Data Mining, CSCI 347, Fall 2018
Exam 1, Sept. 19

1. Input into data mining algorithms can best be described by: (4 pts.)
    a. A wide variety of real world data collected in a wide variety of ways
    b. Structured data consisting of attribute values and relationships
    c. Individual, independent records, which provide values for the same set of attributes
    d. Data normalized to reduce repetition
    e. Data normalized in order to facilitate comparison of attributes

2. The data mining method for classifying a new instance by determining its nearest neighbor and matching the prediction of that neighbor is most likely to be called: (4 pts.)
    a. Clustering
    b. Covering algorithms
    c. Mining association rules
    d. Statistical modeling
    e. Instance based learning

3. Which of the following is most closely based on a "divide and conquer" strategy? (4 pts.)
    a. Trees
    b. Functions
    c. Statistical learning
    d. Rules
    e. Instance based learning

4. Choose the term that best describes a measurement where values are ordered but no distance between values is defined: (4 pts.)
    a. Interval quantity
    b. Ratio quantity
    c. Normalized quantity
    d. Nominal quantity
    e. Ordinal quantity

5. Data integration can best be described as: (4 pts.)
    a. Converting nominal data to numeric
    b. Removing data which is lacking attribute values
    c. Identifying and removing outliers from the data
    d. Combining datasets
    e. All of the above

Short Answer

6. Attributes can be of two types. List these types.

   1. Nominal (3 pts.)

   2. Numeric (3 pts.)

7. Data mining is typically used or two purposes. What are they?

   1. (4 pts.)
      Prediction - given a new instance, predict what the outcome will be

   2. (4 pts.)

      Knowledge – describe a structure that can be used to classify unknown examples

8. Describe the difference between classification and association rules.

   1. Classification rules (4 pts.)
      Predict the classification of an example (i.e. the class value)

   2. Association rules (4 pts.)
      Show strong associations between different attribute values

9. Tell 3 things that people have a right to know when they give personal information.
   (8 pts.)

   1. How will their information be used
   2. Who will use it
   3. For what purpose
   4. How will it be protected
   5. How they can change it if it is wrong

Longer answers

10. Describe the difference between correlation and causation. (10 pts.)

Causation – assume there is a cause/effect relation between the independent variables and the dependent variable. That is, the dependent variable changes solely because the independent variables change.

Correlation – the above may or may not be true. The dependent variable moves with the independent variable, but there isn't necessarily a cause/effect relation. There may be another intervening variable causing the effect.

11. Give three ways to handle missing data. (10 pts.)

- Ignore the instance (usually done when class value is missing)
- Fill in the missing value manually
- Use a global constant to fill in the missing value: e.g., "unknown"
- Imputation: Use the attribute mean to fill in the missing value, or use the attribute mean for all samples belonging to the same class to fill in the missing value

12. Give the 4 levels of measurements and what operations are allowed with each. (20 pts.)

    1. Nominal quantities, only operation is matching, = or ≠
    2. Ordinal quantities, can compare but no arithmetic, =, ≠, , ≥
    3. Interval quantities, can determine the distance between two vales and determine the average, =, ≠, , ≥, + , -, and average
    4. Ratio quantities, can do all operations, =, ≠, <, ≤, >, ≥, + , -, *, /

13. Describe what is meant by a training, testing and validation dataset and why three are needed. (10 pts.)

    - Training data set for selecting the learning algorithm
    - Validation data set for setting parameters on the chosen learning algorithms
    - Testing data set for determining the accuracy

    Three independent data sets are needed because if you test on the training or validation data sets, your results will be overly optimistic.