Assignment 1 (Knowledge Discovery and Data Mining)

No. of problems/points: Two problems for total of 100 points
Out: February 5, 2010
Due: February 17, in class.

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 will be high, and 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 need hints first talk to instructor. If you obtain a solution with help (e.g., through library work, using hint provided by 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: (30 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:
\[ \begin{align*}
x &= 0101010001 \\
y &= 0100011000 
\end{align*} \]

(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.)

Problem 2: (70 points)
Download Adult Database from http://archive.ics.uci.edu/ml/
This database contains a subset of records about the US population collected by the US Census Bureau. The 48842 individuals in this database are described by 8 categorical and 6 continuous attributes (with some missing data). The target column corresponds to annual income below or above $50,000. Use adult dataset for the following data preprocessing experiments:

a) Remove “education” and “native country” attributes and all individuals with missing attribute values. Report fraction of removed with income above $50,000. Compare this to fraction of individuals with income about $50,000 in the entire dataset and conclude if missing values are correlated with target. Plot histograms of all attributes in reduced dataset.

b) In data with removed missing values smooth each continuous attribute by sorting the values and partitioning into equidepth bins of appropriate depth (e.g. in depth k each bin contains k values) followed by replacing the actual values by the median of the corresponding bin. Plot histograms
of reduced attributes. This procedure removed outliers. Summarize findings on outliers removed. Compare to using equiwidth bins, where the interval range of values in each bin is constant.

c) Transform reduced data from subtask b) by mapping each attribute to [0,1] range using min-max normalization such that the relationships among the original data values is preserved. In transformed data find out relevance of each attribute on its own for predicting a person's income. What seems to be the most discriminative attribute? Are there attributes that seem irrelevant or weakly relevant?

d) Report the classification accuracy when using all 12 attributes on transformed data from subtask c) by NB-Tree algorithm from WEKA. You can download WEKA at http://www.cs.waikato.ac.nz/ml/weka/