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.

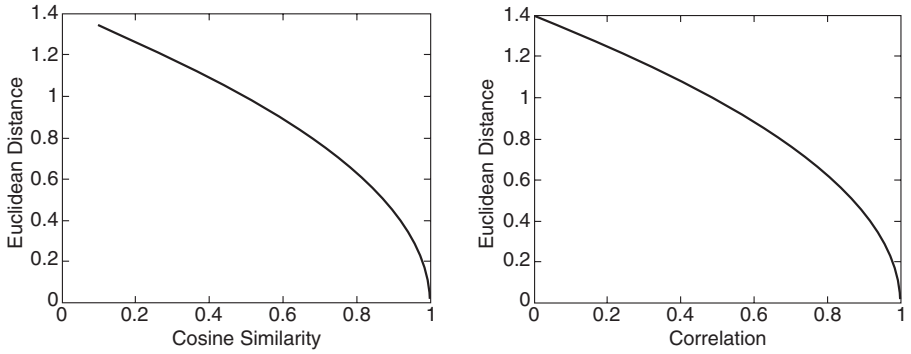
Let $\mathbf{x}$ and $\mathbf{y}$ be two vectors where each vector has an a mean of 0 and a standard deviation of 1. For such vectors, the variance (standard deviation squared) is just n times the sum of its squared attribute values and the correlation between the two vectors is their dot product divided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

Hint: If $\mathbf{z}$ is an arbitrary point of S, then the triangle inequality, $d(\mathbf{x}, \mathbf{y}) \leq d(\mathbf{x}, \mathbf{z}) + d(\mathbf{y}, \mathbf{z})$, can be rewritten as $d(\mathbf{y}, \mathbf{z}) \geq d(\mathbf{x}, \mathbf{y}) - d(\mathbf{x}, \mathbf{z})$.

Another application of the triangle inequality starting with $d(\mathbf{x}, \mathbf{z}) \leq d(\mathbf{x}, \mathbf{y}) + d(\mathbf{y}, \mathbf{z})$, shows that $d(\mathbf{y}, \mathbf{z}) \geq d(\mathbf{x}, \mathbf{z}) - d(\mathbf{x}, \mathbf{y})$. If the lower bound of $d(\mathbf{y}, \mathbf{z})$ obtained from either of these inequalities is greater than ϵ, then $d(\mathbf{y}, \mathbf{z})$ does not need to be calculated. Also, if the upper bound of $d(\mathbf{y}, \mathbf{z})$ obtained from the inequality $d(\mathbf{y}, \mathbf{z}) \leq d(\mathbf{y}, \mathbf{x}) + d(\mathbf{x}, \mathbf{z})$ is less than or equal to ϵ, then $d(\mathbf{x}, \mathbf{z})$ does not need to be calculated.

(b) In general, how would the distance between $\mathbf{x}$ and $\mathbf{y}$ affect the number of distance calculations?

If $\mathbf{x} = \mathbf{y}$ then no calculations are necessary. As $\mathbf{x}$ becomes farther away, typically more distance calculations are needed.

(c) Suppose that you can find a small subset of points S', from the original data set, such that every point in the data set is within a specified distance ε of at least one of the points in S', and that you also have the pairwise distance matrix for S'. Describe a technique that uses this information to compute, with a minimum of distance calculations, the set of all points within a distance of β of a specified point from the data set.

Let $\mathbf{x}$ and $\mathbf{y}$ be the two points and let $\mathbf{x}^*$ and $\mathbf{y}^*$ be the points in S' that are closest to the two points, respectively. If $d(\mathbf{x}^*, \mathbf{y}^*) + 2\epsilon \leq \beta$, then we can safely conclude $d(\mathbf{x}, \mathbf{y}) \leq \beta$. Likewise, if $d(\mathbf{x}^*, \mathbf{y}^*) - 2\epsilon \geq \beta$, then we can safely conclude $d(\mathbf{x}, \mathbf{y}) \geq \beta$. These formulas are derived by considering the cases where $\mathbf{x}$ and $\mathbf{y}$ are as far from $\mathbf{x}^*$ and $\mathbf{y}^*$ as possible and as far or close to each other as possible.

26. Show that 1 minus the Jaccard similarity is a distance measure between two data objects, $\mathbf{x}$ and $\mathbf{y}$, that satisfies the metric axioms given on page 70. Specifically, $d(\mathbf{x}, \mathbf{y}) = 1 - \mathrm{J}(\mathbf{x}, \mathbf{y})$.

