Introduction to Data Mining

Instructor's Solution Manual

Pang-Ning Tan
Michael Steinbach
Vipin Kumar

Contents

1

Introduction

1. Discuss whether or not each of the following activities is a data mining task.

   (a) Dividing the customers of a company according to their gender.
   No. This is a simple database query.

   (b) Dividing the customers of a company according to their profitability.
   No. This is an accounting calculation, followed by the application of a threshold. However, predicting the profitability of a new customer would be data mining.

   (c) Computing the total sales of a company.
   No. Again, this is simple accounting.

   (d) Sorting a student database based on student identification numbers.
   No. Again, this is a simple database query.

   (e) Predicting the outcomes of tossing a (fair) pair of dice.
   No. Since the die is fair, this is a probability calculation. If the die were not fair, and we needed to estimate the probabilities of each outcome from the data, then this is more like the problems considered by data mining. However, in this specific case, solutions to this problem were developed by mathematicians a long time ago, and thus, we wouldn't consider it to be data mining.

   (f) Predicting the future stock price of a company using historical records.
   Yes. We would attempt to create a model that can predict the continuous value of the stock price. This is an example of the

area of data mining known as predictive modelling. We could use regression for this modelling, although researchers in many fields have developed a wide variety of techniques for predicting time series.

(g) Monitoring the heart rate of a patient for abnormalities.
Yes. We would build a model of the normal behavior of heart rate and raise an alarm when an unusual heart behavior occurred. This would involve the area of data mining known as anomaly detection. This could also be considered as a classification problem if we had examples of both normal and abnormal heart behavior.

(h) Monitoring seismic waves for earthquake activities.
Yes. In this case, we would build a model of different types of seismic wave behavior associated with earthquake activities and raise an alarm when one of these different types of seismic activity was observed. This is an example of the area of data mining known as classification.

(i) Extracting the frequencies of a sound wave.
No. This is signal processing.

2. Suppose that you are employed as a data mining consultant for an Internet search engine company. Describe how data mining can help the company by giving specific examples of how techniques, such as clustering, classification, association rule mining, and anomaly detection can be applied.

   The following are examples of possible answers.

   - Clustering can group results with a similar theme and present them to the user in a more concise form, e.g., by reporting the 10 most frequent words in the cluster.
   - Classification can assign results to pre-defined categories such as "Sports," "Politics," etc.
   - Sequential association analysis can detect that that certain queries follow certain other queries with a high probability, allowing for more efficient caching.
   - Anomaly detection techniques can discover unusual patterns of user traffic, e.g., that one subject has suddenly become much more popular. Advertising strategies could be adjusted to take advantage of such developments.

3. For each of the following data sets, explain whether or not data privacy is an important issue.

   (a) Census data collected from 1900–1950. No

   (b) IP addresses and visit times of Web users who visit your Website. Yes

   (c) Images from Earth-orbiting satellites. No

   (d) Names and addresses of people from the telephone book. No

   (e) Names and email addresses collected from the Web. No

2

Data

1. In the initial example of Chapter 2, the statistician says, "Yes, fields 2 and 3 are basically the same." Can you tell from the three lines of sample data that are shown why she says that?

   $\frac{\text{Field 2}}{\text{Field 3}} \approx 7$ for the values displayed. While it can be dangerous to draw conclusions from such a small sample, the two fields seem to contain essentially the same information.

2. Classify the following attributes as binary, discrete, or continuous. Also classify them as qualitative (nominal or ordinal) or quantitative (interval or ratio). Some cases may have more than one interpretation, so briefly indicate your reasoning if you think there may be some ambiguity.

   Example: Age in years. **Answer:** Discrete, quantitative, ratio

   (a) Time in terms of AM or PM. Binary, qualitative, ordinal

   (b) Brightness as measured by a light meter. Continuous, quantitative, ratio

   (c) Brightness as measured by people's judgments. Discrete, qualitative, ordinal

   (d) Angles as measured in degrees between 0° and 360°. Continuous, quantitative, ratio

   (e) Bronze, Silver, and Gold medals as awarded at the Olympics. Discrete, qualitative, ordinal

   (f) Height above sea level. Continuous, quantitative, interval/ratio (depends on whether sea level is regarded as an arbitrary origin)

   (g) Number of patients in a hospital. Discrete, quantitative, ratio

   (h) ISBN numbers for books. (Look up the format on the Web.) Discrete, qualitative, nominal (ISBN numbers do have order information, though)

