Introduction to DM

(Note: you must type in your answers, no handwriting answers will be accepted, one file in pdf format per submission)

1. Discuss whether or not each of the following activities is a data mining task.
   a. Dividing the customers of a company according to their gender.
   b. Predicting the future stock price of a company using historical records.
   c. Monitoring seismic waves for detecting earthquake activities.
   d. Extracting the frequencies of a sound wave.

2. For the following vectors, x and y, calculate the indicated similarity or distance measures.
   a. x = (1, 1, 1, 1), y = (2, 2, 2, 2) cosine, correlation, Euclidean
   b. x = (0, 1, 0, 1), y = (1, 0, 1, 0) cosine, correlation, Euclidean, Jaccard
   c. x = (0,−1, 0, 1), y = (1, 0,−1, 0) cosine, correlation, Euclidean
   d. x = (1, 1, 0, 1, 0, 1), y = (1, 1, 1, 0, 0, 1) cosine, correlation, Jaccard
   e. x = (2,−1, 0, 2, 0,−3), y = (−1, 1,−1, 0, 0,−1) cosine, correlation

3. Using R and R Studio to explore Iris Data.
   a) Download the Iris dataset via the link: http://archive.ics.uci.edu/ml/machine-learning-databases/iris/
   b) Read R manual: http://cran.r-project.org/doc/manuals/r-release/R-intro.pdf
   c) Use R to get the following statistical information: the mean, median, range, variance, percentiles of the sepal length of Iris flowers.
   d) Write R code to do the following visualization:
      1) Print the summary of the data
      2) Generate histograms of four Iris attributes
      3) Generate box plot for Iris attributes
      4) Generate the pie chart of the distribution of the types of Iris flowers
      5) Generate the correlation plots of each pair of four Iris attributes

   (You need to submit the screenshots showing the command line and results generated by R for each question. You can also use R Studio to compile the report (go to File→Compile Report…) and include that in your pdf submission.