Data Mining

Week 6 Exercise Sheet

1. Consider the problem of finding the K nearest neighbours[1] of a data object. A programmer designs for following algorithm for this task:

    1. **for** i=1 to *number of data objects* **do**
    2. Find the distances of the i^{th} objects to all other objects.
    3. Sort these distances in decreasing order. (Keep track of which object is associated with each distance.)
    4. **return** the objects associated with the first K distances of the sorted list
    5. **end for**

    (a) Describe the potential problems with this algorithm if there are duplicate objects in the data set. Assume the distance function will only return a distance of 0 for objects that are the same.

    (b) How would you fix this problem?

2. The following attributes are measured for members of a herd of Asian elephants: *weight, height, tusk length, trunk length, ear area.* Based on these measurements, what sort of similarity measure would you use to compare or group these elephants? Justify your answer and explain any special circumstances.

3. Consider a document-term matrix, where tf_{ij} is the frequency of the i^{th} word (term) in the j^{th} document and m is the number of documents. Consider the variable transformation that is defined by,

$$tf'_{ij} = tf_i \log(\frac{m}{df_i}),$$

    where df_i is the number of documents in which the i^{th} term appears and is known as the *document frequency* of the term. This transformation is known as the *inverse document frequency* transformation.

    (a) What is the effect of this transformation if a term occurs in one document? What about if it occurs in every document?

    (b) What might be the purpose of this transformation?

[1] We will see this in more detail later in the course. For now, this is just to get you thinking.

4. The goal of this exercise is to 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:

   $$x = 0101010001 \quad y = 0100011000.$$

   (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 make-up 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.)

5. For the following vectors $\mathbf{x}$, $\mathbf{y}$, calculate the indicated similarity or distance measures.

   (a) $\mathbf{x} = (1, 1, 1, 1)$, $\mathbf{y} = (2, 2, 2, 2)$: cosine, correlation, Euclidean.

   (b) $\mathbf{x} = (0, 1, 0, 1)$, $\mathbf{y} = (1, 0, 1, 0)$: cosine, correlation, Euclidean, Jaccard.

   (c) $\mathbf{x} = (0, -1, 0, 1)$, $\mathbf{y} = (1, 0, -1, 0)$: cosine, correlation, Euclidean.

   (d) $\mathbf{x} = (1, 1, 0, 1, 0, 1)$, $\mathbf{y} = (1, 1, 1, 0, 0, 1)$: cosine, correlation, Jaccard.

   (e) $\mathbf{x} = (2, -1, 0, 2, 0, -3)$, $\mathbf{y} = (-1, 1, -1, 0, 0, -1)$: cosine, correlation.

6. (a) What is the range of values that are possible for the cosine measure?

   (b) If two objects have a cosine measure of 1, are they identical? Explain.

   (c) What is the relationship of the cosine measure to correlation, if any? (Hint: Look at statistical measures such as mean and standard deviation in cases where cosine and correlation are the same and different.)

   (d) Derive the mathematical relationship between cosine similarity and Euclidean distance when each data object has an L_2 length of 1.

(e) Derive the mathematical relationship between correlation and Euclidean distance when each data point has been standardised by subtracting its mean and dividing by its standard deviation.