Assignment 3 (Knowledge Discovery and Data Mining)

Number of problems/points: Four problems for total of 100 points
Out: February 2, 2018
Due: February 08, 2018 in class (hard copy).

Problem 1: (10 points)
Explain why computing the proximity between two attributes is often simpler than computing the similarity between two objects.

Problem 2: (30 points)
Compute the cosine measure using the raw frequencies between the following two sentences:
(a) “The sly fox jumped over the lazy dog.”
(b) “The dog jumped at the intruder.”

Problem 3: (30 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 4: (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:
\[ x = 0101010001 \]
\[ y = 0100011000 \]

(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.)