1(a). Because $\mathrm{J}(\mathbf{x}, \mathbf{y}) \leq 1, d(\mathbf{x}, \mathbf{y}) \geq 0$.
1(b). Because $\mathrm{J}(\mathbf{x}, \mathbf{x}) = 1, d(\mathbf{x}, \mathbf{x}) = 0$
2. Because $\mathrm{J}(\mathbf{x}, \mathbf{y}) = \mathrm{J}(\mathbf{y}, \mathbf{x}), d(\mathbf{x}, \mathbf{y}) = d(\mathbf{y}, \mathbf{x})$
3. (Proof due to Jeffrey Ullman)
minhash($\mathbf{x}$) is the index of first nonzero entry of $\mathbf{x}$
prob(minhash($\mathbf{x}$) = k) is the probability tha minhash($\mathbf{x}$) = k when $\mathbf{x}$ is randomly permuted.
Note that *prob*(minhash($\mathbf{x}$) = minhash($\mathbf{y}$)) = J($\mathbf{x}, \mathbf{y}$) (minhash lemma)
Therefore, $d(\mathbf{x}, \mathbf{y}) = 1 - prob(\text{minhash}(\mathbf{x}) = \text{minhash}(\mathbf{y})) = prob(\text{minhash}(\mathbf{x}) \neq \text{minhash}(\mathbf{y}))$
We have to show that,
prob(minhash($\mathbf{x}$) $\neq$ minhash($\mathbf{z}$)) $\leq$ *prob*(minhash($\mathbf{x}$) $\neq$ minhash($\mathbf{y}$)) + *prob*(minhash($\mathbf{y}$) $\neq$ minhash($\mathbf{z}$)

However, note that whenever minhash($\mathbf{x}$) $\neq$ minhash($\mathbf{z}$), then at least one of minhash($\mathbf{x}$) $\neq$ minhash($\mathbf{y}$) and minhash($\mathbf{y}$) $\neq$ minhash($\mathbf{z}$) must be true.

27. Show that the distance measure defined as the angle between two data vectors, $\mathbf{x}$ and $\mathbf{y}$, satisfies the metric axioms given on page 70. Specifically, $d(\mathbf{x}, \mathbf{y}) = \arccos(\cos(\mathbf{x}, \mathbf{y}))$.

    Note that angles are in the range 0 to 180°.
    1(a). Because $0 \leq \cos(\mathbf{x}, \mathbf{y}) \leq 1, d(\mathbf{x}, \mathbf{y}) \geq 0$.
    1(b). Because $\cos(\mathbf{x}, \mathbf{x}) = 1, d(\mathbf{x}, \mathbf{x}) = \arccos(1) = 0$
    2. Because $\cos(\mathbf{x}, \mathbf{y}) = \cos(\mathbf{y}, \mathbf{x}), d(\mathbf{x}, \mathbf{y}) = d(\mathbf{y}, \mathbf{x})$
    3. If the three vectors lie in a plane then it is obvious that the angle between $\mathbf{x}$ and $\mathbf{z}$ must be less than or equal to the sum of the angles between $\mathbf{x}$ and $\mathbf{y}$ and $\mathbf{y}$ and $\mathbf{z}$. If $\mathbf{y}'$ is the projection of $\mathbf{y}$ into the plane defined by $\mathbf{x}$ and $\mathbf{z}$, then note that the angles between $\mathbf{x}$ and $\mathbf{y}$ and $\mathbf{y}$ and $\mathbf{z}$ are greater than those between $\mathbf{x}$ and $\mathbf{y}'$ and $\mathbf{y}'$ and $\mathbf{z}$.

28. Explain why computing the proximity between two attributes is often simpler than computing the similarity between two objects.

    In general, an object can be a record whose fields (attributes) are of different types. To compute the overall similarity of two objects in this case, we need to decide how to compute the similarity for each attribute and then combine these similarities. This can be done straightforwardly by using Equations 2.15 or 2.16, but is still somewhat ad hoc, at least compared to proximity measures such as the Euclidean distance or correlation, which are mathematically well-founded. In contrast, the values of an attribute are all of the same type, and thus, if another attribute is of the same type, then the computation of similarity is conceptually and computationally straightforward.

3

Exploring Data

1. Obtain one of the data sets available at the UCI Machine Learning Repository and apply as many of the different visualization techniques described in the chapter as possible. The bibliographic notes and book Web site provide pointers to visualization software.

