

Topic: Cause-and-Effect Modeling

LO1: Use correlation analysis for cause-and-effect modeling

LO2: Identify different business uses for statistics and the major statistical tools businesses use

39) Which of the following would be considered a lagging measure in a restaurant using the cause-and-modeling method of data mining?

A) a manager who is having trouble meeting the daily demands

B) a waiter's behavior toward a customer

C) an accurately billed meal

D) a satisfied customer

Answer: D

Diff: 1

Blooms: Understand

Topic: Cause-and-Effect Modeling

LO1: Use correlation analysis for cause-and-effect modeling

LO2: Identify different business uses for statistics and the major statistical tools businesses use

40) A musical instruments retailer has 10,000 point-of-sale transactions out of which 1500 sales included both items of electric guitars and guitar cases, and out of which 750 had sales of new strings. If the electric guitars are considered A, the guitar cases are considered B, and the strings are considered C, then the associate rule for these sales become "If A and B are purchased, then C is also purchased." Calculate the confidence level, expected confidence level, and lift for this rule, given that total transactions for C is 3000.

Answer:

Total sales = 10,000

Sales of A and B = 1,500

Sales of A, B, and C = 750

The association rule "If A and B are purchased, then C is also purchased" has a support of 750 out of 1500 sales.

confidence = P (consequent | antecedent) = P(antecedent + consequent) / P(antecedent)

Therefore the confidence level = Total support / Total sales = 750 / 1,500 = 50%.

The total number of transactions for C is given as 3000.

Expected confidence is the number of transactions that include the consequent divided by the total number of transactions.

Therefore the expected confidence level = 3,000 / 10,000 = 30%.

The lift is calculated as confidence / expected confidence = 50% / 30% = 1.67.

Diff: 3

Blooms: Apply

AACSB: Analytic Skills

Topic: Association Rule Mining

LO1: Describe association rule mining and its use in market basket analysis.

LO2: Discuss ways in which statistics are used for quality management

41) In cluster analysis, the objects within clusters should exhibit a high amount of dissimilarity.

Answer: FALSE

Diff: 1

Blooms: Remember

Topic: Data Exploration and Reduction

LO1: Explain how cluster analysis is used to explore and reduce data.

LO2: Identify different business uses for statistics and the major statistical tools businesses use

42) In the average linkage clustering, the distance between two clusters is defined as the average of distances between all pairs of objects, where each pair is made up of one object from each group.

Answer: TRUE

Diff: 1

Blooms: Remember

Topic: Data Exploration and Reduction

LO1: Explain how cluster analysis is used to explore and reduce data.

LO2: Identify different business uses for statistics and the major statistical tools businesses use

43) The accuracy of the model on the test data gives a realistic estimate of the performance of the model on completely unseen data.

Answer: TRUE

Diff: 1

Blooms: Remember

Topic: Classification

LO1: Identify different business uses for statistics and the major statistical tools businesses use

LO2: Identify different business uses for statistics and the major statistical tools businesses use

44) Logistic regression cannot be employed when the dependent variable is binary.

Answer: FALSE

Diff: 1

Blooms: Remember

Topic: Classification Techniques

LO1: Explain the purpose of classification methods, how to measure classification performance, and the use of training and validation data.

LO2: Identify different business uses for statistics and the major statistical tools businesses use

45) In association analysis, the antecedent and consequent are sets of items that do not have any items in common.

Answer: TRUE

Diff: 1

Blooms: Remember

Topic: Association Rule Mining

LO1: Describe association rule mining and its use in market basket analysis.

LO2: Identify different business uses for statistics and the major statistical tools businesses use

46) Expected confidence assumes independence between the consequent and the antecedent.

Answer: TRUE

Diff: 1

Blooms: Remember

Topic: Association Rule Mining

LO1: Describe association rule mining and its use in market basket analysis.

LO2: Identify different business uses for statistics and the major statistical tools businesses use

47) The market share of a business would be considered a lagging measure in the cause-and-effect modeling of data mining.

Answer: TRUE

Diff: 1

Blooms: Remember

Topic: Cause-and-Effect Modeling

LO1: Use correlation analysis for cause-and-effect modeling

LO2: Identify different business uses for statistics and the major statistical tools businesses use

48) Lagging and leading measures in cause-and-effect modeling are uncorrelated.

Answer: FALSE

Diff: 1

Blooms: Understand

Topic: Cause-and-Effect Modeling

LO1: Use correlation analysis for cause-and-effect modeling

LO2: Identify different business uses for statistics and the major statistical tools businesses use

49) Briefly explain classification as a data-mining tool with an example.

