Assignment #2

Monday, September 11, 2023 7:58 PM

11136325

Data Mining

Assignment Two — 10 points

Due: Thursday, September 21 @ 11:50 PM

Question 1. (8 points)

(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 = 0000011001

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

(c) 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.)

(d) 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.)

Question 2. (2 points)

Briefly outline how to compute the dissimilarity between objects described by the following:

(a) Nominal attributes
(b) Asymmetric binary attributes
(c) Numeric attributes
(d) Term-frequency vectors

Good luck!

Question 1

Part A:

Hamming Distance

the following two
x = 0101010001
y = 0000011001

Hamming distance = 3

Jaccard Similarity

the following two
x = 0101010001
y = 0000011001

f_{01} = 2 (the number of attributes where **x** was 0 and **y** was 1)
f_{10} = 1 (the number of attributes where **x** was 1 and **y** was 0)
f_{00} = 7 (the number of attributes where **x** was 0 and **y** was 0)
f_{11} = 2 (the number of attributes where **x** was 1 and **y** was 1)

J = number of 11 matches / number of non-zero attributes
= (f_{11}) / (f_{01} + f_{10} + f_{11})
J = (f_{11}) / (f_{01} + f_{01} + f_{11}) = (2) / (1 + 2 + 2) = 0.4

Part B:

Simple Matching Coefficient (SMC)
Jaccard similarity is more similar to the Simple Matching Coefficient. They are both similarity measures that find matches when considering binary data. By doing so, the similarities are finding those commonalities in the two sets of binary data by considering their intersection relative to the total number of elements that consider both the intersection and union for the sets.

Cosine Similarity
Jaccard similarity is also more similar to the Cosine similarity. They are both similarity measures that take into account the way in which the sets of binary data vectors are given. How the sets' data are oriented to one another influences the similarity measures for both of them.

Hamming Distance
On the other hand, determining the proximity of the given sets for the hamming distance is intended to measure how dissimilar the sets are. Not only that, but it only outputs the number of occurrences of which the bits in the vectors differ; such an output is not meaningful enough to make a claim on similarity. Therefore, the fundamental purpose behind it does not give a basis of comparison to the similarity measures.

Part C:

Let's use the sets given to us from the problem earlier to determine which measure is appropriate.

the following two
x = 0101010001
y = 0000011001

Because the question is asking us to consider how similar two organisms of different species are in terms of the number of genes they share, using Jaccard similarity would make more sense as this measure intends to find how similar they are in their genes. If the question asked to see how different their genes were, then the Hamming distance would be appropriate to use. As mentioned earlier, Jaccard similarity intends to find the intersection of two sets of events we are analyzing, which in this instance, is finding the shared genes between two different species of organisms. If we used x and y set to represent the genetic elements of our two species, then we can use the value we found earlier, which was 0.4. A higher Jaccard similarity implies how similar the two species' genes are.

Part D:

Well, the answer to this question depends. Ideally, we would want to choose a measure that can effectively relay the comparisons of two species which are the same. Notice how the question does not imply the reader to consider this as a binary metric. Rather, it gives an opportunity to think of another measure than the Hamming distance or the Jaccard similarity. Therefore, a good metric to consider for comparing the genetic makeup of two species that are the same would be the Euclidean distance. Given that human beings share > 99.9% of the same genes, the number of attributes to consider becomes high in dimensionality. Euclidean distance is known for its ability to deal with high-dimensional data, allowing the genetic variations when comparing the two human beings to be captured by the measure and its reactivity to those variations. The Euclidean distance outputs the difference in these variations quantitatively (otherwise known as its magnitude).

Question 2

Part A:
The choice of what measure to use to understand the proximity of nominal attributes depends on how you intend to provide the analysis. If we go forward with the Hamming distance, we would then be measuring how dissimilar the sets of data are. However, it does not tell us how dissimilar the sets of data are relative to the total elements within them (does not give a probability percentage). Rather, it counts the number of occurrences where the sets of data differ in their bits. Because this metric quantifies the number of bits that are different between two binary vectors, we would then need to convert our nominal attributes into binary vectors. Each category becomes a binary attribute that can be either 0 or 1.

Part B:
An option for comparing objects with asymmetric binary attributes is to use the Jaccard distance. The Jaccard similarity metric considers the intersection of common elements within the sets of data relative to the total elements in those sets. While the Jaccard similarity provides a value between 0 and 1, which can be interpreted as a probability percentage of similarity, the Jaccard distance quantifies dissimilarity by subtracting the Jaccard similarity value from 1. In other words, the Jaccard distance reflects how dissimilar the sets are based on the proportion of elements that are not shared between the sets.

Part C:
We can compute dissimilarity on numerical attributes using the Euclidean distance. Such attributes can be continuous and the metric performs well with them. The basis of Euclidean distance is that it computes the magnitude, in other words, how different the values are between two data points; it quantifies the size of the differences. The interpretation from the results provided by Euclidean metric is that we can measure how dissimilar the numeric attributes are. We can do so if we keep in mind that a smaller Euclidean distance indicates greater similarity between the attributes, and dissimilar if the distances are larger.

Part D:

We can compute dissimilarity on term-frequency vectors as they represent a document as a single vector. The vector contains the frequencies of the words that appear in the document. Cosine similarity could be used for comparing the frequency vectors. It is chosen for its ability to quantify the similarity between these vectors by examining the angle between them. It assesses how similar the word frequencies are without much affect on its output based on variations in document length. Cosine similarity values range between -1 (perfect dissimilarity) and 1 (perfect similarity), with values closer to -1 indicating higher content dissimilarity between documents.