Dmsol 04 09 2020 - Good notes for all students to read and learning and fulfil their targets

Misingi ya elimu (University of Dar es Salaam)

Introduction to Data Mining

Instructor's Solution Manual

Pang-Ning Tan
Michael Steinbach
Anuj Karpatne
Vipin Kumar

Contents

1 Introduction 1

2 Data 5

3 Classification: Basic Concepts and Techniques 19

4 Classification: Alternative Techniques 41

5 Association Analysis: Basic Concepts and Algorithms 65

6 Association Analysis: Advanced Concepts 91

7 Cluster Analysis: Basic Concepts and Algorithms 121

8 Cluster Analysis: Additional Issues and Algorithms 143

9 Anomaly Detection 153

10 Avoiding False Discoveries 161

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 variations, 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 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 orbiting 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 9.)

    (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 guarantees 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, $\text{Euclidean}(\mathbf{x}, \mathbf{y}) = 2$

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

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

(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 $[0, 1]$.

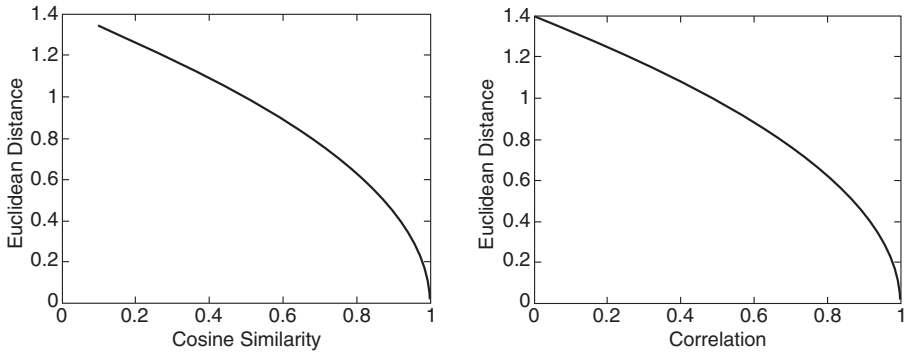
(b) If two objects have a cosine measure of 1, are they identical? Explain.

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?



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

(b) Relationship between Euclidean distance and correlation.

Figure 2.1. Figures for exercise 20.

Since all the 100,000 points fall on the curve, there is a functional relationship between Euclidean distance and cosine similarity for normal-

ized 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, cosine between the two vectors is their dot product.

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

(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 $n - 1$ times the sum of its squared attribute values and the correlation between the two vectors is their dot product

divided by $n - 1$.

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

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 77. A and B are sets and $A - B$ is the set difference.

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 7.2) 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: 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 $d(\mathbf{y}, \mathbf{z})$ 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 77. 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 that 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 77. 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

Classification: Basic Concepts and Techniques

1. Draw the full decision tree for the parity function of four Boolean attributes, A, B, C, and D. Is it possible to simplify the tree?

A	B	C	D	Class
T	T	T	T	T
T	T	T	F	F
T	T	F	T	F
T	T	F	F	T
T	F	T	T	F
T	F	T	F	T
T	F	F	T	T
T	F	F	F	F
F	T	T	T	F
F	T	T	F	T
F	T	F	T	T
F	T	F	F	F
F	F	T	T	T
F	F	T	F	F
F	F	F	T	F
F	F	F	F	T

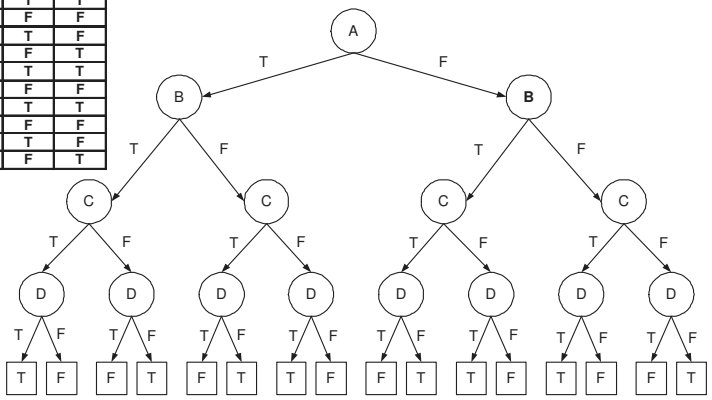


Figure 3.1. Decision tree for parity function of four Boolean attributes.

The preceding tree cannot be simplified.

2. Consider the training examples shown in Table 3.1 for a binary classification problem.

Table 3.1. Data set for Exercise 2.

Customer ID	Gender	Car Type	Shirt Size	Class
1	M	Family	Small	C0
2	M	Sports	Medium	C0
3	M	Sports	Medium	C0
4	M	Sports	Large	C0
5	M	Sports	Extra Large	C0
6	M	Sports	Extra Large	C0
7	F	Sports	Small	C0
8	F	Sports	Small	C0
9	F	Sports	Medium	C0
10	F	Luxury	Large	C0
11	M	Family	Large	C1
12	M	Family	Extra Large	C1
13	M	Family	Medium	C1
14	M	Luxury	Extra Large	C1
15	F	Luxury	Small	C1
16	F	Luxury	Small	C1
17	F	Luxury	Medium	C1
18	F	Luxury	Medium	C1
19	F	Luxury	Medium	C1
20	F	Luxury	Large	C1

(a) Compute the Gini index for the overall collection of training examples.

Answer:

Gini $= 1 - 2 \times 0.5^2 = 0.5$.

(b) Compute the Gini index for the `Customer ID` attribute.

Answer:

The gini for each `Customer ID` value is 0. Therefore, the overall gini for `Customer ID` is 0.

(c) Compute the Gini index for the `Gender` attribute.

Answer:

The gini for `Male` is $1 - 2 \times 0.5^2 = 0.5$. The gini for `Female` is also 0.5. Therefore, the overall gini for `Gender` is $0.5 \times 0.5 + 0.5 \times 0.5 = 0.5$.

(d) Compute the Gini index for the `Car Type` attribute using multiway split.

Answer:

Table 3.2. Data set for Exercise 3.

Instance	a_1	a_2	a_3	Target Class
1	T	T	1.0	+
2	T	T	6.0	+
3	T	F	5.0	−
4	F	F	4.0	+
5	F	T	7.0	−
6	F	T	3.0	−
7	F	F	8.0	−
8	T	F	7.0	+
9	F	T	5.0	−

The gini for `Family` car is 0.375, `Sports` car is 0, and `Luxury` car is 0.2188. The overall gini is 0.1625.

(e) Compute the Gini index for the `Shirt Size` attribute using multiway split.

Answer:

The gini for `Small` shirt size is 0.48, `Medium` shirt size is 0.4898, `Large` shirt size is 0.5, and `Extra Large` shirt size is 0.5. The overall gini for `Shirt Size` attribute is 0.4914.

(f) Which attribute is better, `Gender`, `Car Type`, or `Shirt Size`?

Answer:

`Car Type` because it has the lowest gini among the three attributes.

(g) Explain why `Customer ID` should not be used as the attribute test condition even though it has the lowest Gini.

Answer:

The attribute has no predictive power since new customers are assigned to new `Customer ID`s.

3. Consider the training examples shown in Table 3.2 for a binary classification problem.

(a) What is the entropy of this collection of training examples with respect to the positive class?

Answer:

There are four positive examples and five negative examples. Thus, $P(+) = 4/9$ and $P(-) = 5/9$. The entropy of the training examples is $-4/9 \log_2(4/9) - 5/9 \log_2(5/9) = 0.9911$.

(b) What are the information gains of a_1 and a_2 relative to these training examples?

Answer:

For attribute a_1, the corresponding counts and probabilities are:

a_1	+	-
T	3	1
F	1	4

The entropy for a_1 is

$$\frac{4}{9}\Big[-(3/4)\log_2(3/4)-(1/4)\log_2(1/4)\Big] + \frac{5}{9}\Big[-(1/5)\log_2(1/5)-(4/5)\log_2(4/5)\Big] = 0.7616.$$

Therefore, the information gain for a_1 is $0.9911 - 0.7616 = 0.2294$.

For attribute a_2, the corresponding counts and probabilities are:

a_2	+	-
T	2	3
F	2	2

The entropy for a_2 is

$$\frac{5}{9}\Big[-(2/5)\log_2(2/5)-(3/5)\log_2(3/5)\Big] + \frac{4}{9}\Big[-(2/4)\log_2(2/4)-(2/4)\log_2(2/4)\Big] = 0.9839.$$

Therefore, the information gain for a_2 is $0.9911 - 0.9839 = 0.0072$.

(c) For a_3, which is a continuous attribute, compute the information gain for every possible split.

Answer:

a_3	Class label	Split point	Entropy	Info Gain
1.0	+	2.0	0.8484	0.1427
3.0	-	3.5	0.9885	0.0026
4.0	+	4.5	0.9183	0.0728
5.0	-			
5.0	-	5.5	0.9839	0.0072
6.0	+	6.5	0.9728	0.0183
7.0	+			
7.0	-	7.5	0.8889	0.1022

The best split for a_3 occurs at split point equals to 2.

(d) What is the best split (among a_1, a_2, and a_3) according to the information gain?

Answer:

According to information gain, a_1 produces the best split.

(e) What is the best split (between a_1 and a_2) according to the classification error rate?

Answer:

For attribute a_1: error rate $= 2/9$.

For attribute a_2: error rate $= 4/9$.

Therefore, according to error rate, a_1 produces the best split.

(f) What is the best split (between a_1 and a_2) according to the Gini index?

Answer:

For attribute a_1, the gini index is

$$\frac{4}{9}\left[1-(3/4)^2-(1/4)^2\right]+\frac{5}{9}\left[1-(1/5)^2-(4/5)^2\right]=0.3444.$$

For attribute a_2, the gini index is

$$\frac{5}{9}\left[1-(2/5)^2-(3/5)^2\right]+\frac{4}{9}\left[1-(2/4)^2-(2/4)^2\right]=0.4889.$$

Since the gini index for a_1 is smaller, it produces the better split.

4. Show that the entropy of a node never increases after splitting it into smaller successor nodes.

Answer:

Let $Y=\{y_1,y_2,\cdots,y_c\}$ denote the c classes and $X=\{x_1,x_2,\cdots,x_k\}$ denote the k attribute values of an attribute X. Before a node is split on X, the entropy is:

$$E(Y)=-\sum_{j=1}^{c}P(y_j)\log_2 P(y_j)=\sum_{j=1}^{c}\sum_{i=1}^{k}P(x_i,y_j)\log_2 P(y_j), \tag{3.1}$$

where we have used the fact that $P(y_j)=\sum_{i=1}^{k}P(x_i,y_j)$ from the law of total probability.

After splitting on X, the entropy for each child node $X=x_i$ is:

$$E(Y|x_i)=-\sum_{j=1}^{c}P(y_j|x_i)\log_2 P(y_j|x_i) \tag{3.2}$$

where $P(y_j|x_i)$ is the fraction of examples with $X=x_i$ that belong to class y_j. The entropy after splitting on X is given by the weighted entropy of the

children nodes:

$$\begin{aligned} E(Y|X) &= \sum_{i=1}^{k} P(x_i) E(Y|x_i) \\ &= -\sum_{i=1}^{k}\sum_{j=1}^{c} P(x_i) P(y_j|x_i) \log_2 P(y_j|x_i) \\ &= -\sum_{i=1}^{k}\sum_{j=1}^{c} P(x_i, y_j) \log_2 P(y_j|x_i), \end{aligned} \tag{3.3}$$

where we have used a known fact from probability theory that $P(x_i, y_j) = P(y_j|x_i) \times P(x_i)$. Note that $E(Y|X)$ is also known as the conditional entropy of Y given X.

To answer this question, we need to show that $E(Y|X) \leq E(Y)$. Let us compute the difference between the entropies after splitting and before splitting, i.e., $E(Y|X) - E(Y)$, using Equations 3.1 and 3.3:

$$\begin{aligned} & E(Y|X) - E(Y) \\ = & -\sum_{i=1}^{k}\sum_{j=1}^{c} P(x_i, y_j) \log_2 P(y_j|x_i) + \sum_{i=1}^{k}\sum_{j=1}^{c} P(x_i, y_j) \log_2 P(y_j) \\ = & \sum_{i=1}^{k}\sum_{j=1}^{c} P(x_i, y_j) \log_2 \frac{P(y_j)}{P(y_j|x_i)} \\ = & \sum_{i=1}^{k}\sum_{j=1}^{c} P(x_i, y_j) \log_2 \frac{P(x_i)P(y_j)}{P(x_i, y_j)} \end{aligned} \tag{3.4}$$

To prove that Equation 3.4 is non-positive, we use the following property of a logarithmic function:

$$\sum_{k=1}^{d} a_k \log(z_k) \leq \log\big(\sum_{k=1}^{d} a_k z_k\big), \tag{3.5}$$

subject to the condition that $\sum_{k=1}^{d} a_k = 1$. This property is a special case of a more general theorem involving convex functions (which include the logarithmic function) known as Jensen's inequality.

By applying Jensen's inequality, Equation 3.4 can be bounded as follows:

$$\begin{aligned} E(Y|X) - E(Y) &\leq \log_2 \left[\sum_{i=1}^{k} \sum_{j=1}^{c} P(x_i, y_j) \frac{P(x_i)P(y_j)}{P(x_i, y_j)} \right] \\ &= \log_2 \left[\sum_{i=1}^{k} P(x_i) \sum_{j=1}^{c} P(y_j) \right] \\ &= \log_2(1) \\ &= 0 \end{aligned}$$

Because $E(Y|X) - E(Y) \leq 0$, it follows that entropy never increases after splitting on an attribute.

5. Consider the following data set for a binary class problem.

A	B	Class Label
T	F	+
T	T	+
T	T	+
T	F	−
T	T	+
F	F	−
F	F	−
F	F	−
T	T	−
T	F	−

(a) Calculate the information gain when splitting on A and B. Which attribute would the decision tree induction algorithm choose?

Answer:

The contingency tables after splitting on attributes A and B are:

	$A = T$	$A = F$
+	4	0
−	3	3

	$B = T$	$B = F$
+	3	1
−	1	5

The overall entropy before splitting is:

$$E_{orig} = -0.4 \log 0.4 - 0.6 \log 0.6 = 0.9710$$

The information gain after splitting on A is:

$$\begin{aligned} E_{A=T} &= -\frac{4}{7} \log \frac{4}{7} - \frac{3}{7} \log \frac{3}{7} = 0.9852 \\ E_{A=F} &= -\frac{3}{3} \log \frac{3}{3} - \frac{0}{3} \log \frac{0}{3} = 0 \\ \Delta &= E_{orig} - 7/10 E_{A=T} - 3/10 E_{A=F} = 0.2813 \end{aligned}$$

The information gain after splitting on B is:

$$\begin{aligned} E_{B=T} &= -\frac{3}{4}\log\frac{3}{4} - \frac{1}{4}\log\frac{1}{4} = 0.8113 \\ E_{B=F} &= -\frac{1}{6}\log\frac{1}{6} - \frac{5}{6}\log\frac{5}{6} = 0.6500 \\ \Delta &= E_{orig} - 4/10E_{B=T} - 6/10E_{B=F} = 0.2565 \end{aligned}$$

Therefore, attribute A will be chosen to split the node.

(b) Calculate the gain in the Gini index when splitting on A and B. Which attribute would the decision tree induction algorithm choose?

Answer:

The overall gini before splitting is:

$$G_{orig} = 1 - 0.4^2 - 0.6^2 = 0.48$$

The gain in gini after splitting on A is:

$$\begin{aligned} G_{A=T} &= 1 - \left(\frac{4}{7}\right)^2 - \left(\frac{3}{7}\right)^2 = 0.4898 \\ G_{A=F} &= 1 = \left(\frac{3}{3}\right)^2 - \left(\frac{0}{3}\right)^2 = 0 \\ \Delta &= G_{orig} - 7/10G_{A=T} - 3/10G_{A=F} = 0.1371 \end{aligned}$$

The gain in gini after splitting on B is:

$$\begin{aligned} G_{B=T} &= 1 - \left(\frac{1}{4}\right)^2 - \left(\frac{3}{4}\right)^2 = 0.3750 \\ G_{B=F} &= 1 = \left(\frac{1}{6}\right)^2 - \left(\frac{5}{6}\right)^2 = 0.2778 \\ \Delta &= G_{orig} - 4/10G_{B=T} - 6/10G_{B=F} = 0.1633 \end{aligned}$$

Therefore, attribute B will be chosen to split the node.

(c) Figure 4.13 shows that entropy and the Gini index are both monotonously increasing on the range [0, 0.5] and they are both monotonously decreasing on the range [0.5, 1]. Is it possible that information gain and the gain in the Gini index favor different attributes? Explain.

Answer:

Yes, even though these measures have similar range and monotonous behavior, their respective gains, Δ, which are scaled differences of the measures, do not necessarily behave in the same way, as illustrated by the results in parts (a) and (b).

6. Consider splitting a parent node P into two child nodes, C_1 and C_2, using some attribute test condition. The composition of labeled training instances at every node is summarized in the Table below.

	P	C_1	C_2
Class 0	7	3	4
Class 1	3	0	3

(a) Calculate the Gini index and misclassification error rate of the parent node P.

Answer:

$$\text{Gini of node } P = 1 - \left(\frac{7}{10}\right)^2 - \left(\frac{3}{10}\right)^2 = 0.42$$

$$\text{Error of node } P = \left(\frac{3}{10}\right) = 0.3$$

(b) Calculate the weighted Gini index of the child nodes. Would you consider this attribute test condition if Gini is used as the impurity measure?

Answer:

$$\text{Gini of node } C_1 = 1 - \left(\frac{3}{3}\right)^2 - \left(\frac{0}{3}\right)^2 = 0$$

$$\text{Gini of node } C_2 = 1 - \left(\frac{4}{7}\right)^2 - \left(\frac{3}{7}\right)^2 = 0.49$$

$$\text{Weighted Gini of children} = \left(\frac{3}{10}\right) \times 0 + \left(\frac{7}{10}\right) \times 0.49 = 0.34$$

Based on the drop in Gini measure ($0.42 - 0.34 = 0.08$), we would choose this attribute test condition for splitting.

(c) Calculate the weighted misclassification rate of the child nodes. Would you consider this attribute test condition if misclassification rate is used as the impurity measure?

Answer:

$$\text{Error of node } C_1 = \left(\frac{0}{10}\right) = 0$$

$$\text{Error of node } C_2 = \left(\frac{3}{7}\right)$$

$$\text{Weighted Error of children} = \left(\frac{3}{10}\right) \times 0 + \left(\frac{7}{10}\right) \times \left(\frac{3}{7}\right) = 0.3$$

Since there is no drop in misclassification error rate, we would not consider this attribute test condition for splitting.

7. Consider the following set of training examples.

X	Y	Z	No. of Class C1 Examples	No. of Class C2 Examples
0	0	0	5	40
0	0	1	0	15
0	1	0	10	5
0	1	1	45	0
1	0	0	10	5
1	0	1	25	0
1	1	0	5	20
1	1	1	0	15

(a) Compute a two-level decision tree using the greedy approach described in this chapter. Use the classification error rate as the criterion for splitting. What is the overall error rate of the induced tree?

Answer:

Splitting Attribute at Level 1.

To determine the test condition at the root node, we need to compute the error rates for attributes X, Y, and Z. For attribute X, the corresponding counts are:

X	C1	C2
0	60	60
1	40	40

Therefore, the error rate using attribute X is $(60 + 40)/200 = 0.5$.

For attribute Y, the corresponding counts are:

Y	C1	C2
0	40	60
1	60	40

Therefore, the error rate using attribute Y is $(40 + 40)/200 = 0.4$.

For attribute Z, the corresponding counts are:

Z	C1	C2
0	30	70
1	70	30

Therefore, the error rate using attribute Y is $(30 + 30)/200 = 0.3$.

Since Z gives the lowest error rate, it is chosen as the splitting attribute at level 1.

Splitting Attribute at Level 2.

After splitting on attribute Z, the subsequent test condition may involve either attribute X or Y. This depends on the training examples distributed to the $Z = 0$ and $Z = 1$ child nodes.

For $Z = 0$, the corresponding counts for attributes X and Y are the same, as shown in the table below.

X	C1	C2
0	15	45
1	15	25

Y	C1	C2
0	15	45
1	15	25

The error rate in both cases (X and Y) are $(15 + 15)/100 = 0.3$.

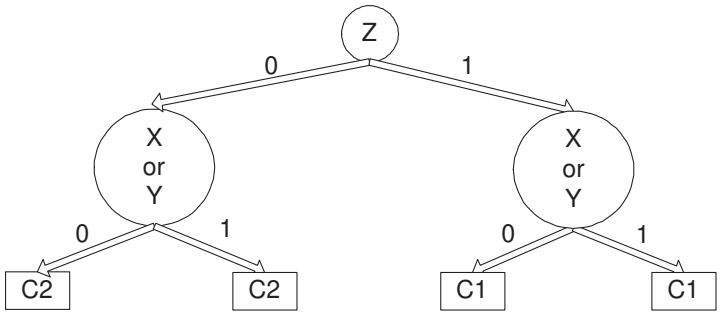
For $Z = 1$, the corresponding counts for attributes X and Y are shown in the tables below.

X	C1	C2
0	45	15
1	25	15

Y	C1	C2
0	25	15
1	45	15

Although the counts are somewhat different, their error rates remain the same, $(15 + 15)/100 = 0.3$.

The corresponding two-level decision tree is shown below.



The overall error rate of the induced tree is $(15+15+15+15)/200 = 0.3$.

(b) Repeat part (a) using X as the first splitting attribute and then choose the best remaining attribute for splitting at each of the two successor nodes. What is the error rate of the induced tree?

Answer:

After choosing attribute X to be the first splitting attribute, the subsequent test condition may involve either attribute Y or attribute Z.

For $X = 0$, the corresponding counts for attributes Y and Z are shown in the table below.

Y	C1	C2
0	5	55
1	55	5

Z	C1	C2
0	15	45
1	45	15

The error rate using attributes Y and Z are 10/120 and 30/120, respectively. Since attribute Y leads to a smaller error rate, it provides a better split.

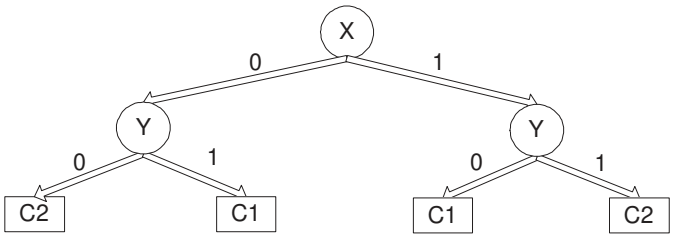
For $X = 1$, the corresponding counts for attributes Y and Z are shown in the tables below.

Y	C1	C2
0	35	5
1	5	35

Z	C1	C2
0	15	25
1	25	15

The error rate using attributes Y and Z are 10/80 and 30/80, respectively. Since attribute Y leads to a smaller error rate, it provides a better split.

The corresponding two-level decision tree is shown below.



The overall error rate of the induced tree is $(10 + 10)/200 = 0.1$.

(c) Compare the results of parts (a) and (b). Comment on the suitability of the greedy heuristic used for splitting attribute selection.

Answer:

From the preceding results, the error rate for part (a) is significantly larger than that for part (b). This examples shows that a greedy heuristic does not always produce an optimal solution.

8. The following table summarizes a data set with three attributes A, B, C and two class labels $+$, $-$. Build a two-level decision tree.

A	B	C	Number of Instances	
			+	−
T	T	T	5	0
F	T	T	0	20
T	F	T	20	0
F	F	T	0	5
T	T	F	0	0
F	T	F	25	0
T	F	F	0	0
F	F	F	0	25

(a) According to the classification error rate, which attribute would be chosen as the first splitting attribute? For each attribute, show the contingency table and the gains in classification error rate.

Answer:

The error rate for the data without partitioning on any attribute is

$$E_{orig} = 1 - \max(\frac{50}{100}, \frac{50}{100}) = \frac{50}{100}.$$

After splitting on attribute A, the gain in error rate is:

	$A=T$	$A=F$
+	25	25
−	0	50

$$E_{A=T} = 1 - \max(\frac{25}{25}, \frac{0}{25}) = \frac{0}{25} = 0$$
$$E_{A=F} = 1 - \max(\frac{25}{75}, \frac{50}{75}) = \frac{25}{75}$$
$$\Delta_A = E_{orig} - \frac{25}{100}E_{A=T} - \frac{75}{100}E_{A=F} = \frac{25}{100}$$

After splitting on attribute B, the gain in error rate is:

	$B=T$	$B=F$
+	30	20
−	20	30

$$E_{B=T} = \frac{20}{50}$$
$$E_{B=F} = \frac{20}{50}$$
$$\Delta_B = E_{orig} - \frac{50}{100}E_{B=T} - \frac{50}{100}E_{B=F} = \frac{10}{100}$$

After splitting on attribute C, the gain in error rate is:

	$C=T$	$C=F$
+	25	25
−	25	25

$$E_{C=T} = \frac{25}{50}$$
$$E_{C=F} = \frac{25}{50}$$
$$\Delta_C = E_{orig} - \frac{50}{100}E_{C=T} - \frac{50}{100}E_{C=F} = \frac{0}{100} = 0$$

The algorithm chooses attribute A because it has the highest gain.

(b) Repeat for the two children of the root node.

Answer:

Because the $A = T$ child node is pure, no further splitting is needed. For the $A = F$ child node, the distribution of training instances is:

B	C	Class label	
		+	−
T	T	0	20
F	T	0	5
T	F	25	0
F	F	0	25

The classification error of the $A = F$ child node is:

$$E_{orig} = \frac{25}{75}$$

After splitting on attribute B, the gain in error rate is:

	$B = T$	$B = F$
+	25	0
−	20	30

$$E_{B=T} = \frac{20}{45}$$
$$E_{B=F} = 0$$
$$\Delta_B = E_{orig} - \frac{45}{75}E_{B=T} - \frac{20}{75}E_{B=F} = \frac{5}{75}$$

After splitting on attribute C, the gain in error rate is:

	$C = T$	$C = F$
+	0	25
−	25	25

$$E_{C=T} = \frac{0}{25}$$
$$E_{C=F} = \frac{25}{50}$$
$$\Delta_C = E_{orig} - \frac{25}{75}E_{C=T} - \frac{50}{75}E_{C=F} = 0$$

The split will be made on attribute B.

(c) How many instances are misclassified by the resulting decision tree?

Answer:

20 instances are misclassified. (The error rate is $\frac{20}{100}$.)

(d) Repeat parts (a), (b), and (c) using C as the splitting attribute.

Answer:

For the $C = T$ child node, the error rate before splitting is:
$E_{orig} = \frac{25}{50}$.
After splitting on attribute A, the gain in error rate is:

	$A = T$	$A = F$
+	25	0
−	0	25

$$E_{A=T} = 0$$
$$E_{A=F} = 0$$
$$\Delta_A = \frac{25}{50}$$

After splitting on attribute B, the gain in error rate is:

	$B = T$	$B = F$
+	5	20
−	20	5

$$E_{B=T} = \frac{5}{25}$$
$$E_{B=F} = \frac{5}{25}$$
$$\Delta_B = \frac{15}{50}$$

Therefore, A is chosen as the splitting attribute.
For the $C = F$ child, the error rate before splitting is: $E_{orig} = \frac{25}{50}$.
After splitting on attribute A, the error rate is:

	$A = T$	$A = F$
+	0	25
−	0	25

$$E_{A=T} = 0$$
$$E_{A=F} = \frac{25}{50}$$
$$\Delta_A = 0$$

After splitting on attribute B, the error rate is:

	$B = T$	$B = F$
+	25	0
−	0	25

$$E_{B=T} = 0$$
$$E_{B=F} = 0$$
$$\Delta_B = \frac{25}{50}$$

Therefore, B is used as the splitting attribute.
The overall error rate of the induced tree is 0.

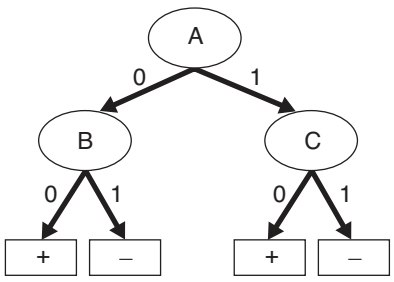
(e) Use the results in parts (c) and (d) to conclude about the greedy nature of the decision tree induction algorithm.
The greedy heuristic does not necessarily lead to the best tree.

9. Consider the decision tree shown in Figure 3.2.

(a) Compute the generalization error rate of the tree using the optimistic approach.
Answer:
According to the optimistic approach, the generalization error rate is $3/10 = 0.3$.

(b) Compute the generalization error rate of the tree using the pessimistic approach. (For simplicity, use the strategy of adding a factor of 0.5 to each leaf node.)
Answer:
According to the pessimistic approach, the generalization error rate is $(3 + 4 \times 0.5)/10 = 0.5$.

(c) Compute the generalization error rate of the tree using the validation set shown above. This approach is known as **reduced error pruning**.
Answer:
According to the reduced error pruning approach, the generalization error rate is $4/5 = 0.8$.



Training:

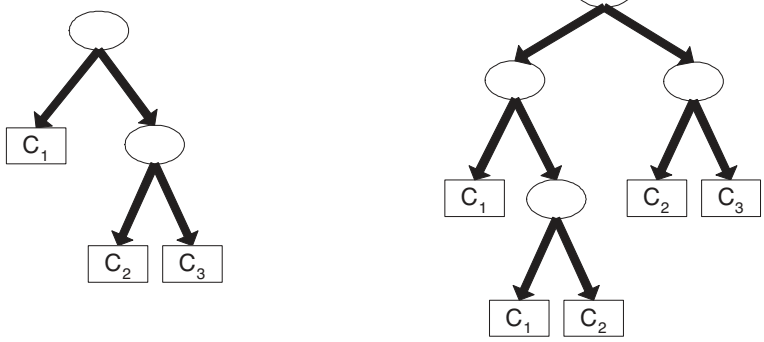
Instance	A	B	C	Class
1	0	0	0	+
2	0	0	1	+
3	0	1	0	+
4	0	1	1	–
5	1	0	0	+
6	1	0	0	+
7	1	1	0	–
8	1	0	1	+
9	1	1	0	–
10	1	1	0	–

Validation:

Instance	A	B	C	Class
11	0	0	0	+
12	0	1	1	+
13	1	1	0	+
14	1	0	1	–
15	1	0	0	+

Figure 3.2. Decision tree and data sets for Exercise 9.

10. Consider the decision trees shown in Figure 3.3. Assume they are generated from a data set that contains 16 binary attributes and 3 classes, C_1, C_2, and C_3. Compute the total description length of each decision tree according to the minimum description length principle.



(a) Decision tree with 7 errors

(b) Decision tree with 4 errors

Figure 3.3. Decision trees for Exercise 10.

- The total description length of a tree is given by:

$$Cost(tree, data) = Cost(tree) + Cost(data|tree).$$

- Each internal node of the tree is encoded by the ID of the splitting attribute. If there are m attributes, the cost of encoding each attribute is $\log_2 m$ bits.
- Each leaf is encoded using the ID of the class it is associated with. If there are k classes, the cost of encoding a class is $\log_2 k$ bits.
- $Cost(tree)$ is the cost of encoding all the nodes in the tree. To simplify the computation, you can assume that the total cost of the tree is obtained by adding up the costs of encoding each internal node and each leaf node.
- $Cost(data|tree)$ is encoded using the classification errors the tree commits on the training set. Each error is encoded by $\log_2 n$ bits, where n is the total number of training instances.

Which decision tree is better, according to the MDL principle?

Answer:

Because there are 16 attributes, the cost for each internal node in the decision tree is:

$$\log_2(m) = \log_2(16) = 4$$

Furthermore, because there are 3 classes, the cost for each leaf node is:

$$\lceil \log_2(k) \rceil = \lceil \log_2(3) \rceil = 2$$

The cost for each misclassification error is $\log_2(n)$.

The overall cost for the decision tree (a) is $2{\times}4{+}3{\times}2{+}7{\times}\log_2 n = 14{+}7\log_2 n$ and the overall cost for the decision tree (b) is $4{\times}4{+}5{\times}2{+}4{\times}5 = 26{+}4\log_2 n$. According to the MDL principle, tree (a) is better than (b) if $n < 16$ and is worse than (b) if $n > 16$.

11. This exercise, inspired by the discussions in [6], highlights one of the known limitations of the leave-one-out model evaluation procedure. Let us consider a data set containing 50 positive and 50 negative instances, where the attributes are purely random and contain no information about the class labels. Hence, the generalization error rate of any classification model learned over this data is expected to be 0.5. Let us consider a classifier that assigns the majority class label of training instances (ties resolved by using the positive label as the default class) to any test instance, irrespective of its attribute values. We can call this approach as the *majority inducer* classifier. Determine the error rate of this classifier using the following methods.

(a) Leave-one-out.

Answer: Let us represent our data set as $\mathcal{D} = \{\mathbf{x}_i\}_{i=1}^{100}$, where $\mathbf{x}_i$ is the i^{th} data instance. We know $\mathcal{D}$ has 50 positives and 50 negatives. In the leave-one-out method, the test error on every instance $\mathbf{x}_i$ is computed by applying a classification model trained on all data instances in $\mathcal{D}$ excluding $\mathbf{x}_i$, denoted as $\mathcal{D}_{-i} = \mathcal{D} \setminus \{\mathbf{x}_i\}$. If we consider the case where $\mathbf{x}_i$ is positive, then $\mathcal{D}_{-i}$ will end up with one less positive than $\mathcal{D}$ (containing 49 positives and 50 negatives). The majority inducer classifier will thus assign $\mathbf{x}_i$ to the majority class, which is negative, and thus incur an error. On the other hand, if we consider $\mathbf{x}_i$ to be negative, then $\mathcal{D}_{-i}$ will contain 49 negatives and 50 positives, and the majority inducer will incorrectly assign $\mathbf{x}_i$ to be positive (the majority class). Hence, the majority inducer would make an error on every data instance using leave-one-out, and thus have an error rate of 1.

(b) 2-fold stratified cross-validation, where the proportion of class labels at every fold is kept same as that of the overall data.

Answer: If we divide the data set into two folds such that both folds have equal number of positives and negatives, the majority inducer trained over any of the two folds will face a tie and thus assign test instances to the default class, which is positive. Since the default class will be correct 50% of times on any fold, the error rate of majority inducer using 2-fold stratified cross-validation will be 0.5.

(c) From the results above, which method provides a more reliable evaluation of the classifier's generalization error rate?

Answer: Cross-validation provides a more reliable estimate of the generalization error rate of the majority inducer classifier on this data set, which is expected to be 0.5. Leave-one-out is quite susceptible to changes in the number of positive and negative instances in the training set, even by a single count, leading to a high error rate of 1 for the majority inducer. As another example, if we consider the minority inducer classifier, which labels every test instance with the minority class in the training set, we would find that the leave-one-out method would result in an error rate of 0 for the minority inducer, which is quite misleading since the attributes contain no information about the classes and any classifier is expected to have an error rate of 0.5.

12. Consider a labeled data set containing 100 data instances, which is randomly partitioned into two sets A and B, each containing 50 instances. We use A as the training set to learn two decision trees, T_{10} with 10 leaf nodes and T_{100} with 100 leaf nodes. The accuracies of the two decision trees on data sets A and B are shown in Table 3.3.

(a) Based on the accuracies shown in Table 3.3, which classification model would you expect to have better performance on unseen instances?

Table 3.3. Comparing the test accuracy of decision trees T_{10} and T_{100}.

