Assignment 1

1. In this problem, you will compare and contrast 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.)

2. Proximity is typically defined between a pair of objects. Define two methods you might use to define the proximity among a group of objects. Specifically discuss how you might use these methods to

a) Define the distance between two sets of points in Euclidean space?

b) Define the proximity between two sets of data objects? (Make no assumption about the data objects, except that a proximity measure is defined between any pair of objects.)

3. In class we discussed visualization using the iris dataset. In this exercise, you will compute the distances between data points for the iris data set (iris.data). Specifically, you will write some code to

a. read the iris data and form a data matrix X of size 150x4

b. Normalize X as $\bar{X}_{,j} = \frac{X_{,j}-\mu_j}{\sigma_j}$, where μ_j is the mean of attribute j and σ_j is the standard deviation of attribute j.

c. Compute the distance matrix between pairs of data in $\bar{X}$ using the following metrics:

    i. Euclidean distance

    ii. Cosine distance = 1-Cosine_similarity

    iii. Mahalanobis Distance

d. Plot the above distance matrices (In python, you can use the matplotlib package and matshow to plot a matrix)

e. Using the proximity measure you developed in 2(b) compute the proximity between data belonging to the same class. Express the proximity as a symmetric 3x3 matrix (for distances between the three classes). Do this using all three distance metrics in (c). Discuss your results.

****Submit using homework using dropbox. Remember to include plots as well as any code you developed with your submission.