   MATLAB and R have excellent facilities for visualization. Most of the figures in this chapter were created using MATLAB. R is freely available from `http://www.r-project.org/`.

2. Identify at least two advantages and two disadvantages of using color to visually represent information.

   Advantages: Color makes it much easier to visually distinguish visual elements from one another. For example, three clusters of two-dimensional points are more readily distinguished if the markers representing the points have different colors, rather than only different shapes. Also, figures with color are more interesting to look at.

   Disadvantages: Some people are color blind and may not be able to properly interpret a color figure. Grayscale figures can show more detail in some cases. Color can be hard to use properly. For example, a poor color scheme can be garish or can focus attention on unimportant elements.

3. What are the arrangement issues that arise with respect to three-dimensional plots?

   It would have been better to state this more generally as "What are the issues . . . ," since selection, as well as arrangement plays a key issue in displaying a three-dimensional plot.

   The key issue for three dimensional plots is how to display information so that as little information is obscured as possible. If the plot is of a two-dimensional surface, then the choice of a viewpoint is critical. However, if the plot is in electronic form, then it is sometimes possible to interactively change

the viewpoint to get a complete view of the surface. For three dimensional solids, the situation is even more challenging. Typically, portions of the information must be omitted in order to provide the necessary information. For example, a slice or cross-section of a three dimensional object is often shown. In some cases, transparency can be used. Again, the ability to change the arrangement of the visual elements interactively can be helpful.

4. Discuss the advantages and disadvantages of using sampling to reduce the number of data objects that need to be displayed. Would simple random sampling (without replacement) be a good approach to sampling? Why or why not?

   Simple random sampling is not the best approach since it will eliminate most of the points in sparse regions. It is better to undersample the regions where data objects are too dense while keeping most or all of the data objects from sparse regions.

5. Describe how you would create visualizations to display information that describes the following types of systems.

   Be sure to address the following issues:

   - **Representation.** How will you map objects, attributes, and relationships to visual elements?
   - **Arrangement.** Are there any special considerations that need to be taken into account with respect to how visual elements are displayed? Specific examples might be the choice of viewpoint, the use of transparency, or the separation of certain groups of objects.
   - **Selection.** How will you handle a large number of attributes and data objects?

   The following solutions are intended for illustration.

   (a) Computer networks. Be sure to include both the static aspects of the network, such as connectivity, and the dynamic aspects, such as traffic.

   The connectivity of the network would best be represented as a graph, with the nodes being routers, gateways, or other communications devices and the links representing the connections. The bandwidth of the connection could be represented by the width of the links. Color could be used to show the percent usage of the links and nodes.

   (b) The distribution of specific plant and animal species around the world for a specific moment in time.

   The simplest approach is to display each species on a separate map of the world and to shade the regions of the world where the species occurs. If several species are to be shown at once, then icons for each species can be placed on a map of the world.

(c) The use of computer resources, such as processor time, main memory, and disk, for a set of benchmark database programs.

The resource usage of each program could be displayed as a bar plot of the three quantities. Since the three quantities would have different scales, a proper scaling of the resources would be necessary for this to work well. For example, resource usage could be displayed as a percentage of the total. Alternatively, we could use three bar plots, one for type of resource usage. On each of these plots there would be a bar whose height represents the usage of the corresponding program. This approach would not require any scaling. Yet another option would be to display a line plot of each program's resource usage. For each program, a line would be constructed by (1) considering processor time, main memory, and disk as different x locations, (2) letting the percentage resource usage of a particular program for the three quantities be the y values associated with the x values, and then (3) drawing a line to connect these three points. Note that an ordering of the three quantities needs to be specified, but is arbitrary. For this approach, the resource usage of all programs could be displayed on the same plot.

(d) The change in occupation of workers in a particular country over the last thirty years. Assume that you have yearly information about each person that also includes gender and level of education.

For each gender, the occupation breakdown could be displayed as an array of pie charts, where each row of pie charts indicates a particular level of education and each column indicates a particular year. For convenience, the time gap between each column could be 5 or ten years.

Alternatively, we could order the occupations and then, for each gender, compute the cumulative percent employment for each occupation. If this quantity is plotted for each gender, then the area between two successive lines shows the percentage of employment for this occupation. If a color is associated with each occupation, then the area between each set of lines can also be colored with the color associated with each occupation. A similar way to show the same information would be to use a sequence of stacked bar graphs.

6. Describe one advantage and one disadvantage of a stem and leaf plot with respect to a standard histogram.

   A stem and leaf plot shows you the actual distribution of values. On the other hand, a stem and leaf plot becomes rather unwieldy for a large number of values.