	Accuracy	
Data Set	T_{10}	T_{100}
A	0.86	0.97
B	0.84	0.77

Answer:

Tree T_{10} is expected to show better generalization performance on unseen instances than T_{100}. We can see that the training accuracy of T_{100} on dataset A is very high, but its test accuracy on dataset B that was not used for training is low. This gap between training and test accuracies implies that T_{100} suffers from overfitting. Hence, even if the training accuracy of T_{100} is very high on dataset A, it is not representative of generalization performance on unseen instances in dataset B. On the other hand, tree T_{10} has moderately low training accuracy on dataset A, but the test accuracy of T_{10} on dataset B is not very different. This implies that T_{10} does not suffer from overfitting and the training performance is indeed indicative of generalization performance.

(b) Now, you tested T_{10} and T_{100} on the entire data set $(A+B)$ and found that the classification accuracy of T_{10} on data set $(A+B)$ is 0.85, whereas the classification accuracy of T_{100} on the data set $(A+B)$ is 0.87. Based on this new information and your observations from Table 3.3, which classification model would you finally choose for classification?

Answer:

We would still choose T_{10} over T_{100} for classification. The high accuracy of T_{100} on dataset $(A+B)$ can be attributed to the high training accuracy of T_{100} on dataset A, which is an artifact of overfitting. Note that the performance of a classifier on the combined dataset $(A+B)$ cannot be viewed as an estimate of generalization performance, since it contains instances used for training (from dataset A). Hence, the final decision of choosing T_{10} over T_{100} is still motivated solely by its superior accuracy on unseen test instances in B.

13. Consider the following approach for testing whether a classifier A beats another classifier B. Let N be the size of a given data set, p_A be the accuracy of classifier A, p_B be the accuracy of classifier B, and $p = (p_A + p_B)/2$ be the average accuracy for both classifiers. To test whether classifier A is significantly

better than B, the following Z-statistic is used:

$$Z = \frac{p_A - p_B}{\sqrt{\frac{2p(1-p)}{N}}}.$$

Classifier A is assumed to be better than classifier B if Z > 1.96.

Table 3.4 compares the accuracies of three different classifiers, decision tree classifiers, naïve Bayes classifiers, and support vector machines, on various data sets. (The latter two classifiers are described in Chapter 5.)

Table 3.4. Comparing the accuracy of various classification methods.

Data Set	Size (N)	Decision Tree (%)	naïve Bayes (%)	Support vector machine (%)
Anneal	898	92.09	79.62	87.19
Australia	690	85.51	76.81	84.78
Auto	205	81.95	58.05	70.73
Breast	699	95.14	95.99	96.42
Cleve	303	76.24	83.50	84.49
Credit	690	85.80	77.54	85.07
Diabetes	768	72.40	75.91	76.82
German	1000	70.90	74.70	74.40
Glass	214	67.29	48.59	59.81
Heart	270	80.00	84.07	83.70
Hepatitis	155	81.94	83.23	87.10
Horse	368	85.33	78.80	82.61
Ionosphere	351	89.17	82.34	88.89
Iris	150	94.67	95.33	96.00
Labor	57	78.95	94.74	92.98
Led7	3200	73.34	73.16	73.56
Lymphography	148	77.03	83.11	86.49
Pima	768	74.35	76.04	76.95
Sonar	208	78.85	69.71	76.92
Tic-tac-toe	958	83.72	70.04	98.33
Vehicle	846	71.04	45.04	74.94
Wine	178	94.38	96.63	98.88
Zoo	101	93.07	93.07	96.04

Answer:

A summary of the relative performance of the classifiers is given below:

win-loss-draw	Decision tree	Naïve Bayes	Support vector machine
Decision tree	0 - 0 - 23	9 - 3 - 11	2 - 7- 14
Naïve Bayes	3 - 9 - 11	0 - 0 - 23	0 - 8 - 15
Support vector machine	7 - 2 - 14	8 - 0 - 15	0 - 0 - 23

14. Let X be a binomial random variable with mean Np and variance $Np(1-p)$. Show that the ratio X/N also has a binomial distribution with mean p and variance $p(1-p)/N$.

Answer: Let $r = X/N$. Since X has a binomial distribution, r also has the same distribution. The mean and variance for r can be computed as follows:

$$\text{Mean}, E[r] = E[X/N] = E[X]/N = (Np)/N = p;$$

$$\begin{aligned} \text{Variance}, E[(r - E[r])^2] &= E[(X/N - E[X/N])^2] \\ &= E[(X - E[X])^2]/N^2 \\ &= Np(1-p)/N^2 \\ &= p(1-p)/N \end{aligned}$$

4

Classification: Alternative Techniques

1. Consider a binary classification problem with the following set of attributes and attribute values:

   - Air Conditioner = {Working, Broken}
   - Engine = {Good, Bad}
   - Mileage = {High, Medium, Low}
   - Rust = {Yes, No}

   Suppose a rule-based classifier produces the following rule set:

Mileage = High $\longrightarrow$ Value = Low
Mileage = Low $\longrightarrow$ Value = High
Air Conditioner = Working, Engine = Good $\longrightarrow$ Value = High
Air Conditioner = Working, Engine = Bad $\longrightarrow$ Value = Low
Air Conditioner = Broken $\longrightarrow$ Value = Low

   (a) Are the rules mutually exclustive?
   Answer: No

   (b) Is the rule set exhaustive?
   Answer: Yes

   (c) Is ordering needed for this set of rules?
   Answer: Yes because a test instance may trigger more than one rule.

   (d) Do you need a default class for the rule set?
   Answer: No because every instance is guaranteed to trigger at least one rule.

2. The RIPPER algorithm (by Cohen [1]) is an extension of an earlier algorithm called IREP (by Fürnkranz and Widmer [3]). Both algorithms apply the **reduced-error pruning** method to determine whether a rule needs to be pruned. The reduced error pruning method uses a validation set to estimate the generalization error of a classifier. Consider the following pair of rules:

$$R_1: \quad A \longrightarrow C$$
$$R_2: \quad A \wedge B \longrightarrow C$$

R_2 is obtained by adding a new conjunct, B, to the left-hand side of R_1. For this question, you will be asked to determine whether R_2 is preferred over R_1 from the perspectives of rule-growing and rule-pruning. To determine whether a rule should be pruned, IREP computes the following measure:

$$v_{IREP} = \frac{p + (N - n)}{P + N},$$

where P is the total number of positive examples in the validation set, N is the total number of negative examples in the validation set, p is the number of positive examples in the validation set covered by the rule, and n is the number of negative examples in the validation set covered by the rule. v_{IREP} is actually similar to classification accuracy for the validation set. IREP favors rules that have higher values of v_{IREP}. On the other hand, RIPPER applies the following measure to determine whether a rule should be pruned:

$$v_{RIPPER} = \frac{p - n}{p + n}.$$

(a) Suppose R_1 is covered by 350 positive examples and 150 negative examples, while R_2 is covered by 300 positive examples and 50 negative examples. Compute the FOIL's information gain for the rule R_2 with respect to R_1.

Answer:

For this problem, $p_0 = 350$, $n_0 = 150$, $p_1 = 300$, and $n_1 = 50$. Therefore, the FOIL's information gain for R_2 with respect to R_1 is:

$$Gain = 300 \times \left[\log_2 \frac{300}{350} - \log_2 \frac{350}{500} \right] = 87.65$$

(b) Consider a validation set that contains 500 positive examples and 500 negative examples. For R_1, suppose the number of positive examples covered by the rule is 200, and the number of negative examples covered by the rule is 50. For R_2, suppose the number of positive examples covered by the rule is 100 and the number of negative examples is 5. Compute v_{IREP} for both rules. Which rule does IREP prefer?

Answer:

For this problem, $P = 500$, and $N = 500$.
For rule $R1$, $p = 200$ and $n = 50$. Therefore,

$$V_{IREP}(R1) = \frac{p + (N - n)}{P + N} = \frac{200 + (500 - 50)}{1000} = 0.65$$

For rule $R2$, $p = 100$ and $n = 5$.

$$V_{IREP}(R2) = \frac{p + (N - n)}{P + N} = \frac{100 + (500 - 5)}{1000} = 0.595$$

Thus, IREP prefers rule R1.

(c) Compute v_{RIPPER} for the previous problem. Which rule does RIPPER prefer?
Answer:

$$V_{RIPPER}(R1) = \frac{p - n}{p + n} = \frac{150}{250} = 0.6$$

$$V_{RIPPER}(R2) = \frac{p - n}{p + n} = \frac{95}{105} = 0.9$$

Thus, RIPPER prefers the rule R2.

3. C4.5rules is an implementation of an indirect method for generating rules from a decision tree. RIPPER is an implementation of a direct method for generating rules directly from data.

(a) Discuss the strengths and weaknesses of both methods.
Answer:
The C4.5 rules algorithm generates classification rules from a global perspective. This is because the rules are derived from decision trees, which are induced with the objective of partitioning the feature space into homogeneous regions, without focusing on any classes. In contrast, RIPPER generates rules one-class-at-a-time. Thus, it is more biased towards the classes that are generated first.

(b) Consider a data set that has a large difference in the class size (i.e., some classes are much bigger than others). Which method (between C4.5rules and RIPPER) is better in terms of finding high accuracy rules for the small classes?
Answer:
The class-ordering scheme used by C4.5rules has an easier interpretation than the scheme used by RIPPER.

4. Consider a training set that contains 100 positive examples and 400 negative examples. For each of the following candidate rules,

R_1: $A \longrightarrow +$ (covers 4 positive and 1 negative examples),
R_2: $B \longrightarrow +$ (covers 30 positive and 10 negative examples),
R_3: $C \longrightarrow +$ (covers 100 positive and 90 negative examples),

determine which is the best and worst candidate rule according to:

(a) Rule accuracy.

Answer:

The accuracies of the rules are 80% (for R_1), 75% (for R_2), and 52.6% (for R_3), respectively. Therefore R_1 is the best candidate and R_3 is the worst candidate according to rule accuracy.

(b) FOIL's information gain.

Answer:

Assume the initial rule is $\emptyset \longrightarrow +$. This rule covers $p_0 = 100$ positive examples and $n_0 = 400$ negative examples.

The rule R_1 covers $p_1 = 4$ positive examples and $n_1 = 1$ negative example. Therefore, the FOIL's information gain for this rule is

$$4 \times \left(\log_2 \frac{4}{5} - \log_2 \frac{100}{500} \right) = 8.$$

The rule R_2 covers $p_1 = 30$ positive examples and $n_1 = 10$ negative example. Therefore, the FOIL's information gain for this rule is

$$30 \times \left(\log_2 \frac{30}{40} - \log_2 \frac{100}{500} \right) = 57.2.$$

The rule R_3 covers $p_1 = 100$ positive examples and $n_1 = 90$ negative example. Therefore, the FOIL's information gain for this rule is

$$100 \times \left(\log_2 \frac{100}{190} - \log_2 \frac{100}{500} \right) = 139.6.$$

Therefore, R_3 is the best candidate and R_1 is the worst candidate according to FOIL's information gain.

(c) The likelihood ratio statistic.

Answer:

For R_1, the expected frequency for the positive class is $5 \times 100/500 = 1$ and the expected frequency for the negative class is $5 \times 400/500 = 4$. Therefore, the likelihood ratio for R_1 is

$$2 \times \left[4 \times \log_2(4/1) + 1 \times \log_2(1/4) \right] = 12.$$

For R_2, the expected frequency for the positive class is $40 \times 100/500 = 8$ and the expected frequency for the negative class is $40 \times 400/500 = 32$. Therefore, the likelihood ratio for R_2 is

$$2 \times \left[30 \times \log_2(30/8) + 10 \times \log_2(10/32)\right] = 80.85$$

For R_3, the expected frequency for the positive class is $190 \times 100/500 = 38$ and the expected frequency for the negative class is $190 \times 400/500 = 152$. Therefore, the likelihood ratio for R_3 is

$$2 \times \left[100 \times \log_2(100/38) + 90 \times \log_2(90/152)\right] = 143.09$$

Therefore, R_3 is the best candidate and R_1 is the worst candidate according to the likelihood ratio statistic.

(d) The Laplace measure.

Answer:

The Laplace measure of the rules are 71.43% (for R_1), 73.81% (for R_2), and 52.6% (for R_3), respectively. Therefore R_2 is the best candidate and R_3 is the worst candidate according to the Laplace measure.

(e) The m-estimate measure (with $k = 2$ and $p_+ = 0.2$).

Answer:

The m-estimate measure of the rules are 62.86% (for R_1), 73.38% (for R_2), and 52.3% (for R_3), respectively. Therefore R_2 is the best candidate and R_3 is the worst candidate according to the m-estimate measure.

5. Figure 4.1 illustrates the coverage of the classification rules $R1$, $R2$, and $R3$. Determine which is the best and worst rule according to:

(a) The likelihood ratio statistic.

Answer:

There are 29 positive examples and 21 negative examples in the data set. $R1$ covers 12 positive examples and 3 negative examples. The expected frequency for the positive class is $15 \times 29/50 = 8.7$ and the expected frequency for the negative class is $15 \times 21/50 = 6.3$. Therefore, the likelihood ratio for $R1$ is

$$2 \times \left[12 \times \log_2(12/8.7) + 3 \times \log_2(3/6.3)\right] = 4.71.$$

$R2$ covers 7 positive examples and 3 negative examples. The expected frequency for the positive class is $10 \times 29/50 = 5.8$ and the expected frequency for the negative class is $10 \times 21/50 = 4.2$. Therefore, the likelihood ratio for $R2$ is

$$2 \times \left[7 \times \log_2(7/5.8) + 3 \times \log_2(3/4.2)\right] = 0.89.$$

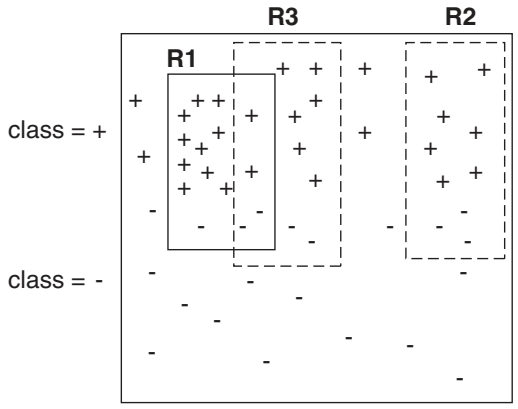


Figure 4.1. Elimination of training records by the sequential covering algorithm. $R1$, $R2$, and $R3$ represent regions covered by three different rules.

$R3$ covers 8 positive examples and 4 negative examples. The expected frequency for the positive class is $12 \times 29/50 = 6.96$ and the expected frequency for the negative class is $12 \times 21/50 = 5.04$. Therefore, the likelihood ratio for $R3$ is

$$2 \times \left[8 \times \log_2(8/6.96) + 4 \times \log_2(4/5.04) \right] = 0.5472.$$

$R1$ is the best rule and $R3$ is the worst rule according to the likelihood ratio statistic.

(b) The Laplace measure.

Answer:

The Laplace measure for the rules are 76.47% (for $R1$), 66.67% (for $R2$), and 64.29% (for $R3$), respectively. Therefore $R1$ is the best rule and $R3$ is the worst rule according to the Laplace measure.

(c) The m-estimate measure (with $k = 2$ and $p_+ = 0.58$).

Answer:

The m-estimate measure for the rules are 77.41% (for $R1$), 68.0% (for $R2$), and 65.43% (for $R3$), respectively. Therefore $R1$ is the best rule and $R3$ is the worst rule according to the m-estimate measure.

(d) The rule accuracy after $R1$ has been discovered, where none of the examples covered by $R1$ are discarded).

Answer:

If the examples for $R1$ are not discarded, then $R2$ will be chosen because it has a higher accuracy (70%) than $R3$ (66.7%).

(e) The rule accuracy after $R1$ has been discovered, where only the positive examples covered by $R1$ are discarded).

Answer:

If the positive examples covered by $R1$ are discarded, the new accuracies for $R2$ and $R3$ are 70% and 60%, respectively. Therefore $R2$ is preferred over $R3$.

(f) The rule accuracy after $R1$ has been discovered, where both positive and negative examples covered by $R1$ are discarded.

Answer:

If the positive and negative examples covered by $R1$ are discarded, the new accuracies for $R2$ and $R3$ are 70% and 75%, respectively. In this case, $R3$ is preferred over $R2$.

6. (a) Suppose the fraction of undergraduate students who smoke is 15% and the fraction of graduate students who smoke is 23%. If one-fifth of the college students are graduate students and the rest are undergraduates, what is the probability that a student who smokes is a graduate student?

Answer:

Given $P(S|UG) = 0.15$, $P(S|G) = 0.23$, $P(G) = 0.2$, $P(UG) = 0.8$. We want to compute $P(G|S)$.

According to Bayesian Theorem,

$$P(G|S) = \frac{0.23 \times 0.2}{0.15 \times 0.8 + 0.23 \times 0.2} = 0.277. \tag{4.1}$$

(b) Given the information in part (a), is a randomly chosen college student more likely to be a graduate or undergraduate student?

Answer:

An undergraduate student, because $P(UG) > P(G)$.

(c) Repeat part (b) assuming that the student is a smoker.

Answer:

An undergraduate student because $P(UG|S) > P(G|S)$.

(d) Suppose 30% of the graduate students live in a dorm but only 10% of the undergraduate students live in a dorm. If a student smokes and lives in the dorm, is he or she more likely to be a graduate or undergraduate student? You can assume independence between students who live in a dorm and those who smoke.

Answer:

First, we need to estimate all the probabilities.

$P(D|UG) = 0.1$, $P(D|G) = 0.3$.

$P(D) = P(UG).P(D|UG) + P(G).P(D|G) = 0.8 * 0.1 + 0.2 * 0.3 = 0.14$.

$P(S) = P(S|UG)P(UG) + P(S|G)P(G) = 0.15 * 0.8 + 0.23 * 0.2 = 0.166$.

$P(DS|G) = P(D|G) \times P(S|G) = 0.3 \times 0.23 = 0.069$ (using conditional independent assumption)
$P(DS|UG) = P(D|UG) \times P(S|UG) = 0.1 \times 0.15 = 0.015$.
We need to compute $P(G|DS)$ and $P(UG|DS)$.

$$P(G|DS) = \frac{0.069 \times 0.2}{P(DS)} = \frac{0.0138}{P(DS)}$$

$$P(UG|DS) = \frac{0.015 \times 0.8}{P(DS)} = \frac{0.012}{P(DS)}$$

Since $P(G|DS) > P(UG|DS)$, he/she is more likely to be a graduate student.

7. Consider the data set shown in Table 4.1

Table 4.1. Data set for Exercise 7.

Record	A	B	C	Class
1	0	0	0	+
2	0	0	1	−
3	0	1	1	−
4	0	1	1	−
5	0	0	1	+
6	1	0	1	+
7	1	0	1	−
8	1	0	1	−
9	1	1	1	+
10	1	0	1	+

(a) Estimate the conditional probabilities for $P(A|+)$, $P(B|+)$, $P(C|+)$, $P(A|-)$, $P(B|-)$, and $P(C|-)$.

Answer:

$P(A = 1|-) = 2/5 = 0.4$, $P(B = 1|-) = 2/5 = 0.4$,
$P(C = 1|-) = 1$, $P(A = 0|-) = 3/5 = 0.6$,
$P(B = 0|-) = 3/5 = 0.6, P(C = 0|-) = 0$; $P(A = 1|+) = 3/5 = 0.6$,
$P(B = 1|+) = 1/5 = 0.2$, $P(C = 1|+) = 2/5 = 0.4$,
$P(A = 0|+) = 2/5 = 0.4$, $P(B = 0|+) = 4/5 = 0.8$,
$P(C = 0|+) = 3/5 = 0.6$.

(b) Use the estimate of conditional probabilities given in the previous question to predict the class label for a test sample $(A = 0, B = 1, C = 0)$ using the naïve Bayes approach.

Answer:

Let $P(A = 0, B = 1, C = 0) = K$.

$$\begin{aligned} & P(+|A = 0, B = 1, C = 0) \\ = & \frac{P(A = 0, B = 1, C = 0|+) \times P(+)}{P(A = 0, B = 1, C = 0)} \\ = & \frac{P(A = 0|+)P(B = 1|+)P(C = 0|+) \times P(+)}{K} \\ = & 0.4 \times 0.2 \times 0.6 \times 0.5/K \\ = & 0.024/K. \end{aligned}$$

$$\begin{aligned} & P(-|A = 0, B = 1, C = 0) \\ = & \frac{P(A = 0, B = 1, C = 0|-) \times P(-)}{P(A = 0, B = 1, C = 0)} \\ = & \frac{P(A = 0|-) \times P(B = 1|-) \times P(C = 0|-) \times P(-)}{K} \\ = & 0/K \end{aligned}$$

The class label should be '+'.

(c) Estimate the conditional probabilities using the m-estimate approach, with $p = 1/2$ and $m = 4$.

Answer:

$P(A = 0|+) = (2 + 2)/(5 + 4) = 4/9$,
$P(A = 0|-) = (3 + 2)/(5 + 4) = 5/9$,
$P(B = 1|+) = (1 + 2)/(5 + 4) = 3/9$,
$P(B = 1|-) = (2 + 2)/(5 + 4) = 4/9$,
$P(C = 0|+) = (3 + 2)/(5 + 4) = 5/9$,
$P(C = 0|-) = (0 + 2)/(5 + 4) = 2/9$.

(d) Repeat part (b) using the conditional probabilities given in part (c).

Answer:

Let $P(A = 0, B = 1, C = 0) = K$

$$\begin{aligned} & P(+|A = 0, B = 1, C = 0) \\ = & \frac{P(A = 0, B = 1, C = 0|+) \times P(+)}{P(A = 0, B = 1, C = 0)} \\ = & \frac{P(A = 0|+)P(B = 1|+)P(C = 0|+) \times P(+)}{K} \\ = & \frac{(4/9) \times (3/9) \times (5/9) \times 0.5}{K} \\ = & 0.0412/K \end{aligned}$$

$$
\begin{aligned}
& P(-|A=0,B=1,C=0) \\
= & \frac{P(A=0,B=1,C=0|-)\times P(-)}{P(A=0,B=1,C=0)} \\
= & \frac{P(A=0|-)\times P(B=1|-)\times P(C=0|-)\times P(-)}{K} \\
= & \frac{(5/9)\times(4/9)\times(2/9)\times 0.5}{K} \\
= & \ 0.0274/K
\end{aligned}
$$

The class label should be '+'.

(e) Compare the two methods for estimating probabilities. Which method is better and why?

Answer:

When one of the conditional probability is zero, the estimate for conditional probabilities using the m-estimate probability approach is better, since we don't want the entire expression becomes zero.

8. Consider the data set shown in Table 4.2.

Table 4.2. Data set for Exercise 8.

Instance	A	B	C	Class
1	0	0	1	−
2	1	0	1	+
3	0	1	0	−
4	1	0	0	−
5	1	0	1	+
6	0	0	1	+
7	1	1	0	−
8	0	0	0	−
9	0	1	0	+
10	1	1	1	+

(a) Estimate the conditional probabilities for $P(A=1|+)$, $P(B=1|+)$, $P(C=1|+)$, $P(A=1|-)$, $P(B=1|-)$, and $P(C=1|-)$ using the same approach as in the previous problem.

Answer:

$P(A=1|+)=0.6$, $P(B=1|+)=0.4$, $P(C=1|+)=0.8$, $P(A=1|-)=0.4$, $P(B=1|-)=0.4$, and $P(C=1|-)=0.2$

(b) Use the conditional probabilities in part (a) to predict the class label for a test sample $(A=1,B=1,C=1)$ using the naïve Bayes approach.

Answer:

Let $R : (A = 1, B = 1, C = 1)$ be the test record. To determine its class, we need to compute $P(+|R)$ and $P(-|R)$. Using Bayes theorem, $P(+|R) = P(R|+)P(+)/P(R)$ and $P(-|R) = P(R|-)P(-)/P(R)$. Since $P(+) = P(-) = 0.5$ and $P(R)$ is constant, R can be classified by comparing $P(+|R)$ and $P(-|R)$.

For this question,

$$\begin{aligned} P(R|+) &= P(A = 1|+) \times P(B = 1|+) \times P(C = 1|+) = 0.192 \\ P(R|-) &= P(A = 1|-) \times P(B = 1|-) \times P(C = 1|-) = 0.032 \end{aligned}$$

Since $P(R|+)$ is larger, the record is assigned to $(+)$ class.

(c) Compare $P(A = 1)$, $P(B = 1)$, and $P(A = 1, B = 1)$. State the relationships between A and B.

Answer:

$P(A = 1) = 0.5$, $P(B = 1) = 0.4$ and $P(A = 1, B = 1) = P(A) \times P(B) = 0.2$. Therefore, A and B are independent.

(d) Repeat the analysis in part (c) using $P(A = 1)$, $P(B = 0)$, and $P(A = 1, B = 0)$.

Answer:

$P(A = 1) = 0.5$, $P(B = 0) = 0.6$, and $P(A = 1, B = 0) = P(A = 1) \times P(B = 0) = 0.3$. A and B are still independent.

(e) Compare $P(A = 1, B = 1|Class = +)$ against $P(A = 1|Class = +)$ and $P(B = 1|Class = +)$. Are the variables conditionally independent given the class?

Answer:

Compare $P(A = 1, B = 1|+) = 0.2$ against $P(A = 1|+) = 0.6$ and $P(B = 1|Class = +) = 0.4$. Since the product between $P(A = 1|+)$ and $P(A = 1|-)$ are not the same as $P(A = 1, B = 1|+)$, A and B are not conditionally independent given the class.

9. (a) Explain how naïve Bayes performs on the data set shown in Figure 4.2.

Answer:

NB will not do well on this data set because the conditional probabilities for each distinguishing attribute given the class are the same for both class A and class B.

(b) If each class is further divided such that there are four classes ($A1$, $A2$, $B1$, and $B2$), will naïve Bayes perform better?

Answer:

The performance of NB will improve on the subclasses because the product of conditional probabilities among the distinguishing attributes will be different for each subclass.

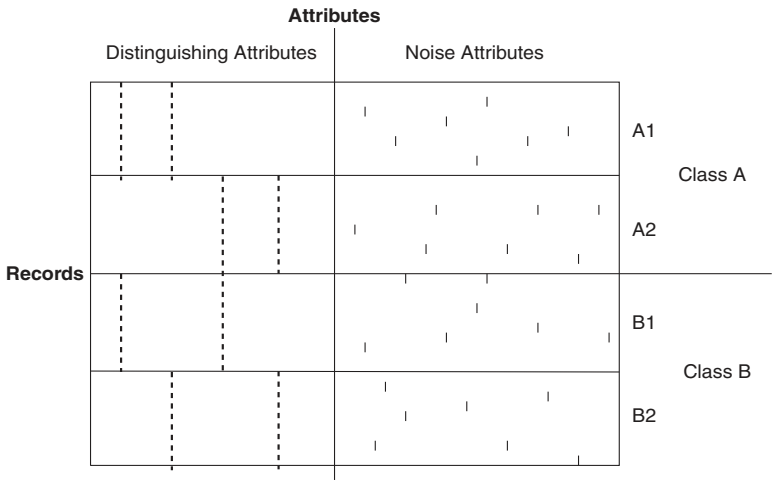


Figure 4.2. Data set for Exercise 9.

(c) How will a decision tree perform on this data set (for the two-class problem)? What if there are four classes?

Answer:

For the two-class problem, decision tree will not perform well because the entropy will not improve after splitting the data using the distinguishing attributes. If there are four classes, then decision tree will improve considerably.

10. Figure 4.3 illustrates the Bayesian belief network for the data set shown in Table 4.3. (Assume that all the attributes are binary).

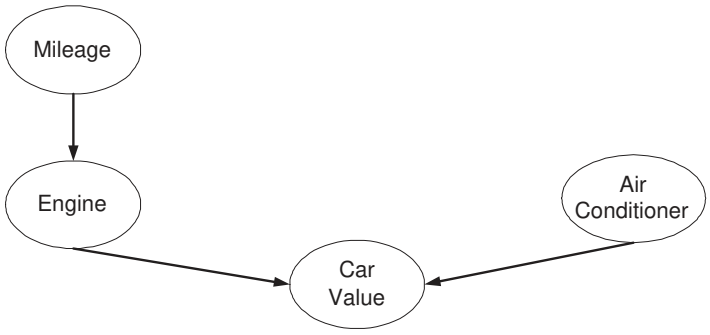


Figure 4.3. Bayesian belief network.

Table 4.3. Data set for Exercise 10.

Mileage	Engine	Air Conditioner	Number of Records with Car Value=Hi	Number of Records with Car Value=Lo
Hi	Good	Working	3	4
Hi	Good	Broken	1	2
Hi	Bad	Working	1	5
Hi	Bad	Broken	0	4
Lo	Good	Working	9	0
Lo	Good	Broken	5	1
Lo	Bad	Working	1	2
Lo	Bad	Broken	0	2

(a) Draw the probability table for each node in the network.
P(Mileage=Hi) = 0.5
P(Air Cond=Working) = 0.625
P(Engine=Good|Mileage=Hi) = 0.5
P(Engine=Good|Mileage=Lo) = 0.75
P(Value=High|Engine=Good, Air Cond=Working) = 0.750
P(Value=High|Engine=Good, Air Cond=Broken) = 0.667
P(Value=High|Engine=Bad, Air Cond=Working) = 0.222
P(Value=High|Engine=Bad, Air Cond=Broken) = 0

(b) Use the Bayesian network to compute P(Engine = Bad, Air Conditioner = Broken).

$$\begin{aligned} & P(Engine = Bad, Air\ Cond = Broken) \\ = & \sum_{\alpha\beta} P(Engine = Bad, Air\ Cond = Broken, Mileage = \alpha, Value = \beta) \\ = & \sum_{\alpha\beta} P(Value = \beta | Engine = Bad, Air\ Cond = Broken) \\ & \times P(Engine = Bad | Mileage = \alpha) P(Mileage = \alpha) P(Air\ Cond = Broken) \\ = & \ 0.1453. \end{aligned}$$

11. Given the Bayesian network shown in Figure 4.4, compute the following probabilities:

(a) P(B = good, F = empty, G = empty, S = yes).
Answer:

$$\begin{aligned} & P(B = good, F = empty, G = empty, S = yes) \\ = & \ P(B = good) \times P(F = empty) \times P(G = empty | B = good, F = empty) \\ & \times P(S = yes | B = good, F = empty) \\ = & \ 0.9 \times 0.2 \times 0.8 \times 0.2 = 0.0288. \end{aligned}$$

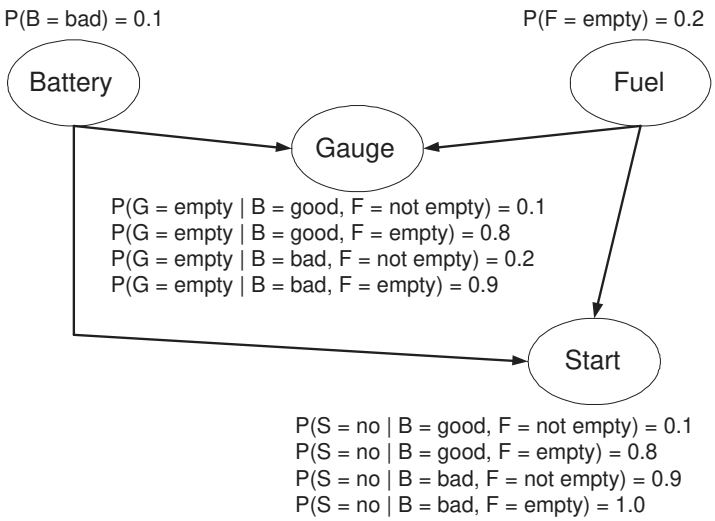


Figure 4.4. Bayesian belief network for Exercise 11.

(b) $P(\text{B = bad, F = empty, G = not empty, S = no})$.
Answer:

$$\begin{aligned} & P(B = bad, F = empty, G = not\ empty, S = no) \\ = \ & P(B = bad) \times P(F = empty) \times P(G = not\ empty | B = bad, F = empty) \\ & \times P(S = no | B = bad, F = empty) \\ = \ & 0.1 \times 0.2 \times 0.1 \times 1.0 = 0.002. \end{aligned}$$

(c) Given that the battery is bad, compute the probability that the car will start.
Answer:

$$\begin{aligned} & P(S = yes | B = bad) \\ = \ & \sum_{\alpha} P(S = yes | B = bad, F = \alpha) P(B = bad) P(F = \alpha) \\ = \ & 0.1 \times 0.1 \times 0.8 \\ = \ & 0.008 \end{aligned}$$

12. Consider the one-dimensional data set shown in Table 4.4.

(a) Classify the data point $x = 5.0$ according to its 1-, 3-, 5-, and 9-nearest neighbors (using majority vote).
Answer:
1-nearest neighbor: +,

Table 4.4. Data set for Exercise 12.

x	0.5	3.0	4.5	4.6	4.9	5.2	5.3	5.5	7.0	9.5
y	−	−	+	+	+	−	−	+	−	−

3-nearest neighbor: −,

5-nearest neighbor: +,

9-nearest neighbor: −.

(b) Repeat the previous analysis using the distance-weighted voting approach described in Section 4.3.1.

Answer:

1-nearest neighbor: +,

3-nearest neighbor: +,

5-nearest neighbor: +,

9-nearest neighbor: +.

13. The nearest-neighbor algorithm described in Section 4.3 can be extended to handle nominal attributes. A variant of the algorithm called PEBLS (Parallel Examplar-Based Learning System) by Cost and Salzberg [2] measures the distance between two values of a nominal attribute using the modified value difference metric (MVDM). Given a pair of nominal attribute values, V_1 and V_2, the distance between them is defined as follows:

$$d(V_1, V_2) = \sum_{i=1}^{k} \left| \frac{n_{i1}}{n_1} - \frac{n_{i2}}{n_2} \right|, \tag{4.2}$$

where n_{ij} is the number of examples from class i with attribute value V_j and n_j is the number of examples with attribute value V_j.

Consider the training set for the loan classification problem shown in Figure 4.9. Use the MVDM measure to compute the distance between every pair of attribute values for the `Home Owner` and `Marital Status` attributes.

Answer:

The training set shown in Figure 4.8 can be summarized for the `Home Owner` and `Marital Status` attributes as follows.

	Marital Status		
Class	Single	Married	Divorced
Yes	2	0	1
No	2	4	1

	Home Owner	
Class	Yes	No
Yes	0	3
No	3	4

d(Single, Married) = 1

d(Single, Divorced) = 0

d(Married, Divorced) = 1

d(Refund=Yes, Refund=No) = 6/7

14. For each of the Boolean functions given below, state whether the problem is linearly separable.

    (a) A AND B AND C
    Answer: Yes

    (b) NOT A AND B
    Answer: Yes

    (c) (A OR B) AND (A OR C)
    Answer: Yes

    (d) (A XOR B) AND (A OR B)
    Answer: No

15. (a) Demonstrate how the perceptron model can be used to represent the AND and OR functions between a pair of Boolean variables.
    Answer:
    Let x_1 and x_2 be a pair of Boolean variables and y be the output. For AND function, a possible perceptron model is:

$$y = sgn\Big[x_1 + x_2 - 1.5\Big].$$

    For OR function, a possible perceptron model is:

$$y = sgn\Big[x_1 + x_2 - 0.5\Big].$$

    (b) Comment on the disadvantage of using linear functions as activation functions for multilayer neural networks.
    Answer:
    Multilayer neural networks is useful for modeling nonlinear relationships between the input and output attributes. However, if linear functions are used as activation functions (instead of sigmoid or hyperbolic tangent function), the output is still a linear combination of its input attributes. Such a network is just as expressive as a perceptron.