Answer: Classification is the process of analyzing data to predict how to classify a new data element. An example of classification is spam filtering in an e-mail client. By examining textual characteristics of a message, the message is classified as junk or not. Classification methods can help predict whether a credit-card transaction may be fraudulent, whether a loan applicant is high risk, or whether a consumer will respond to an advertisement.

Diff: 1

Blooms: Remember

Topic: The Scope of Data Mining

LO1: Explain the purpose of classification methods, how to measure classification performance, and the use of training and validation data.

LO2: Identify different business uses for statistics and the major statistical tools businesses use

50) How are objects clustered in agglomerative hierarchical clustering?

Answer: An agglomerative hierarchical clustering procedure produces a series of partitions of the data,*Pn*, *Pn*-1, ...,*P*1 . *Pn* consists of *n* single-object clusters, and consists of a single group containing all *n* observations. At each particular stage, the method joins together the two clusters that are closest together. At the first stage, this consists of simply joining together the two objects that are closest together. The most commonly used measure of distance between objects is Euclidean distance. Some clustering methods use the squared Euclidean distance (that is, without the square root) because it speeds up the calculations.

Diff: 2

Blooms: Remember

Topic: Classification

LO1: Explain how cluster analysis is used to explore and reduce data.

LO2: Identify different business uses for statistics and the major statistical tools businesses use

51) Exemplify the *k*-Nearest Neighbor algorithm of classification.

Answer: Suppose we have a record X that we want to classify. If we use the *k*-Nearest Neighbor (*k*-NN) method, the nearest neighbor to that record in the training data set is the one that has the smallest distance from it. The 1-NN rule then classifies record X in the same category as its nearest neighbor. We can extend this idea to a *k*-NN rule by finding the *k*-Nearest Neighbors in the training data set to each record we want to classify and then assigning the classification as the classification of majority of the *k*-Nearest Neighbors. The choice of *k* is somewhat arbitrary. If *k* is too small, the classification of a record is very sensitive to the classification of the single record to which it is closest. A larger *k* reduces this variability, but making *k* too large introduces bias into the classification decisions. For example, if *k* is the count of the entire training dataset, all records will be classified the same way.

Diff: 2

Blooms: Remember

Topic: Classification Techniques

LO1: Explain the purpose of classification methods, how to measure classification performance, and the use of training and validation data.

LO2: Identify different business uses for statistics and the major statistical tools businesses use

52) How is the strength of an association measured?

Answer: To measure the strength of association, an association rule has two numbers that express the degree of uncertainty about the rule. The first number is called the support for the association rule. The support is simply the number of transactions that include all items in the antecedent and consequent parts of the rule. The second number is the confidence of the association rule. Confidence is the ratio of the number of transactions that include all items in the consequent as well as the antecedent to the number of transactions that include all items in the antecedent. Another measure of the strength of an association rule is lift, which is defined as the ratio of confidence to expected confidence. Expected confidence is the number of transactions that include the consequent divided by the total number of transactions. The higher the lift ratio, the stronger the association rule; a value greater than 1.0 is usually a good minimum.

Diff: 2

Blooms: Remember

Topic: Association Rule Mining

LO1: Describe association rule mining and its use in market basket analysis.

LO2: Identify different business uses for statistics and the major statistical tools businesses use

53) Explain how data-mining using lagging and leading measures of the cause-and-effect model can help managers make business decisions.

Answer: Managers are always interested in results, such as profit, customer satisfaction and retention, production yield, and so on. These results are divided as either lagging measures or leading measures. Customer satisfaction results in regard to sales or service transaction would be lagging measures; employee satisfaction, sales representative behavior, billing accuracy, and so on, would be examples of leading measures that might influence customer satisfaction. If employees are not satisfied, their behavior toward customers could be negatively affected, and customer satisfaction could be low. If this can be explained using business analytics, managers can take steps to improve employee satisfaction, leading to improved customer satisfaction. Therefore, it is important to understand what controllable factors significantly influence key business performance measures that managers cannot directly control. Correlation analysis can help to identify these influences and lead to the development of cause-and-effect models that can help managers make better decisions today that will influence results tomorrow.

Diff: 2

Blooms: Remember

Topic: Cause-and-Effect Modeling

LO1: Use correlation analysis for cause-and-effect modeling

LO2: Identify different business uses for statistics and the major statistical tools businesses use