(i) Ability to pass light in terms of the following values: opaque, translucent, transparent. Discrete, qualitative, ordinal

(j) Military rank. Discrete, qualitative, ordinal

(k) Distance from the center of campus. Continuous, quantitative, interval/ratio (depends)

(l) Density of a substance in grams per cubic centimeter. Discrete, quantitative, ratio

(m) Coat check number. (When you attend an event, you can often give your coat to someone who, in turn, gives you a number that you can use to claim your coat when you leave.) Discrete, qualitative, nominal

3. You are approached by the marketing director of a local company, who believes that he has devised a foolproof way to measure customer satisfaction. He explains his scheme as follows: "It's so simple that I can't believe that no one has thought of it before. I just keep track of the number of customer complaints for each product. I read in a data mining book that counts are ratio attributes, and so, my measure of product satisfaction must be a ratio attribute. But when I rated the products based on my new customer satisfaction measure and showed them to my boss, he told me that I had overlooked the obvious, and that my measure was worthless. I think that he was just mad because our best-selling product had the worst satisfaction since it had the most complaints. Could you help me set him straight?"

   (a) Who is right, the marketing director or his boss? If you answered, his boss, what would you do to fix the measure of satisfaction?

   The boss is right. A better measure is given by

$$\text{Satisfaction(product)} = \frac{\text{number of complaints for the product}}{\text{total number of sales for the product}}.$$

   (b) What can you say about the attribute type of the original product satisfaction attribute?

   Nothing can be said about the attribute type of the original measure. For example, two products that have the same level of customer satisfaction may have different numbers of complaints and vice-versa.

4. A few months later, you are again approached by the same marketing director as in Exercise 3. This time, he has devised a better approach to measure the extent to which a customer prefers one product over other, similar products. He explains, "When we develop new products, we typically create several variations and evaluate which one customers prefer. Our standard procedure is to give our test subjects all of the product variations at one time and then

ask them to rank the product variations in order of preference. However, our test subjects are very indecisive, especially when there are more than two products. As a result, testing takes forever. I suggested that we perform the comparisons in pairs and then use these comparisons to get the rankings. Thus, if we have three product variations, we have the customers compare variations 1 and 2, then 2 and 3, and finally 3 and 1. Our testing time with my new procedure is a third of what it was for the old procedure, but the employees conducting the tests complain that they cannot come up with a consistent ranking from the results. And my boss wants the latest product evaluations, yesterday. I should also mention that he was the person who came up with the old product evaluation approach. Can you help me?"

(a) Is the marketing director in trouble? Will his approach work for generating an ordinal ranking of the product variations in terms of customer preference? Explain.

Yes, the marketing director is in trouble. A customer may give inconsistent rankings. For example, a customer may prefer 1 to 2, 2 to 3, but 3 to 1.

(b) Is there a way to fix the marketing director's approach? More generally, what can you say about trying to create an ordinal measurement scale based on pairwise comparisons?

One solution: For three items, do only the first two comparisons. A more general solution: Put the choice to the customer as one of ordering the product, but still only allow pairwise comparisons. In general, creating an ordinal measurement scale based on pairwise comparison is difficult because of possible inconsistencies.

(c) For the original product evaluation scheme, the overall rankings of each product variation are found by computing its average over all test subjects. Comment on whether you think that this is a reasonable approach. What other approaches might you take?

First, there is the issue that the scale is likely not an interval or ratio scale. Nonetheless, for practical purposes, an average may be good enough. A more important concern is that a few extreme ratings might result in an overall rating that is misleading. Thus, the median or a trimmed mean (see Chapter 3) might be a better choice.

5. Can you think of a situation in which identification numbers would be useful for prediction?

One example: Student IDs are a good predictor of graduation date.

6. An educational psychologist wants to use association analysis to analyze test results. The test consists of 100 questions with four possible answers each.

(a) How would you convert this data into a form suitable for association analysis?

Association rule analysis works with binary attributes, so you have to convert original data into binary form as follows:

$Q_1 = A$	$Q_1 = B$	$Q_1 = C$	$Q_1 = D$	...	$Q_{100} = A$	$Q_{100} = B$	$Q_{100} = C$	$Q_{100} = D$
1	0	0	0	...	1	0	0	0
0	0	1	0	...	0	1	0	0

(b) In particular, what type of attributes would you have and how many of them are there?

400 asymmetric binary attributes.