16. You are asked to evaluate the performance of two classification models, M_1 and M_2. The test set you have chosen contains 26 binary attributes, labeled as A through Z.

    Table 4.5 shows the posterior probabilities obtained by applying the models to the test set. (Only the posterior probabilities for the positive class are shown). As this is a two-class problem, $P(-) = 1 - P(+)$ and $P(-|A, \ldots, Z) = 1 - P(+|A, \ldots, Z)$. Assume that we are mostly interested in detecting instances from the positive class.

Table 4.5. Posterior probabilities for Exercise 16.

Instance	True Class	$P(+\|A, \ldots, Z, M_1)$	$P(+\|A, \ldots, Z, M_2)$
1	+	0.73	0.61
2	+	0.69	0.03
3	−	0.44	0.68
4	−	0.55	0.31
5	+	0.67	0.45
6	+	0.47	0.09
7	−	0.08	0.38
8	−	0.15	0.05
9	+	0.45	0.01
10	−	0.35	0.04

(a) Plot the ROC curve for both M_1 and M_2. (You should plot them on the same graph.) Which model do you think is better? Explain your reasons.

Answer:

The ROC curve for $M1$ and $M2$ are shown in the Figure 4.5.

$M1$ is better, since its area under the ROC curve is larger than the area under ROC curve for $M2$.

(b) For model M_1, suppose you choose the cutoff threshold to be $t = 0.5$. In other words, any test instances whose posterior probability is greater than t will be classified as a positive example. Compute the precision, recall, and F-measure for the model at this threshold value.

When $t = 0.5$, the confusion matrix for $M1$ is shown below.

		+	-
Actual	+	3	2
	-	1	4

Precision $= 3/4 = 75\%$.

Recall $= 3/5 = 60\%$.

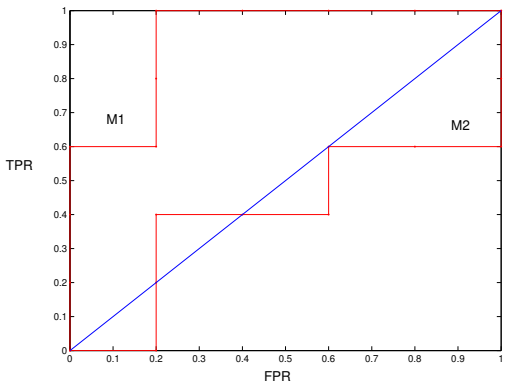


Figure 4.5. ROC curve.

F-measure $= (2 \times .75 \times .6)/(.75 + .6) = 0.667$.

(c) Repeat the analysis for part (c) using the same cutoff threshold on model M_2. Compare the F-measure results for both models. Which model is better? Are the results consistent with what you expect from the ROC curve?

Answer:

When $t = 0.5$, the confusion matrix for $M2$ is shown below.

		+	-
Actual	+	1	4
	-	1	4

Precision $= 1/2 = 50\%$.

Recall $= 1/5 = 20\%$.

F-measure $= (2 \times .5 \times .2)/(.5 + .2) = 0.2857$.

Based on F-measure, $M1$ is still better than $M2$. This result is consistent with the ROC plot.

(d) Repeat part (c) for model M_1 using the threshold $t = 0.1$. Which threshold do you prefer, $t = 0.5$ or $t = 0.1$? Are the results consistent with what you expect from the ROC curve?

Answer:

When $t = 0.1$, the confusion matrix for $M1$ is shown below.

		+	-
Actual	+	5	0
	-	4	1

Precision $= 5/9 = 55.6\%$.

Recall $= 5/5 = 100\%$.

F-measure $= (2 \times .556 \times 1)/(.556 + 1) = 0.715$.

According to F-measure, $t = 0.1$ is better than $t = 0.5$.

When $t = 0.1$, $FPR = 0.8$ and $TPR = 1$. On the other hand, when $t = 0.5$, $FPR = 0.2$ and $TRP = 0.6$. Since (0.2, 0.6) is closer to the point (0, 1), we favor $t = 0.5$. This result is inconsistent with the results using F-measure. We can also show this by computing the area under the ROC curve

For $t = 0.5$, area $= 0.6 \times (1 - 0.2) = 0.6 \times 0.8 = 0.48$.

For $t = 0.1$, area $= 1 \times (1 - 0.8) = 1 \times 0.2 = 0.2$.

Since the area for $t = 0.5$ is larger than the area for $t = 0.1$, we prefer $t = 0.5$.

17. Following is a data set that contains two attributes, X and Y, and two class labels, "+" and "−". Each attribute can take three different values: $0, 1$, or 2.

X	Y	Number of Instances	
		+	−
0	0	0	100
1	0	0	0
2	0	0	100
0	1	10	100
1	1	10	0
2	1	10	100
0	2	0	100
1	2	0	0
2	2	0	100

The concept for the "+" class is $Y = 1$ and the concept for the "−" class is $X = 0 \vee X = 2$.

(a) Build a decision tree on the data set. Does the tree capture the "+" and "−" concepts?

Answer:

There are 30 positive and 600 negative examples in the data. Therefore, at the root node, the error rate is

$$E_{orig} = 1 - \max(30/630, 600/630) = 30/630.$$

If we split on X, the gain in error rate is:

	$X=0$	$X=1$	$X=2$
+	10	10	10
−	300	0	300

$$E_{X=0} = 10/310$$
$$E_{X=1} = 0$$
$$E_{X=2} = 10/310$$

$$\Delta_X = E_{orig} - \frac{310}{630}\frac{10}{310} - \frac{10}{630}0 - \frac{310}{630}\frac{10}{310} = 10/630.$$

If we split on Y, the gain in error rate is:

	$Y=0$	$Y=1$	$Y=2$
+	0	30	0
−	200	200	200

$$E_{Y=0} = 0$$
$$E_{Y=1} = 30/230$$
$$E_{Y=2} = 0$$

$$\Delta_Y = E_{orig} - \frac{230}{630}\frac{30}{230} = 0.$$

Therefore, X is chosen to be the first splitting attribute. Since the $X = 1$ child node is pure, it does not require further splitting. We may use attribute Y to split the impure nodes, $X = 0$ and $X = 2$, as follows:

- The $Y = 0$ and $Y = 2$ nodes contain 100 − instances.
- The $Y = 1$ node contains 100 − and 10 + instances.

In all three cases for Y, the child nodes are labeled as −. The resulting concept is

$$\text{class} = \begin{cases} +, & \text{X = 1;} \\ -, & \text{otherwise.} \end{cases}$$

(b) What are the accuracy, precision, recall, and F_1-measure of the decision tree? (Note that precision, recall, and F_1-measure are defined with respect to the "+" class.)

Answer: The confusion matrix on the training data:

		Predicted +	Predicted −
Actual	+	10	20
	−	0	600

$$\text{accuracy} : \frac{610}{630} = 0.9683$$
$$\text{precision} : \frac{10}{10} = 1.0$$
$$\text{recall} : \frac{10}{30} = 0.3333$$
$$\text{F} - \text{measure} : \frac{2 * 0.3333 * 1.0}{1.0 + 0.3333} = 0.5$$

(c) Build a new decision tree with the following cost function:

$$C(i,j) = \begin{cases} 0, & \text{if } i = j; \\ 1, & \text{if } i = +,\ j = -; \\ \frac{\text{Number of } - \text{ instances}}{\text{Number of } + \text{ instances}}, & \text{if } i = -,\ j = +. \end{cases}$$

(Hint: only the leaves of the old decision tree need to be changed.) Does the decision tree capture the "+" concept?

Answer:

The cost matrix can be summarized as follows:

		Predicted +	Predicted −
Actual	+	0	600/30=20
	−	1	0

The decision tree in part (a) has 7 leaf nodes, $X = 1$, $X = 0 \wedge Y = 0$, $X = 0 \wedge Y = 1$, $X = 0 \wedge Y = 2$, $X = 2 \wedge Y = 0$, $X = 2 \wedge Y = 1$, and $X = 2 \wedge Y = 2$. Only $X = 0 \wedge Y = 1$ and $X = 2 \wedge Y = 1$ are impure nodes. The cost of misclassifying these impure nodes as positive class is:

$$10 * 0 + 1 * 100 = 100$$

while the cost of misclassifying them as negative class is:

$$10 * 20 + 0 * 100 = 200.$$

These nodes are therefore labeled as +.

The resulting concept is

$$\text{class} = \begin{cases} +, & X = 1 \vee (X = 0 \wedge Y = 1) \vee (X = 2 \wedge Y = 2); \\ -, & \text{otherwise.} \end{cases}$$

(d) What are the accuracy, precision, recall, and F_1-measure of the new decision tree?

Answer:

The confusion matrix of the new tree

		Predicted +	Predicted −
Actual	+	30	0
	−	200	400

$$\begin{aligned} \text{accuracy} &: \frac{430}{630} = 0.6825 \\ \text{precision} &: \frac{30}{230} = 0.1304 \\ \text{recall} &: \frac{30}{30} = 1.0 \\ \text{F} - \text{measure} &: \frac{2 * 0.1304 * 1.0}{1.0 + 0.1304} = 0.2307 \end{aligned}$$

18. Consider the task of building a classifier from random data, where the attribute values are generated randomly irrespective of the class labels. Assume the data set contains records from two classes, "+" and "−." Half of the data set is used for training while the remaining half is used for testing.

    (a) Suppose there are an equal number of positive and negative records in the data and the decision tree classifier predicts every test record to be positive. What is the expected error rate of the classifier on the test data?
    Answer: 50%.

    (b) Repeat the previous analysis assuming that the classifier predicts each test record to be positive class with probability 0.8 and negative class with probability 0.2.
    Answer: 50%.

    (c) Suppose two-thirds of the data belong to the positive class and the remaining one-third belong to the negative class. What is the expected error of a classifier that predicts every test record to be positive?
    Answer: 33%.

    (d) Repeat the previous analysis assuming that the classifier predicts each test record to be positive class with probability 2/3 and negative class with probability 1/3.
    Answer: 44.4%.

19. Derive the dual Lagrangian for the linear SVM with nonseparable data where the objective function is

$$f(\mathbf{w}) = \frac{\|\mathbf{w}\|^2}{2} + C\big(\sum_{i=1}^{N} \xi_i\big)^2.$$

    Answer:

$$L_D = \sum_{i=1}^{N} \lambda_i - \frac{1}{2} \sum_{i,j} \lambda_i \lambda_j y_i y_j \mathbf{x_i} \cdot \mathbf{x_j} - C\left(\sum_i \xi_i\right)^2.$$

    Notice that the dual Lagrangian depends on the slack variables ξ_i's.

20. Consider the XOR problem where there are four training points:

$$(1, 1, -), (1, 0, +), (0, 1, +), (0, 0, -).$$

    Transform the data into the following feature space:

$$\Phi = (1, \sqrt{2}x_1, \sqrt{2}x_2, \sqrt{2}x_1x_2, x_1^2, x_2^2).$$

    Find the maximum margin linear decision boundary in the transformed space.

    Answer:

    The decision boundary is $f(x_1, x_2) = x_1x_2$.

21. Given the data sets shown in Figures 4.6, explain how the decision tree, naïve Bayes, and k-nearest neighbor classifiers would perform on these data sets.

Answer:

(a) Both decision tree and NB will do well on this data set because the distinguishing attributes have better discriminating power than noise attributes in terms of entropy gain and conditional probability. k-NN will not do as well due to relatively large number of noise attributes.

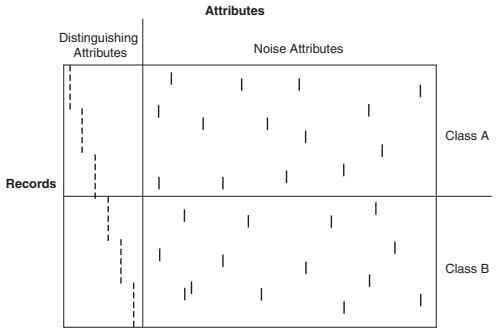
(b) NB will not work at all with this data set due to attribute dependency. Other schemes will do better than NB.

(c) NB will do very well in this data set, because each discriminating attribute has higher conditional probability in one class over the other and the overall classification is done by multiplying these individual conditional probabilities. Decision tree will not do as well, due to the relatively large number of distinguishing attributes. It will have an overfitting problem. k-NN will do reasonably well.

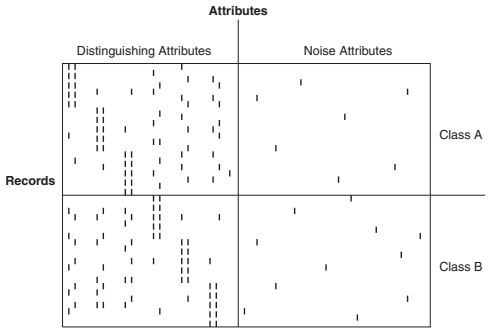
(d) k-NN will do well on this data set. Decision trees will also work, but will result in a fairly large decision tree. The first few splits will be quite random, because it may not find a good initial split at the beginning. NB will not perform quite as well due to the attribute dependency.

(e) k-NN will do well on this data set. Decision trees will also work, but will result in a large decision tree. If decision tree uses an oblique split instead of just vertical and horizontal splits, then the resulting decision tree will be more compact and highly accurate. NB will not perform quite as well due to attribute dependency.

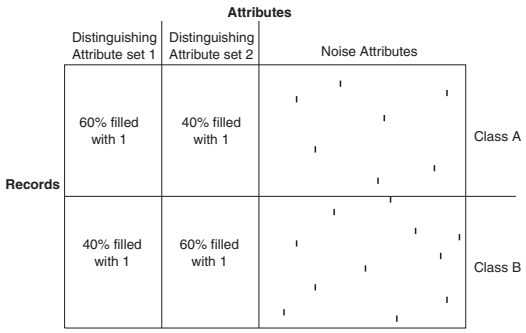
(f) kNN works the best. NB does not work well for this data set due to attribute dependency. Decision tree will have a large tree in order to capture the circular decision boundaries.



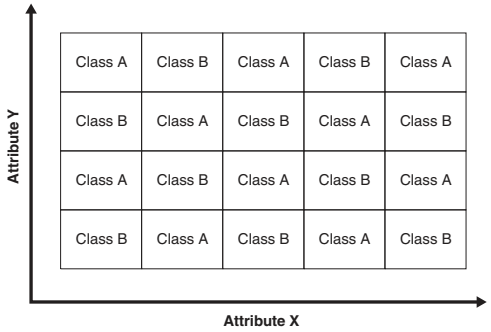
(a) Synthetic data set 1.



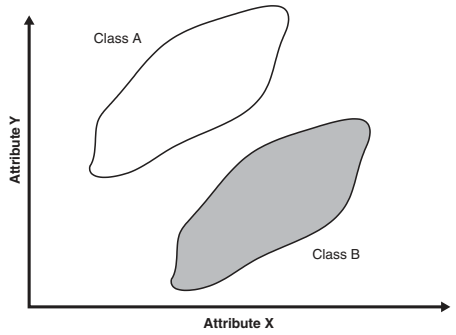
(b) Synthetic data set 2.



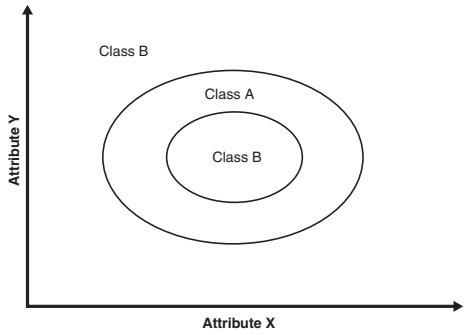
(c) Synthetic data set 3.



(d) Synthetic data set 4



(e) Synthetic data set 5.



(f) Synthetic data set 6.

Figure 4.6. Data set for Exercise 21.

5

Association Analysis: Basic Concepts and Algorithms

1. For each of the following questions, provide an example of an association rule from the market basket domain that satisfies the following conditions. Also, describe whether such rules are subjectively interesting.

   (a) A rule that has high support and high confidence.
   Answer: Milk $\longrightarrow$ Bread. Such obvious rule tends to be uninteresting.

   (b) A rule that has reasonably high support but low confidence.
   Answer: Milk $\longrightarrow$ Tuna. While the sale of tuna and milk may be higher than the support threshold, not all transactions that contain milk also contain tuna. Such low-confidence rule tends to be uninteresting.

   (c) A rule that has low support and low confidence.
   Answer: Cooking oil $\longrightarrow$ Laundry detergent. Such low confidence rule tends to be uninteresting.

   (d) A rule that has low support and high confidence.
   Answer: Vodka $\longrightarrow$ Caviar. Such rule tends to be interesting.

2. Consider the data set shown in Table 5.1.

   (a) Compute the support for itemsets $\{e\}$, $\{b, d\}$, and $\{b, d, e\}$ by treating each transaction ID as a market basket.
   Answer:

Table 5.1. Example of market basket transactions.

Customer ID	Transaction ID	Items Bought
1	0001	$\{a, d, e\}$
1	0024	$\{a, b, c, e\}$
2	0012	$\{a, b, d, e\}$
2	0031	$\{a, c, d, e\}$
3	0015	$\{b, c, e\}$
3	0022	$\{b, d, e\}$
4	0029	$\{c, d\}$
4	0040	$\{a, b, c\}$
5	0033	$\{a, d, e\}$
5	0038	$\{a, b, e\}$

$$\begin{aligned} s(\{e\}) &= \frac{8}{10} = 0.8 \\ s(\{b,d\}) &= \frac{2}{10} = 0.2 \\ s(\{b,d,e\}) &= \frac{2}{10} = 0.2 \end{aligned} \tag{5.1}$$

(b) Use the results in part (a) to compute the confidence for the association rules $\{b, d\} \longrightarrow \{e\}$ and $\{e\} \longrightarrow \{b, d\}$. Is confidence a symmetric measure?

Answer:

$$\begin{aligned} c(bd \longrightarrow e) &= \frac{0.2}{0.2} = 100\% \\ c(e \longrightarrow bd) &= \frac{0.2}{0.8} = 25\% \end{aligned}$$

No, confidence is not a symmetric measure.

(c) Repeat part (a) by treating each customer ID as a market basket. Each item should be treated as a binary variable (1 if an item appears in at least one transaction bought by the customer, and 0 otherwise.)

Answer:

$$\begin{aligned} s(\{e\}) &= \frac{4}{5} = 0.8 \\ s(\{b,d\}) &= \frac{5}{5} = 1 \\ s(\{b,d,e\}) &= \frac{4}{5} = 0.8 \end{aligned}$$

(d) Use the results in part (c) to compute the confidence for the association rules $\{b, d\} \longrightarrow \{e\}$ and $\{e\} \longrightarrow \{b, d\}$.

Answer:

$$c(bd \longrightarrow e) = \frac{0.8}{1} = 80\%$$
$$c(e \longrightarrow bd) = \frac{0.8}{0.8} = 100\%$$

(e) Suppose s_1 and c_1 are the support and confidence values of an association rule r when treating each transaction ID as a market basket. Also, let s_2 and c_2 be the support and confidence values of r when treating each customer ID as a market basket. Discuss whether there are any relationships between s_1 and s_2 or c_1 and c_2.

Answer:

There are no apparent relationships between s_1, s_2, c_1, and c_2.

3. (a) What is the confidence for the rules $\emptyset \longrightarrow A$ and $A \longrightarrow \emptyset$?

Answer:

$c(\emptyset \longrightarrow A) = s(\emptyset \longrightarrow A)$.

$c(A \longrightarrow \emptyset) = 100\%$.

(b) Let c_1, c_2, and c_3 be the confidence values of the rules $\{p\} \longrightarrow \{q\}$, $\{p\} \longrightarrow \{q, r\}$, and $\{p, r\} \longrightarrow \{q\}$, respectively. If we assume that c_1, c_2, and c_3 have different values, what are the possible relationships that may exist among c_1, c_2, and c_3? Which rule has the lowest confidence?

Answer:

$c_1 = \frac{s(p \cup q)}{s(p)}$

$c_2 = \frac{s(p \cup q \cup r)}{s(p)}$

$c_3 = \frac{s(p \cup q \cup r)}{s(p \cup r)}$

Considering $s(p) \geq s(p \cup q) \geq s(p \cup q \cup r)$

Thus: $c1 \geq c2$ & $c3 \geq c2$.

Therefore $c2$ has the lowest confidence.

(c) Repeat the analysis in part (b) assuming that the rules have identical support. Which rule has the highest confidence?

Answer:

Considering $s(p \cup q) = s(p \cup q \cup r)$

but $s(p) \geq s(p \cup r)$

Thus: $c3 \geq (c1 = c2)$

Either all rules have the same confidence or c3 has the highest confidence.

(d) Transitivity: Suppose the confidence of the rules $A \longrightarrow B$ and $B \longrightarrow C$ are larger than some threshold, $minconf$. Is it possible that $A \longrightarrow C$ has a confidence less than $minconf$?

Answer:

Yes, It depends on the support of items A, B, and C.

For example:

s(A,B) = 60% s(A) = 90%
s(A,C) = 20% s(B) = 70%
s(B,C) = 50% s(C) = 60%

Let $minconf = 50\%$ Therefore:

$c(A \rightarrow B) = 66\% > minconf$
$c(B \rightarrow C) = 71\% > minconf$
But $c(A \rightarrow C) = 22\% < minconf$

4. For each of the following measures, determine whether it is monotone, anti-monotone, or non-monotone (i.e., neither monotone nor anti-monotone).

Example: Support, $s = \frac{\sigma(X)}{|T|}$ is anti-monotone because $s(X) \geq s(Y)$ whenever $X \subset Y$.

(a) A characteristic rule is a rule of the form $\{p\} \longrightarrow \{q_1, q_2, \ldots, q_n\}$, where the rule antecedent contains only a single item. An itemset of size k can produce up to k characteristic rules. Let ζ be the minimum confidence of all characteristic rules generated from a given itemset:

$$\begin{aligned} \zeta(\{p_1, p_2, \ldots, p_k\}) &= \min\big[\, c\big(\{p_1\} \longrightarrow \{p_2, p_3, \ldots, p_k\}\big), \ldots \\ &\qquad c\big(\{p_k\} \longrightarrow \{p_1, p_3 \ldots, p_{k-1}\}\big)\,\big] \end{aligned}$$

Is ζ monotone, anti-monotone, or non-monotone?

Answer:

ζ is an anti-monotone measure because

$$\zeta(\{A_1, A_2, \cdots, A_k\}) \geq \zeta(\{A_1, A_2, \cdots, A_k, A_{k+1}\}) \tag{5.2}$$

For example, we can compare the values of ζ for $\{A, B\}$ and $\{A, B, C\}$.

$$\begin{aligned} \zeta(\{A, B\}) &= \min\big(c(A \longrightarrow B), c(B \longrightarrow A)\big) \\ &= \min\big(\frac{s(A,B)}{s(A)}, \frac{s(A,B)}{s(B)}\big) \\ &= \frac{s(A,B)}{\max(s(A), s(B))} \end{aligned} \tag{5.3}$$

$$\begin{aligned}\zeta(\{A,B,C\}) &= \min\big(c(A \longrightarrow BC), c(B \longrightarrow AC), c(C \longrightarrow AB)\big) \\ &= \min\big(\frac{s(A,B,C)}{s(A)}, \frac{s(A,B,C)}{s(B)}, \frac{s(A,B,C)}{s(C)}\big) \\ &= \frac{s(A,B,C)}{\max(s(A), s(B), s(C))} \qquad (5.4)\end{aligned}$$

Since $s(A,B,C) \leq s(A,B)$ and $\max(s(A), s(B), s(C)) \geq \max(s(A), s(B))$, therefore $\zeta(\{A,B\}) \geq \zeta(\{A,B,C\})$.

(b) A discriminant rule is a rule of the form $\{p_1, p_2, \ldots, p_n\} \longrightarrow \{q\}$, where the rule consequent contains only a single item. An itemset of size k can produce up to k discriminant rules. Let η be the minimum confidence of all discriminant rules generated from a given itemset:

$$\begin{aligned}\eta(\{p_1, p_2, \ldots, p_k\}) = \min\big[\, & c\big(\{p_2, p_3, \ldots, p_k\} \longrightarrow \{p_1\}\big), \ldots \\ & c\big(\{p_1, p_2, \ldots p_{k-1}\} \longrightarrow \{p_k\}\big)\,\big]\end{aligned}$$

Is η monotone, anti-monotone, or non-monotone?

Answer:

η is non-monotone. We can show this by comparing $\eta(\{A,B\})$ against $\eta(\{A,B,C\})$.

$$\begin{aligned}\eta(\{A,B\}) &= \min\big(c(A \longrightarrow B), c(B \longrightarrow A)\big) \\ &= \min\big(\frac{s(A,B)}{s(A)}, \frac{s(A,B)}{s(B)}\big) \\ &= \frac{s(A,B)}{\max(s(A), s(B))} \qquad (5.5)\end{aligned}$$

$$\begin{aligned}\eta(\{A,B,C\}) &= \min\big(c(AB \longrightarrow C), c(AC \longrightarrow B), c(BC \longrightarrow A)\big) \\ &= \min\big(\frac{s(A,B,C)}{s(A,B)}, \frac{s(A,B,C)}{s(A,C)}, \frac{s(A,B,C)}{s(B,C)}\big) \\ &= \frac{s(A,B,C)}{\max(s(A,B), s(A,C), s(B,C))} \qquad (5.6)\end{aligned}$$

Since $s(A,B,C) \leq s(A,B)$ and $\max(s(A,B), s(A,C), s(B,C)) \leq \max(s(A), s(B))$, therefore $\eta(\{A,B,C\})$ can be greater than or less than $\eta(\{A,B\})$. Hence, the measure is non-monotone.

(c) Repeat the analysis in parts (a) and (b) by replacing the min function with a max function.

Answer:

Let

$$\zeta'(\{A_1, A_2, \cdots, A_k\}) = \max(\quad c(A_1 \longrightarrow A_2, A_3, \cdots, A_k), \cdots \\ c(A_k \longrightarrow A_1, A_3 \cdots, A_{k-1}))$$

$$\begin{aligned} \zeta'(\{A, B\}) &= \max\big(c(A \longrightarrow B), c(B \longrightarrow A)\big) \\ &= \max\big(\frac{s(A,B)}{s(A)}, \frac{s(A,B)}{s(B)}\big) \\ &= \frac{s(A,B)}{\min(s(A), s(B))} \end{aligned} \tag{5.7}$$

$$\begin{aligned} \zeta'(\{A, B, C\}) &= \max\big(c(A \longrightarrow BC), c(B \longrightarrow AC), c(C \longrightarrow AB)\big) \\ &= \max\big(\frac{s(A,B,C)}{s(A)}, \frac{s(A,B,C)}{s(B)}, \frac{s(A,B,C)}{s(C)}\big) \\ &= \frac{s(A,B,C)}{\min(s(A), s(B), s(C))} \end{aligned} \tag{5.8}$$

Since $s(A, B, C) \leq s(A, B)$ and $\min(s(A), s(B), s(C)) \leq \min(s(A), s(B))$, $\zeta'(\{A, B, C\})$ can be greater than or less than $\zeta'(\{A, B\})$. Therefore, the measure is non-monotone.

Let

$$\eta'(\{A_1, A_2, \cdots, A_k\}) = \max(\quad c(A_2, A_3, \cdots, A_k \longrightarrow A_1), \cdots \\ c(A_1, A_2, \cdots A_{k-1} \longrightarrow A_k))$$

$$\begin{aligned} \eta'(\{A, B\}) &= \max\big(c(A \longrightarrow B), c(B \longrightarrow A)\big) \\ &= \max\big(\frac{s(A,B)}{s(A)}, \frac{s(A,B)}{s(B)}\big) \\ &= \frac{s(A,B)}{\min(s(A), s(B))} \end{aligned} \tag{5.9}$$

$$\begin{aligned} \eta(\{A, B, C\}) &= \max\big(c(AB \longrightarrow C), c(AC \longrightarrow B), c(BC \longrightarrow A)\big) \\ &= \max\big(\frac{s(A,B,C)}{s(A,B)}, \frac{s(A,B,C)}{s(A,C)}, \frac{s(A,B,C)}{s(B,C)}\big) \\ &= \frac{s(A,B,C)}{\min(s(A,B), s(A,C), s(B,C))} \end{aligned} \tag{5.10}$$

Since $s(A, B, C) \leq s(A, B)$ and $\min(s(A, B), s(A, C), s(B, C)) \leq \min(s(A), s(B), s(C))$ $\leq \min(s(A), s(B))$, $\eta'(\{A, B, C\})$ can be greater than or less than $\eta'(\{A, B\})$. Hence, the measure is non-monotone.

5. Prove Equation 5.3. (Hint: First, count the number of ways to create an itemset that forms the left hand side of the rule. Next, for each size k itemset selected for the left-hand side, count the number of ways to choose the remaining $d - k$ items to form the right-hand side of the rule.)

Answer:

Suppose there are d items. We first choose k of the items to form the left-hand side of the rule. There are $\binom{d}{k}$ ways for doing this. After selecting the items for the left-hand side, there are $\binom{d-k}{i}$ ways to choose the remaining items to form the right hand side of the rule, where $1 \leq i \leq d - k$. Therefore the total number of rules (R) is:

$$\begin{aligned} R &= \sum_{k=1}^{d} \binom{d}{k} \sum_{i=1}^{d-k} \binom{d-k}{i} \\ &= \sum_{k=1}^{d} \binom{d}{k} \left(2^{d-k} - 1\right) \\ &= \sum_{k=1}^{d} \binom{d}{k} 2^{d-k} - \sum_{k=1}^{d} \binom{d}{k} \\ &= \sum_{k=1}^{d} \binom{d}{k} 2^{d-k} - \left[2^d + 1\right], \end{aligned}$$

where

$$\sum_{i=1}^{n} \binom{n}{i} = 2^n - 1.$$

Since

$$(1 + x)^d = \sum_{i=1}^{d} \binom{d}{i} x^{d-i} + x^d,$$

substituting $x = 2$ leads to:

$$3^d = \sum_{i=1}^{d} \binom{d}{i} 2^{d-i} + 2^d.$$

Therefore, the total number of rules is:

$$R = 3^d - 2^d - \left[2^d + 1\right] = 3^d - 2^{d+1} + 1.$$

Table 5.2. Market basket transactions.

Transaction ID	Items Bought
1	{Milk, Beer, Diapers}
2	{Bread, Butter, Milk}
3	{Milk, Diapers, Cookies}
4	{Bread, Butter, Cookies}
5	{Beer, Cookies, Diapers}
6	{Milk, Diapers, Bread, Butter}
7	{Bread, Butter, Diapers}
8	{Beer, Diapers}
9	{Milk, Diapers, Bread, Butter}
10	{Beer, Cookies}

6. Consider the market basket transactions shown in Table 5.2.

   (a) What is the maximum number of association rules that can be extracted from this data (including rules that have zero support)?

   Answer: There are six items in the data set. Therefore the total number of rules is 602.

   (b) What is the maximum size of frequent itemsets that can be extracted (assuming $minsup > 0$)?

   Answer: Because the longest transaction contains 4 items, the maximum size of frequent itemset is 4.

   (c) Write an expression for the maximum number of size-3 itemsets that can be derived from this data set.

   Answer: $\binom{6}{3} = 20$.

   (d) Find an itemset (of size 2 or larger) that has the largest support.

   Answer: {Bread, Butter}.

   (e) Find a pair of items, a and b, such that the rules $\{a\} \longrightarrow \{b\}$ and $\{b\} \longrightarrow \{a\}$ have the same confidence.

   Answer: (Beer, Cookies) or (Bread, Butter).

7. Show that if a candidate k-itemset X has a subset of size less than $k-1$ that is infrequent, then at least one of the $(k-1)$-size subsets of X is necessarily infrequent.

   Answer: We are given a k-itemset $X = \{x_1, x_2, \ldots, x_k\}$. Let the subset of X that is infrequent and of size less than $k-1$ be denoted by $Y = \{x_{y_1}, x_{y_2}, \ldots, x_{y_m}\}$, where $m < k-1$. The set of items in X that are not part of Y can then be represented as $Z = \{x_{z_1}, x_{z_2}, \ldots, x_{z_{k-m}}\}$, where Z contains at least one item (note that $X = Y \cup Z$). Using the anti-monotone property of the support measure, if Y is infrequent then any superset of Y will also

be infrequent. Hence, the $(k-1)$-size itemset, $V = Y \cup Z \setminus \{x_{z_1}\}$ will be infrequent since it is a superset of Y by construction. Further, note that V is a subset of X of size $k-1$. This proves that there is at least one $(k-1)$-size subset of X that is infrequent if any subset of X of size less than $k-1$ is infrequent. Because of this result, we only need to check the $(k-1)$-subsets of every candidate k-itemset during candidate pruning.

8. Consider the following set of frequent 3-itemsets:

$$\{1,2,3\}, \{1,2,4\}, \{1,2,5\}, \{1,3,4\}, \{1,3,5\}, \{2,3,4\}, \{2,3,5\}, \{3,4,5\}.$$

Assume that there are only five items in the data set.

(a) List all candidate 4-itemsets obtained by a candidate generation procedure using the $F_{k-1} \times F_1$ merging strategy.

Answer:

$\{1,2,3,4\}$,$\{1,2,3,5\}$,$\{1,2,3,6\}$.
$\{1,2,4,5\}$,$\{1,2,4,6\}$,$\{1,2,5,6\}$.
$\{1,3,4,5\}$,$\{1,3,4,6\}$,$\{2,3,4,5\}$.
$\{2,3,4,6\}$,$\{2,3,5,6\}$.

(b) List all candidate 4-itemsets obtained by the candidate generation procedure in *Apriori*.

Answer:

$\{1,2,3,4\}$, $\{1,2,3,5\}$, $\{1,2,4,5\}$, $\{2,3,4,5\}$, $\{2,3,4,6\}$.

(c) List all candidate 4-itemsets that survive the candidate pruning step of the *Apriori* algorithm.

Answer:

$\{1,2,3,4\}$

9. The *Apriori* algorithm uses a generate-and-count strategy for deriving frequent itemsets. Candidate itemsets of size $k+1$ are created by joining a pair of frequent itemsets of size k (this is known as the candidate generation step). A candidate is discarded if any one of its subsets is found to be infrequent during the candidate pruning step. Suppose the *Apriori* algorithm is applied to the data set shown in Table 5.3 with $minsup = 30\%$, i.e., any itemset occurring in less than 3 transactions is considered to be infrequent.

(a) Draw an itemset lattice representing the data set given in Table 5.3. Label each node in the lattice with the following letter(s):

- **N**: If the itemset is not considered to be a candidate itemset by the *Apriori* algorithm. There are two reasons for an itemset not to be considered as a candidate itemset: (1) it is not generated at all

Table 5.3. Example of market basket transactions.

Transaction ID	Items Bought
1	$\{a, b, d, e\}$
2	$\{b, c, d\}$
3	$\{a, b, d, e\}$
4	$\{a, c, d, e\}$
5	$\{b, c, d, e\}$
6	$\{b, d, e\}$
7	$\{c, d\}$
8	$\{a, b, c\}$
9	$\{a, d, e\}$
10	$\{b, d\}$

during the candidate generation step, or (2) it is generated during the candidate generation step but is subsequently removed during the candidate pruning step because one of its subsets is found to be infrequent.

- **F**: If the candidate itemset is found to be frequent by the *Apriori* algorithm.
- **I**: If the candidate itemset is found to be infrequent after support counting.

