Homework Assignment 1

Due Date: Thursday, June 21 Before Midnight

(10 points)

1. **(2 points)** In your own words, define *data mining*. Describe the steps involved in data mining when viewed as a process of knowledge discovery.

2. **(2 points)** Explain the difference between discrimination and classification, characterization and clustering, and classification and regression. Support your answers with clear examples.

3. **(2 points)** Suppose that the data for analysis includes the feature *age*. The age values for the data tuples are: 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 36, 40, 45, 46, 52, 70.

    a. Define, then provide values for, *mean*, *median*, *mode,* and *mid-range*.

    b. Define, then provide values for, the *five-number summary*.

    c. Show a boxplot of the data (hand-drawn or programmed).

4. **(2 points)** Briefly outline how to compute the dissimilarity between objects described by the following (include mathematical notation):

    a. Nominal Features

    b. Asymmetric Binary Features

    c. Numeric Features

    d. Term-Frequency Vectors

5. **(2 points)** Given two objects represented by the tuples (22, 1, 41, 10) and (20, 0, 36, 8), compute distance between objects using the following:

    a. Euclidean Distance

    b. Manhattan Distance

    c. Minkowski Distance, using q = 3

    d. Supremum Distance