7. Which of the following quantities is likely to show more temporal autocorrelation: daily rainfall or daily temperature? Why?

   A feature shows spatial auto-correlation if locations that are closer to each other are more similar with respect to the values of that feature than locations that are farther away. It is more common for physically close locations to have similar temperatures than similar amounts of rainfall since rainfall can be very localized;, i.e., the amount of rainfall can change abruptly from one location to another. Therefore, daily temperature shows more spatial autocorrelation then daily rainfall.

8. Discuss why a document-term matrix is an example of a data set that has asymmetric discrete or asymmetric continuous features.

   The ij^{th} entry of a document-term matrix is the number of times that term j occurs in document i. Most documents contain only a small fraction of all the possible terms, and thus, zero entries are not very meaningful, either in describing or comparing documents. Thus, a document-term matrix has asymmetric discrete features. If we apply a TFIDF normalization to terms and normalize the documents to have an L_2 norm of 1, then this creates a term-document matrix with continuous features. However, the features are still asymmetric because these transformations do not create non-zero entries for any entries that were previously 0, and thus, zero entries are still not very meaningful.

9. Many sciences rely on observation instead of (or in addition to) designed experiments. Compare the data quality issues involved in observational science with those of experimental science and data mining.

   Observational sciences have the issue of not being able to completely control the quality of the data that they obtain. For example, until Earth orbit-

ing satellites became available, measurements of sea surface temperature relied on measurements from ships. Likewise, weather measurements are often taken from stations located in towns or cities. Thus, it is necessary to work with the data available, rather than data from a carefully designed experiment. In that sense, data analysis for observational science resembles data mining.

10. Discuss the difference between the precision of a measurement and the terms single and double precision, as they are used in computer science, typically to represent floating-point numbers that require 32 and 64 bits, respectively.

    The precision of floating point numbers is a maximum precision. More explicity, precision is often expressed in terms of the number of significant digits used to represent a value. Thus, a single precision number can only represent values with up to 32 bits, ≈ 9 decimal digits of precision. However, often the precision of a value represented using 32 bits (64 bits) is far less than 32 bits (64 bits).

11. Give at least two advantages to working with data stored in text files instead of in a binary format.

    (1) Text files can be easily inspected by typing the file or viewing it with a text editor.
    (2) Text files are more portable than binary files, both across systems and programs.
    (3) Text files can be more easily modified, for example, using a text editor or perl.

12. Distinguish between noise and outliers. Be sure to consider the following questions.

    (a) Is noise ever interesting or desirable? Outliers?
    No, by definition. Yes. (See Chapter 10.)

    (b) Can noise objects be outliers?
    Yes. Random distortion of the data is often responsible for outliers.

    (c) Are noise objects always outliers?
    No. Random distortion can result in an object or value much like a normal one.

    (d) Are outliers always noise objects?
    No. Often outliers merely represent a class of objects that are different from normal objects.

    (e) Can noise make a typical value into an unusual one, or vice versa?
    Yes.