Answer:

The lattice structure is shown below.

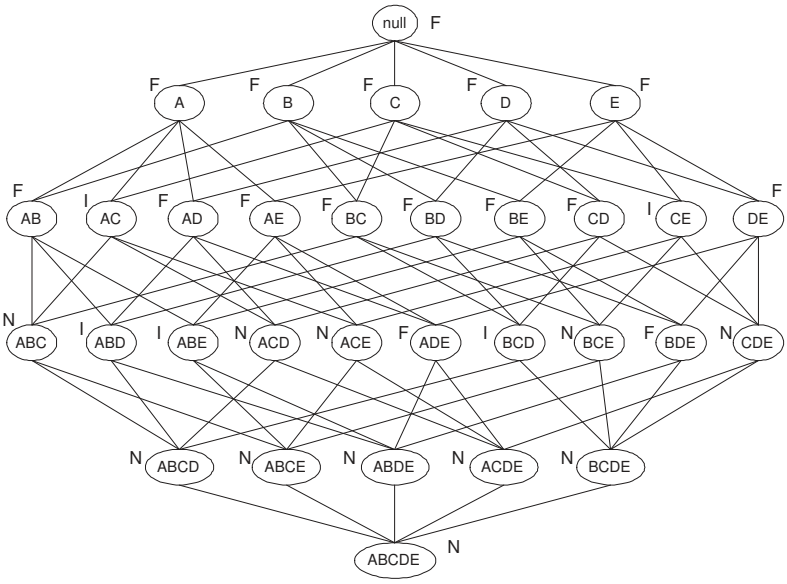


Figure 5.1. Solution.

(b) What is the percentage of frequent itemsets (with respect to all itemsets in the lattice)?

Answer:

Percentage of frequent itemsets = 16/32 = 50.0% (including the null set).

(c) What is the pruning ratio of the *Apriori* algorithm on this data set? (Pruning ratio is defined as the percentage of itemsets not considered to be a candidate because (1) they are not generated during candidate generation or (2) they are pruned during the candidate pruning step.)

Answer:

Pruning ratio is the ratio of N to the total number of itemsets. Since the count of $N = 11$, therefore pruning ratio is $11/32 = 34.4\%$.

(d) What is the false alarm rate (i.e, percentage of candidate itemsets that are found to be infrequent after performing support counting)?

Answer:

False alarm rate is the ratio of I to the total number of itemsets. Since the count of $I = 5$, therefore the false alarm rate is $5/32 = 15.6\%$.

10. The *Apriori* algorithm uses a hash tree data structure to efficiently count the support of candidate itemsets. Consider the hash tree for candidate 3-itemsets shown in Figure 5.2.

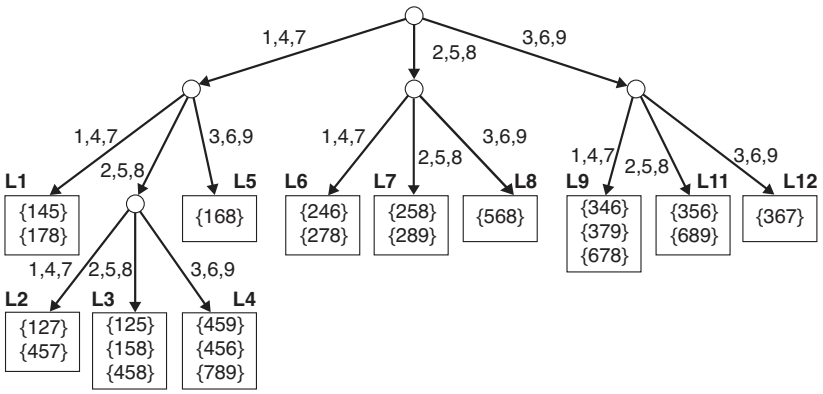


Figure 5.2. An example of a hash tree structure.

(a) Given a transaction that contains items $\{1, 3, 4, 5, 8\}$, which of the hash tree leaf nodes will be visited when finding the candidates of the transaction?

Answer:

The leaf nodes visited are L1, L3, L5, L9, and L11.

(b) Use the visited leaf nodes in part (b) to determine the candidate itemsets that are contained in the transaction $\{1, 3, 4, 5, 8\}$.

Answer:

The candidates contained in the transaction are $\{1, 4, 5\}$, $\{1, 5, 8\}$, and $\{4, 5, 8\}$.

11. Consider the following set of candidate 3-itemsets:

$$\{1, 2, 3\}, \{1, 2, 6\}, \{1, 3, 4\}, \{2, 3, 4\}, \{2, 4, 5\}, \{3, 4, 6\}, \{4, 5, 6\}$$

(a) Construct a hash tree for the above candidate 3-itemsets. Assume the tree uses a hash function where all odd-numbered items are hashed to the left child of a node, while the even-numbered items are hashed to the right child. A candidate k-itemset is inserted into the tree by hashing on each successive item in the candidate and then following the appropriate branch of the tree according to the hash value. Once a leaf node is reached, the candidate is inserted based on one of the following conditions:

Condition 1: If the depth of the leaf node is equal to k (the root is assumed to be at depth 0), then the candidate is inserted regardless of the number of itemsets already stored at the node.

Condition 2: If the depth of the leaf node is less than k, then the candidate can be inserted as long as the number of itemsets stored at the node is less than $maxsize$. Assume $maxsize = 2$ for this question.

Condition 3: If the depth of the leaf node is less than k and the number of itemsets stored at the node is equal to $maxsize$, then the leaf node is converted into an internal node. New leaf nodes are created as children of the old leaf node. Candidate itemsets previously stored in the old leaf node are distributed to the children based on their hash values. The new candidate is also hashed to its appropriate leaf node.

Answer:

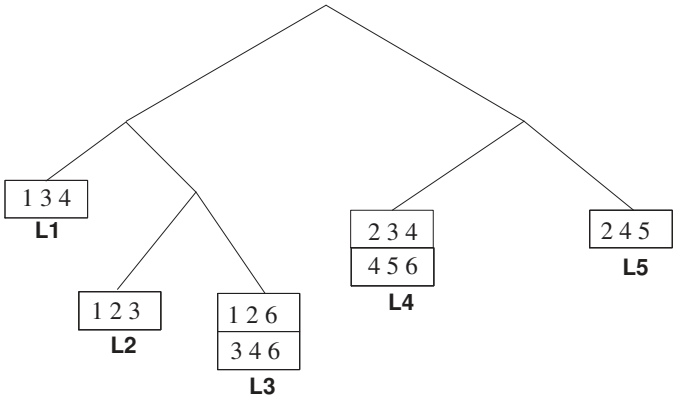


Figure 5.3. Hash tree for Exercise 11.

(b) How many leaf nodes are there in the candidate hash tree? How many internal nodes are there?

Answer: There are 5 leaf nodes and 4 internal nodes.

(c) Consider a transaction that contains the following items: $\{1, 2, 3, 5, 6\}$. Using the hash tree constructed in part (a), which leaf nodes will be checked against the transaction? What are the candidate 3-itemsets contained in the transaction?

Answer: The leaf nodes L1, L2, L3, and L4 will be checked against the transaction. The candidate itemsets contained in the transaction include {1,2,3} and {1,2,6}.

12. Given the lattice structure shown in Figure 5.4 and the transactions given in Table 5.3, label each node with the following letter(s):

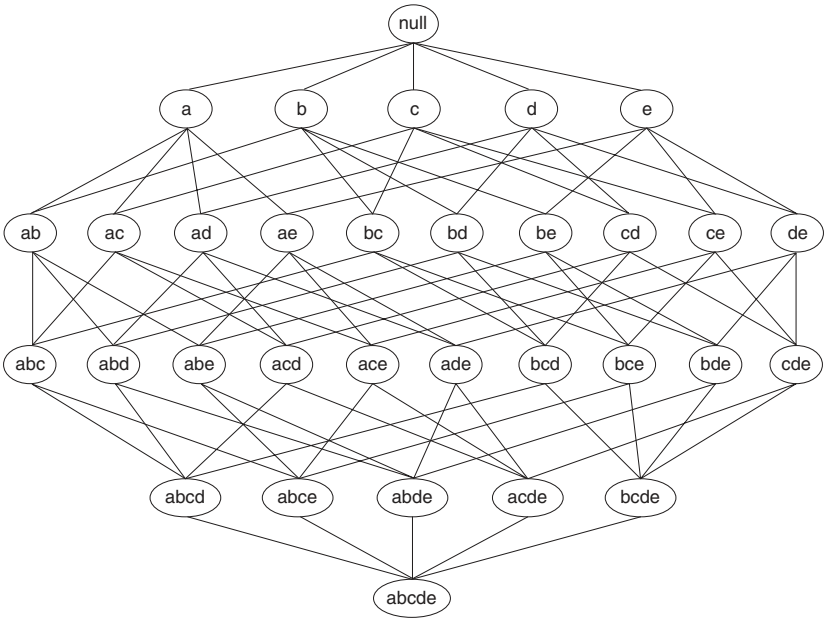


Figure 5.4. An itemset lattice

- M if the node is a maximal frequent itemset,
- C if it is a closed frequent itemset,
- N if it is frequent but neither maximal nor closed, and
- I if it is infrequent.

Assume that the support threshold is equal to 30%.

Answer:

The lattice structure is shown below.

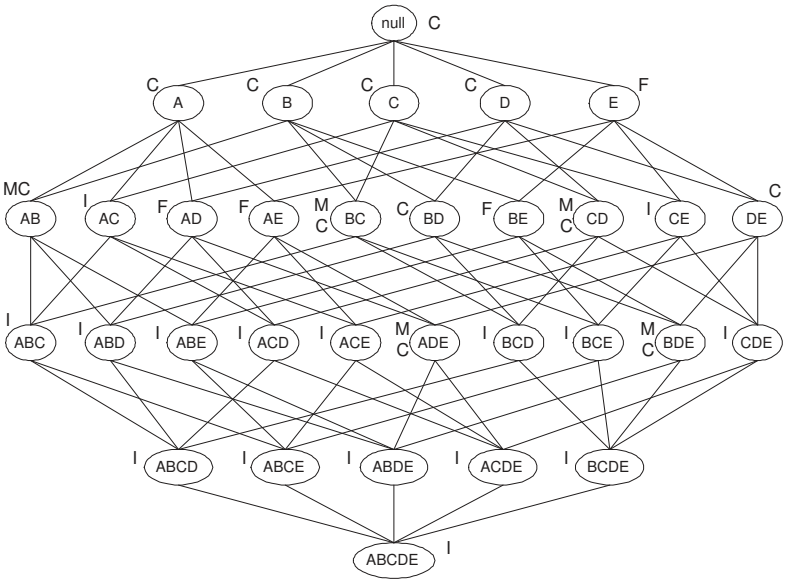


Figure 5.5. Solution for Exercise 12.

13. The original association rule mining formulation uses the support and confidence measures to prune uninteresting rules.

    (a) Draw a contingency table for each of the following rules using the transactions shown in Table 5.4.

Table 5.4. Example of market basket transactions.

Transaction ID	Items Bought
1	$\{a, b, d, e\}$
2	$\{b, c, d\}$
3	$\{a, b, d, e\}$
4	$\{a, c, d, e\}$
5	$\{b, c, d, e\}$
6	$\{b, d, e\}$
7	$\{c, d\}$
8	$\{a, b, c\}$
9	$\{a, d, e\}$
10	$\{b, d\}$

Rules: $\{b\} \longrightarrow \{c\}$, $\{a\} \longrightarrow \{d\}$, $\{b\} \longrightarrow \{d\}$, $\{e\} \longrightarrow \{c\}$, $\{c\} \longrightarrow \{a\}$.

Answer:

	c	$\overline{c}$
b	3	4
$\overline{b}$	2	1

	d	$\overline{d}$
a	4	1
$\overline{a}$	5	0

	d	$\overline{d}$
b	6	1
$\overline{b}$	3	0

	c	$\overline{c}$
e	2	4
$\overline{e}$	3	1

	a	$\overline{a}$
c	2	3
$\overline{c}$	3	2

(b) Use the contingency tables in part (a) to compute and rank the rules in decreasing order according to the following measures.

i. Support.
Answer:

Rules	Support	Rank
$b \longrightarrow c$	0.3	3
$a \longrightarrow d$	0.4	2
$b \longrightarrow d$	0.6	1
$e \longrightarrow c$	0.2	4
$c \longrightarrow a$	0.2	4

ii. Confidence.
Answer:

Rules	Confidence	Rank
$b \longrightarrow c$	3/7	3
$a \longrightarrow d$	4/5	2
$b \longrightarrow d$	6/7	1
$e \longrightarrow c$	2/6	5
$c \longrightarrow a$	2/5	4

iii. Interest$(X \longrightarrow Y) = \frac{P(X,Y)}{P(X)}P(Y)$.
Answer:

Rules	Interest	Rank
$b \longrightarrow c$	0.214	3
$a \longrightarrow d$	0.72	2
$b \longrightarrow d$	0.771	1
$e \longrightarrow c$	0.167	5
$c \longrightarrow a$	0.2	4

iv. IS$(X \longrightarrow Y) = \frac{P(X,Y)}{\sqrt{P(X)P(Y)}}$.
Answer:

Rules	IS	Rank
$b \longrightarrow c$	0.507	3
$a \longrightarrow d$	0.596	2
$b \longrightarrow d$	0.756	1
$e \longrightarrow c$	0.365	5
$c \longrightarrow a$	0.4	4

v. Klosgen$(X \longrightarrow Y) = \sqrt{P(X,Y)} \times (P(Y|X) - P(Y))$, where $P(Y|X) = \frac{P(X,Y)}{P(X)}$.

Answer:

Rules	Klosgen	Rank
$b \longrightarrow c$	-0.039	2
$a \longrightarrow d$	-0.063	4
$b \longrightarrow d$	-0.033	1
$e \longrightarrow c$	-0.075	5
$c \longrightarrow a$	-0.045	3

vi. Odds ratio$(X \longrightarrow Y) = \frac{P(X,Y)P(\overline{X},\overline{Y})}{P(X,\overline{Y})P(\overline{X},Y)}$.

Answer:

Rules	Odds Ratio	Rank
$b \longrightarrow c$	0.375	2
$a \longrightarrow d$	0	4
$b \longrightarrow d$	0	4
$e \longrightarrow c$	0.167	3
$c \longrightarrow a$	0.444	1

14. Given the rankings you had obtained in Exercise 13, compute the correlation between the rankings of confidence and the other five measures. Which measure is most highly correlated with confidence? Which measure is least correlated with confidence?

Answer:

Correlation(Confidence, Support) = 0.97.

Correlation(Confidence, Interest) = 1.

Correlation(Confidence, IS) = 1.

Correlation(Confidence, Klosgen) = 0.7.

Correlation(Confidence, Odds Ratio) = -0.606.

Interest and IS are the most highly correlated with confidence, while odds ratio is the least correlated.

15. Answer the following questions using the data sets shown in Figure 5.6. Note that each data set contains 1000 items and 10,000 transactions. Dark cells indicate the presence of items and white cells indicate the absence of

items. We will apply the *Apriori* algorithm to extract frequent itemsets with $minsup = 10\%$ (i.e., itemsets must be contained in at least 1000 transactions)?

(a) Which data set(s) will produce the most number of frequent itemsets?
Answer: Data set (e) because it has to generate the longest frequent itemset along with its subsets.

(b) Which data set(s) will produce the fewest number of frequent itemsets?
Answer: Data set (d) which does not produce any frequent itemsets at 10% support threshold.

(c) Which data set(s) will produce the longest frequent itemset?
Answer: Data set (e).

(d) Which data set(s) will produce frequent itemsets with highest maximum support?
Answer: Data set (b).

(e) Which data set(s) will produce frequent itemsets containing items with wide-varying support levels (i.e., items with mixed support, ranging from less than 20% to more than 70%).
Answer: Data set (e).

16. (a) Prove that the ϕ coefficient is equal to 1 if and only if $f_{11} = f_{1+} = f_{+1}$.
Answer:
Instead of proving $f_{11} = f_{1+} = f_{+1}$, we will show that $P(A,B) = P(A) = P(B)$, where $P(A,B) = f_{11}/N$, $P(A) = f_{1+}/N$, and $P(B) = f_{+1}/N$. When the ϕ-coefficient equals to 1:

$$\phi = \frac{P(A,B) - P(A)P(B)}{\sqrt{P(A)P(B)\big[1-P(A)\big]\big[1-P(B)\big]}} = 1$$

The preceding equation can be simplified as follows:

$$\begin{aligned}
\Big[P(A,B) - P(A)P(B)\Big]^2 &= P(A)P(B)\big[1-P(A)\big]\big[1-P(B)\big] \\
P(A,B)^2 - 2P(A,B)P(A)P(B) &= P(A)P(B)\big[1-P(A)-P(B)\big] \\
P(A,B)^2 &= P(A)P(B)\big[1-P(A)-P(B)+2P(A,B)\big]
\end{aligned}$$

We may rewrite the equation in terms of $P(B)$ as follows:

$$P(A)P(B)^2 - P(A)\big[1-P(A)+2P(A,B)\big]P(B) + P(A,B)^2 = 0$$

The solution to the quadratic equation in $P(B)$ is:

$$P(B) = \frac{P(A)\beta - \sqrt{P(A)^2\beta^2 - 4P(A)P(A,B)^2}}{2P(A)},$$

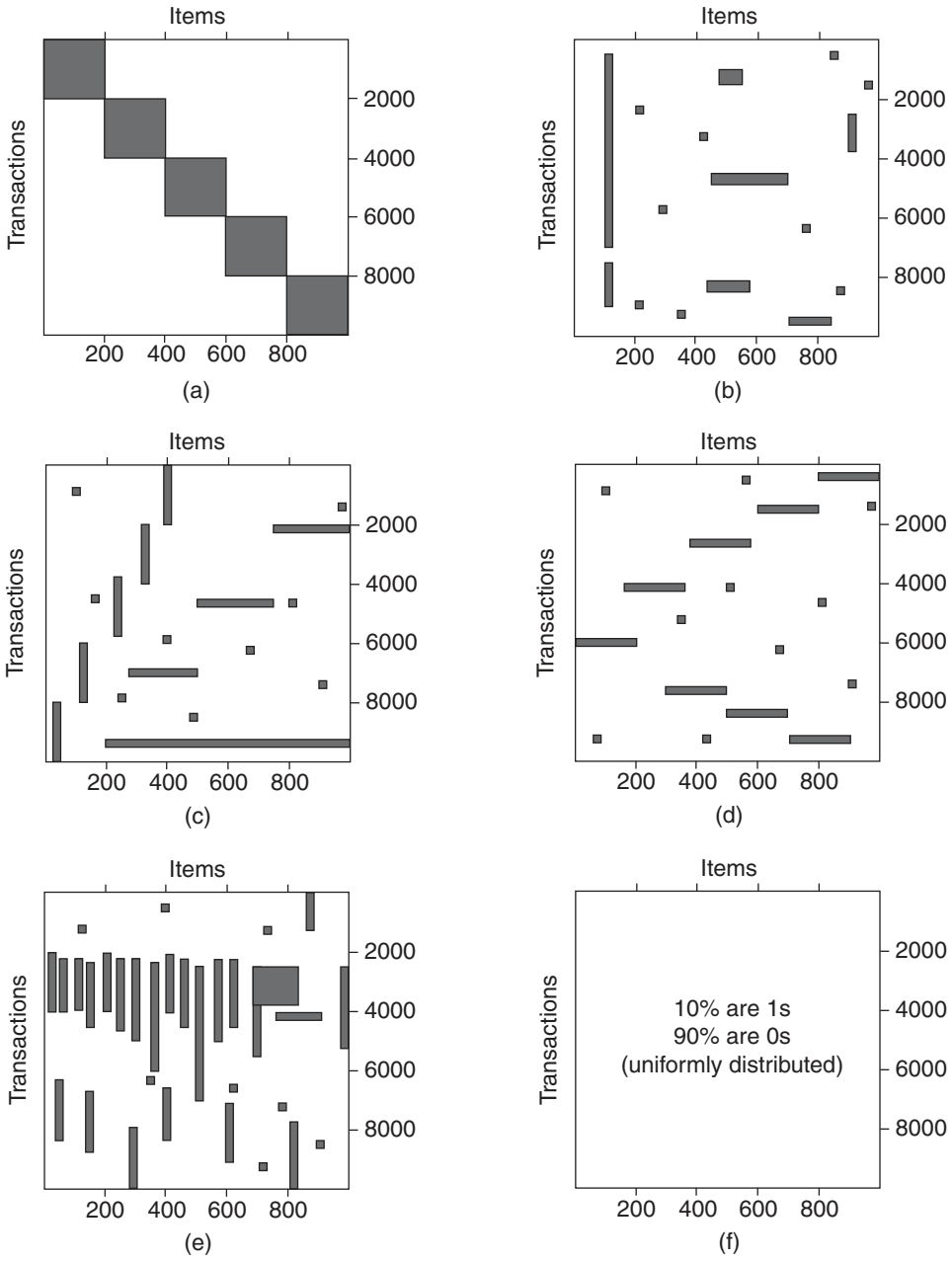


Figure 5.6. Figures for Exercise 15.

where $\beta = 1 - P(A) + 2P(A,B)$. Note that the second solution, in which the second term on the left hand side is positive, is not a feasible solution because it corresponds to $\phi = -1$. Furthermore, the solution for $P(B)$ must satisfy the following constraint: $P(B) \geq P(A,B)$. It can be shown that:

$$\begin{aligned} & P(B) - P(A,B) \\ = & \frac{1-P(A)}{2} - \frac{\sqrt{(1-P(A))^2 + 4P(A,B)(1-P(A))(1-P(A,B)/P(A))}}{2} \\ \leq & 0 \end{aligned}$$

Because of the constraint, $P(B) = P(A,B)$, which can be achieved by setting $P(A,B) = P(A)$.

(b) Show that if A and B are independent, then $P(A,B) \times P(A,\overline{B}) = P(A,\overline{B}) \times P(\overline{A},B)$.

Answer:

When A and B are independent, $P(A,B) = P(A) \times P(B)$ or equivalently:

$$\begin{aligned} & P(A,B) - P(A)P(B) = 0 \\ & P(A,B) - [P(A,B) + P(A,\overline{B})][P(A,B) + P(\overline{A},B)] = 0 \\ & P(A,B)[1 - P(A,B) - P(A,\overline{B}) - P(\overline{A},B)] - P(\overline{A},B)P(A,\overline{B}) = 0 \\ & P(A,B)P(\overline{A},\overline{B}) - P(\overline{A},B)P(A,\overline{B}) = 0. \end{aligned}$$

(c) Show that Yule's Q and Y coefficients

$$\begin{aligned} Q &= \left[\frac{f_{11}f_{00} - f_{10}f_{01}}{f_{11}f_{00} + f_{10}f_{01}}\right] \\ Y &= \left[\frac{\sqrt{f_{11}f_{00}} - \sqrt{f_{10}f_{01}}}{\sqrt{f_{11}f_{00}} + \sqrt{f_{10}f_{01}}}\right] \end{aligned}$$

are normalized versions of the odds ratio.

Answer:

Odds ratio can be written as:

$$\alpha = \frac{f_{11}f_{00}}{f_{10}f_{01}}.$$

We can express Q and Y in terms of α as follows:

$$\begin{aligned} Q &= \frac{\alpha - 1}{\alpha + 1} \\ Y &= \frac{\sqrt{\alpha} - 1}{\sqrt{\alpha} + 1} \end{aligned}$$

In both cases, Q and Y increase monotonically with α. Furthermore, when $\alpha = 0$, $Q = Y = -1$ to represent perfect negative correlation. When $\alpha = 1$, which is the condition for attribute independence, $Q = Y = 1$. Finally, when $\alpha = \infty$, $Q = Y = +1$. This suggests that Q and Y are normalized versions of α.

(d) Write a simplified expression for the value of each measure shown in Tables 6.11 and 6.12 when the variables are statistically independent.

Answer:

Measure	Value under independence
ϕ-coefficient	0
Odds ratio	1
Kappa κ	0
Interest	1
Cosine, IS	$\sqrt{P(A,B)}$
Piatetsky-Shapiro's	0
Collective strength	1
Jaccard	$0 \cdots 1$
Conviction	1
Certainty factor	0
Added value	0

17. Consider the interestingness measure, $M = \frac{P(B|A)-P(B)}{1-P(B)}$, for an association rule $A \longrightarrow B$.

(a) What is the range of this measure? When does the measure attain its maximum and minimum values?

Answer:

The range of the measure is from 0 to 1. The measure attains its maximum value when $P(B|A) = 1$ and its minimum value when $P(B|A) = P(B)$.

(b) How does M behave when $P(A, B)$ is increased while $P(A)$ and $P(B)$ remain unchanged?

Answer:

The measure can be rewritten as follows:

$$\frac{P(A,B) - P(A)P(B)}{P(A)(1-P(B))}.$$

It increases when $P(A, B)$ is increased.

(c) How does M behave when $P(A)$ is increased while $P(A, B)$ and $P(B)$ remain unchanged?

Answer:

The measure decreases with increasing $P(A)$.

(d) How does M behave when $P(B)$ is increased while $P(A, B)$ and $P(A)$ remain unchanged?

Answer:

The measure decreases with increasing $P(B)$.

(e) Is the measure symmetric under variable permutation?

Answer: No.

(f) What is the value of the measure when A and B are statistically independent?

Answer: 0.

(g) Is the measure null-invariant?

Answer: No.

(h) Does the measure remain invariant under row or column scaling operations?

Answer: No.

(i) How does the measure behave under the inversion operation?

Answer: Asymmetric.

18. Suppose we have market basket data consisting of 100 transactions and 20 items. If the support for item a is 25%, the support for item b is 90% and the support for itemset $\{a, b\}$ is 20%. Let the support and confidence thresholds be 10% and 60%, respectively.

(a) Compute the confidence of the association rule $\{a\} \rightarrow \{b\}$. Is the rule interesting according to the confidence measure?

Answer:

Confidence is $0.2/0.25 = 80\%$. The rule is interesting because it exceeds the confidence threshold.

(b) Compute the interest measure for the association pattern $\{a, b\}$. Describe the nature of the relationship between item a and item b in terms of the interest measure.

Answer:

The interest measure is $0.2/(0.25 \times 0.9) = 0.889$. The items are negatively correlated according to interest measure.

(c) What conclusions can you draw from the results of parts (a) and (b)?

Answer:

High confidence rules may not be interesting.

(d) Prove that if the confidence of the rule $\{a\} \longrightarrow \{b\}$ is less than the support of $\{b\}$, then:

i. $c(\{\overline{a}\} \longrightarrow \{b\}) > c(\{a\} \longrightarrow \{b\})$,

ii. $c(\{\overline{a}\} \longrightarrow \{b\}) > s(\{b\})$,

where $c(\cdot)$ denote the rule confidence and $s(\cdot)$ denote the support of an itemset.

Answer:

Let

$$c(\{a\} \longrightarrow \{b\}) = \frac{P(\{a,b\})}{P(\{a\})} < P(\{b\}),$$

which implies that

$$P(\{a\})P(\{b\}) > P(\{a,b\}).$$

Furthermore,

$$c(\{\overline{a}\} \longrightarrow \{b\}) = \frac{P(\{\overline{a},b\})}{P(\{\overline{a}\})} = \frac{P(\{b\}) - P(\{a,b\})}{1 - P(\{a\})}$$

i. Therefore, we may write

$$\begin{aligned} c(\{\overline{a}\} \longrightarrow \{b\}) - c(\{a\} \longrightarrow \{b\}) &= \frac{P(\{b\}) - P(\{a,b\})}{1 - P(\{a\})} - \frac{P(\{a,b\})}{P(\{a\})} \\ &= \frac{P(\{a\})P(\{b\}) - P(\{a,b\})}{P(\{a\})(1 - P(\{a\}))} \end{aligned}$$

which is positive because $P(\{a\})P(\{b\}) > P(\{a,b\})$.

ii. We can also show that

$$\begin{aligned} c(\{\overline{a}\} \longrightarrow \{b\}) - s(\{b\}) &= \frac{P(\{b\}) - P(\{a,b\})}{1 - P(\{a\})} - P(\{b\}) \\ &= \frac{P(\{a\})P(\{b\}) - P(\{a,b\})}{1 - P(\{a\})} \end{aligned}$$

is always positive because $P(\{a\})P(\{b\}) > P(\{a,b\})$.

19. Table 5.5 shows a $2 \times 2 \times 2$ contingency table for the binary variables A and B at different values of the control variable C.

(a) Compute the ϕ coefficient for A and B when $C = 0$, $C = 1$, and $C = 0$ or 1. Note that $\phi(\{A,B\}) = \frac{P(A,B) - P(A)P(B)}{\sqrt{P(A)P(B)(1-P(A))(1-P(B))}}$.

Answer:

i. When $C = 0$, $\phi(A,B) = -1/3$.
ii. When $C = 1$, $\phi(A,B) = 1$.
iii. When $C = 0$ or $C = 1$, $\phi = 0$.

(b) What conclusions can you draw from the above result?

Answer:

The result shows that some interesting relationships may disappear if the confounding factors are not taken into account.

Table 5.5. A Contingency Table.

			A	
			1	0
C = 0	B	1	0	15
		0	15	30
C = 1	B	1	5	0
		0	0	15

Table 5.6. Contingency tables for Exercise 20.

	B	$\overline{B}$
A	9	1
$\overline{A}$	1	89

(a) Table I.

	B	$\overline{B}$
A	89	1
$\overline{A}$	1	9

(b) Table II.

20. Consider the contingency tables shown in Table 5.6.

    (a) For table I, compute support, the interest measure, and the ϕ correlation coefficient for the association pattern {A, B}. Also, compute the confidence of rules $A \rightarrow B$ and $B \rightarrow A$.

    Answer:

    $s(A) = 0.1$, $s(B) = 0.9$, $s(A, B) = 0.09$.

    $I(A, B) = 9$, $\phi(A, B) = 0.89$.

    $c(A \longrightarrow B) = 0.9$, $c(B \longrightarrow A) = 0.9$.

    (b) For table II, compute support, the interest measure, and the ϕ correlation coefficient for the association pattern {A, B}. Also, compute the confidence of rules $A \rightarrow B$ and $B \rightarrow A$.

    Answer:

    $s(A) = 0.9$, $s(B) = 0.9$, $s(A, B) = 0.89$.

    $I(A, B) = 1.09$, $\phi(A, B) = 0.89$.

    $c(A \longrightarrow B) = 0.98$, $c(B \longrightarrow A) = 0.98$.

    (c) What conclusions can you draw from the results of (a) and (b)?

    Answer:

    Interest, support, and confidence are non-invariant while the ϕ-coefficient is invariant under the inversion operation. This is because ϕ-coefficient

takes into account the absence as well as the presence of an item in a transaction.

21. Consider the relationship between customers who buy high-definition televisions and exercise machines as shown in Tables 5.17 and 5.18.

    (a) Compute the odds ratios for both tables.

    Answer:

    For Table 6.19, odds ratio = 1.4938.

    For Table 6.20, the odds ratios are 0.8333 and 0.98.

    (b) Compute the ϕ-coefficient for both tables.

    Answer:

    For table 6.19, $\phi = 0.098$.

    For Table 6.20, the ϕ-coefficients are -0.0233 and -0.0047.

    (c) Compute the interest factor for both tables.

    Answer:

    For Table 6.19, $I = 1.0784$.

    For Table 6.20, the interest factors are 0.88 and 0.9971.

    For each of the measures given above, describe how the direction of association changes when data is pooled together instead of being stratified.

    Answer:

    The direction of association changes sign (from negative to positive correlated) when the data is pooled together.

6

Association Analysis: Advanced Concepts

1. Consider the traffic accident data set shown in Table 6.1.

Table 6.1. Traffic accident data set.

Weather Condition	Driver's Condition	Traffic Violation	Seat Belt	Crash Severity
Good	Alcohol-impaired	Exceed speed limit	No	Major
Bad	Sober	None	Yes	Minor
Good	Sober	Disobey stop sign	Yes	Minor
Good	Sober	Exceed speed limit	Yes	Major
Bad	Sober	Disobey traffic signal	No	Major
Good	Alcohol-impaired	Disobey stop sign	Yes	Minor
Bad	Alcohol-impaired	None	Yes	Major
Good	Sober	Disobey traffic signal	Yes	Major
Good	Alcohol-impaired	None	No	Major
Bad	Sober	Disobey traffic signal	No	Major
Good	Alcohol-impaired	Exceed speed limit	Yes	Major
Bad	Sober	Disobey stop sign	Yes	Minor

(a) Show a binarized version of the data set.
Answer: See Table 6.2.

(b) What is the maximum width of each transaction in the binarized data?
Answer: 5

(c) Assuming that support threshold is 30%, how many candidate and frequent itemsets will be generated?

Table 6.2. Traffic accident data set.

Good	Bad	Alcohol	Sober	Exceed speed	None	Disobey stop	Disobey traffic	Belt = No	Belt = Yes	Major	Minor
1	0	1	0	1	0	0	0	1	0	1	0
0	1	0	1	0	1	0	0	0	1	0	1
1	0	0	1	0	0	1	0	0	1	0	1
1	0	0	1	1	0	0	0	0	1	1	0
0	1	0	1	0	0	0	1	1	0	1	0
1	0	1	0	0	0	1	0	0	1	0	1
0	1	1	0	0	1	0	0	0	1	1	0
1	0	0	1	0	0	0	1	0	1	1	0
1	0	1	0	0	1	0	0	1	0	1	0
0	1	0	1	0	0	0	1	1	0	1	0
1	0	1	0	1	0	0	0	0	1	1	0
0	1	0	1	0	0	1	0	0	1	0	1

Answer: 5

The number of candidate itemsets from size 1 to size 3 is $10+28+3 = 41$.

The number of frequent itemsets from size 1 to size 3 is $8+10+0 = 18$.

(d) Create a data set that contains only the following asymmetric binary attributes: (`Weather` = `Bad`, `Driver's condition` = `Alcohol-impaired`, `Traffic violation` = `Yes`, `Seat Belt` = `No`, `Crash Severity` = `Major`). For `Traffic violation`, only `None` has a value of 0. The rest of the attribute values are assigned to 1. Assuming that support threshold is 30%, how many candidate and frequent itemsets will be generated?

Answer:

The binarized data is shown in Table 6.3.

Table 6.3. Traffic accident data set.

Bad	Alcohol Impaired	Traffic violation	Belt = No	Major
0	1	1	1	1
1	0	0	0	0
0	0	1	0	0
0	0	1	0	1
1	0	1	1	1
0	1	1	0	0
1	1	0	0	1
0	0	1	0	1
0	1	0	1	1
1	0	1	1	1
0	1	1	0	1
1	0	1	0	0

The number of candidate itemsets from size 1 to size 3 is $5+10+0 = 15$.

The number of frequent itemsets from size 1 to size 3 is $5 + 3 + 0 = 8$.

(e) Compare the number of candidate and frequent itemsets generated in parts (c) and (d).

Answer:

The second method produces less number of candidate and frequent itemsets.

2. (a) Consider the data set shown in Table 6.4. Suppose we apply the following discretization strategies to the continuous attributes of the data set.

D1: Partition the range of each continuous attribute into 3 equal-sized bins.

D2: Partition the range of each continuous attribute into 3 bins; where each bin contains an equal number of transactions

For each strategy, answer the following questions:

i. Construct a binarized version of the data set.

ii. Derive all the frequent itemsets having support $\geq 30\%$.

Table 6.4. Data set for Exercise 2.

