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, which indicates the number of bits that differ between two binary vectors = 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} = 0 (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)
The Simple Matching Coefficient (SMC) is similar to the Hamming distance because it can be expressed as the Hamming distance (the number of bits that differ) divided by the total number of bits (the union of both binary vectors). Both SMC and Hamming distance are related to binary data. SMC measures similarity by finding matches when considering binary data and determining the proportion of matching bits relative to the total number of bits.

Cosine Similarity
Jaccard similarity is 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. Both disregard matches where both bits from each vector are zero. This is because matches where both terms are zero indicate that the two items do not share any features, which is less informative for similarity assessment. Both measures emphasize the presence or absence of features rather than their exact values, which can be particularly useful in scenarios where the presence of certain features is more important than their absence.

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). However, if it was only between Hamming distance or Jaccard similarity, the former would be best in this context. Two of the same species with near identical genetic makeup makes it plausible for the Hamming distance to be appropriate as it measures how the species are different. That would be a more effective metric.

Question 2

Part A:

Attribute Type	Dissimilarity
Nominal	$d = \begin{cases} 0 & \text{if } x = y \\ 1 & \text{if } x \neq y \end{cases}$

We would need to convert our nominal attributes into binary vectors. Each category becomes a binary attribute that can be either 0 or 1. The number of successes would be bits that do not match, which using the figure above, would be quantified to be one. Then, the bits that do match would be given a value of 0. Look at this for instance:
x = 0101010001
y = 0000011001
The Hamming distance can be use here to dictate how much do our two binary vectors differ. We had that out to be 3, which would indicate that we had 3 successes (which translates to three 1 bits) that different bit values between the two sets of vectors.

Part B:
The Jaccard distance is a suitable method for quantifying dissimilarity between objects when working with asymmetric binary attributes. It focuses on the common elements within sets of data relative to the total elements in those sets, yielding a similarity value between 0 and 1. However, in cases of asymmetric binary attributes, certain elements are more relevant for dissimilarity assessment. Specifically, we are concerned with the presence of attributes in one object but their absence in another. To measure this dissimilarity effectively, we compute the Jaccard distance.

The Jaccard distance is calculated by considering the number of attributes that differ between the two objects and dividing it by the total number of attributes under consideration. It reflects the dissimilarity between sets based on the proportion of elements that are not shared between them. This approach aligns with the idea that the number of negative matches is not considered and is not included in the calculation.

$$d(i,j) = \frac{r+s}{q+r+s}.$$

Image credits

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. The results are never negative, rather positive or zero due to the formula squaring the differences as seen below. We can measure how dissimilar the numeric attributes are if we keep in mind that a smaller Euclidean distance indicates greater similarity between the attributes, and dissimilar if the distances are larger.

Euclidean Distance

$$d(\mathbf{x},\mathbf{y}) = \sqrt{\sum_{k=1}^{n}(x_k - y_k)^2}$$

where n is the number of dimensions (attributes) and x_k and y_k are, respectively, the k^{th} attributes (components) or data objects $\mathbf{x}$ and $\mathbf{y}$.

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 and 1, with values closer to -1 indicating higher content dissimilarity between documents. As seen below, to compute cosine similarity, you would need to find the dot product of the two sets of vectors, then computing their individual magnitudes. Lastly, divide the dot product of the set by the product of the two magnitudes.

- If $\mathbf{d}_1$ and $\mathbf{d}_2$ are two document vectors, then

$$\cos(\mathbf{d}_1, \mathbf{d}_2) = \langle \mathbf{d}_1, \mathbf{d}_2 \rangle / \|\mathbf{d}_1\| \|\mathbf{d}_2\|,$$

where $\langle \mathbf{d}_1, \mathbf{d}_2 \rangle$ indicates inner product or vector dot product of vectors, $\mathbf{d}_1$ and $\mathbf{d}_2$, and $\|\mathbf{d}\|$ is the length of vector $\mathbf{d}$.