7. How might you address the problem that a histogram depends on the number and location of the bins?

The best approach is to estimate what the actual distribution function of the data looks like using kernel density estimation. This branch of data analysis is relatively well-developed and is more appropriate if the widely available, but simplistic approach of a histogram is not sufficient.

8. Describe how a box plot can give information about whether the value of an attribute is symmetrically distributed. What can you say about the symmetry of the distributions of the attributes shown in Figure 3.11?

   (a) If the line representing the median of the data is in the middle of the box, then the data is symmetrically distributed, at least in terms of the 75% of the data between the first and third quartiles. For the remaining data, the length of the whiskers and outliers is also an indication, although, since these features do not involve as many points, they may be misleading.

   (b) Sepal width and length seem to be relatively symmetrically distributed, petal length seems to be rather skewed, and petal width is somewhat skewed.

9. Compare sepal length, sepal width, petal length, and petal width, using Figure 3.12.

   For Setosa, sepal length > sepal width > petal length > petal width. For Versicolour and Virginiica, sepal length > sepal width and petal length > petal width, but although sepal length > petal length, petal length > sepal width.

10. Comment on the use of a box plot to explore a data set with four attributes: age, weight, height, and income.

    A great deal of information can be obtained by looking at (1) the box plots for each attribute, and (2) the box plots for a particular attribute across various categories of a second attribute. For example, if we compare the box plots of age for different categories of ages, we would see that weight increases with age.

11. Give a possible explanation as to why most of the values of petal length and width fall in the buckets along the diagonal in Figure 3.9.

    We would expect such a distribution if the three species of Iris can be ordered according to their size, and if petal length and width are both correlated to the size of the plant and each other.

12. Use Figures 3.14 and 3.15 to identify a characteristic shared by the petal width and petal length attributes.

There is a relatively flat area in the curves of the Empirical CDF's and the percentile plots for both petal length and petal width. This indicates a set of flowers for which these attributes have a relatively uniform value.

13. Simple line plots, such as that displayed in Figure 2.12 on page 56, which shows two time series, can be used to effectively display high-dimensional data. For example, in Figure 56 it is easy to tell that the frequencies of the two time series are different. What characteristic of time series allows the effective visualization of high-dimensional data?

    The fact that the attribute values are ordered.

14. Describe the types of situations that produce sparse or dense data cubes. Illustrate with examples other than those used in the book.

    Any set of data for which all combinations of values are unlikely to occur would produce sparse data cubes. This would include sets of continuous attributes where the set of objects described by the attributes doesn't occupy the entire data space, but only a fraction of it, as well as discrete attributes, where many combinations of values don't occur.

    A dense data cube would tend to arise, when either almost all combinations of the categories of the underlying attributes occur, or the level of aggregation is high enough so that all combinations are likely to have values. For example, consider a data set that contains the type of traffic accident, as well as its location and date. The original data cube would be very sparse, but if it is aggregated to have categories consisting single or multiple car accident, the state of the accident, and the month in which it occurred, then we would obtain a dense data cube.

15. How might you extend the notion of multidimensional data analysis so that the target variable is a qualitative variable? In other words, what sorts of summary statistics or data visualizations would be of interest?

    A summary statistics that would be of interest would be the frequencies with which values or combinations of values, target and otherwise, occur. From this we could derive conditional relationships among various values. In turn, these relationships could be displayed using a graph similar to that used to display Bayesian networks.

16. Construct a data cube from Table 3.1. Is this a dense or sparse data cube? If it is sparse, identify the cells that are empty.

    The data cube is shown in Table 3.2. It is a dense cube; only two cells are empty.

Table 3.1. Fact table for Exercise 16.

Product ID	Location ID	Number Sold
1	1	10
1	3	6
2	1	5
2	2	22

Table 3.2. Data cube for Exercise 16.

Product ID	Location ID			
	1	2	3	Total
1	10	0	6	16
2	5	22	0	27
Total	15	22	6	43

17. Discuss the differences between dimensionality reduction based on aggregation and dimensionality reduction based on techniques such as PCA and SVD.

    The dimensionality of PCA or SVD can be viewed as a projection of the data onto a reduced set of dimensions. In aggregation, groups of dimensions are combined. In some cases, as when days are aggregated into months or the sales of a product are aggregated by store location, the aggregation can be viewed as a change of scale. In contrast, the dimensionality reduction provided by PCA and SVD do not have such an interpretation.