TID	Temperature	Pressure	Alarm 1	Alarm 2	Alarm 3
1	95	1105	0	0	1
2	85	1040	1	1	0
3	103	1090	1	1	1
4	97	1084	1	0	0
5	80	1038	0	1	1
6	100	1080	1	1	0
7	83	1025	1	0	1
8	86	1030	1	0	0
9	101	1100	1	1	1

Answer:

Table 6.5 shows the discretized data using D1, where the discretized intervals are:

- X1: Temperature between 80 and 87,
- X2: Temperature between 88 and 95,
- X3: Temperature between 96 and 103,
- Y1: Pressure between 1025 and 1051,
- Y2: Pressure between 1052 and 1078,
- Y3: Pressure between 1079 and 1105.

Table 6.5. Discretized data using D1.

TID	X1	X2	X3	Y1	Y2	Y3	Alarm1	Alarm2	Alarm3
1	0	1	0	0	0	1	0	0	1
2	1	0	0	1	0	0	1	1	0
3	0	0	1	0	0	1	1	1	1
4	0	0	1	0	0	1	1	0	0
5	1	0	0	1	0	0	0	1	1
6	0	0	1	0	0	1	1	1	0
7	1	0	0	1	0	0	1	0	1
8	1	0	0	1	0	0	1	0	0
9	0	0	1	0	0	1	1	1	1

Table 6.6. Discretized data using D2.

TID	X1	X2	X3	Y1	Y2	Y3	Alarm1	Alarm2	Alarm3
1	0	1	0	0	0	1	0	0	1
2	1	0	0	0	1	0	1	1	0
3	0	0	1	0	0	1	1	1	1
4	0	1	0	0	1	0	1	0	0
5	1	0	0	1	0	0	0	1	1
6	0	0	1	0	1	0	1	1	0
7	1	0	0	1	0	0	1	0	1
8	0	1	0	1	0	0	1	0	0
9	0	0	1	0	0	1	1	1	1

Table 6.6 shows the discretized data using D1, where the discretized intervals are:

- X1: Temperature between 80 and 85,
- X2: Temperature between 86 and 97,
- X3: Temperature between 100 and 103,
- Y1: Pressure between 1025 and 1038,
- Y2: Pressure between 1039 and 1084,
- Y3: Pressure between 1085 and 1105.

For D1, there are 7 frequent 1-itemset, 12 frequent 2-itemset, and 5 frequent 3-itemset.

For D2, there are 9 frequent 1-itemset, 7 frequent 2-itemset, and 1 frequent 3-itemset.

(b) The continuous attribute can also be discretized using a clustering approach.

i. Plot a graph of temperature versus pressure for the data points shown in Table 6.4.

Answer:
The graph of Temperature and Pressure is shown below.

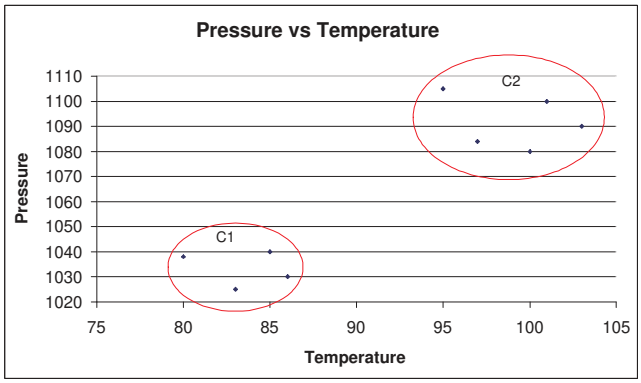


Figure 6.1. Temperature versus Pressure.

ii. How many natural clusters do you observe from the graph? Assign a label (C_1, C_2, etc.) to each cluster in the graph.
Answer: There are two natural clusters in the data.

iii. What type of clustering algorithm do you think can be used to identify the clusters? State your reasons clearly.
Answer: K-means algorithm.

iv. Replace the temperature and pressure attributes in Table 6.4 with asymmetric binary attributes C_1, C_2, etc. Construct a transaction matrix using the new attributes (along with attributes Alarm1, Alarm2, and Alarm3).
Answer:

Table 6.7. Example of numeric data set.

TID	C1	C2	Alarm1	Alarm2	Alarm3
1	0	1	0	0	1
2	1	0	1	1	0
3	0	1	1	1	1
4	0	1	1	0	0
5	1	0	0	1	1
6	0	1	1	1	0
7	1	0	1	0	1
8	1	0	1	0	0
9	0	1	1	1	1

v. Derive all the frequent itemsets having support $\geq 30\%$ from the binarized data.
Answer:
There are 5 frequent 1-itemset, 7 frequent 2-itemset, and 1 frequent 3-itemset.

3. Consider the data set shown in Table 6.8. The first attribute is continuous, while the remaining two attributes are asymmetric binary. A rule is considered to be strong if its support exceeds 15% and its confidence exceeds 60%. The data given in Table 6.8 supports the following two strong rules:

(i) $\{(1 \leq A \leq 2), B = 1\} \rightarrow \{C = 1\}$
(ii) $\{(5 \leq A \leq 8), B = 1\} \rightarrow \{C = 1\}$

Table 6.8. Data set for Exercise 3.

A	B	C
1	1	1
2	1	1
3	1	0
4	1	0
5	1	1
6	0	1
7	0	0
8	1	1
9	0	0
10	0	0
11	0	0
12	0	1

(a) Compute the support and confidence for both rules.
Answer:
$s(\{(1 \leq A \leq 2), B = 1\} \rightarrow \{C = 1\}) = 1/6$
$c(\{(1 \leq A \leq 2), B = 1\} \rightarrow \{C = 1\}) = 1$
$s(\{(5 \leq A \leq 8), B = 1\} \rightarrow \{C = 1\}) = 1/6$
$c(\{(5 \leq A \leq 8), B = 1\} \rightarrow \{C = 1\}) = 1$

(b) To find the rules using the traditional *Apriori* algorithm, we need to discretize the continuous attribute A. Suppose we apply the equal width binning approach to discretize the data, with *bin-width* $= 2, 3, 4$. For each *bin-width*, state whether the above two rules are discovered by the *Apriori* algorithm. (Note that the rules may not be in the same exact form as before because it may contain wider or narrower intervals

for A.) For each rule that corresponds to one of the above two rules, compute its support and confidence.

Answer:
When $bin - width = 2$:

Table 6.9. A Synthetic Data set

A1	A2	A3	A4	A5	A6	B	C
1	0	0	0	0	0	1	1
1	0	0	0	0	0	1	1
0	1	0	0	0	0	1	0
0	1	0	0	0	0	1	0
0	0	1	0	0	0	1	1
0	0	1	0	0	0	0	1
0	0	0	1	0	0	0	0
0	0	0	1	0	0	1	1
0	0	0	0	1	0	0	0
0	0	0	0	1	0	0	0
0	0	0	0	0	1	0	0
0	0	0	0	0	1	0	1

Where

$$A1 = 1 \leq A \leq 2;\ A2 = 3 \leq A \leq 4;$$
$$A3 = 5 \leq A \leq 6;\ A4 = 7 \leq A \leq 8;$$
$$A5 = 9 \leq A \leq 10;\ A6 = 11 \leq A \leq 12;$$

For the first rule, there is one corresponding rule:

$$\{A1 = 1, B = 1\} \rightarrow \{C = 1\}$$

$s(A1 = 1, B = 1\} \rightarrow \{C = 1\}) = 1/6$
$c(A1 = 1, B = 1\} \rightarrow \{C = 1\}) = 1$

Since the support and confidence are greater than the thresholds, the rule can be discovered.

For the second rule, there are two corresponding rules:

$$\{A3 = 1, B = 1\} \rightarrow \{C = 1\}$$
$$\{A4 = 1, B = 1\} \rightarrow \{C = 1\}$$

For both rules, the support is 1/12 and the confidence is 1. Since the support is less than the threshold (15%), these rules cannot be generated.

When $bin - width = 3$:

Table 6.10. A Synthetic Data set

A1	A2	A3	A4	B	C
1	0	0	0	1	1
1	0	0	0	1	1
1	0	0	0	1	0
0	1	0	0	1	0
0	1	0	0	1	1
0	1	0	0	0	1
0	0	1	0	0	0
0	0	1	0	1	1
0	0	1	0	0	0
0	0	0	1	0	0
0	0	0	1	0	0
0	0	0	1	0	1

Where

$$A1 = 1 \leq A \leq 3;\ A2 = 4 \leq A \leq 6;$$
$$A3 = 7 \leq A \leq 9;\ A4 = 10 \leq A \leq 12;$$

For the first rule, there is one corresponding rule:

$$\{A1 = 1, B = 1\} \rightarrow \{C = 1\}$$

$s(A1 = 1, B = 1\} \rightarrow \{C = 1\}) = 1/6$
$c(A1 = 1, B = 1\} \rightarrow \{C = 1\}) = 2/3$

Since the support and confidence are greater than the thresholds, the rule can be discovered. The discovered rule is in general form than the original rule.

For the second rule, there are two corresponding rules:

$$\{A2 = 1, B = 1\} \rightarrow \{C = 1\}$$
$$\{A3 = 1, B = 1\} \rightarrow \{C = 1\}$$

For both rules, the support is 1/12 and the confidence is 1. Since the support is less than the threshold (15%), these rules cannot be generated.

When $bin - width = 4$:

Table 6.11. A Synthetic Data set

A1	A2	A3	B	C
1	0	0	1	1
1	0	0	1	1
1	0	0	1	0
1	0	0	1	0
0	1	0	1	1
0	1	0	0	1
0	1	0	0	0
0	1	0	1	1
0	0	1	0	0
0	0	1	0	0
0	0	1	0	0
0	0	1	0	1

Where

$$A1 = 1 \leq A \leq 4;\ A2 = 5 \leq A \leq 8;$$
$$A3 = 9 \leq A \leq 12;$$

For the first rule, there is one correspomding rule:

$$\{A1 = 1, B = 1\} \rightarrow \{C = 1\}$$

$s(A1 = 1, B = 1\} \rightarrow \{C = 1\}) = 1/6$
$c(A1 = 1, B = 1\} \rightarrow \{C = 1\}) = 1/2$

Since the confidence is less than the threshold (60%), then the rule cannot be generated.

For the second rule, there is one corresponding rule:

$$\{A2 = 1, B = 1\} \rightarrow \{C = 1\}$$

$s(A2 = 1, B = 1\} \rightarrow \{C = 1\}) = 1/6$
$c(A2 = 1, B = 1\} \rightarrow \{C = 1\}) = 1$

Since the support and threshold are greater than thresholds, the the rule can be discovered.

(c) Comment on the effectiveness of using the equal width approach for classifying the above data set. Is there a *bin*-width that allows you to

find both rules satisfactorily? If not, what alternative approach can you take to ensure that you will find both rules?

Answer:

None of the discretization methods can effectively find both rules. One approach to ensure that you can find both rules is to start with bin width equals to 2 and consider all possible mergings of the adjacent intervals. For example, the discrete intervals are:

$1 <= A <= 2$, $3 <= A <= 4$, $5 <= A <= 6$, $\cdot$, $11 <= A <= 12$
$1 <= A <= 4$, $5 <= A <= 8$, $9 <= A <= 12$

4. Consider the data set shown in Table 6.12.

Table 6.12. Data set for Exercise 4.

Age (A)	Number of Hours Online per Week (B)				
	0 – 5	5 – 10	10 – 20	20 – 30	30 – 40
10 – 15	2	3	5	3	2
15 – 25	2	5	10	10	3
25 – 35	10	15	5	3	2
35 – 50	4	6	5	3	2

(a) For each combination of rules given below, specify the rule that has the highest confidence.

i. $15 < A < 25 \longrightarrow 10 < B < 20$, $10 < A < 25 \longrightarrow 10 < B < 20$, and $15 < A < 35 \longrightarrow 10 < B < 20$.
Answer:
Both $15 < A < 25 \longrightarrow 10 < B < 20$ and $10 < A < 25 \longrightarrow 10 < B < 20$ have confidence 33.3%.

ii. $15 < A < 25 \longrightarrow 10 < B < 20$, $15 < A < 25 \longrightarrow 5 < B < 20$, and $15 < A < 25 \longrightarrow 5 < B < 30$.
Answer:
The rule $15 < A < 25 \longrightarrow 5 < B < 30$ has the highest confidence (83.3%).

iii. $15 < A < 25 \longrightarrow 10 < B < 20$ and $10 < A < 35 \longrightarrow 5 < B < 30$.
Answer:
The rule $10 < A < 35 \longrightarrow 5 < B < 30$ has the highest confidence (73.8%).

(b) Suppose we are interested in finding the average number of hours spent online per week by Internet users between the age of 15 and 35. Write the corresponding statistics-based association rule to characterize the segment of users. To compute the average number of hours spent online,

approximate each interval by its midpoint value (e.g., use $B = 7.5$ to represent the interval $5 < B < 10$).

Answer:

There are 65 people whose average age is between 15 and 35.

The average number of hours spent online is

$$2.5 \times 12/65 + 7.5 \times 20/65 + 15 \times 15/65 + 25 \times 13/65 + 35 \times 5/65 = 13.92.$$

Therefore the statistics-based association rule is:

$$15 \leq A < 35 \longrightarrow B : \mu = 13.82.$$

(c) Test whether the quantitative association rule given in part (b) is statistically significant by comparing its mean against the average number of hours spent online by other users who do not belong to the age group.

For other users, the average number of hours spent online is:

$$2.5 \times 6/35 + 7.5 \times 9/35 + 15 \times 10/35 + 25 \times 6/65 + 35 \times 4/65 = 14.93.$$

The standard deviations for the two groups are 9.786 ($15 \leq$ Age < 35) and 10.203 (Age < 15 or Age ≥ 35), respectively.

$$Z = \frac{14.93 - 13.82}{\sqrt{\frac{9.786^2}{65} + \frac{10.203^2}{35}}} = 0.476 < 1.64$$

The difference is not significant at 95% confidence level.

5. For the data set with the attributes given below, describe how you would convert it into a binary transaction data set appropriate for association analysis. Specifically, indicate for each attribute in the original data set

   (a) How many binary attributes it would correspond to in the transaction data set,

   (b) How the values of the original attribute would be mapped to values of the binary attributes, and

   (c) If there is any hierarchical structure in the data values of an attribute that could be useful for grouping the data into fewer binary attributes.

   The following is a list of attributes for the data set along with their possible values. Assume that all attributes are collected on a per-student basis:

   - **Year** : Freshman, Sophomore, Junior, Senior, Graduate:Masters, Graduate:PhD, Professional

     Answer:

(a) Each attribute value can be represented using an asymmetric binary attribute. Therefore, there are altogether 7 binary attributes.

(b) There is a one-to-one mapping between the original attribute values and the asymmetric binary attributes.

(c) We have a hierarchical structure involving the following high-level concepts: Undergraduate, Graduate, Professional.

- **Zip code** : zip code for the home address of a U.S. student, zip code for the local address of a non-U.S. student

  Answer:

  (a) Each attribute value is represented by an asymmetric binary attribute. Therefore, we have as many asymmetric binary attributes as the number of distinct zipcodes.

  (b) There is a one-to-one mapping between the original attribute values and the asymmetric binary attributes.

  (c) We can have a hierarchical structure based on geographical regions (e.g., zipcodes can be grouped according to their corresponding states).

- **College** : Agriculture, Architecture, Continuing Education, Education, Liberal Arts, Engineering, Natural Sciences, Business, Law, Medical, Dentistry, Pharmacy, Nursing, Veterinary Medicine

  Answer:

  (a) Each attribute value is represented by an asymmetric binary attribute. Therefore, we have as many asymmetric binary attributes as the number of distinct colleges.

  (b) There is a one-to-one mapping between the original attribute values and the asymmetric binary attributes.

  (c) We can have a hierarchical structure based on the type of school. For example, colleges of Medical and Medical might be grouped together as Medical school while Engineering and Natural Sciences might be grouped together into the same school.

- **On Campus** : 1 if the student lives on campus, 0 otherwise

  Answer:

  (a) This attribute can be mapped to one binary attribute.

  (b) There is no hierarchical structure.

- Each of the following is a separate attribute that has a value of 1 if the person speaks the language and a value of 0, otherwise.
  - Arabic
  - Bengali
  - Chinese Mandarin
  - English

– Portuguese
– Russian
– Spanish

Answer:

(a) Each attribute value can be represented by an asymmetric binary attribute. Therefore, we have as many asymmetric binary attributes as the number of distinct dialects.

(b) There is a one-to-one mapping between the original attribute values and the asymmetric binary attributes.

(c) We can have a hierarchical structure based on the region in which the languages are spoken (e.g., Asian, European, etc.)

6. Consider the data set shown in Table 6.13. Suppose we are interested in extracting the following association rule:

$$\{\alpha_1 \leq \text{Age} \leq \alpha_2, \text{Play Piano} = \text{Yes}\} \longrightarrow \{\text{Enjoy Classical Music} = \text{Yes}\}$$

Table 6.13. Data set for Exercise 6.

Age	Play Piano	Enjoy Classical Music
9	Yes	Yes
11	Yes	Yes
14	Yes	No
17	Yes	No
19	Yes	Yes
21	No	No
25	No	No
29	Yes	Yes
33	No	No
39	No	Yes
41	No	No
47	No	Yes

To handle the continuous attribute, we apply the equal-frequency approach with 3, 4, and 6 intervals. Categorical attributes are handled by introducing as many new asymmetric binary attributes as the number of categorical values. Assume that the support threshold is 10% and the confidence threshold is 70%.

(a) Suppose we discretize the Age attribute into 3 equal-frequency intervals. Find a pair of values for α_1 and α_2 that satisfy the minimum support and minimum confidence requirements.

Answer:

$(\alpha_1 = 19, \alpha_2 = 29)$: $s = 16.7\%, c = 100\%$.

(b) Repeat part (a) by discretizing the Age attribute into 4 equal-frequency intervals. Compare the extracted rules against the ones you had obtained in part (a).

Answer:

No rule satisfies the support and confidence thresholds.

(c) Repeat part (a) by discretizing the Age attribute into 6 equal-frequency intervals. Compare the extracted rules against the ones you had obtained in part (a).

Answer:

$(\alpha_1 = 9, \alpha_2 = 11)$: $s = 16.7\%, c = 100\%$.

(d) From the results in part (a), (b), and (c), discuss how the choice of discretization intervals will affect the rules extracted by association rule mining algorithms.

If the discretization interval is too wide, some rules may not have enough confidence to be detected by the algorithm. If the discretization interval is too narrow, the rule in part (a) will be lost.

7. Consider the transactions shown in Table 6.14, with an item taxonomy given in Figure 6.25.

Table 6.14. Example of market basket transactions.

Transaction ID	Items Bought
1	Chips, Cookies, Regular Soda, Ham
2	Chips, Ham, Boneless Chicken, Diet Soda
3	Ham, Bacon, Whole Chicken, Regular Soda
4	Chips, Ham, Boneless Chicken, Diet Soda
5	Chips, Bacon, Boneless Chicken
6	Chips, Ham, Bacon, Whole Chicken, Regular Soda
7	Chips, Cookies, Boneless Chicken, Diet Soda

(a) What are the main challenges of mining association rules with item taxonomy?

Answer:

Difficulty of deciding the right support and confidence thresholds. Items residing at higher levels of the taxonomy have higher support than those residing at lower levels of the taxonomy. Many of the rules may also be redundant.

(b) Consider the approach where each transaction t is replaced by an extended transaction t' that contains all the items in t as well as their respective ancestors. For example, the transaction $t = \{$ `Chips`, `Cookies`$\}$ will be replaced by $t' = \{$`Chips`, `Cookies`, `Snack Food`, `Food`$\}$. Use this approach to derive all frequent itemsets (up to size 4) with support $\geq$ 70%.

Answer:

There are 8 frequent 1-itemsets, 25 frequent 2-itemsets, 34 frequent 3-itemsets and 20 frequent 4-itemsets. The frequent 4-itemsets are:

{Food, Snack Food, Meat, Soda}	{Food, Snack Food, Meat, Chips}
{Food, Snack Food, Meat, Pork}	{Food, Snack Food, Meat, Chicken}
{Food, Snack Food, Soda, Chips}	{Food, Snack Food, Chips, Pork}
{Food, Snack Food, Chips, Chicken}	{Food, Meat, Soda, Chips}
{Food, Meat, Soda, Pork}	{Food, Meat, Soda, Chicken}
{Food, Meat, Soda, Ham}	{Food, Meat, Chips, Pork}
{Food, Meat, Chips, Chicken}	{Food, Meat, Pork, Chicken}
{Food, Meat, Pork, Ham}	{Food, Soda, Pork, Ham}
{Snack Food, Meat, Soda, Chips}	{Snack Food, Meat, Chips, Pork}
{Snack Food, Meat, Chips, Chicken}	{Meat, Soda, Pork, Ham}

(c) Consider an alternative approach where the frequent itemsets are generated one level at a time. Initially, all the frequent itemsets involving items at the highest level of the hierarchy are generated. Next, we use the frequent itemsets discovered at the higher level of the hierarchy to generate candidate itemsets involving items at the lower levels of the hierarchy. For example, we generate the candidate itemset {`Chips`, `Diet Soda`} only if {`Snack Food`, `Soda`} is frequent. Use this approach to derive all frequent itemsets (up to size 4) with support $\geq$ 70%.

Answer:

There are 8 frequent 1-itemsets, 6 frequent 2-itemsets, and 1 frequent 3-itemset. The frequent 2-itemsets and 3-itemsets are:

{Snack Food, Meat}	{Snack Food, Soda}
{Meat, Soda}	{Chips, Pork}
{Chips, Chicken}	{Pork, Chicken}
{Snack Food, Meat, Soda}	

(d) Compare the frequent itemsets found in parts (b) and (c). Comment on the efficiency and completeness of the algorithms.

Answer:

The method in part (b) is more complete but less efficient compared to the method in part (c). The method in part (c) is more efficient but may lose some frequent itemsets.

8. The following questions examine how the support and confidence of an association rule may vary in the presence of a concept hierarchy.

(a) Consider an item x in a given concept hierarchy. Let $\overline{x}_1$, $\overline{x}_2$, ..., $\overline{x}_k$ denote the k children of x in the concept hierarchy. Show that $s(x) \leq \sum_{i=1}^{k} s(\overline{x}_i)$, where $s(\cdot)$ is the support of an item. Under what conditions will the inequality become an equality?

Answer:

If no transaction contains more than one child of x, then $s(x) = \sum_{i=1}^{k} s(\overline{x}_i)$.

(b) Let p and q denote a pair of items, while $\hat{p}$ and $\hat{q}$ are their corresponding parents in the concept hierarchy. If $s(\{p, q\}) > minsup$, which of the following itemsets are guaranteed to be frequent? (i) $s(\{\hat{p}, q\})$, (ii) $s(\{p, \hat{q}\})$, and (iii) $s(\{\hat{p}, \hat{q}\})$.

Answer:

All three itemsets are guaranteed to be frequent.

(c) Consider the association rule $\{p\} \longrightarrow \{q\}$. Suppose the confidence of the rule exceeds $minconf$. Which of the following rules are guaranteed to have confidence higher than $minconf$? (i) $\{p\} \longrightarrow \{\hat{q}\}$, (ii) $\{\hat{p}\} \longrightarrow \{q\}$, and (iii) $\{\hat{p}\} \longrightarrow \{\hat{q}\}$.

Answer:

Only $\{p\} \longrightarrow \{\hat{q}\}$ is guaranteed to have confidence higher than $minconf$.

9. (a) List all the 4-subsequences contained in the following data sequence:

$$< \{1,3\}\ \{2\}\ \{2,3\}\ \{4\} >,$$

assuming no timing constraints.

Answer:

$< \{1,3\}\ \{2\}\ \{2\} >$	$< \{1,3\}\ \{2\}\ \{3\} >$
$< \{1,3\}\ \{2\}\ \{4\} >$	$< \{1,3\}\{2,3\} >$
$< \{1,3\}\ \{3\}\ \{4\} >$	$< \{1\}\ \{2\}\ \{2,3\} >$
$< \{1\}\ \{2\}\ \{2\}\ \{4\} >$	$< \{1\}\ \{2\}\ \{3\}\ \{4\} >$
$< \{1\}\ \{2,3\}\ \{4\} >$	$< \{3\}\ \{2\}\ \{2,3\} >$
$< \{3\}\ \{2\}\ \{2\}\ \{4\} >$	$< \{3\}\ \{2\}\ \{3\}\ \{4\} >$
$< \{3\}\ \{2,3\}\ \{4\} >$	$< \{2\}\ \{2,3\}\ \{4\} >$

(b) List all the 3-element subsequences contained in the data sequence for part (a) assuming that no timing constraints are imposed.

Answer:

$< \{1,3\}\ \{2\}\ \{2,3\} >$	$< \{1,3\}\ \{2\}\ \{4\} >$
$< \{1,3\}\ \{3\}\ \{4\} >$	$< \{1,3\}\ \{2\}\ \{2\} >$
$< \{1,3\}\ \{2\}\ \{3\} >$	$< \{1,3\}\ \{2,3\}\ \{4\} >$
$< \{1\}\ \{2\}\ \{2,3\} >$	$< \{1\}\ \{2\}\ \{4\} >$
$< \{1\}\ \{3\}\ \{4\} >$	$< \{1\}\ \{2\}\ \{2\} >$
$< \{1\}\ \{2\}\ \{3\} >$	$< \{1\}\ \{2,3\}\ \{4\} >$
$< \{3\}\ \{2\}\ \{2,3\} >$	$< \{3\}\ \{2\}\ \{4\} >$
$< \{3\}\ \{3\}\ \{4\} >$	$< \{3\}\ \{2\}\ \{2\} >$
$< \{3\}\ \{2\}\ \{3\} >$	$< \{3\}\ \{2,3\}\ \{4\} >$

(c) List all the 4-subsequences contained in the data sequence for part (a) (assuming the timing constraints are flexible).

Answer:

This will include all the subsequences in part (a) as well as the following:

$< \{1,2,3,4\} >$
$< \{1,2,3\}\ \{2\} >$
$< \{1,2,3\}\ \{3\} >$
$< \{1,2,3\}\ \{4\} >$
$< \{1,3\}\ \{2,4\} >$
$< \{1,3\}\ \{3,4\} >$
$< \{1\}\ \{2\}\ \{2,4\} >$
$< \{1\}\ \{2\}\ \{3,4\} >$
$< \{3\}\ \{2\}\ \{2,4\} >$
$< \{3\}\ \{2\}\ \{3,4\} >$
$< \{1,2\}\ \{2,3\} >$
$< \{1,2\}\ \{2,4\} >$
$< \{1,2\}\ \{3,4\} >$
$< \{1,2\}\ \{2\}\{4\} >$
$< \{1,2\}\ \{3\}\ \{4\} >$
$< \{2,3\}\ \{2,3\} >$
$< \{2,3\}\ \{2,4\} >$
$< \{2,3\}\ \{3,4\} >$
$< \{2,3\}\ \{2\}\{4\} >$
$< \{2,3\}\ \{3\}\ \{4\} >$
$< \{1\}\ \{2,3,4\} >$
$< \{1\}\ \{2\}\ \{2,4\} >$
$< \{1\}\ \{2\}\ \{3,4\} >$
$< \{3\}\ \{2,3,4\} >$
$< \{3\}\ \{2\}\ \{2,4\} >$
$< \{3\}\ \{2\}\ \{3,4\} >$
$< \{2\}\ \{2,3,4\} >$

(d) List all the 3-element subsequences contained in the data sequence for part (a) (assuming the timing constraints are flexible).

Answer:

This will include all the subsequences in part (b) as well as the following:

$< \{1,2,3\}\ \{2\}\ \{4\} >$
$< \{1,2,3\}\ \{3\}\ \{4\} >$
$< \{1,2,3\}\ \{2,3\}\ \{4\} >$
$< \{1,2\}\ \{2\}\ \{4\} >$
$< \{1,2\}\ \{3\}\ \{4\} >$
$< \{1,2\}\ \{2,3\}\ \{4\} >$
$< \{2,3\}\ \{2\}\ \{4\} >$
$< \{2,3\}\ \{3\}\ \{4\} >$
$< \{2,3\}\ \{2,3\}\ \{4\} >$
$< \{1\}\ \{2\}\ \{2,4\} >$
$< \{1\}\ \{2\}\ \{3,4\} >$
$< \{1\}\ \{2\}\ \{2,3,4\} >$
$< \{3\}\ \{2\}\ \{2,4\} >$
$< \{3\}\ \{2\}\ \{3,4\} >$
$< \{3\}\ \{2\}\ \{2,3,4\} >$
$< \{1,3\}\ \{2\}\ \{2,4\} >$
$< \{1,3\}\ \{2\}\ \{3,4\} >$
$< \{1,3\}\ \{2\}\ \{2,3,4\} >$

10. Find all the frequent subsequences with support $\geq 50\%$ given the sequence database shown in Table 6.15. Assume that there are no timing constraints imposed on the sequences.

Answer:

$< \{A\} >, < \{B\} >, < \{C\} >, < \{D\} >, < \{E\} >$
$< \{A\}\ \{C\} >, < \{A\}\ \{D\} >, < \{A\}\ \{E\} >, < \{B\}\ \{C\} >,$
$< \{B\}\ \{D\} >, < \{B\}\ \{E\} >, < \{C\}\ \{D\} >, < \{C\}\ \{E\} >, < \{D,E\} >$

11. (a) For each of the sequences $w =< e_1 e_2 \ldots e_i \ldots e_{i+1} \ldots e_{last} >$ given below, determine whether they are subsequences of the sequence

$$< \{1,2,3\}\{2,4\}\{2,4,5\}\{3,5\}\{6\} >$$

Table 6.15. Example of event sequences generated by various sensors.

Sensor	Timestamp	Events
S1	1	A, B
	2	C
	3	D, E
	4	C
S2	1	A, B
	2	C, D
	3	E
S3	1	B
	2	A
	3	B
	4	D, E
S4	1	C
	2	D, E
	3	C
	4	E
S5	1	B
	2	A
	3	B, C
	4	A, D

subjected to the following timing constraints:

mingap = 0	(interval between last event in e_i and first event in e_{i+1} is > 0)
maxgap = 3	(interval between first event in e_i and last event in e_{i+1} is ≤ 3)
maxspan = 5	(interval between first event in e_1 and last event in e_{last} is ≤ 5)
$ws = 1$	(time between first and last events in e_i is ≤ 1)

- $w =< \{1\}\{2\}\{3\} >$
  Answer: Yes.
- $w =< \{1,2,3,4\}\{5,6\} >$
  Answer: No.
- $w =< \{2,4\}\{2,4\}\{6\} >$
  Answer: Yes.
- $w =< \{1\}\{2,4\}\{6\} >$
  Answer: Yes.
- $w =< \{1,2\}\{3,4\}\{5,6\} >$
  Answer: No.

(b) Determine whether each of the subsequences w given in the previous question are contiguous subsequences of the following sequences s.

- $s =< \{1,2,3,4,5,6\}\{1,2,3,4,5,6\}\{1,2,3,4,5,6\} >$
  - $w =< \{1\}\{2\}\{3\} >$
    Answer: Yes.
  - $w =< \{1,2,3,4\}\{5,6\} >$
    Answer: Yes.
  - $w =< \{2,4\}\{2,4\}\{6\} >$
    Answer: Yes.
  - $w =< \{1\}\{2,4\}\{6\} >$
    Answer: Yes.
  - $w =< \{1,2\}\{3,4\}\{5,6\} >$
    Answer: Yes.
- $s =< \{1,2,3,4\}\{1,2,3,4,5,6\}\{3,4,5,6\} >$
  - $w =< \{1\}\{2\}\{3\} >$
    Answer: Yes.
  - $w =< \{1,2,3,4\}\{5,6\} >$
    Answer: Yes.
  - $w =< \{2,4\}\{2,4\}\{6\} >$
    Answer: Yes.
  - $w =< \{1\}\{2,4\}\{6\} >$
    Answer: Yes.
  - $w =< \{1,2\}\{3,4\}\{5,6\} >$
    Answer: Yes.
- $s =< \{1,2\}\{1,2,3,4\}\{3,4,5,6\}\{5,6\} >$
  - $w =< \{1\}\{2\}\{3\} >$
    Answer: Yes.
  - $w =< \{1,2,3,4\}\{5,6\} >$
    Answer: Yes.
  - $w =< \{2,4\}\{2,4\}\{6\} >$
    Answer: No.
  - $w =< \{1\}\{2,4\}\{6\} >$
    Answer: Yes.
  - $w =< \{1,2\}\{3,4\}\{5,6\} >$
    Answer: Yes.
- $s =< \{1,2,3\}\{2,3,4,5\}\{4,5,6\} >$
  - $w =< \{1\}\{2\}\{3\} >$
    Answer: No.
  - $w =< \{1,2,3,4\}\{5,6\} >$
    Answer: No.
  - $w =< \{2,4\}\{2,4\}\{6\} >$
    Answer: No.
  - $w =< \{1\}\{2,4\}\{6\} >$
    Answer: Yes.

– $w =< \{1,2\}\{3,4\}\{5,6\} >$
Answer: Yes.

12. For each of the sequence $w = \langle e_1, \ldots, e_{last} \rangle$ below, determine whether they are subsequences of the following data sequence:

$$\langle \{A,B\}\{C,D\}\{A,B\}\{C,D\}\{A,B\}\{C,D\} \rangle$$

subjected to the following timing constraints:

mingap = 0	(interval between last event in e_i and first event in e_{i+1} is > 0)
maxgap = 2	(interval between first event in e_i and last event in e_{i+1} is ≤ 2)
maxspan = 6	(interval between first event in e_1 and last event in e_{last} is ≤ 6)
$ws = 1$	(time between first and last events in e_i is ≤ 1)

(a) $w = \langle \{A\}\{B\}\{C\}\{D\} \rangle$
Answer: Yes.

(b) $w = \langle \{A\}\{B,C,D\}\{A\} \rangle$
Answer: No.

(c) $w = \langle \{A\}\{B,C,D\}\{A\} \rangle$
Answer: No.

(d) $w = \langle \{B,C\}\{A,D\}\{B,C\} \rangle$
Answer: No.

(e) $w = \langle \{A,B,C,D\}\{A,B,C,D\} \rangle$
Answer: No.

13. Consider the following frequent 3-sequences:

$< \{1,2,3\} >$, $< \{1,2\}\{3\} >$, $< \{1\}\{2,3\} >$, $< \{1,2\}\{4\} >$, $< \{1,3\}\{4\} >$, $< \{1,2,4\} >$, $< \{2,3\}\{3\} >$, $< \{2,3\}\{4\} >$, $< \{2\}\{3\}\{3\} >$, and $< \{2\}\{3\}\{4\} >$.

(a) List all the candidate 4-sequences produced by the candidate generation step of the GSP algorithm.
Answer:

$< \{1,2,3\}\ \{3\} >$, $< \{1,2,3\}\ \{4\} >$, $< \{1,2\}\ \{3\}\ \{3\} >$, $< \{1,2\}\ \{3\}\ \{4\} >$, $< \{1\}\ \{2,3\}\ \{3\} >$, $< \{1\}\ \{2,3\}\ \{4\} >$.

(b) List all the candidate 4-sequences pruned during the candidate pruning step of the GSP algorithm (assuming no timing constraints).
Answer:

When there is no timing constraints, all subsequences of a candidate must be frequent. Therefore, the pruned candidates are:

$< \{1,2,3\}\ \{3\} >$, $< \{1,2\}\ \{3\}\ \{3\} >$, $< \{1,2\}\ \{3\}\ \{4\} >$,
$< \{1\}\ \{2,3\}\ \{3\} >$, $< \{1\}\ \{2,3\}\ \{4\} >$.

