Assignment #1 Solutions

13. Consider the problem of finding the *K* nearest neighbors of a data object. A programmer designs Algorithm 2.1 for this task.

Algorithm 2.1 Algorithm for finding *K* nearest neighbors.

---

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

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.

Answer: The main problem is that the nearest neighbor list of an object O may primarily consist of duplicate objects that are different than O, or may primarily consist of duplicates of O.

How would you fix this problem?

Answer: There are various approaches depending on the situation. One approach is to preprocess the data and keep only one object for each group of duplicate objects. In this case, each neighbor can represent either a single object or a group of duplicate objects.

14. The following attributes are measured for members of a herd of Asian elephants: *weight, height, tusk length, trunk length,* and *ear area*. Based on these measurements, what sort of similarity measure from Section 2.4 would you use to compare or group these elephants? Justify your answer and ex-plain any special circumstances.

Answer: These attributes are all numerical, but can have widely varying ranges of values, depending on the scale used to measure them. Furthermore, the attributes are not asymmetric and the magnitude of an attribute matters. These latter two facts eliminate the cosine and correlation measure. Euclidean distance, applied after standardizing the attributes to have a mean of 0 and a standard deviation of 1, would be appropriate.

15. You are given a set of *m* objects that is divided into *K* groups, where the i^{th} group is of size m_i. If the goal is to obtain a sample of size $n < m$, what is the difference between the following two sampling schemes? (Assume sampling with replacement.)

(a) We randomly select $n*m_i/m$ elements from each group.
(b) We randomly select *n* elements from the data set, without regard for the group to which an object belongs.

Answer: The scheme referred to in (a) is stratified sampling, which is proportionate – that is, the sample from each group is proportional to its size relative to the total number of objects. The second scheme (b) is a simple random sampling scheme. Stratified sampling in general has two advantages in the case when the objects in each group are homogeneous. The first is that we are assured of a sample from each group, which can be used to estimate various statistical parameters of that group whenever the corresponding sample size is large enough. The second advantage is that the variance of the stratified sample is always smaller than the variance of the simple random sampling since the latter has also to include the variance between the different groups. Hence stratified sampling is in general more accurate than simple random sampling.

18. This exercise compares and contrasts 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

Answer: Hamming distance = number of different bits = 3
Jaccard Similarity = number of 1-1 matches /(number of bits - number 0-0 matches) = 2 / 5 = 0.4

(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.)

Answer: The Hamming distance is similar to the SMC. In fact, SMC = Hamming distance / number of bits.

The Jaccard measure is similar to the cosine measure because both ignore 0-0 matches.

(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.)

Answer: Jaccard is more appropriate for comparing the genetic makeup of two organisms; since we want to see how many genes these two organisms share.

(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.)

Answer: Two human beings share >99.9% of the same genes. If we want to compare the genetic makeup of two human beings, we should focus on their differences. Thus, the Hamming distance is more appropriate in this situation.

24. Proximity is typically defined between a pair of objects.

(a) Define two ways in which you might define the proximity among a group of objects.

Answer: Two examples are the following: (i) based on pairwise proximity, i.e., minimum pairwise similarity or maximum pairwise dissimilarity, or (ii) for points in Euclidean space compute a centroid (the mean of all the points—see Section 8.2) and then compute the sum or average of the distances of the points to the centroid.

(b) How might you define the distance between two sets of points in Euclidean space?

Answer: One approach is to compute the distance between the centroids of the two sets of points. The centroid is simply the vector sum of all the Euclidean vectors scaled by their number.

(c) How might you 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.)

Answer: One approach is to compute the average pairwise proximity of objects in one group of objects with those objects in the other group. Other approaches are to take the minimum or maximum proximity of all such pairs.