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.

```
1: for i = 1 to number of data objects do
2:    Find the distances of the i^th object 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.

There are several problems. First, the order of duplicate objects on a nearest neighbor list will depend on details of the algorithm and the order of objects in the data set. Second, if there are enough duplicates, the nearest neighbor list may consist only of duplicates. Third, an object may not be its own nearest neighbor.

(b) How would you fix this problem?

There are various approaches depending on the situation. One approach is to to 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 explain any special circumstances.

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.

The first scheme is guaranteed to get the same number of objects from each group, while for the second scheme, the number of objects from each group will vary. More specifically, the second scheme only guarantes that, on average, the number of objects from each group will be $n * m_i/m$.

16. 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

$$tf'_{ij} = tf_{ij} * \log \frac{m}{df_i}, \tag{2.1}$$

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? In every document?

Terms that occur in every document have 0 weight, while those that occur in one document have maximum weight, i.e., $\log m$.

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

This normalization reflects the observation that terms that occur in every document do not have any power to distinguish one document from another, while those that are relatively rare do.

17. Assume that we apply a square root transformation to a ratio attribute x to obtain the new attribute x^*. As part of your analysis, you identify an interval (a, b) in which x^* has a linear relationship to another attribute y.

(a) What is the corresponding interval (a, b) in terms of x? (a^2, b^2)

(b) Give an equation that relates y to x. In this interval, $y = x^2$.

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.

$\mathbf{x} = 0101010001$
$\mathbf{y} = 0100011000$

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.)

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.)

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.)

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.

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

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

$\cos(\mathbf{x}, \mathbf{y}) = 1$, $\text{corr}(\mathbf{x}, \mathbf{y}) = 0/0$ (undefined), $\text{Euclidean}(\mathbf{x}, \mathbf{y}) = 2$

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

$\cos(\mathbf{x}, \mathbf{y}) = 0$, $\text{corr}(\mathbf{x}, \mathbf{y}) = -1$, $\text{Euclidean}(\mathbf{x}, \mathbf{y}) = 2$, $\text{Jaccard}(\mathbf{x}, \mathbf{y}) = 0$

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

$\cos(\mathbf{x}, \mathbf{y}) = 0$, $\text{corr}(\mathbf{x}, \mathbf{y})$=0, Euclidean$(\mathbf{x}, \mathbf{y}) = 2$

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

$\cos(\mathbf{x}, \mathbf{y}) = 0.75$, $\text{corr}(\mathbf{x}, \mathbf{y}) = 0.25$, Jaccard$(\mathbf{x}, \mathbf{y}) = 0.6$

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

$\cos(\mathbf{x}, \mathbf{y}) = 0$, $\text{corr}(\mathbf{x}, \mathbf{y}) = 0$

20. Here, we further explore the cosine and correlation measures.

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

$[-1, 1]$. Many times the data has only positive entries and in that case the range is $[0, 1]$.

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

Not necessarily. All we know is that the values of their attributes differ by a constant factor.

(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.)

For two vectors, $\mathbf{x}$ and $\mathbf{y}$ that have a mean of 0, $\text{corr}(\mathbf{x}, \mathbf{y}) = \cos(\mathbf{x}, \mathbf{y})$.

(d) Figure 2.1(a) shows the relationship of the cosine measure to Euclidean distance for 100,000 randomly generated points that have been normalized to have an L2 length of 1. What general observation can you make about the relationship between Euclidean distance and cosine similarity when vectors have an L2 norm of 1?

Since all the 100,000 points fall on the curve, there is a functional relationship between Euclidean distance and cosine similarity for normalized data. More specifically, there is an inverse relationship between cosine similarity and Euclidean distance. For example, if two data points are identical, their cosine similarity is one and their Euclidean distance is zero, but if two data points have a high Euclidean distance, their cosine value is close to zero. Note that all the sample data points were from the positive quadrant, i.e., had only positive values. This means that all cosine (and correlation) values will be positive.