(c) List all the candidate 4-sequences pruned during the candidate pruning step of the GSP algorithm (assuming $maxgap = 1$).

Answer:

With timing constraint, only contiguous subsequences of a candidate must be frequent. Therefore, the pruned candidates are:

$< \{1,2,3\}\ \{3\} >$, $< \{1,2\}\ \{3\}\ \{3\} >$, $< \{1,2\}\ \{3\}\ \{4\} >$,
$< \{1\}\ \{2,3\}\ \{3\} >$, $< \{1\}\ \{2,3\}\ \{4\} >$.

14. Consider the data sequence shown in Table 6.16 for a given object. Count the number of occurrences for the sequence $\langle \{p\}\{q\}\{r\} \rangle$ according to the following counting methods:

Assume that $ws = 0$, $mingap = 0$, $maxgap = 3$, $maxspan = 5$).

Table 6.16. Example of event sequence data for Exercise 14.

Timestamp	Events
1	p, q
2	r
3	s
4	p, q
5	r, s
6	p
7	q, r
8	q, s
9	p
10	q, r, s

(a) COBJ (one occurrence per object).
Answer: 1.

(b) CWIN (one occurrence per sliding window).
Answer: 2.

(c) CMINWIN (number of minimal windows of occurrence).
Answer: 2.

(d) CDIST_O (distinct occurrences with possibility of event-timestamp overlap).
Answer: 3.

(e) CDIST (distinct occurrences with no event timestamp overlap allowed).
Answer: 2.

15. Describe the types of modifications necessary to adapt the frequent subgraph mining algorithm to handle:

    (a) Directed graphs

    (b) Unlabeled graphs

    (c) Acyclic graphs

    (d) Disconnected graphs

    For each type of graph given above, describe which step of the algorithm will be affected (candidate generation, candidate pruning, and support counting), and any further optimization that can help improve the efficiency of the algorithm.

    Answer:

    (a) Adjacency matrix may not be symmetric, which affects candidate generation using vertex growing approach.

    (b) An unlabeled graph is equivalent to a labeled graph where all the vertices have identical labels.

    (c) No effect on algorithm. If the graph is a rooted labeled tree, more efficient techniques can be developed to encode the tree (see: M.J. Zaki, Efficiently Mining Frequent Trees in a Forest, In Proc. of the Eighth ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining, 2002).

16. Draw all candidate subgraphs obtained from joining the pair of graphs shown in Figure 6.2. Assume the edge-growing method is used to expand the subgraphs.

    Answer: See Figure 6.3.

17. Draw all the candidate subgraphs obtained by joining the pair of graphs shown in Figure 6.4. Assume the edge-growing method is used to expand the subgraphs.

    Answer: See Figure 6.5.

18. Show that the candidate generation procedure introduced in Section 6.5.3 for frequent subgraph mining is complete, i.e., no frequent k-subgraph can be missed from being generated if every pair of frequent $(k-1)$-subgraphs is considered for merging.

    Answer: We will employ proof by contradiction to show that the candidate generation procedure for frequent subgraph mining is complete. Let us assume

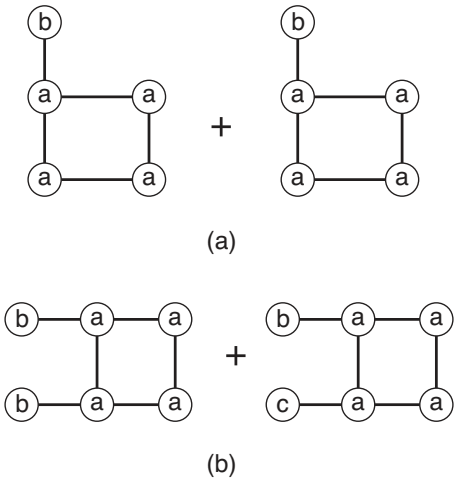


Figure 6.2. Graphs for Exercise 16.

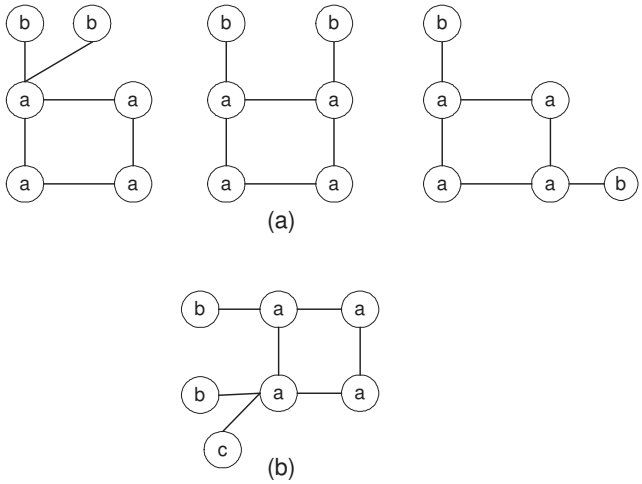


Figure 6.3. Solution for Exercise 16.

that there is a frequent k-subgraph G that got missed from being generated using the procedure described in Section 6.5.3 for merging frequent $(k-1)$-subgraphs. Let the set of edges in G be denoted as $E = \{e_1, e_2, \ldots, e_k\}$.

We construct k subgraphs of G of size $k-1$ by removing a single edge from E. Specifically, let G_{-i} be the $(k-1)$-subgraph constructed by removing edge e_i from E (with edges $E_{-i} = E \setminus \{e_i\}$). Since G is frequent, every one of its $(k-1)$-subgraphs is also frequent. Further, notice that there exists at least

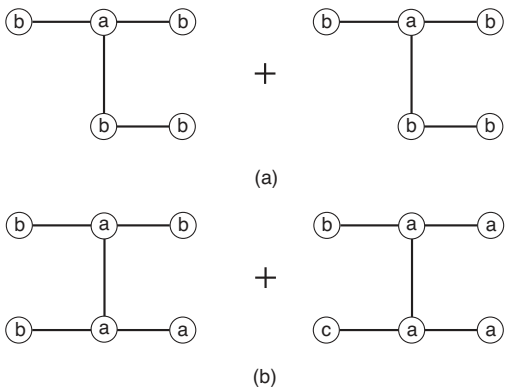


Figure 6.4. Graphs for Exercise 17.

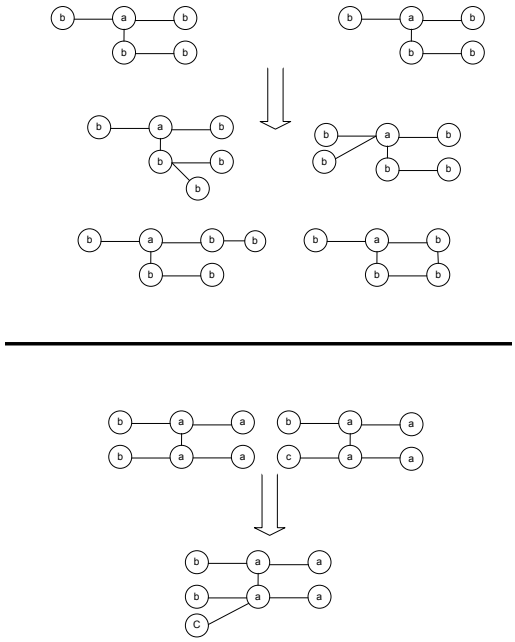


Figure 6.5. Solution for Exercise 17.

two $(k-1)$-subgraphs of G that are not disconnected. This is because even in the extreme case where G is a chain of edges, removing the first edge e_1 or removing the last edge e_k still results in connected graphs. Hence, let us consider a pair of connected $(k-1)$-subgraphs, G_{-i} and G_{-j}, obtained by removing e_i and e_j from E respectively. For notational convenience, let us

denote the set of edges E as $E = E_c \cup \{e_i\} \cup \{e_j\}$, where E_c is the common set of (core) edges in both G_{-i} and G_{-j}. In particular,

$$E_{-i} = E_c \cup e_j$$
$$E_{-j} = E_c \cup e_i$$

Further, let us denote the edge e_i as (u, u') and edge e_j as (v, v'), where u, u', v, v' are vertices of G. Since we have assumed that G was missed by our candidate generation procedure, it suffices to show that we would obtain G by merging G_{-i} and G_{-j} to arrive at a contradiction. To do this, let us consider all possible configurations of (u, u') and (v, v'). First, if $u = v$, then $u' \neq v'$, so that e_i and e_j are distinct edges. In this case, merging G_{-i} and G_{-j} using the procedure described in Figure 6.17(a) would result in G. On the other hand, if $u \neq v$, then we will have two different cases: either $u' = v'$ or $u' \neq v'$. In both these cases, we would obtain graph G by using the subgraph merging procedure described in Figure 6.17(b). Hence, for all possible configurations of e_i and e_j, we would obtain G by merging at least one pair of frequent $(k-1)$-subgraphs of G. Hence, using proof by contradiction, the candidate generation procedure described in Section 6.5.3 is complete and does not miss the generation of any frequent k-subgraph.

19. (a) If support is defined in terms of induced subgraph relationship, show that the confidence of the rule $g_1 \longrightarrow g_2$ can be greater than 1 if g_1 and g_2 are allowed to have overlapping vertex sets.

Answer:

We illustrate this with an example. Consider the five graphs, G_1, G_2, $\cdots$, G_5, shown in Figure 6.6. The graph g_1 shown on the top-right hand diagram is a subgraph of G_1, G_3, G_4, and G_5. Therefore, $s(g_1) = 4/5 = 80\%$. Similarly, we can show that $s(g_2) = 60\%$ because g_2 is a subgraph of G_1, G_2, and G_3, while $s(g_3) = 40\%$ because g_3 is a subgraph of G_1 and G_3.

Consider the association rule, $g_2 \longrightarrow g_1$. Using the standard definition of confidence as the ratio between the support of $g_2 \cup g_1 \equiv g_3$ to the support of g_2, we obtain a confidence value greater than 1 because $s(g_3) > s(g_2)$.

(b) What is the time complexity needed to determine the canonical label of a graph that contains $|V|$ vertices?

Answer:

A naïve approach requires $|V|!$ computations to examine all possible permutations of the canonical label.

(c) The core of a subgraph can have multiple automorphisms. This will increase the number of candidate subgraphs obtained after merging two frequent subgraphs that share the same core. Determine the maximum number of candidate subgraphs obtained due to automorphism of a core of size k.

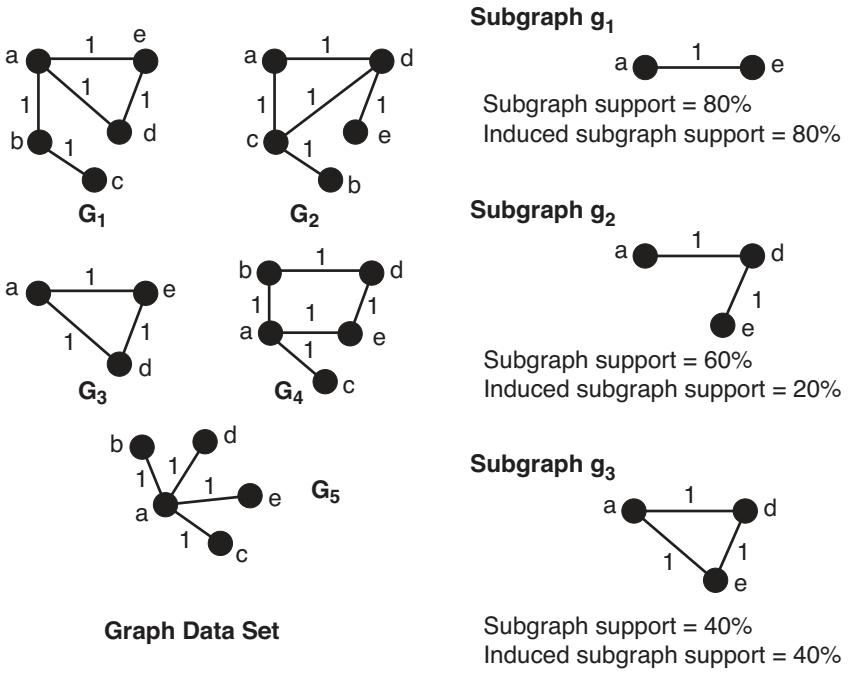


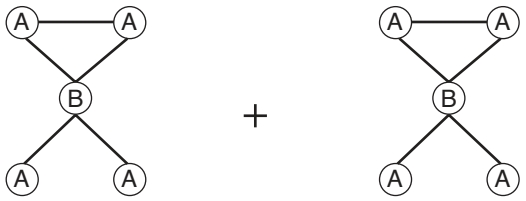
Figure 6.6. Computing the support of a subgraph from a set of graphs.

Answer: k.

(d) Two frequent subgraphs of size k may share multiple cores. Determine the maximum number of cores that can be shared by the two frequent subgraphs.

Answer: $k - 1$.

20. (a) Consider the two graphs shown below.



i. Draw all the distinct cores obtained when merging the two subgraphs.

Answer: See Figure 6.7.

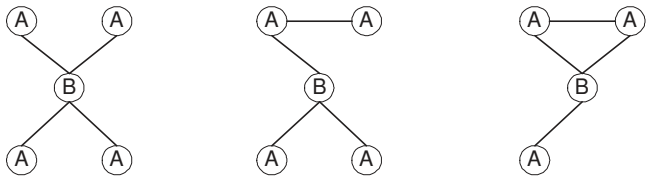
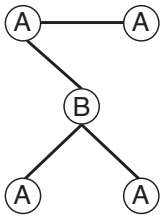


Figure 6.7. Solution to Exercise 20.

ii. How many candidates are generated using the following core?



Answer: No candidate $k+1$-subgraph can be generated from the core.

21. The original association rule mining framework considers only presence of items together in the same transaction. There are situations in which itemsets that are infrequent may also be informative. For instance, the itemset TV, DVD, ¬ VCR suggests that many customers who buy TVs and DVDs do not buy VCRs.

    In this problem, you are asked to extend the association rule framework to negative itemsets (i.e., itemsets that contain both presence and absence of items). We will use the negation symbol (¬) to refer to absence of items.

    (a) A naïve way for deriving negative itemsets is to extend each transaction to include absence of items as shown in Table 6.17.

Table 6.17. Example of numeric data set.

TID	TV	¬TV	DVD	¬DVD	VCR	¬VCR	...
1	1	0	0	1	0	1	...
2	1	0	0	1	0	1	...

i. Suppose the transaction database contains 1000 distinct items. What is the total number of positive itemsets that can be gener-

ated from these items? (Note: A positive itemset does not contain any negated items).
Answer: $2^{1000} - 1$.

ii. What is the maximum number of frequent itemsets that can be generated from these transactions? (Assume that a frequent itemset may contain positive, negative, or both types of items)
Answer: $2^{2000} - 1$.

iii. Explain why such a naïve method of extending each transaction with negative items is not practical for deriving negative itemsets.
Answer: The number of candidate itemsets is too large, many of the them are also redundant and useless (e.g., an itemset that contains both items x and $\overline{x}$).

(b) Consider the database shown in Table 6.14. What are the support and confidence values for the following negative association rules involving regular and diet soda?

i. ¬Regular $\longrightarrow$ Diet.
Answer: $s = 42.9\%$, $c = 100\%$.

ii. Regular $\longrightarrow$ ¬Diet.
Answer: $s = 42.9\%$, $c = 100\%$.

iii. ¬Diet $\longrightarrow$ Regular.
Answer: $s = 42.9\%$, $c = 100\%$.

iv. Diet $\longrightarrow$ ¬Regular.
Answer: $s = 42.9\%$, $c = 100\%$.

22. Suppose we would like to extract positive and negative itemsets from a data set that contains d items.

(a) Consider an approach where we introduce a new variable to represent each negative item. With this approach, the number of items grows from d to $2d$. What is the total size of the itemset lattice, assuming that an itemset may contain both positive and negative items of the same variable?
Answer: 2^{2d}.

(b) Assume that an itemset must contain positive or negative items of different variables. For example, the itemset $\{a, \overline{a}, b, \overline{c}\}$ is invalid because it contains both positive and negative items for variable a. What is the total size of the itemset lattice?
Answer: $\sum_{k=1}^{d} \binom{d}{k} \sum_{i=0}^{k} \binom{k}{i} = \sum_{k=1}^{d} \binom{d}{k} 2^k = 3^d - 1$.

23. For each type of pattern defined below, determine whether the support measure is monotone, anti-monotone, or non-monotone (i.e., neither monotone nor anti-monotone) with respect to increasing itemset size.

(a) Itemsets that contain both positive and negative items such as $\{a, b, \overline{c}, \overline{d}\}$. Is the support measure monotone, anti-monotone, or non-monotone when applied to such patterns?

Answer: Anti-monotone.

(b) Boolean logical patterns such as {(a $\vee$ b $\vee$ c), d, e}, which may contain both disjunctions and conjunctions of items. Is the support measure monotone, anti-monotone, or non-monotone when applied to such patterns?

Answer: Non-monotone.

24. Many association analysis algorithms rely on an *Apriori*-like approach for finding frequent patterns. The overall structure of the algorithm is given below.

Algorithm 6.1 *Apriori*-like algorithm.

1: $k = 1$.
2: $F_k = \{ i \mid i \in I \wedge \frac{\sigma(\{i\})}{N} \geq minsup\}$. {Find frequent 1-patterns.}
3: **repeat**
4: $k = k + 1$.
5: $C_k = \text{genCandidate}(F_{k-1})$. {Candidate Generation}
6: $C_k = \text{pruneCandidate}(C_k, F_{k-1})$. {Candidate Pruning}
7: $C_k = \text{count}(C_k, D)$. {Support Counting}
8: $F_k = \{ c \mid c \in C_k \wedge \frac{\sigma(c)}{N} \geq minsup\}$. {Extract frequent patterns}
9: **until** $F_k = \emptyset$
10: Answer $= \bigcup F_k$.

Suppose we are interested in finding boolean logical rules such as

$$\{a \vee b\} \longrightarrow \{c, d\},$$

which may contain both disjunctions and conjunctions of items. The corresponding itemset can be written as $\{(a \vee b), c, d\}$.

(a) Does the *Apriori* principle still hold for such itemsets?

(b) How should the candidate generation step be modified to find such patterns?

(c) How should the candidate pruning step be modified to find such patterns?

(d) How should the support counting step be modified to find such patterns?

Answer:

Refer to R. Srikant, Q. Vu, R. Agrawal: Mining Association Rules with Item Constraints. In Proc of the Third Int'l Conf on Knowledge Discovery and Data Mining, 1997.

7

Cluster Analysis: Basic Concepts and Algorithms

1. Consider a data set consisting of 2^{20} data vectors, where each vector has 32 components and each component is a 4-byte value. Suppose that vector quantization is used for compression and that 2^{16} prototype vectors are used. How many bytes of storage does that data set take before and after compression and what is the compression ratio?

   Before compression, the data set requires $4 \times 32 \times 2^{20} = 134,217,728$ bytes. After compression, the data set requires $4 \times 32 \times 2^{16} = 8,388,608$ bytes for the prototype vectors and $2 \times 2^{20} = 2,097,152$ bytes for vectors, since identifying the prototype vector associated with each data vector requires only two bytes. Thus, after compression, 10,485,760 bytes are needed to represent the data. The compression ratio is 12.8.

2. Find all well-separated clusters in the set of points shown in Figure 7.1.

   The solutions are also indicated in Figure 7.1.

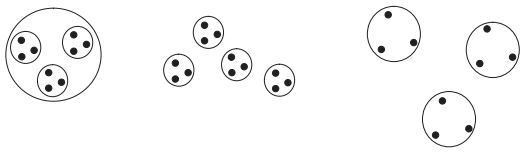


Figure 7.1. Points for Exercise 2.

3. Many partitional clustering algorithms that automatically determine the number of clusters claim that this is an advantage. List two situations in which this is not the case.

   (a) When there is hierarchical structure in the data. Most algorithms that automatically determine the number of clusters are partitional, and thus, ignore the possibility of subclusters.

   (b) When clustering for utility. If a certain reduction in data size is needed, then it is necessary to specify how many clusters (cluster centroids) are produced.

4. Given K equally sized clusters, the probability that a randomly chosen initial centroid will come from any given cluster is $1/K$, but the probability that each cluster will have exactly one initial centroid is much lower. (It should be clear that having one initial centroid in each cluster is a good starting situation for K-means.) In general, if there are K clusters and each cluster has n points, then the probability, p, of selecting in a sample of size K one initial centroid from each cluster is given by Equation 7.1. (This assumes sampling with replacement.) From this formula we can calculate, for example, that the chance of having one initial centroid from each of four clusters is $4!/4^4 = 0.0938$.

$$p = \frac{\text{number of ways to select one centroid from each cluster}}{\text{number of ways to select } K \text{ centroids}} = \frac{K!n^K}{(Kn)^K} = \frac{K!}{K^K} \quad (7.1)$$

   (a) Plot the probability of obtaining one point from each cluster in a sample of size K for values of K between 2 and 100.

   The solution is shown in Figure 4. Note that the probability is essentially 0 by the time $K = 10$.

   (b) For K clusters, $K = 10, 100$, and 1000, find the probability that a sample of size $2K$ contains at least one point from each cluster. You can use either mathematical methods or statistical simulation to determine the answer.

   We used simulation to compute the answer. Respectively, the probabilities are 0.21, $< 10^{-6}$, and $< 10^{-6}$.

   Proceeding analytically, the probability that a point doesn't come from a particular cluster is, $1-\frac{1}{K}$, and thus, the probability that all $2K$ points don't come from a particular cluster is $(1-\frac{1}{K})^{2K}$. Hence, the probability that at least one of the 200 points comes from a particular cluster is $1 - (1 - \frac{1}{K})^{2K}$. If we assume independence (which is too optimistic, but becomes approximately true for larger values of K), then an upper bound for the probability that all clusters are represented in the final sample is given by $(1 - (1 - \frac{1}{K})^{2K})^K$. The values given by this bound are 0.27, 5.7e-07, and 8.2e-64, respectively.

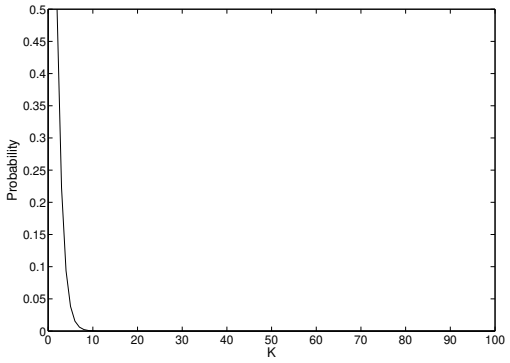
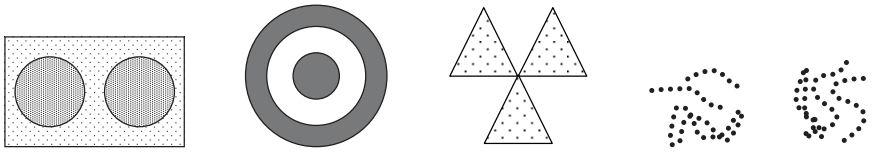


Figure 7.2. Probability of at least one point from each cluster. Exercise 4.

5. Identify the clusters in Figure 7.3 using the center-, contiguity-, and density-based definitions. Also indicate the number of clusters for each case and give a brief indication of your reasoning. Note that darkness or the number of dots indicates density. If it helps, assume center-based means K-means, contiguity-based means single link, and density-based means DBSCAN.



(a) (b) (c) (d)

Figure 7.3. Clusters for Exercise 5.

(a) **center-based** 2 clusters. The rectangular region will be split in half. Note that the noise is included in the two clusters.
contiguity-based 1 cluster because the two circular regions will be joined by noise.
density-based 2 clusters, one for each circular region. Noise will be eliminated.

(b) **center-based** 1 cluster that includes both rings.
contiguity-based 2 clusters, one for each rings.
density-based 2 clusters, one for each ring.

(c) **center-based** 3 clusters, one for each triangular region. One cluster is also an acceptable answer.
contiguity-based 1 cluster. The three triangular regions will be joined together because they touch.
density-based 3 clusters, one for each triangular region. Even though the three triangles touch, the density in the region where they touch is lower than throughout the interior of the triangles.

(d) **center-based** 2 clusters. The two groups of lines will be split in two.
contiguity-based 5 clusters. Each set of lines that intertwines becomes a cluster.
density-based 2 clusters. The two groups of lines define two regions of high density separated by a region of low density.

6. For the following sets of two-dimensional points, (1) provide a sketch of how they would be split into clusters by K-means for the given number of clusters and (2) indicate approximately where the resulting centroids would be. Assume that we are using the squared error objective function. If you think that there is more than one possible solution, then please indicate whether each solution is a global or local minimum. Note that the label of each diagram in Figure 7.4 matches the corresponding part of this question, e.g., Figure 7.4(a) goes with part (a).

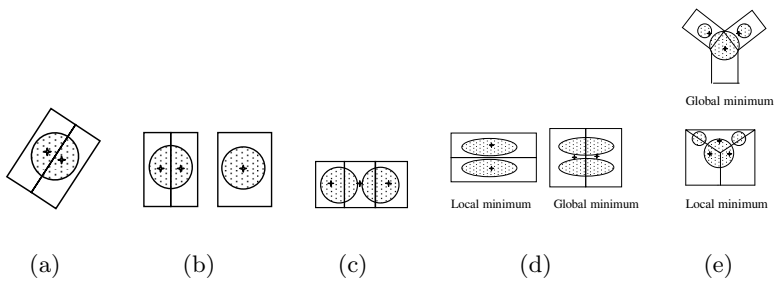


Figure 7.4. Diagrams for Exercise 6.

(a) $K = 2$. Assuming that the points are uniformly distributed in the circle, how many possible ways are there (in theory) to partition the points into two clusters? What can you say about the positions of the two centroids? (Again, you don't need to provide exact centroid locations, just a qualitative description.)

In theory, there are an infinite number of ways to split the circle into two clusters - just take any line that bisects the circle. This line can

make any angle $0° \leq \theta \leq 180°$ with the x axis. The centroids will lie on the perpendicular bisector of the line that splits the circle into two clusters and will be symmetrically positioned. All these solutions will have the same, globally minimal, error.

(b) $K = 3$. The distance between the edges of the circles is slightly greater than the radii of the circles.

If you start with initial centroids that are real points, you will necessarily get this solution because of the restriction that the circles are more than one radius apart. Of course, the bisector could have any angle, as above, and it could be the other circle that is split. All these solutions have the same globally minimal error.

(c) $K = 3$. The distance between the edges of the circles is much less than the radii of the circles.

The three boxes show the three clusters that will result in the realistic case that the initial centroids are actual data points.

(d) $K = 2$.

In both case, the rectangles show the clusters. In the first case, the two clusters are only a local minimum while in the second case the clusters represent a globally minimal solution.

(e) $K = 3$. Hint: Use the symmetry of the situation and remember that we are looking for a rough sketch of what the result would be.

For the solution shown in the top figure, the two top clusters are enclosed in two boxes, while the third cluster is enclosed by the regions defined by a triangle and a rectangle. (The two smaller clusters in the drawing are supposed to be symmetrical.) I believe that the second solution—suggested by a student—is also possible, although it is a local minimum and might rarely be seen in practice for this configuration of points. Note that while the two pie shaped cuts out of the larger circle are shown as meeting at a point, this is not necessarily the case—it depends on the exact positions and sizes of the circles. There could be a gap between the two pie shaped cuts which is filled by the third (larger) cluster. (Imagine the small circles on opposite sides.) Or the boundary between the two pie shaped cuts could actually be a line segment.

7. Suppose that for a data set

- there are m points and K clusters,
- half the points and clusters are in "more dense" regions,
- half the points and clusters are in "less dense" regions, and
- the two regions are well-separated from each other.

For the given data set, which of the following should occur in order to minimize the squared error when finding K clusters:

(a) Centroids should be equally distributed between more dense and less dense regions.

(b) More centroids should be allocated to the less dense region.

(c) More centroids should be allocated to the denser region.

Note: Do not get distracted by special cases or bring in factors other than density. However, if you feel the true answer is different from any given above, justify your response.

The correct answer is (c). Less dense regions require more centroids if the squared error is to be minimized.

8. Consider the mean of a cluster of objects from a binary transaction data set. What are the minimum and maximum values of the components of the mean? What is the interpretation of components of the cluster mean? Which components most accurately characterize the objects in the cluster?

   (a) The components of the mean range between 0 and 1.

   (b) For any specific component, its value is the fraction of the objects in the cluster that have a 1 for that component. If we have asymmetric binary data, such as market basket data, then this can be viewed as the probability that, for example, a customer in group represented by the the cluster buys that particular item.

   (c) This depends on the type of data. For binary asymmetric data, the components with higher values characterize the data, since, for most clusters, the vast majority of components will have values of zero. For regular binary data, such as the results of a true-false test, the significant components are those that are unusually high or low with respect to the entire set of data.

9. Give an example of a data set consisting of three natural clusters, for which (almost always) K-means would likely find the correct clusters, but bisecting K-means would not.

   Consider a data set that consists of three circular clusters, that are identical in terms of the number and distribution of points, and whose centers lie on a line and are located such that the center of the middle cluster is equally distant from the other two. Bisecting K-means would always split the middle cluster during its first iteration, and thus, could never produce the correct set of clusters. (Postprocessing could be applied to address this.)

10. Would the cosine measure be the appropriate similarity measure to use with K-means clustering for time series data? Why or why not? If not, what similarity measure would be more appropriate?

Time series data is dense high-dimensional data, and thus, the cosine measure would not be appropriate since the cosine measure is appropriate for sparse data. If the magnitude of a time series is important, then Euclidean distance would be appropriate. If only the shapes of the time series are important, then correlation would be appropriate. Note that if the comparison of the time series needs to take in account that one time series might lead or lag another or only be related to another during specific time periods, then more sophisticated approaches to modeling time series similarity must be used.

11. Total SSE is the sum of the SSE for each separate attribute. What does it mean if the SSE for one variable is low for all clusters? Low for just one cluster? High for all clusters? High for just one cluster? How could you use the per variable SSE information to improve your clustering?

    (a) If the SSE of one attribute is low for all clusters, then the variable is essentially a constant and of little use in dividing the data into groups.

    (b) if the SSE of one attribute is relatively low for just one cluster, then this attribute helps define the cluster.

    (c) If the SSE of an attribute is relatively high for all clusters, then it could well mean that the attribute is noise.

    (d) If the SSE of an attribute is relatively high for one cluster, then it is at odds with the information provided by the attributes with low SSE that define the cluster. It could merely be the case that the clusters defined by this attribute are different from those defined by the other attributes, but in any case, it means that this attribute does not help define the cluster.

    (e) The idea is to eliminate attributes that have poor distinguishing power between clusters, i.e., low or high SSE for all clusters, since they are useless for clustering. Note that attributes with high SSE for all clusters are particularly troublesome if they have a relatively high SSE with respect to other attributes (perhaps because of their scale) since they introduce a lot of noise into the computation of the overall SSE.

12. The leader algorithm (Hartigan [4]) represents each cluster using a point, known as a *leader*, and assigns each point to the cluster corresponding to the closest leader, unless this distance is above a user-specified threshold. In that case, the point becomes the leader of a new cluster.

    Note that the algorithm described here is not quite the leader algorithm described in Hartigan, which assigns a point to the first leader that is within the threshold distance. The answers apply to the algorithm as stated in the problem.

    (a) What are the advantages and disadvantages of the leader algorithm as compared to K-means?

The leader algorithm requires only a single scan of the data and is thus more computationally efficient since each object is compared to the final set of centroids at most once. Although the leader algorithm is order dependent, for a fixed ordering of the objects, it always produces the same set of clusters. However, unlike K-means, it is not possible to set the number of resulting clusters for the leader algorithm, except indirectly. Also, the K-means algorithm almost always produces better quality clusters as measured by SSE.

(b) Suggest ways in which the leader algorithm might be improved.

Use a sample to determine the distribution of distances between the points. The knowledge gained from this process can be used to more intelligently set the value of the threshold.

The leader algorithm could be modified to cluster for several thresholds during a single pass.

13. The Voronoi diagram for a set of K points in the plane is a partition of all the points of the plane into K regions, such that every point (of the plane) is assigned to the closest point among the K specified points. (See Figure 7.5.) What is the relationship between Voronoi diagrams and K-means clusters? What do Voronoi diagrams tell us about the possible shapes of K-means clusters?

(a) If we have K K-means clusters, then the plane is divided into K Voronoi regions that represent the points closest to each centroid.

(b) The boundaries between clusters are piecewise linear. It is possible to see this by drawing a line connecting two centroids and then drawing a perpendicular to the line halfway between the centroids. This perpendicular line splits the plane into two regions, each containing points that are closest to the centroid the region contains.

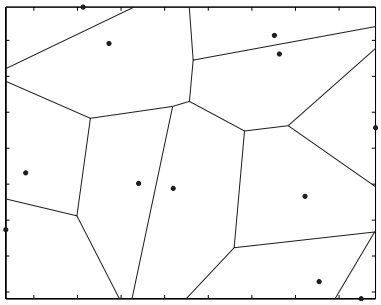


Figure 7.5. Voronoi diagram for Exercise 13.

14. You are given a data set with 100 records and are asked to cluster the data. You use K-means to cluster the data, but for all values of K, $1 \leq K \leq 100$, the K-means algorithm returns only one non-empty cluster. You then apply an incremental version of K-means, but obtain exactly the same result. How is this possible? How would single link or DBSCAN handle such data?

    (a) The data consists completely of duplicates of one object.

    (b) Single link (and many of the other agglomerative hierarchical schemes) would produce a hierarchical clustering, but which points appear in which cluster would depend on the ordering of the points and the exact algorithm. However, if the dendrogram were plotted showing the proximity at which each object is merged, then it would be obvious that the data consisted of duplicates. DBSCAN would find that all points were core points connected to one another and produce a single cluster.

15. Traditional agglomerative hierarchical clustering routines merge two clusters at each step. Does it seem likely that such an approach accurately captures the (nested) cluster structure of a set of data points? If not, explain how you might postprocess the data to obtain a more accurate view of the cluster structure.

    (a) Such an approach does not accurately capture the nested cluster structure of the data. For example, consider a set of three clusters, each of which has two, three, and four subclusters, respectively. An ideal hierarchical clustering would have three branches from the root—one to each of the three main clusters—and then two, three, and four branches from each of these clusters, respectively. A traditional agglomerative approach cannot produce such a structure.

    (b) The simplest type of postprocessing would attempt to flatten the hierarchical clustering by moving clusters up the tree.

16. Use the similarity matrix in Table 7.1 to perform single and complete link hierarchical clustering. Show your results by drawing a dendrogram. The dendrogram should clearly show the order in which the points are merged.

    The solutions are shown in Figures 7.6(a) and 7.6(b).

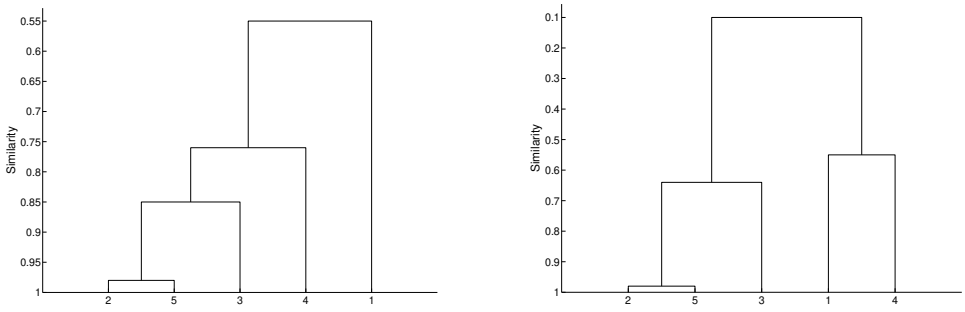
17. Hierarchical clustering is sometimes used to generate K clusters, $K > 1$ by taking the clusters at the K^{th} level of the dendrogram. (Root is at level 1.) By looking at the clusters produced in this way, we can evaluate the behavior of hierarchical clustering on different types of data and clusters, and also compare hierarchical approaches to K-means.

