**Lab 4**

**Objective**

Compare and contrast some similarity and distance measures

**Assignments**

* 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 = 010101001

Y = 010011000

Answer:

Hamming Distance of given binary data: 6

Jaccard similarity: 0.4

* Which approach, Jaccard or Hamming distance, is more similar to the Simple Matching Coefficient, and which approach is more similar to the cosine measure? Please explain. Note: The Hamming measure is a distance, while the other three measures are similarities, but don’t let this confuse you.)

Answer:

Jaccard is more akin to cosine similarity, and the hamming distance is similar to

simple matching coefficent. The Jaccard coefficient is a similar method of comparison to the Cosine Similarity due to how both methods compare one type of attribute distributed among all data. Hamming and SMC look at the whole data and looks for data points that are similar or dissimilar to one a another.

* 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 organism. Please 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)

X = (0,0,0,1,1,1,0,0,1,0)

Y = (1,0,0,1,0,0,1,0,1,1)

Answer:

You want to compare the genes that are alike so the method to use is Jaccard coefficent. The genes we care about are the genes that are alike. Genes that are not present does not support that the two organisms are similar.

* For the following vectors, x and y, calculate the indicated similarity or distance.

X = (1,1,1,1)

Y = (2,2,2,2)

What is the cosine similarity, correlation, and Euclidean distance?

Answer:

Cosine similarity: 1.0

Euclidean distance: 2.0

Correlation can not be indeicted since the std(x) and std(y) are 0.