(e) Figure 2.1(b) shows the relationship of correlation to Euclidean distance for 100,000 randomly generated points that have been standardized to have a mean of 0 and a standard deviation of 1. What general observation can you make about the relationship between Euclidean distance and correlation when the vectors have been standardized to have a mean of 0 and a standard deviation of 1?

Same as previous answer, but with correlation substituted for cosine.

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

Let $\mathbf{x}$ and $\mathbf{y}$ be two vectors where each vector has an L_2 length of 1. For such vectors, the variance is just n times the sum of its squared attribute values and the correlation between the two vectors is their dot product divided by n.

$$\begin{aligned} d(\mathbf{x}, \mathbf{y}) &= \sqrt{\sum_{k=1}^{n} (x_k - y_k)^2} \\ &= \sqrt{\sum_{k=1}^{n} x_k^2 - 2x_k y_k + y_k^2} \\ &= \sqrt{1 - 2\cos(\mathbf{x}, \mathbf{y}) + 1} \\ &= \sqrt{2(1 - \cos(\mathbf{x}, \mathbf{y}))} \end{aligned}$$

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

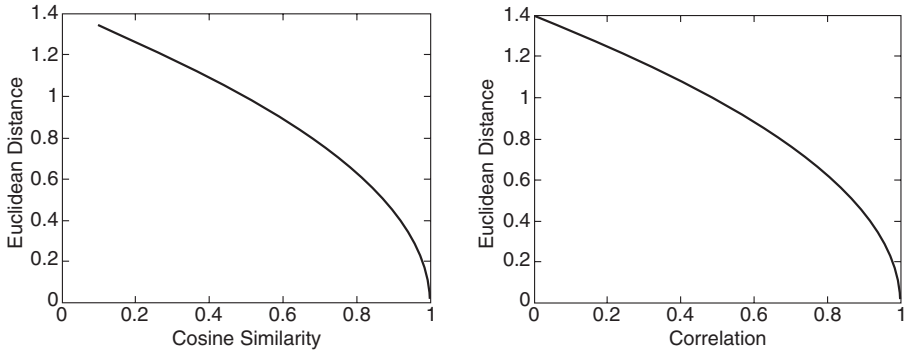
Let $\mathbf{x}$ and $\mathbf{y}$ be two vectors where each vector has an a mean of 0 and a standard deviation of 1. For such vectors, the variance (standard deviation squared) is just n times the sum of its squared attribute values and the correlation between the two vectors is their dot product divided by n.

$$\begin{aligned} d(\mathbf{x}, \mathbf{y}) &= \sqrt{\sum_{k=1}^{n} (x_k - y_k)^2} \\ &= \sqrt{\sum_{k=1}^{n} x_k^2 - 2x_k y_k + y_k^2} \\ &= \sqrt{n - 2n\text{corr}(\mathbf{x}, \mathbf{y}) + n} \\ &= \sqrt{2n(1 - \text{corr}(\mathbf{x}, \mathbf{y}))} \end{aligned}$$

21. Show that the set difference metric given by

$$d(A, B) = size(A - B) + size(B - A)$$

satisfies the metric axioms given on page 70. A and B are sets and $A - B$ is the set difference.



(a) Relationship between Euclidean distance and the cosine measure.

(b) Relationship between Euclidean distance and correlation.

Figure 2.1. Figures for exercise 20.

1(a). Because the size of a set is greater than or equal to 0, $d(\mathbf{x}, \mathbf{y}) \geq 0$.
1(b). if $A = B$, then $A - B = B - A =$ empty set and thus $d(\mathbf{x}, \mathbf{y}) = 0$
2. $d(A, B) = size(A-B)+size(B-A) = size(B-A)+size(A-B) = d(B, A)$
3. First, note that $d(A, B) = size(A) + size(B) - 2size(A \cap B)$.
$\therefore d(A, B)+d(B, C) = size(A)+size(C)+2size(B)-2size(A \cap B)-2size(B \cap C)$
Since $size(A \cap B) \leq size(B)$ and $size(B \cap C) \leq size(B)$,
$d(A, B) + d(B, C) \geq size(A) + size(C) + 2size(B) - 2size(B) = size(A) + size(C) \geq size(A) + size(C) - 2size(A \cap C) = d(A, C)$
$\therefore d(A, C) \leq d(A, B) + d(B, C)$

22. 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: clustering time series and predicting the behavior of one time series given another.

For time series clustering, time series with relatively high positive correlation should be put together. For this purpose, the following transformation would be appropriate:

$$sim = \begin{cases} corr & \text{if } corr \geq 0 \\ 0 & \text{if } corr < 0 \end{cases}$$

For predicting the behavior of one time series from another, it is necessary to consider strong negative, as well as strong positive, correlation. In this case, the following transformation, $sim = |corr|$ might be appropriate. Note that this assumes that you only want to predict magnitude, not direction.

23. Given a similarity measure with values in the interval [0,1] describe two ways to transform this similarity value into a dissimilarity value in the interval [0,∞].

    $d = \frac{1-s}{s}$ and $d = -\log s$.

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.

    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?

    One approach is to compute the distance between the centroids of the two sets of points.

    (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.)

    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.

    Note that the cohesion of a cluster is related to the notion of the proximity of a group of objects among themselves and that the separation of clusters is related to concept of the proximity of two groups of objects. (See Section 8.4.) Furthermore, the proximity of two clusters is an important concept in agglomerative hierarchical clustering. (See Section 8.2.)

25. You are given a set of points S in Euclidean space, as well as the distance of each point in S to a point $\mathbf{x}$. (It does not matter if $\mathbf{x} \in S$.)

    (a) If the goal is to find all points within a specified distance ε of point $\mathbf{y}$, $\mathbf{y} \neq \mathbf{x}$, explain how you could use the triangle inequality and the already calculated distances to $\mathbf{x}$ to potentially reduce the number of distance calculations necessary? Hint: The triangle inequality, $d(\mathbf{x}, \mathbf{z}) \leq d(\mathbf{x}, \mathbf{y}) + d(\mathbf{y}, \mathbf{x})$, can be rewritten as $d(\mathbf{x}, \mathbf{y}) \geq d(\mathbf{x}, \mathbf{z}) - d(\mathbf{y}, \mathbf{z})$.

    Unfortunately, there is a typo and a lack of clarity in the hint. The hint should be phrased as follows: