Chapter 3

Multiple Choice Questions

*(3.1)*

* 1. A data mining algorithm is *unstable* if

1. test set accuracy depends on the ordering of test set instances.
2. the algorithm builds models unable to classify outliers.
3. the algorithm is highly sensitive to small changes in the training data.
4. test set accuracy depends on the choice of input attributes.
   1. Which statement is true about the decision tree attribute selection process described in your book?
5. A categorical attribute may appear in a tree node several times but a numeric attribute may appear at most once.
6. A numeric attribute may appear in several tree nodes but a categorical attribute may appear at most once.
7. Both numeric and categorical attributes may appear in several tree nodes.
8. Numeric and categorical attributes may appear in at most one tree node.

*(3.2)*

* 1. Given a rule of the form IF X THEN Y, rule *confidence* is defined as the conditional probability that

1. Y is true when X is known to be true.
2. X is true when Y is known to be true.
3. Y is false when X is known to be false.
4. X is false when Y is known to be false.
   1. Association rule *support* is defined as
5. the percentage of instances that contain the antecendent conditional items listed in the association rule.
6. the percentage of instances that contain the consequent conditions listed in the association rule.
7. the percentage of instances that contain all items listed in the association rule.
8. the percentage of instances in the database that contain at least one of the antecendent conditional items listed in the association rule.

Use these tables to answer questions 5 and 6.

|  |  |
| --- | --- |
| **Single Item Sets** | **Number of Items** |
| Magazine Promo = Yes | 7 |
| Watch Promo = No | 6 |
| Life Ins Promo = Yes | 5 |
| Life Ins Promo = No | 5 |
| Card Insurance = No | 8 |
| Sex = Male | 6 |

|  |  |
| --- | --- |
| **Two Item Sets** | **Number of Items** |
| Magazine Promo = Yes & Watch Promo = No | 4 |
| Magazine Promo = Yes & Life Ins Promo = Yes | 5 |
| Magazine Promo = Yes & Card Insurance = No | 5 |
| Watch Promo = No & Card Insurance = No | 5 |

* 1. One two-item set rule that can be generated from the tables above is:

If Magazine Promo = Yes Then Life Ins promo = Yes

The confidence for this rule is:

1. 5 / 7
2. 5 / 12
3. 7 / 12
4. 1
   1. Based on the two-item set table, which of the following is *not* a possible two-item set rule?
5. IF Life Ins Promo = Yes THEN Magazine Promo = Yes
6. IF Watch Promo = No THEN Magazine Promo = Yes
7. IF Card Insurance = No THEN Magazine Promo = Yes
8. IF Life Ins Promo = No THEN Card Insurance = No

(*3.3)*

* 1. Which statement is true about the K-Means algorithm?
  2. All attribute values must be categorical.
  3. The output attribute must be cateogrical.
  4. Attribute values may be either categorical or numeric.
  5. All attributes must be numeric.
  6. The K-Means algorithm terminates when

1. a user-defined minimum value for the summation of squared error differences between instances and their corresponding cluster center is seen.
2. the cluster centers for the current iteration are identical to the cluster centers for the previous iteration.
3. the number of instances in each cluster for the current iteration is identical to the number of instances in each cluster of the previous iteration.
4. the number of clusters formed for the current iteration is identical to the number of clusters formed in the previous iteration.

*(3.4)*

* 1. A genetic learning operation that creates new population elements by combining parts of two or more existing elements.

1. selection
2. crossover
3. mutation
4. absorption
   1. An evolutionary approach to data mining.
5. backpropagation learning
6. genetic learning
7. decision tree learning
8. linear regression
   1. The computational complexity as well as the explanation offered by a genetic algorithm is largely determined by the
9. fitness function
10. techniques used for crossover and mutation
11. training data
12. population of elements

*(3.5)*

* 1. This approach is best when we are interested in finding all possible interactions among a set of attributes.

1. decision tree
2. association rules
3. K-Means algorithm
4. genetic learning

Computational Questions

* 1. Construct a decision tree with root node *Type* from the data in the table below. The first row contains attribute names. Each row after the first represents the values for one data instance. The output attribute is *Class*.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Scale** | **Type** | Shade | **Texture** | **Class** |
| One | One | Light | Thin | A |
| Two | One | Light | Thin | A |
| Two | Two | Light | Thin | B |
| Two | Two | Dark | Thin | B |
| Two | One | Dark | Thin | C |
| One | One | Dark | Thin | C |
| One | Two | Light | Thin | C |

Answers to Chapter 3 Questions

Multiple Choice Questions

* 1. c
  2. b
  3. a
  4. c
  5. a
  6. d
  7. d
  8. b
  9. b
  10. b
  11. a
  12. b

**Computational Questions**

1. Here is the tree.

