1. **Problem 1:** Explain why computing the proximity between two attributes is often simpler than computing the similarity between two objects.

   In data mining, computing the similarity measure is sometimes a redundant step if we are already computing dissimilarity. For example, if we were hoping to use a Euclidean distance (sometimes referred to as Euclidean dissimilarity) to compute a similarity measure, we would have to transform the dissimilarity score to be a similarity measure using the following formula $s = 1 - d$, where s is the similarity score and d is the Euclidean distance. Interpreting the dissimilarity measure would be simpler because the two move inversely from one another. The distinction of similarity and dissimilarity remains repetitive in the discussion of proximity because it accounts for both.

2. **Problem 2:** Compute the cosine measure using the raw frequencies between the following two sentences:

   (a) "The sly fox jumped over the lazy dog."

   (b) "The dog jumped at the intruder."

	The	sly	fox	jumped	over	lazy	dog	at	intruder
a	2	1	1	1	1	1	1	0	0
b	2	0	0	1	0	0	1	1	1

$$\mathbf{a} = \langle 2, 1, 1, 1, 1, 1, 1, 0, 0 \rangle$$
$$\mathbf{b} = \langle 2, 0, 0, 1, 0, 0, 1, 1, 1 \rangle$$
$$\langle \mathbf{a}, \mathbf{b} \rangle = (2*2 + 1*0 + 1*0 + 1*1 + 1*0 + 1*0 + 1*1 + 0*1 + 0*1) = 6$$
$$\|\mathbf{a}\| = (2*2 + 1*1 + 1*1 + 1*1 + 1*1 + 1*1 + 1*1 + 0*0 + 0*0)^{0.5} = \sqrt{10}$$
$$\|\mathbf{b}\| = (2*2 + 0*0 + 0*0 + 1*1 + 0*0 + 0*0 + 1*1 + 1*1 + 1*1)^{0.5} = \sqrt{8}$$

$$cos(\mathbf{a}, \mathbf{b}) = \frac{\langle \mathbf{a}, \mathbf{b} \rangle}{\|\mathbf{a}\| \|\mathbf{b}\|} = \frac{6}{\sqrt{80}} = \frac{6}{4\sqrt{5}} = \frac{3}{2\sqrt{5}} \approx 0.671$$

3. **Problem 3:**Discuss how you might map correlation values from the interval [-1,1] to the interval [0,1]. Note that the type of transformation that you use might depend on the application that you have in mind. Thus, consider two applications:

   (a) clustering time series

   If I were clustering time series, I would want to cluster time series that move together (i.e., $r \geq 0$) and discard time series that move in opposite directions, so I would most likely use the following transformation:

$$r = \begin{cases} r & r \geq 0 \\ 0 & r < 0 \end{cases}$$

   (b) predicting the behavior of one time series given another

If I were predicting time series behaviors, I would square the [-1,1] interval to minimize the impacts of lowly correlated values (about $-0.25 < r < 0.25$) and maximize the impact of highly correlated values (about $r > 0.5$) because any number between [-1,1] (save 0) will become smaller when squared. But the rate of transformation for numbers with greater magnitudes is smaller than those for lesser magnitudes. This allows us to isolate those time series that have similar magnitudes

4. **Problem 4:** 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:
   $\mathbf{x} = 0101010001$
   $\mathbf{y} = 0100011000$

   i.
   $$HammingDistance = 3$$

   ii.
   $$JaccardSimilarity = \frac{2}{5} = 0.4$$

   (b) 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.)

   Because we are looking at the number of genes these two organisms *share*, I think the best approach would be Jaccard similarity. In our calculations for Jaccard similarity, we do not take into account any genes that both organisms do not have, allowing us to discover the level of overlap between each of these two organism's genetic makeup.

   (c) 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 $\geq 99.9\%$ of the same genes.)

   Since our data is binary, using either the Simple Matching Coefficient (SMC), Jaccard Coefficient, or the Hamming distance makes most sense. Taking into

account the magnitude of similarity between the two organisms, assessing how *different* the sequences are would be more insightful. Thus, I would use the Hamming distance to compare these organisms.

5. **Problem 5:** (10 points) Donor data consists of 11 records in the following format: Name Age Salary Donor(Y/N). Donor training dataset:

Nancy	21	37,000	N
Jim	27	41,000	N
Allen	43	61,000	Y
Jane	38	55,000	N
Steve	44	30,000	N
Peter	51	56,000	Y
Sayani	53	70,000	Y
Lata	56	74,000	Y
Mary	59	25,000	N
Victor	61	68,000	Y
Dale	63	51,000	Y

Compute the Gini index for the entire Donor data set, with respect to the two classes.

$$GINI_{2class} = 1 - (\frac{6}{11})^2 - (\frac{5}{11})^2 = \frac{60}{121} \approx 0.496$$

Compute the Gini index for the portion of the data set with age at least 50.

$$GINI_{age \geq 50} = 1 - (\frac{5}{6})^2 - (\frac{1}{6})^2 = \frac{10}{36} = \frac{5}{18} \approx 0.278$$

6. Compute the Entropy index for the entire Donor data set, with respect to the two classes.

$$Entropy_{2class} = -(\frac{6}{11})log_2(\frac{6}{11}) - (\frac{5}{11})log_2(\frac{5}{11}) \approx 0.994$$

Compute the Gini index for the portion of the data set with age at least 50.

$$Entropy_{age \geq 50} = -(\frac{5}{6})log_2(\frac{5}{6}) - (\frac{1}{6})log_2(\frac{1}{6}) \approx 0.650$$