    The following is a set of one-dimensional points: $\{6, 12, 18, 24, 30, 42, 48\}$.

    (a) For each of the following sets of initial centroids, create two clusters by assigning each point to the nearest centroid, and then calculate the

Table 7.1. Similarity matrix for Exercise 16.

	p1	p2	p3	p4	p5
p1	1.00	0.10	0.41	0.55	0.35
p2	0.10	1.00	0.64	0.47	0.98
p3	0.41	0.64	1.00	0.44	0.85
p4	0.55	0.47	0.44	1.00	0.76
p5	0.35	0.98	0.85	0.76	1.00



(a) Single link. (b) Complete link.

Figure 7.6. Dendrograms for Exercise 16.

total squared error for each set of two clusters. Show both the clusters and the total squared error for each set of centroids.

i. $\{18, 45\}$
First cluster is 6, 12, 18, 24, 30.
Error = 360.
Second cluster is 42, 48.
Error = 18.
Total Error = 378

ii. $\{15, 40\}$ First cluster is 6, 12, 18, 24 .
Error = 180.
Second cluster is 30, 42, 48.
Error = 168.
Total Error = 348.

(b) Do both sets of centroids represent stable solutions; i.e., if the K-means algorithm was run on this set of points using the given centroids as the starting centroids, would there be any change in the clusters generated?

Yes, both centroids are stable solutions.

(c) What are the two clusters produced by single link?

The two clusters are {6, 12, 18, 24, 30} and {42, 48}.

(d) Which technique, K-means or single link, seems to produce the "most natural" clustering in this situation? (For K-means, take the clustering with the lowest squared error.)

MIN produces the most natural clustering.

(e) What definition(s) of clustering does this natural clustering correspond to? (Well-separated, center-based, contiguous, or density.)

MIN produces contiguous clusters. However, density is also an acceptable answer. Even center-based is acceptable, since one set of centers gives the desired clusters.

(f) What well-known characteristic of the K-means algorithm explains the previous behavior?

K-means is not good at finding clusters of different sizes, at least when they are not well separated. The reason for this is that the objective of minimizing squared error causes it to "break" the larger cluster. Thus, in this problem, the low error clustering solution is the "unnatural" one.

18. Suppose we find K clusters using Ward's method, bisecting K-means, and ordinary K-means. Which of these solutions represents a local or global minimum? Explain.

Although Ward's method picks a pair of clusters to merge based on minimizing SSE, there is no refinement step as in regular K-means. Likewise, bisecting K-means has no overall refinement step. Thus, unless such a refinement step is added, neither Ward's method nor bisecting K-means produces a local minimum. Ordinary K-means produces a local minimum, but like the other two algorithms, it is not guaranteed to produce a global minimum.

19. Hierarchical clustering algorithms require $O(m^2 \log(m))$ time, and consequently, are impractical to use directly on larger data sets. One possible technique for reducing the time required is to sample the data set. For example, if K clusters are desired and $\sqrt{m}$ points are sampled from the m points, then a hierarchical clustering algorithm will produce a hierarchical clustering in roughly $O(m)$ time. K clusters can be extracted from this hierarchical clustering by taking the clusters on the K^{th} level of the dendrogram. The remaining points can then be assigned to a cluster in linear time, by using various strategies. To give a specific example, the centroids of the K clusters can be computed, and then each of the $m - \sqrt{m}$ remaining points can be assigned to the cluster associated with the closest centroid.

For each of the following types of data or clusters, discuss briefly if (1) sampling will cause problems for this approach and (2) what those problems are. Assume that the sampling technique randomly chooses points from the total set of m points and that any unmentioned characteristics of the data or clusters are as optimal as possible. In other words, focus only on problems caused by the particular characteristic mentioned. Finally, assume that K is very much less than m.

(a) Data with very different sized clusters.

This can be a problem, particularly if the number of points in a cluster is small. For example, if we have a thousand points, with two clusters, one of size 900 and one of size 100, and take a 5% sample, then we will, on average, end up with 45 points from the first cluster and 5 points from the second cluster. Five points is much easier to miss or cluster improperly than 50. Also, the second cluster will sometimes be represented by fewer than 5 points, just by the nature of random samples.

(b) High-dimensional data.

This can be a problem because data in high-dimensional space is typically sparse and more points may be needed to define the structure of a cluster in high-dimensional space.

(c) Data with outliers, i.e., atypical points.

By definition, outliers are not very frequent and most of them will be omitted when sampling. Thus, if finding the correct clustering depends on having the outliers present, the clustering produced by sampling will likely be misleading. Otherwise, it is beneficial.

(d) Data with highly irregular regions.

This can be a problem because the structure of the border can be lost when sampling unless a large number of points are sampled.

(e) Data with globular clusters.

This is typically not a problem since not as many points need to be sampled to retain the structure of a globular cluster as an irregular one.

(f) Data with widely different densities.

In this case the data will tend to come from the denser region. Note that the effect of sampling is to reduce the density of all clusters by the sampling factor, e.g., if we take a 10% sample, then the density of the clusters is decreased by a factor of 10. For clusters that aren't very dense to begin with, this may means that they are now treated as noise or outliers.

(g) Data with a small percentage of noise points.

Sampling will not cause a problem. Actually, since we would like to exclude noise, and since the amount of noise is small, this may be beneficial.

(h) Non-Euclidean data.

This has no particular impact.

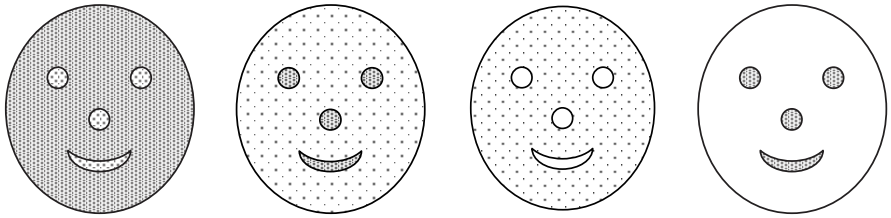
(i) Euclidean data.

This has no particular impact.

(j) Data with many and mixed attribute types.

Many attributes was discussed under high-dimensionality. Mixed attributes have no particular impact.

20. Consider the following four faces shown in Figure 7.7. Again, darkness or number of dots represents density. Lines are used only to distinguish regions and do not represent points.



(a) (b) (c) (d)

Figure 7.7. Figure for Exercise 20.

(a) For each figure, could you use single link to find the patterns represented by the nose, eyes, and mouth? Explain.

Only for (b) and (d). For (b), the points in the nose, eyes, and mouth are much closer together than the points between these areas. For (d) there is only space between these regions.

(b) For each figure, could you use K-means to find the patterns represented by the nose, eyes, and mouth? Explain.

Only for (b) and (d). For (b), K-means would find the nose, eyes, and mouth, but the lower density points would also be included. For (d), K-

means would find the nose, eyes, and mouth straightforwardly as long as the number of clusters was set to 4.

(c) What limitation does clustering have in detecting all the patterns formed by the points in Figure 7.7(c)?

Clustering techniques can only find patterns of points, not of empty spaces.

21. Compute the entropy and purity for the confusion matrix in Table 7.2.

Table 7.2. Confusion matrix for Exercise 21.

Cluster	Entertainment	Financial	Foreign	Metro	National	Sports	Total	Entropy	Purity
#1	1	1	0	11	4	676	693	0.20	0.98
#2	27	89	333	827	253	33	1562	1.84	0.53
#3	326	465	8	105	16	29	949	1.70	0.49
Total	354	555	341	943	273	738	3204	1.44	0.61

22. You are given two sets of 100 points that fall within the unit square. One set of points is arranged so that the points are uniformly spaced. The other set of points is generated from a uniform distribution over the unit square.

(a) Is there a difference between the two sets of points?

Yes. The random points will have regions of lesser or greater density, while the uniformly distributed points will, of course, have uniform density throughout the unit square.

(b) If so, which set of points will typically have a smaller SSE for K=10 clusters?

The random set of points will have a lower SSE.

(c) What will be the behavior of DBSCAN on the uniform data set? The random data set?

DBSCAN will merge all points in the uniform data set into one cluster or classify them all as noise, depending on the threshold. There might be some boundary issues for points at the edge of the region. However, DBSCAN can often find clusters in the random data, since it does have some variation in density.

23. Using the data in Exercise 24, compute the silhouette coefficient for each point, each of the two clusters, and the overall clustering.

Cluster 1 contains {P1, P2}, Cluster 2 contains {P3, P4}. The dissimilarity matrix that we obtain from the similarity matrix is the following:

Table 7.3. Table of distances for Exercise 23

	P1	P2	P3	P4
P1	0	0.10	0.65	0.55
P2	0.10	0	0.70	0.60
P3	0.65	0.70	0	0.30
P4	0.55	0.60	0.30	0

.

Let a indicate the average distance of a point to other points in its cluster. Let b indicate the minimum of the average distance of a point to points in another cluster.

Point P1: SC = 1- a/b = 1 - 0.1/((0.65+0.55)/2)= 5/6 = 0.833
Point P2: SC = 1- a/b = 1 - 0.1/((0.7+0.6)/2) = 0.846
Point P2: SC = 1- a/b = 1 - 0.3/((0.65+0.7)/2) = 0.556
Point P2: SC = 1- a/b = 1 - 0.3/((0.55+0.6)/2) = 0.478

Cluster 1 Average SC = (0.833+0.846)/2 = 0.84
Cluster 2 Average SC = (0.556+0.478)/2 = 0.52
Overall Average SC = (0.840+0.517)/2 = 0.68

24. Given the set of cluster labels and similarity matrix shown in Tables 7.4 and 7.5, respectively, compute the correlation between the similarity matrix and the ideal similarity matrix, i.e., the matrix whose ij^{th} entry is 1 if two objects belong to the same cluster, and 0 otherwise.

Table 7.4. Table of cluster labels for Exercise 24.

Point	Cluster Label
P1	1
P2	1
P3	2
P4	2

Table 7.5. Similarity matrix for Exercise 24.

Point	P1	P2	P3	P4
P1	1	0.8	0.65	0.55
P2	0.8	1	0.7	0.6
P3	0.65	0.7	1	0.9
P4	0.55	0.6	0.9	1

We need to compute the correlation between the vector $\mathbf{x} =< 1, 0, 0, 0, 0, 1 >$ and the vector $\mathbf{y} =< 0.8, 0.65, 0.55, 0.7, 0.6, 0.3 >$, which is the correlation between the off-diagonal elements of the distance matrix and the ideal similarity matrix.

We get:
Standard deviation of the vector $\mathbf{x} : \sigma_x = 0.5164$
Standard deviation of the vector $\mathbf{y} : \sigma_y = 0.1703$
Covariance of $\mathbf{x}$ and $\mathbf{y}$: $\text{cov}(\mathbf{x}, \mathbf{y}) = -0.200$

Therefore, $\text{corr}(\mathbf{x}, \mathbf{y}) = \text{cov}(\mathbf{x}, \mathbf{y})/\sigma_x\sigma_y = -0.227$

25. Compute the hierarchical F-measure for the eight objects {p1, p2, p3, p4, p5, p6, p7, p8} and hierarchical clustering shown in Figure 7.8. Class A contains points p1, p2, and p3, while p4, p5, p6, p7, and p8 belong to class B.

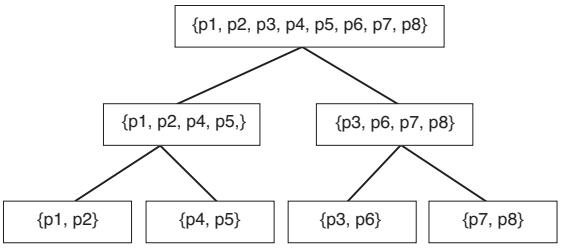


Figure 7.8. Hierarchical clustering for Exercise 25.

Let $R(i,j) = n_{ij}/n_i$ indicate the recall of class i with respect to cluster j.
Let $P(i,j) = n_{ij}/n_j$ indicate the precision of class i with respect to cluster j.
$F(i,j) = 2R(i,j) \times P(i,j)/(R(i,j) + P(i,j))$ is the F-measure for class i and cluster j.

For cluster $\#1 = \{p1, p2, p3, p4, p5, p6, p7, p8\}$:
Class = A:
$R(A,1) = 3/3 = 1,\ P(A,1) = 3/8 = 0.375$
$F(A,1) = 2 \times 1 \times 0.375/(1 + 0.375) = 0.55$
Class = B:
$R(B,1) = 5/5 = 1,\ P(A,1) = 5/8 = 0.625,\ F(A,1) = 0.77$

For cluster #2= {p1,p2,p4,p5}
Class = A:
$R(A,2) = 2/3, P(A,2) = 2/4, F(A,2) = 0.57$
Class = B:
$R(B,2) = 2/5, P(B,2) = 2/4, F(B,2) = 0.44$

For cluster #3= {p3, p6, p7, p8}
Class = A:
$R(A,3) = 1/3,\ P(A,3) = 1/4,\ F(A,3) = 0.29$
Class =B:
$R(B,3) = 3/5, P(B,3) = 3/4, F(B,3) = 0.67$

For cluster #4={p1, p2}
Class = A:

$R(A,4) = 2/3,\ P(A,4) = 2/2,\ F(A,4) = 0.8$
Class =B:
$R(B,4) = 0/5,\ P(B,4) = 0/2,\ F(B,4) = 0$

For cluster #5 = {p4, p5}
Class = A:
$R(A,5) = 0,\ P(A,5) = 0,\ F(A,5) = 0$
Class =B:
$R(B,5) = 2/5,\ P(B,5) = 2/2,\ F(B,5) = 0.57$

For cluster #6 = {p3, p6}
Class = A:
$R(A,6) = 1/3,\ P(A,6) = 1/2,\ F(A,6) = 0.4$
Class =B:
$R(B,6) = 1/5,\ P(B,6) = 1/2,\ F(B,6) = 0.29$

For cluster #7 = {p7, p8}
Class = A:
$R(A,7) = 0,\ P(A,7) = 1,\ F(A,7) = 0$
Class = B:
$R(B,7) = 2/5,\ P(B,7) = 2/2,\ F(B,7) = 0.57$

Class A: $F(A) = \max\{F(A,j)\} = \max\{0.55, 0.57, 0.29, 0.8, 0, 0.4, 0\} = 0.8$
Class B: $F(B) = \max\{F(B,j)\} = \max\{0.77, 0.44, 0.67, 0, 0.57, 0.29, 0.57\} = 0.77$

Overall Clustering: $F = \sum_1^2 \frac{n_i}{n} \max_i F(i,j) = 3/8 * F(A) + 5/8 * F(B) = 0.78$

26. Compute the cophenetic correlation coefficient for the hierarchical clusterings in Exercise 16. (You will need to convert the similarities into dissimilarities.)

    This can be easily computed using a package, e.g., MATLAB. The answers are single link, 0.8116, and complete link, 0.7480.

27. Prove Equation 7.14.

$$\begin{aligned}
\frac{1}{2|C_i|}\sum_{x\in C_i}\sum_{y\in C_i}(x-y)^2 &= \frac{1}{2|C_i|}\sum_{x\in C_i}\sum_{y\in C_i}((x-c_i)-(y-c_i))^2 \\
&= \frac{1}{2|C_i|}\Big(\sum_{x\in C_i}\sum_{y\in C_i}(x-c_i)^2 - 2\sum_{x\in C_i}\sum_{y\in C_i}(x-c_i)(y-c_i) \\
&+ \sum_{x\in C_i}\sum_{y\in C_i}(y-c_i)^2\Big) \\
&= \frac{1}{2|C_i|}\Big(\sum_{x\in C_i}\sum_{y\in C_i}(x-c_i)^2 + \sum_{x\in C_i}\sum_{y\in C_i}(y-c_i)^2\Big) \\
&= \frac{1}{|C_i|}\sum_{x\in C_i}|C_i|(x-c_i)^2 \\
&= \text{SSE}
\end{aligned}$$

The cross term $\sum_{x\in C_i}\sum_{y\in C_i}(x-c_i)(y-c_i)$ is 0.

28. Prove Equation 7.16.

$$\begin{aligned}
\frac{1}{K}\sum_{i=1}^{K}\sum_{j=1}^{K}|C_i|(c_j-c_i)^2 &= \frac{1}{2K}\sum_{i=1}^{K}\sum_{j=1}^{K}|C_i|((m-c_i)-(m-c_j))^2 \\
&= \frac{1}{2K}\Big(\sum_{i=1}^{K}\sum_{j=1}^{K}|C_i|(m-c_i)^2 - 2\sum_{i=1}^{K}\sum_{j=1}^{K}|C_i|(m-c_i)(m-c_j) \\
&+ \sum_{i=1}^{K}\sum_{j=1}^{K}|C_i|(m-c_j)^2\Big) \\
&= \frac{1}{2K}\Big(\sum_{i=1}^{K}\sum_{j=1}^{K}|C_i|(m-c_i)^2 + \sum_{i=1}^{K}\sum_{j=1}^{K}|C_i|(m-c_j)^2\Big) \\
&= \frac{1}{K}\sum_{i=1}^{K}K|C_i|(m-c_i)^2 \\
&= \text{SSB}
\end{aligned}$$

Again, the cross term cancels.

29. Prove that $\sum_{i=1}^{K}\sum_{x\in C_i}(x-m_i)(m-m_i)=0$. This fact was used in the proof that TSS = SSE + SSB in Section 7.5.2.

$$\begin{aligned}
\sum_{i=1}^{K}\sum_{x\in C_i}(x-c_i)(c-c_i) &= \sum_{i=1}^{K}\sum_{x\in C_i}(xc - c_ic - xc_i + c_i^2) \\
&= \sum_{i=1}^{K}\sum_{x\in C_i} xc - \sum_{i=1}^{K}\sum_{x\in C_i} c_ic - \sum_{i=1}^{K}\sum_{x\in C_i} xc_i + \sum_{i=1}^{K}\sum_{x\in C_i} c_i^2 \\
&= m_ic_ic - m_ic_ic - m_ic_i^2 + m_ic_i^2 \\
&= 0
\end{aligned}$$

30. Clusters of documents can be summarized by finding the top terms (words) for the documents in the cluster, e.g., by taking the most frequent k terms, where k is a constant, say 10, or by taking all terms that occur more frequently than a specified threshold. Suppose that K-means is used to find clusters of both documents and words for a document data set.

    (a) How might a set of term clusters defined by the top terms in a document cluster differ from the word clusters found by clustering the terms with K-means?

        First, the top words clusters could, and likely would, overlap somewhat. Second, it is likely that many terms would not appear in any of the clusters formed by the top terms. In contrast, a K-means clustering of the terms would cover all the terms and would not be overlapping.

    (b) How could term clustering be used to define clusters of documents?

        An obvious approach would be to take the top documents for a term cluster; i.e., those documents that most frequently contain the terms in the cluster.

31. We can represent a data set as a collection of object nodes and a collection of attribute nodes, where there is a link between each object and each attribute, and where the weight of that link is the value of the object for that attribute. For sparse data, if the value is 0, the link is omitted. Bipartite clustering attempts to partition this graph into disjoint clusters, where each cluster consists of a set of object nodes and a set of attribute nodes. The objective is to maximize the weight of links between the object and attribute nodes of a cluster, while minimizing the weight of links between object and attribute links in different clusters. This type of clustering is also known as **co-clustering** since the objects and attributes are clustered at the same time.

    (a) How is bipartite clustering (co-clustering) different from clustering the sets of objects and attributes separately?

        In regular clustering, only one set of constraints, related either to objects or attributes, is applied. In co-clustering both sets of constraints are

applied simultaneously. Thus, partitioning the objects and attributes independently of one another typically does not produce the same results.

(b) Are there any cases in which these approaches yield the same clusters?

Yes. For example, if a set of attributes is associated only with the objects in one particular cluster, i.e., has 0 weight for objects in all other clusters, and conversely, the set of objects in a cluster has 0 weight for all other attributes, then the clusters found by co-clustering will match those found by clustering the objects and attributes separately. To use documents as an example, this would correspond to a document data set that consists of groups of documents that only contain certain words and groups of words that only appear in certain documents.

(c) What are the strengths and weaknesses of co-clustering as compared to ordinary clustering?

Co-clustering automatically provides a description of a cluster of objects in terms of attributes, which can be more useful than a description of clusters as a partitioning of objects. However, the attributes that distinguish different clusters of objects, may overlap significantly, and in such cases, co-clustering will not work well.

32. In Figure 7.9, match the similarity matrices, which are sorted according to cluster labels, with the sets of points. Differences in shading and marker shape distinguish between clusters, and each set of points contains 100 points and three clusters. In the set of points labeled 2, there are three very tight, equal-sized clusters.

Answers: 1 - D, 2 - C, 3 - A, 4 - B

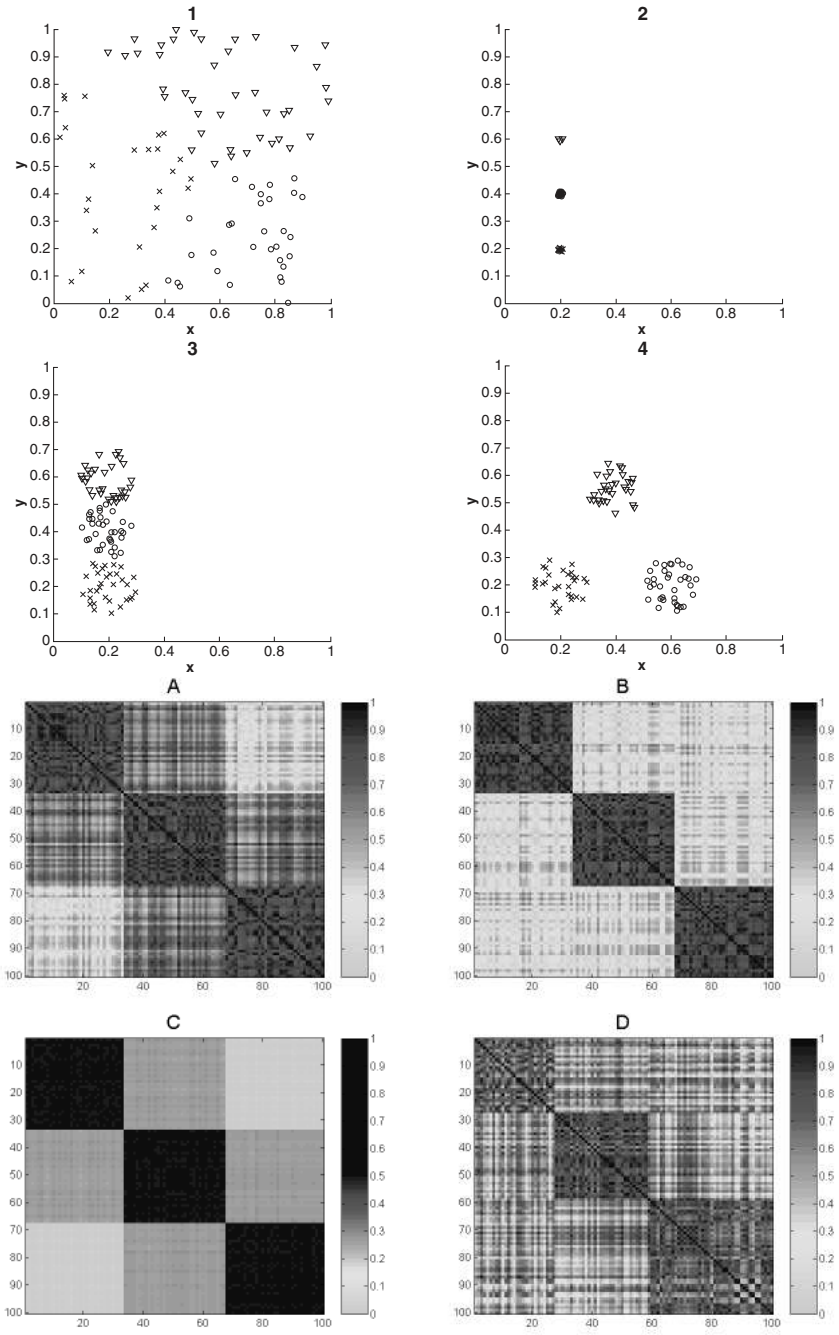


Figure 7.9. Points and similarity matrices for Exercise 32.

8

Cluster Analysis: Additional Issues and Algorithms

1. For sparse data, discuss why considering only the presence of non-zero values might give a more accurate view of the objects than considering the actual magnitudes of values. When would such an approach not be desirable?

   Consider document data. Intuitively, two documents are similar if they contain many of the same words. Although we can also include the frequency with which those words occur in the similarity computation, this can sometimes give a less reliable assessment of similarity. In particular, if one of the words in a document occurs rather frequently compared to other words, then this word can dominate the similarity comparison when magnitudes are taken into account. In that case, the document will only be highly similar to other documents that also contain the same word with a high frequency. While this may be appropriate in many or even most cases, it may lead to the wrong conclusion if the word can appear in different contexts, that can only be distinguished by other words. For instance, the word, 'game,' appears frequently in discussions of sports and video games.

2. Describe the change in the time complexity of K-means as the number of clusters to be found increases.

   As the number of clusters increases, the complexity of K-means approaches $O(m^2)$.

3. Consider a set of documents. Assume that all documents have been normalized to have unit length of 1. What is the "shape" of a cluster that consists of all documents whose cosine similarity to a centroid is greater than some specified constant? In other words, $\cos(d, c) \geq \delta$, where $0 < \delta \leq 1$.

Once document vectors have been normalized, they lie on am n-dimensional hypershpere. The constraint that all documents have a minimum cosine similarity with respect to a centroids is a constraint that the document vectors lie within a cone, whose intersection with the sphere is a circle on the surface of the sphere.

4. Discuss the advantages and disadvantages of treating clustering as an optimization problem. Among other factors, consider efficiency, non-determinism, and whether an optimization-based approach captures all types of clusterings that are of interest.

   Two key advantage to treating clustering as an optimization problem are that (1) it provides a clear definition of what the clustering process is doing, and (2) it allows the use of powerful optimization techniques that have been developed in a wide variety of fields. Unfortunately, most of these optimization techniques have a high time complexity. Furthermore, it can be shown that many optimization problems are NP hard, and therefore, it is necessary to use heuristic optimization approaches that can only guarantee a locally optimal solution. Often such techniques work best when used with random initialization, and thus, the solution found can vary from one run to another. Another problem with optimization approaches is that the objective functions they use tend to favor large clusters at the expense of smaller ones.

5. What is the time and space complexity of fuzzy c-means? Of SOM? How do these complexities compare to those of K-means?

   The time complexity of K-means $O(I * K * m * n)$, where I is the number of iterations required for convergence, K is the number of clusters, m is the number of points, and n is the number of attributes. The time required by fuzzy c-means is basically the same as K-means, although the constant is much higher. The time complexity of SOM is also basically the same as K-means because it consists of multiple passes in which objects are assigned to centroids and the centroids are updated. However, because the surrounding centroids are also updated and the number of passes can be large, SOM will typically be slower than K-means.

6. Traditional K-means has a number of limitations, such as sensitivity to outliers and difficulty in handling clusters of different sizes and densities, or with non-globular shapes. Comment on the ability of fuzzy c-means to handle these situations.

   Fuzzy c-means has all the limitations of traditional K-means, except that it does not make a hard assignment of an object to a cluster.

7. For the fuzzy c-means algorithm described in this book, the sum of the membership degree of any point over all clusters is 1. Instead, we could only require that the membership degree of a point in a cluster be between 0 and 1. What are the advantages and disadvantages of such an approach?

The main advantage of this approach occurs when a point is an outlier and does not really belong very strongly to any cluster, since in that situation, the point can have low membership in all clusters. However, this approach is often harder to initialize properly and can perform poorly when the clusters are not are not distinct. In that case, several cluster centers may merge together, or a cluster center may vary significantly from one iteration to another, instead of changing only slightly, as in ordinary K-means or fuzzy c-means.

8. Explain the difference between likelihood and probability.

   Probability is, according to one common statistical definition, the frequency with which an event occurs as the number of experiments tends to infinity. Probability is defined by a probability density function which is a function of the attribute values of an object. Typically, a probability density function depends on some parameters. Considering probability density function to be a function of the parameters yields the likelihood function.

9. Equation 8.12 gives the likelihood for a set of points from a Gaussian distribution as a function of the mean μ and the standard deviation σ. Show mathematically that the maximum likelihood estimate of μ and σ are the sample mean and the sample standard deviation, respectively.

   First, we solve for μ.

$$\begin{aligned}\frac{\partial \ell((\mu,\sigma)|\mathcal{X})}{\partial \mu} &= \frac{\partial}{\partial \mu}\left(-\sum_{i=1}^{m}\frac{(x_i-\mu)^2}{2\sigma^2} - 0.5m\log 2\pi - m\log\sigma\right)\\ &= -\sum_{i=1}^{m}\frac{2(x_i-\mu)}{2\sigma^2}\end{aligned}$$

   Setting this equal to 0 and solving, we get $\mu = \frac{1}{m}\sum_{i=1}^{m} x_i$.

   Likewise, we can solve for σ.

$$\begin{aligned}\frac{\partial \ell((\mu,\sigma)|\mathcal{X})}{\partial \sigma} &= \frac{\partial}{\partial \sigma}\left(-\sum_{i=1}^{m}\frac{(x_i-\mu)^2}{2\sigma^2} - 0.5m\log 2\pi - m\log\sigma\right)\\ &= \sum_{i=1}^{m}\frac{2(x_i-\mu)^2}{2\sigma^3} - \frac{m}{\sigma}\end{aligned}$$

   Setting this equal to 0 and solving, we get $\sigma^2 = \frac{1}{m}\sum_{i=1}^{m}(x_i-\mu)^2$.

10. We take a sample of adults and measure their heights. If we record the gender of each person, we can calculate the average height and the variance of the height, separately, for men and women. Suppose, however, that this information was not recorded. Would it be possible to still obtain this information? Explain.

The height of men and women will have separate Gaussian distributions with different means and perhaps different variances. By using a mixture model approach, we can obtain estimates of the mean and variance of the two height distributions. Given a large enough sample size, the estimated parameters should be close to those that could be computed if we knew the gender of each person.

11. Compare the membership weights and probabilities of Figures 8.1 and 8.4, which come, respectively, from applying fuzzy and EM clustering to the same set of data points. What differences do you detect, and how might you explain these differences?

    The fuzzy clustering approach only assigns very high weights to those points in the center of the clusters. Those points that are close to two or three clusters have relatively low weights. The points that are on the far edges of the clusters, away from other clusters also have lower weights than the center points, but not as low as points that are near two or three clusters.

    The EM clustering approach assigns high weights both to points in the center of the clusters and those on the far edges. The points that are near two or three clusters have lower weights, but not as much so as with the fuzzy clustering procedure.

    The main difference between the approaches is that as a point on the far edge of a cluster gets further away from the center of the cluster, the weight with which is belongs to a cluster becomes more equal among clusters for the fuzzy clustering approach, but for the EM approach the point tends to belong more strongly to the cluster to which it is closest.

12. Figure 8.1 shows a clustering of a two-dimensional point data set with two clusters: The leftmost cluster, whose points are marked by asterisks, is somewhat diffuse, while the rightmost cluster, whose points are marked by circles, is compact. To the right of the compact cluster, there is a single point (marked by an arrow) that belongs to the diffuse cluster, whose center is farther away than that of the compact cluster. Explain why this is possible with EM clustering, but not K-means clustering.

    In EM clustering, we compute the probability that a point belongs to a cluster. In turn, this probability depends on both the distance from the cluster center and the spread (variance) of the cluster. Hence, a point that is closer to the centroid of one cluster than another can still have a higher probability with respect to the more distant cluster if that cluster has a higher spread than the closer cluster. K-means only takes into account the distance to the closest cluster when assigning points to clusters. This is equivalent to an EM approach where all clusters are assumed to have the same variance.

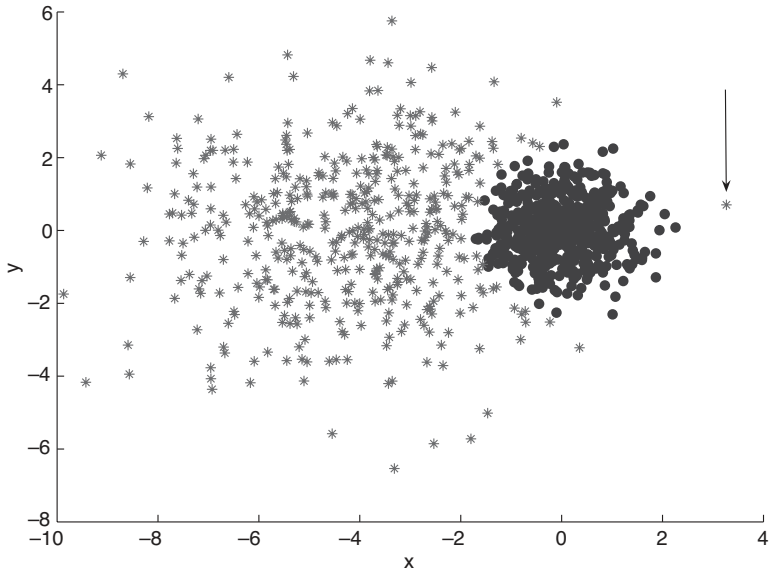


Figure 8.1. Data set for Exercise 12. EM clustering of a two-dimensional point set with two clusters of differing density.

13. Show that the MST clustering technique of Section 8.4.2 produces the same clusters as single link. To avoid complications and special cases, assume that all the pairwise similarities are distinct.

    In single link, we start with with clusters of individual points and then successively join two clusters that have the pair of points that are closest together. Conceptually, we can view the merging of the clusters as putting an edge between the two closest points of the two clusters. Note that if both clusters are currently connected, then the resulting cluster will also be connected. However, since the clusters are formed from disjoint sets of points, and edges are only placed between points in different clusters, no cycle can be formed. From these observations and by noting that we start with clusters (graphs) of size one that are vacuously connected, we can deduce by induction that at any stage in single link clustering process, each cluster consists of a connected set of points without any cycles. Thus, when the last two clusters are merged to form a cluster containing all the points, we also have a connected graph of all the points that is a spanning tree of the graph. Furthermore, since each point in the graph is connected to its nearest point, the spanning tree must be a minimum spanning tree. All that remains to establish the equivalence of MST and single link is to note that MST essentially reverses the process by which single link built the minimum spanning tree; i.e., by

breaking edges beginning with the longest and proceeding until the smallest. Thus, it generates the same clusters as single link, but in reverse order.

14. One way to sparsify a proximity matrix is the following: For each object (row in the matrix), set all entries to 0 except for those corresponding to the objects k-nearest neighbors. However, the sparsified proximity matrix is typically not symmetric.

    (a) If object a is among the k-nearest neighbors of object b, why is b not guaranteed to be among the k-nearest neighbors of a?

    Consider a dense set of $k+1$ objects and another object, an outlier, that is farther from any of the objects than they are from each other. None of the objects in the dense set will have the outlier on their k-nearest neighbor list, but the outlier will have k of the objects from the dense set on its k-nearest neighbor list.

