Assignment 1 (Knowledge Discovery and Data Mining)

Number of problems/points: Four problems for total of 100 points
Out: January 28, 2016
Due: February 04, 2016 in class (hard copy).

Homework Policies (applicable for all 9664 assignments):
1. You are required to do the homework problems in order to pass.
2. Understandability of the solution is as desired as correctness.
3. Penalty for late homework assignments submissions is 20% per day. So, do it on time.
4. Solutions are expected to be your own work. Group work is not allowed unless explicitly approved for a particular problem. If you obtained a hint with help (e.g., through library work, discussion with another person, etc.) acknowledge your source, and write up the solution on your own. Plagiarism and other anti-intellectual behavior will be dealt with severely.

Problem 1: (10 points)
Solve exercises 8 and 9 in section 1.9 in the textbook (page 26).

Problem 2: (20 points)
a) Write a program (in any programing language) to take as parameters: k the number of samples to create, and n the sample size. Submit your source code and provide some validation evidence that it works correctly for small k and small n.
b) Download Adult Database from http://archive.ic.uci.edu/ml/ and remove all categorical attributes. Use your program to generate 50 samples (without replacement) of size 500 from Adult Database. In each of obtained 50 samples remove all individuals with missing attribute values. Report fraction of removed with income above $50,000 in each sample. Compare this to fraction of individuals with income about $50,000 in the entire dataset. Plot histograms of all attributes in the reduced dataset.
c) Compute the mean and covariance matrix of each of 50 samples. Then compute the mean of the means and covariances over the 50 runs and plot those mean values.

Problem 3: (20 points)
You are given a set of m objects that is divided into K groups, where the i-th group is of size \( m_i \). If the goal is to obtain a sample of size \( n < m \), what is the difference between the following two sampling schemes? (Assume sampling with replacement.)

(a) We randomly select \( n \ast m_i / m \) elements from each group.

(b) We randomly select \( n \) elements from the data set, without regard for the group to which an object belongs.

Problem 4: (20 points)
In arithmetic progressive sampling the size of the i-th sample is \( n_0 + (i-1) n_1 \) where \( n_0 \) is the size of the initial sample and \( n_1 \) is the size of the increment.
In geometric progressive sampling the i-th sample is of size \( a^k n_0 \) where \( a \) is a constant and \( n_0 \) is the size of the initial sample.

(a) Prove that if computational complexity of an inductive learning algorithm is \( n^2 \) or worse, then geometric sampling is better than arithmetic as long as convergence can be detected with samples smaller than \( N/2 \) where \( N \) is the size of the database.

(b) Then show that if computational complexity of a given inductive learning algorithm is linear, then geometric sampling is better as long as convergence can be detected with samples smaller than \( N/6 \).

**Problem 5: (15 points)**
Discuss how you might map correlation values from the interval \([-1, 1]\) to the interval \([0, 1]\). Note that the type of transformation that you use might depend on the application that you have in mind.

Thus, consider two applications:
(a) clustering time series and
(b) predicting the behavior of one time series given another.

**Problem 6: (15 points)**
This exercise compares and contrasts some similarity and distance measures.

(a) For binary data, the L1 distance corresponds to the Hamming distance; that is, the number of bits that are different between two binary vectors. The Jaccard similarity is a measure of the similarity between two binary vectors. Compute the Hamming distance and the Jaccard similarity between the following two binary vectors:

\[
x = 0101010001 \\
y = 0100011000
\]

(b) Suppose that you are comparing how similar two organisms of different species are in terms of the number of genes they share. Describe which measure, Hamming or Jaccard, you think would be more appropriate for comparing the genetic makeup of two organisms. Explain. (Assume that each animal is represented as a binary vector, where each attribute is 1 if a particular gene is present in the organism and 0 otherwise.)

(c) If you wanted to compare the genetic makeup of two organisms of the same species, e.g., two human beings, would you use the Hamming distance, the Jaccard coefficient, or a different measure of similarity or distance? Explain. (Note that two human beings share > 99.9% of the same genes.)