    (b) Suggest at least two approaches that could be used to make the sparsified proximity matrix symmetric.

    One approach is to set the ij^{th} entry to 0 if the ji^{th} entry is 0, or vice versa. Another approach is to set the ij^{th} entry to 1 if the ji^{th} entry is 1, or vice versa.

15. Give an example of a set of clusters in which merging based on the closeness of clusters leads to a more natural set of clusters than merging based on the strength of connection (interconnectedness) of clusters.

    An example of this is given in the Chameleon paper that can be found at http://www.cs.umn.edu/ karypis/publications/Papers/PDF/chameleon.pdf. The example consists of two narrow rectangles of points that are side by side. The top rectangle is split into two clusters, one much smaller than the other. Even though the two rectangles on the top are close, they are not strongly connected since the strong links between them are across a small area. On the other hand, the largest rectangle on the top and the rectangle on the bottom are strongly connected. Each individual connection is not as strong, because these two rectangles are not as close, but there are more of them because the area of connection is large. Thus, an approach based on connectivity will merge the largest rectangle on top with the bottom rectangle.

16. Table 8.1 lists the two nearest neighbors of four points.

    Calculate the SNN similarity between each pair of points using the definition of SNN similarity defined in Algorithm 8.11.

    The following is the SNN similarity matrix.

17. For the definition of SNN similarity provided by Algorithm 8.11, the calculation of SNN distance does not take into account the position of shared

Table 8.1. Two nearest neighbors of four points.

Point	First Neighbor	Second Neighbor
1	4	3
2	3	4
3	4	2
4	3	1

Table 8.2. Two nearest neighbors of four points.

Point	1	2	3	4
1	2	0	0	1
2	0	2	1	0
3	0	1	2	0
4	1	0	0	2

neighbors in the two nearest neighbor lists. In other words, it might be desirable to give higher similarity to two points that share the same nearest neighbors in the same or roughly the same order.

(a) Describe how you might modify the definition of SNN similarity to give higher similarity to points whose shared neighbors are in roughly the same order.

This can be done by assigning weights to the points based on their position in the nearest neighbor list. For example, we can weight the i^{th} point in the nearest neighbor list by $n - i + 1$. For each point, we then take the sum or product of its rank on both lists. These values are then summed to compute the similarity between the two objects. This approach was suggested by Jarvis and Patrick [5].

(b) Discuss the advantages and disadvantages of such a modification.

Such an approach is more complex. However, it is advantageous if it is the case that two objects are more similar if the shared neighbors are roughly of the same rank. Furthermore, it may also help to compensate for arbitrariness in the choice of k.

18. Name at least one situation in which you would *not* want to use clustering based on SNN similarity or density.

When you wish to cluster based on absolute density or distance.

19. Grid-clustering techniques are different from other clustering techniques in that they partition space instead of sets of points.

(a) How does this affect such techniques in terms of the description of the resulting clusters and the types of clusters that can be found?

In grid-based clustering, the clusters are described in terms of collections of adjacent cells. In some cases, as in CLIQUE, a more compact description is generated. In any case, the description of the clusters is given in terms of a region of space, not a set of objects. (However, such a description can easily be generated.) Because it is necessary to work in terms of rectangular regions, the shapes of non-rectangular clusters can only be approximated. However, the groups of adjacent cells can have holes.

(b) What kind of cluster can be found with grid-based clusters that cannot be found by other types of clustering approaches? (Hint: See Exercise 20 in Chapter 7, page 608.)

Typically, grid-based clustering techniques only pay attention to dense regions. However, such techniques could also be used to identify sparse or empty regions and thus find patterns of the absence of points. Note, however, that this would not be appropriate for a sparse data space.

20. In CLIQUE, the threshold used to find cluster density remains constant, even as the number of dimensions increases. This is a potential problem since density drops as dimensionality increases; i.e., to find clusters in higher dimensions the threshold has to be set at a level that may well result in the merging of low-dimensional clusters. Comment on whether you feel this is truly a problem and, if so, how you might modify CLIQUE to address this problem.

This is a real problem. A similar problem exists in association analysis. In particular, the support of association patterns with a large number of items is often low. To find such patterns using an algorithm such as Apriori is difficult because the low support threshold required results in a large number of association patterns with few items that are of little interest. In other words, association patterns with many items (patterns in higher-dimensional space) are interesting at support levels (densities) that do not make for interesting patterns when the size of the association pattern (number of dimensions) is low. One approach is to decrease the support threshold (density threshold) as the number of items (number of dimensions) is increased.

21. Given a set of points in Euclidean space, which are being clustered using the K-means algorithm with Euclidean distance, the triangle inequality can be used in the assignment step to avoid calculating all the distances of each point to each cluster centroid. Provide a general discussion of how this might work.

Charles Elkan presented the following theorem in his keynote speech at the Workshop on Clustering High-Dimensional Data at SIAM 2004.

Lemma 1: Let x be a point, and let b and c be centers.
If $d(b, c) \geq 2d(x, b)$ then $d(x, c) \geq d(x, b)$.

Proof:
We know $d(b,c) \leq d(b,x) + d(x,c)$.
So $d(b,c) - d(x,b) \leq d(x,c)$.
Now $d(b,c) - d(x,b) \geq 2d(x,b) - d(x,b) = d(x,b)$.
So $d(x,b) \leq d(x,c)$.

This theorem can be used to eliminate a large number of unnecessary distance calculations.

22. Instead of using the formula derived in CURE—see Equation 8.21—we could run a Monte Carlo simulation to directly estimate the probability that a sample of size s would contain at least a certain fraction of the points from a cluster. Using a Monte Carlo simulation compute the probability that a sample of size s contains 50% of the elements of a cluster of size 100, where the total number of points is 1000, and where s can take the values 100, 200, or 500.

    This question should have said "contains *at least* 50%."

    The results of our simulation consisting of 100,000 trials was 0, 0, and 0.54 for the sample sizes 100, 200, and 500 respectively.

9

Anomaly Detection

1. Compare and contrast the different techniques for anomaly detection that were presented in sections 9.3-9.8. In particular, try to identify circumstances in which the definitions of anomalies used in the different techniques might be equivalent or situations in which one might make sense, but another would not. Be sure to consider different types of data.

   First, note that the proximity- and density-based anomaly detection techniques are related. Specifically, high density in the neighborhood of a point implies that many points are close to it, and vice-versa. In practice, density is often defined in terms of distance, although it can also be defined using a grid-based approach.

   The model-based approach can be used with virtually any underlying technique that fits a model to the data. However, note that a particular model, statistical or otherwise, must be assumed. Consequently, model-based approaches are restricted in terms of the data to which they can be applied. For example, if the model assumes a Gaussian distribution, then it cannot be applied to data with a non-Gaussian distribution.

   On the other hand, the proximity- and density-based approaches do not make any particular assumption about the data, although the definition of an anomaly does vary from one proximity- or density-based technique to another. Proximity-based approaches can be used for virtually any type of data, although the proximity metric used must be chosen appropriately. For example, Euclidean distance is typically used for dense, low-dimensional data, while the cosine similarity measure is used for sparse, high-dimensional data. Since density is typically defined in terms of proximity, density-based approaches can also be used for virtually any type of data. However, the meaning of density is less clear in a non-Euclidean data space.

   Proximity- and density-based anomaly detection techniques can often produce similar results, although there are significant differences between tech-

niques that do not account for the variations in density throughout a data set or that use different proximity measures for the same data set. Model-based methods will often differ significantly from one another and from proximity- and density-based approaches.

2. Consider the following definition of an anomaly: An anomaly is an object that is unusually influential in the creation of a data model.

   (a) Compare this definition to that of the standard model-based definition of an anomaly.

   The standard model-based definition labels objects that don't fit the model very well as anomalies. Although these object often are unusually influential in the model, it can also be the case that an unusually influential object can fit the model very well.

   (b) For what sizes of data sets (small, medium, or large) is this definition appropriate?

   This definition is typically more appropriate for smaller data sets, at least if we are talking about one very influential object. However, a relatively small group highly influential objects can have a significant impact on a model, but still fit it well, even for medium or large data sets.

3. In one approach to anomaly detection, objects are represented as points in a multidimensional space, and the points are grouped into successive shells, where each shell represents a layer around a grouping of points, such as a convex hull. An object is an anomaly if it lies in one of the outer shells.

   (a) To which of the definitions of an anomaly in sections 9.3-9.8 is this definition most closely related?

   This definition is most closely related to the distance-based approach.

   (b) Name two problems with this definition of an anomaly.

      i. For the convex hull approach, the distance of the points in a convex hull from the center of the points can vary significantly if the distribution of points in not symmetrical.

      ii. This approach does not lend itself to assigning meaningful numerical anomaly scores.

4. Association analysis can be used to find anomalies as follows. Find strong association patterns, which involve some minimum number of objects. Anomalies are those objects that do not belong to any such patterns. To make this more concrete, we note that the hyperclique association pattern discussed in Section 5.8 is particularly suitable for such an approach. Specifically, given a user-selected h-confidence level, maximal hyperclique patterns of objects are

found. All objects that do not appear in a maximal hyperclique pattern of at least size three are classified as outliers.

(a) Does this technique fall into any of the categories discussed in this chapter? If so, which one?

In a hyperclique, all pairs of objects have a guaranteed cosine similarity of the h-confidence or higher. Thus, this approach can be viewed as a proximity-based approach. However, rather than a condition on the proximity of objects with respect to a particular object, there is a requirement on the pairwise proximities of all objects in a group.

(b) Name one potential strength and one potential weakness of this approach.

Strengths of this approach are that (1) the objects that do not belong to any size 3 hyperclique are not strongly connected to other objects and are likely anomalous and (2) it is computationally efficient. Potential weaknesses are (1) this approach does not assign a numerical anomaly score, but simply classifies an object as normal or an anomaly, (2) it is not possible to directly control the number of objects classified as anomalies because the only parameters are the h-confidence and support threshold, and (3) the data needs to be discretized.

5. Discuss techniques for combining multiple anomaly detection techniques to improve the identification of anomalous objects. Consider both supervised and unsupervised cases.

In the supervised case, we could use ensemble classification techniques. In these approaches, the classification of an object is determined by combining the classifications of a number of classifiers, e.g., by voting. In the unsupervised approach, a voting approach could also be used. Note that this assumes that we have a binary assignment of an object as an anomaly. If we have anomaly scores, then the scores could be combined in some manner, e.g., an average or minimum, to yield an overall score.

6. Describe the potential time complexity of anomaly detection approaches based on the following approaches: model-based using clustering, proximity-based, and density. No knowledge of specific techniques is required. Rather, focus on the basic computational requirements of each approach, such as the time required to compute the density of each object.

If K-means clustering is used, then the complexity is dominated by finding the clusters. This requires time proportional to the number of objects, i.e., $O(m)$. The distance and density based approaches, usually require computing all the pairwise proximities, and thus the complexity is often $O(m^2)$. In some cases, such as low dimensional data, special techniques, such as the R* tree

or k-d trees can be used to compute the nearest neighbors of an object more efficiently, i.e., $O(m \log m)$, and this can reduce the overall complexity when the technique is based only on nearest neighbors. Also, a grid-based approach to computing density can reduce the complexity of density-based anomaly detection to $O(m)$, but such techniques are not as accurate and are only effective for low dimensions.

7. The Grubbs' test, which is described by Algorithm 9.2, is a more statistically robust procedure for detecting outliers. It is iterative and also takes into account the fact that the z-score does not have a normal distribution. This algorithm computes the z-score of each value based on the sample mean and standard deviation of the current set of values. The value with the largest magnitude z-score is discarded if its z-score is larger than g_c, the critical value of the test for an outlier at significance level α. This process is repeated until no objects are eliminated. Note that the sample mean, standard deviation, and g_c are updated at each iteration.

Algorithm 9.2 Grubbs' approach for outlier elimination.

1: Input the values and α
   {m is the number of values, α is a parameter, and t_c is a value chosen so that $\alpha = prob(x \geq t_c)$ for a t distribution with $m - 2$ degrees of freedom.}
2: **repeat**
3: Compute the sample mean ($\overline{x}$) and standard deviation (s_x).
4: Compute a value g_c so that $prob(|z| \geq g_c) = \alpha$.
   (In terms of t_c and m, $g_c = \frac{m-1}{\sqrt{m}} \sqrt{\frac{t_c^2}{m-2+t_c^2}}$.)
5: Compute the z-score of each value, i.e., $z = (x - \overline{x})/s_x$.
6: Let $g = \max |z|$, i.e., find the z-score of largest magnitude and call it g.
7: **if** $g > g_c$ **then**
8: Eliminate the value corresponding to g.
9: $m \leftarrow m - 1$
10: **end if**
11: **until** No objects are eliminated.

(a) What is the limit of the value $\frac{m-1}{\sqrt{m}} \sqrt{\frac{t_c^2}{m-2+t_c^2}}$ used for Grubbs' test as m approaches infinity? Use a significance level of 0.05.

The value of this expression approaches t_c, but strictly speaking this is not a limit as t_c depends on m.

$$\lim_{m \to \infty} \frac{m-1}{\sqrt{m}} \sqrt{\frac{t_c^2}{m-2+t_c^2}} = \lim_{m \to \infty} \frac{m-1}{\sqrt{m(m-2+t_c^2)}} \times t_c = 1 \times t_c = t_c.$$

Also, the value of t_c will continue to increase with m, although slowly. For $m = 10^{20}$, $t_c = 93$ for a significance value of 0.05.

(b) Describe, in words, the meaning of the previous result.

The distribution of g is becomes a t distribution as m increases.

8. Many statistical tests for outliers were developed in an environment in which a few hundred observations was a large data set. We explore the limitations of such approaches.

(a) For a set of 1,000,000 values, how likely are we to have outliers according to the test that says a value is an outlier if it is more than three standard deviations from the average? (Assume a normal distribution.)

This question should have asked how many outliers we would have since the object of this question is to show that even a small probability of an outlier yields a large number of outliers for a large data set. The probability is unaffected by the number of objects.

The probability is either 0.00135 for a single sided deviation of 3 standard deviations or 0.0027 for a double-sided deviation. Thus, the number of anomalous objects will be either 1,350 or 2,700.

(b) Does the approach that states an outlier is an object of unusually low probability need to be adjusted when dealing with large data sets? If so, how?

There are thousands of outliers (under the specified definition) in a million objects. We may choose to accept these objects as outliers or prefer to increase the threshold so that fewer outliers result.

9. The probability density of a point $\mathbf{x}$ with respect to a multivariate normal distribution having a mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ is given by the equation

$$prob(\mathbf{x}) = \frac{1}{(\sqrt{2\pi})^m |\Sigma|^{1/2}} \; e^{-\frac{(\mathbf{x}-\boldsymbol{\mu})\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})^T}{2}}. \tag{9.10}$$

Using the sample mean $\overline{\mathbf{x}}$ and covariance matrix $\mathbf{S}$ as estimates of the mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, respectively, show that the $\log(prob(\mathbf{x}))$ is equal to the Mahalanobis distance between a data point $\mathbf{x}$ and the sample mean $\overline{\mathbf{x}}$ plus a constant that does not depend on $\mathbf{x}$.

$$\log prob(\mathbf{x}) = -\log\left((\sqrt{2\pi})^m |\Sigma|^{1/2}\right) - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})^T.$$

If we use the sample mean and covariance as estimates of $\boldsymbol{\mu}$ and Σ, respectively, then

$$\log prob(\mathbf{x}) = -\log\left((\sqrt{2\pi})^m |S|^{1/2}\right) - \frac{1}{2}(\mathbf{x}-\overline{\mathbf{x}})\mathbf{S}^{-1}(\mathbf{x}-\overline{\mathbf{x}})^T$$

The constant and the constant factor do not affect the ordering of this quantity, only their magnitude. Thus, if we want to base a distance on this quantity, we can keep only the variable part, which is the Mahalanobis distance.

10. Compare the following two measures of the extent to which an object belongs to a cluster: (1) distance of an object from the centroid of its closest cluster and (2) the silhouette coefficient described in Section 7.5.2.

    The first measure is somewhat limited since it disregards that fact that the object may also be close to another cluster. The silhouette coefficient takes into account both the distance of an object to its cluster and its distance to other clusters. Thus, it can be more informative about how strongly an object belongs to the cluster to which it was assigned.

11. Consider the (relative distance) K-means scheme for outlier detection described in Section 9.5 and the accompanying figure, Figure 9.10.

    (a) The points at the bottom of the compact cluster shown in Figure 9.10 have a somewhat higher outlier score than those points at the top of the compact cluster. Why?

        The mean of the points is pulled somewhat upward from the center of the compact cluster by point D.

    (b) Suppose that we choose the number of clusters to be much larger, e.g., 10. Would the proposed technique still be effective in finding the most extreme outlier at the top of the figure? Why or why not?

        No. This point would become a cluster by itself.

    (c) The use of relative distance adjusts for differences in density. Give an example of where such an approach might lead to the wrong conclusion.

        If absolute distances are important. For example, consider heart rate monitors for patients. If the heart rate goes above or below a specified range of values, then this has an physical meaning. It would be incorrect not to identify any patient outside that range as abnormal, even though there may be a group of patients that are relatively similar to each other and all have abnormal heart rates.

12. If the probability that a normal object is classified as an anomaly is 0.01 and the probability that an anomalous object is classified as anomalous is 0.99, then what is the false alarm rate and detection rate if 99% of the objects are normal? (Use the definitions given below.)

$$\text{detection rate} = \frac{\text{number of anomalies detected}}{\text{total number of anomalies}} \tag{9.11}$$

$$\text{false alarm rate} = \frac{\text{number of false anomalies}}{\text{number of objects classified as anomalies}} \tag{9.12}$$

The detection rate is simply 99%.
The false alarm rate $= 0.99m \times 0.01/(0.99m \times 0.01 + 0.01m \times 0.99) = 0.50 = 50\%$.

13. When a comprehensive training set is available, a supervised anomaly detection technique can typically outperform an unsupervised anomaly technique when performance is evaluated using measures such as the detection and false alarm rate. However, in some cases, such as fraud detection, new types of anomalies are always developing. Performance can be evaluated according to the detection and false alarm rates, because it is usually possible to determine, upon investigation, whether an object (transaction) is anomalous. Discuss the relative merits of supervised and unsupervised anomaly detection under such conditions.

    When new anomalies are to be detected, an unsupervised anomaly detection scheme must be used. However, supervised anomaly detection techniques are still important for detecting known types of anomalies. Thus, both supervised and unsupervised anomaly detection methods should be used. A good example of such a situation is network intrusion detection. Profiles or signatures can be created for well-known types of intrusions, but cannot detect new types of intrusions.

14. Consider a group of documents that has been selected from a much larger set of diverse documents so that the selected documents are as dissimilar from one another as possible. If we consider documents that are not highly related (connected, similar) to one another as being anomalous, then all of the documents that we have selected might be classified as anomalies. Is it possible for a data set to consist only of anomalous objects or is this an abuse of the terminology?

    The connotation of anomalous is that of rarity and many of the definitions of an anomaly incorporate this notion to some extent. However, there are situations in which an anomaly typically does not occur very often, e.g., a network failure, but has a very concrete definition. This makes it possible to distinguish an anomaly in an absolute sense and for a situation to arise where the majority of objects are anomalous. Also, in providing mathematical or algorithmic definitions of an anomaly, it can happen that these definitions produce situations in which many or all objects of a data set can be classified as anomalies. Another viewpoint might say that if it is impossible to define a meaningful normal situation, then all objects are anomalous. ("Unique" is the term typically used in this context.) In summary, this can be regarded as a philosophical or semantic question. A good argument (although likely not an uncontested one) can be made that it is possible to have a collection of objects that are mostly or all anomalies.

15. Consider a set of points, where most points are in regions of low density, but a few points are in regions of high density. If we define an anomaly as a point in a region of low density, then most points will be classified as anomalies. Is this an appropriate use of the density-based definition of an anomaly or should the definition be modified in some way?

    If the density has an absolute meaning, such as assigned by the domain, then it may be perfectly reasonable to consider most of the points as anomalous. (See the answer to the previous exercise.) However, in many circumstances, the appropriate approach would be to use an anomaly detection technique that takes the relative density into account.

16. Consider a set of points that are uniformly distributed on the interval [0,1]. Is the statistical notion of an outlier as an infrequently observed value meaningful for this data?

    Not really. The traditional statistical notion of an outlier relies on the idea that an object with a relatively low probability is suspect. With a uniform distribution, no such distinction can be made.

17. An analyst applies an anomaly detection algorithm to a data set and finds a set of anomalies. Being curious, the analyst then applies the anomaly detection algorithm to the set of anomalies.

    (a) Discuss the behavior of each of the anomaly detection techniques described in this chapter. (If possible, try this for real data sets and algorithms.)

    (b) What do you think the behavior of an anomaly detection algorithm should be when applied to a set of anomalous objects?

    In some cases, such as the statistically-based anomaly detection techniques, it would not be valid to apply the technique a second time, since the assumptions would no longer hold. This could also be true for other model-based approaches. The behavior of the proximity- and density-based approaches would depend on the particular techniques. Interestingly, the approaches that use absolute thresholds of distance or density would likely classify the set of anomalies as anomalies, at least if the original parameters were kept. The relative approaches would likely classify most of the anomalies as normal and some as anomalies.

    Whether an object is anomalous depends on the the entire group of objects, and thus, it is probably unreasonable to expect that an anomaly detection technique will identify a set of anomalies as such in the absence of the original data set.

10

Avoiding False Discoveries

1. Statistical testing proceeds in a manner analogous to the mathematical proof technique, proof by contradiction, which proves a statement by assuming it is false and then deriving a contradiction. Compare and contrast statistical testing and proof by contradiction.

   Answer: Statistical testing assumes a result is consistent with the null hypothesis which is typically the opposite of what we would like to establish, and then rejects this assertion if the p-value of the result is lower than a specified threshold. Unlike proof by contradiction, this conclusion can be erroneous since it is based on probability. If no contradiction can be derived, proof by contradiction does not assume the original statement is true, and analogously, if the p-value of the result is high, then statistical testing does not, or at least, should not, assume that this proves the null hypothesis is true.

2. Which of the following are suitable null hypotheses. If not, explain why.

   (a) *Comparing two groups* Consider comparing the average blood pressure of a group of subjects, both before and after they are placed on a low salt diet. In this case, the null hypothesis is that a low salt diet does reduce blood pressure, i.e., that the average blood pressure of the subjects is the same before and after the change in diet.

   (b) *Classification* Assume there are two classes, labeled + and -, where we are most interested in the positive class, e.g., the presence of a disease. H_0 is the statement that the class of an object is negative, i.e., that the patient does not have the disease.

   (c) *Association Analysis* For frequent patterns, the null hypothesis is that the items are independent and thus, any pattern that we detect is spurious.

(d) *Clustering* The null hypothesis is that there is cluster structure in the data beyond what might occur at random.

(e) *Anomaly Detection* Our assumption, H_0, is that an object is not anomalous.

Answer: (b), (c) and (e) are appropriate null hypotheses since they assume the opposite of what we would like to prove. By contrast, (a) and (d) take the null hypotheses to be what we would like to prove.

3. Consider once again the coffee-tea example, presented in Example 10.9. The following two tables are the same as the one presented in Example 10.9 except that each entry has been divided by 10 (left table) or multiplied by 10 (right table).

Table 10.1. Beverage preferences among a group of 100 people (left) and 10,000 people (right).

	Coffee	$\overline{Coffee}$	
Tea	15	5	20
$\overline{Tea}$	65	15	80
	80	20	100

	Coffee	$\overline{Coffee}$	
Tea	1500	500	2000
$\overline{Tea}$	6500	1500	8000
	8000	2000	10000

(a) Compute the p-value of the observed support count for each table, i.e., for 15 and 1500. What pattern do you observe as the sample size increases?

(b) Compute the odds ratio and interest factor for the two contingency tables presented in this problem and the original table of Example 10.9. (See Section 5.7.1 for definitions of these two measures.) What pattern do you observe?

(c) The odds ratio and interest factor are measures of effect size. Are these two effect sizes significant from a practical point of view?

(d) What would you conclude about the relationship between p-values and effect size for this situation?

Answer:

(a) In order of increasing number of people, the p-values are 0.3647, 0.0321, 5.4929e-10. Even though the percentage of people who drink both tea and coffee remains constant, the p-value decreases, i.e., the result becomes more significant.

(b) The odds ratio is 0.6923 for all cases, while the interest factor is 0.9375 for all cases.

(c) These values are not typically significant from a practical point of view, although such a judgment does depend on the particular application.

(d) The p-values and effect sizes are independent in this case.

4. Consider the different combinations of effect size and p-value applied to an experiment where we want to determine the efficacy of a new drug.
(i) effect size small, p-value small
(ii) effect size small, p-value large
iii) effect size large, p-value small
(iv) effect size large, p-value large

Whether effect size is small or large depends on the domain, which in this case is medical. For this problem consider a small p-value to be less than 0.001, while a large p-value is above 0.05. Assume that the sample size is relatively large, e.g., thousands of patients with the condition that the drug hopes to treat.

(a) Which combination(s) would very likely be of interest?

(b) Which combinations(s) would very likely **not** be of interest?

(c) If the sample size were small, would that change your answers?

Answer:

(a) (iii) would almost certainly be of interest since the p-value indicates that the effect is likely not a statistical fluke and effect size indicates that the effect of the new drug is worthwhile from a medical point of view.

(b) (i) and (ii) are likely not interesting from a practical point of view since the effect size is small. (iv) is not interesting since it has a large p-value and may not be trustworthy.

(c) Yes. For a small sample, results are much less trustworthy, i.e., it is much more likely that an ineffective drug may show up as effective and an effective drug may show no or little effect.

5. For Neyman-Pearson hypothesis testing, we need to balance the tradeoff between α, the probability of a type I error and power, i.e., $1 - \beta$, where β is the probability of a type II error. Compute α, β, and the power for the cases given below, where we specify the null and alternative distributions and the accompanying critical region. All distributions are Gaussian with some specified mean u and standard deviation σ, i.e., $\mathcal{N}(\mu, \sigma)$. Let T be the test statistic.

(a) H_0: $\mathcal{N}(0, 1)$, $H1 : \mathcal{N}(3, 1)$, critical region: $T > 2$.

(b) H_0: $\mathcal{N}(0, 1)$, $H1 : \mathcal{N}(3, 1)$, critical region: $|T| > 2$.

(c) H_0: $\mathcal{N}(-1, 1)$, $H1 : \mathcal{N}(3, 1)$, critical region: $T > 1$.

(d) H_0: $\mathcal{N}(-1, 1)$, $H1 : \mathcal{N}(3, 1)$, critical region: $|T| > 1$.

(e) H_0: $\mathcal{N}(-1, 0.5)$, $H1 : \mathcal{N}(3, 0.5)$, critical region: $T > 1$.

(f) H_0: $\mathcal{N}(-1, 0.5)$, $H1 : \mathcal{N}(3, 0.5)$, critical region: $|T| > 1$.

Answer:

(a) $\alpha = 0.0228$, $\beta = 0.1587$, power $= 0.8413$

(b) $\alpha = 0.0455$, $\beta = 0.1587$, power $= 0.8413$

(c) $\alpha = 0.0228$, $\beta = 0.0228$, power $= 0.9772$

(d) $\alpha = 0.0455$, $\beta = 0.1587$, power $= 0.8413$

6. A p-value measures the probability of the result given that the null hypothesis is true. However, many people who calculate p-values have used it as the probability of the null hypothesis given the result, which is erroneous. A Bayesian approach to this problem is summarized by Equation 10.1.

$$\begin{aligned} \frac{\mathrm{P}(H_1|x_{obs})}{\mathrm{P}(H_0|x_{obs})} &= \frac{f(H_1|x_{obs})}{f(H_0|x_{obs})} \times \frac{\mathrm{P}(H_1)}{\mathrm{P}(H_0)} \\ \text{posterior odds of } H_1 &= \text{Bayes Factor} \times \text{prior odds of } H_1 \end{aligned} \tag{10.1}$$

This approach computes the ratio of the probability of the alternative and null hypotheses (H_1 and H_0, respectively) given the observed outcome, x_{obs}. In turn, this quantity is expressed as the product of two factors: the Bayes factor and the prior odds. The prior odds is the ratio of the probability of H_1 to the probability of H_0 based on prior information about how likely we believe each hypothesis is. Usually, the prior odds is estimated directly based on experience, For example, in drug testing in the laboratory, it may be known that most drugs do not produce potentially therapeutic effects. The Bayes factor is the ratio of the probability or probability density of the observed outcome, x_{obs}, under H_1 and H_0. This quantity is computed and represents a measure of how much more or less likely the observed result is under the alternative hypothesis than the null hypothesis. Conceptually, the higher it is, the more we would tend to prefer the alternative to the null. The higher the Bayes factor, the stronger the evidence provided by the data for H_1. More generally, this approach can be applied to assess the evidence for any hypothesis versus another. Thus, the roles of H_0 can be (and often are) reversed in Equation 10.1.

(a) Suppose that the Bayes factor is 20, which is very strong, but the prior odds are 0.01. Would you be inclined to prefer the alternative or null hypothesis?

(b) Suppose the prior odds are 0.25, the null distribution is Gaussian with density given by $f_0(x) = \mathcal{N}(0, 2)$, and the alternative distribution is given by $f_1(x) = \mathcal{N}(3, 1)$. Compute the Bayes factor and posterior odds of H_1 for the following values of x_{obs}: 2, 2.5, 3, 3.5, 4, 4.5, 5. Explain the pattern that you see in both quantities.

Answer:

(a) We would still prefer the null hypothesis since the alternative is so unlikely, but the high Bayes factor would likely encourage us to investigate further.

(b) Bayes factor: 2.00, 3.85, 6.16, 8.16, 8.96, 8.16, 6.16
posterior odds: 0.50, 0.96, 1.54, 2.04, 2.24, 2.04, 1.54
A Bayes factors of 6 or more would indicate a moderate preference for H_1 but given that H_1 is not as likely as H_0, this weakens the evidence provided by x_{obs} for H_1, as indicated by the posterior odds. An interesting pattern is that the Bayes factor and posterior odds increase with x_{obs} but then decrease. This is because the null distribution has a higher variance and thus, more of its probability lies in its tails whereas the alternate distribution has a smaller variance and its probability is more concentrated around its mean.

7. Consider the problem of determining whether a coin is a fair one, i.e., P(heads) = P(tails) = 0.5, by flipping the coin 10 times. Use the binomial theorem and basic probability to answer the following questions.

(a) A coin is flipped ten times and it comes up heads every time. What is the probability of getting 10 heads in a row and what would you conclude about whether the coin is fair?

(b) Suppose 10,000 coins are each flipped 10 times in a row and the flips of 10 coins result in all heads, can you confidently say that these coins are not fair?

(c) What can you conclude about results when evaluated individually versus in a group?

(d) Suppose that you flip each coin 20 times and then evaluate 10,000 coins. Can you now confidently say that any coin which yields all heads is not fair?

Answer:

(a) The probability of 10 heads in a row is $0.5^10 \approx 0.001$. It seems very unlikely that the coin is fair.

(b) We would expect roughly 10 such coins in our tests just by random chance. No, we should not conclude these coins are not fair.

(c) Rare events can happen frequently by random chance when tests are repeated a large number of times and so the threshold for concluding that a result is incompatible with the null hypothesis—in this case, that the coin is fair—must be much more stringent.

(d) Yes. The probability of such an occurrence is $0.5^20 \approx 10^{-6}$ or one chance in a million. Thus, even in 10,000 tests, a string of 20 heads is unlikely to occur by chance.

8. Algorithm 10.1 on page 773 provides a method for calculating the false discovery rate using the method advocated by Benjamini and Hochberg. The description in the text is presented in terms of ordering the p-values and adjusting the significance level to assess whether a p-value is significant. Another way to interpret this method is in terms of ordering the p-values, smallest to largest, and computing adjusted" p-values, $p'_i = p_i * m/i$, where i identifies the i^{th} smallest p-value and m is the number of p-values. The statistical significance is determined based on whether $p'_i \leq \alpha$, where α is the desired false discovery rate.

(a) Compute the adjusted p-values for the p-values in Table 10.2. Note that the adjusted p-values may not be monotonic. In that case, an adjusted p-value that is larger than its successor is changed to have the same value as its successor.

(b) If the desired FDR is 20%, i.e., $\alpha = 0.20$, then for which p-values is H_0 rejected?

(c) Suppose that we use the Bonferroni procedure instead. For different values of α, namely 0.01, 0.05, and 0.10, compute the modified p-value threshold, $\alpha^* = \alpha/10$, that the Bonferroni procedure will use to evaluate p-values. Then determine, for each value of α^*, for which p-values, H_0 will be rejected. (If a p-value equals the threshold, it is rejected.)

	1	2	3	4	5	6	7	8	9	10
original p-values	0.001	0.005	0.05	0.065	0.15	0.21	0.25	0.3	0.45	0.5

Table 10.2. Ordered Collection of p-values.

Answer:

(a) See Table 10.3.

(b) H_0 is rejected for the first 4 p-values, i.e., 0.001, 0.005, 0.05, and 0.065.

(c) $\alpha^* = 0.001$, reject p-value 0.001. $\alpha^* = 0.005$, reject p-values 0.001 and 0.005. $\alpha^* = 0.01$, reject p-values 0.001 and 0.005.

	1	2	3	4	5	6	7	8	9	10
original p-values	0.001	0.005	0.05	0.065	0.15	0.21	0.25	0.3	0.45	0.5
p'_i	0.01	0.025	0.1667	0.1625	0.3	0.35	0.357	0.375	0.5	0.5
adjusted p-values	0.01	0.025	0.1625	0.1625	0.3	0.35	0.357	0.375	0.5	0.5

Table 10.3. Ordered Collection of p-values and corresponding adjusted p-values.

9. The positive false discovery rate (pFDR) is similar to the false discovery rate defined in Section 10.1.3 but assumes that the number of true positives is greater than 0. Calculation of the pFDR is similar to that of FDR, but requires an assumption on the value of m_0, the number of results that satisfy the null hypothesis. The pFDR is less conservative than FDR, but more complicated to compute.

   The positive false discovery rate also allows the definition of an FDR analogue of the p-value. The q-value is the expected fraction of hypotheses that will be false if the given hypothesis is accepted. Specifically, the q-value associated with a p-value is the expected proportion of false positives among all hypotheses that are more extreme, i.e., have a lower p-value. Thus, the q-value associated with a p-value is the positive false discovery rate that would result if the p-value was used as the threshold for rejection.

   Below we show 50 p-values, their Benjamini-Hochberg adjusted p-values, and their q-values.

   p-values
   0.0000 0.0000 0.0002 0.0004 0.0004 0.0010 0.0089 0.0089 0.0288 0.0479
   0.0755 0.0755 0.0755 0.1136 0.1631 0.2244 0.2964 0.3768 0.3768 0.3768
   0.4623 0.4623 0.4623 0.5491 0.5491 0.6331 0.7107 0.7107 0.7107 0.7107
   0.7107 0.8371 0.9201 0.9470 0.9470 0.9660 0.9660 0.9660 0.9790 0.9928
   0.9928 0.9928 0.9928 0.9960 0.9960 0.9989 0.9989 0.9995 0.9999 1.0000
   BH adjusted p-values
   0.0000 0.0000 0.0033 0.0040 0.0040 0.0083 0.0556 0.0556 0.1600 0.2395
   0.2904 0.2904 0.2904 0.4057 0.5437 0.7012 0.8718 0.9420 0.9420 0.9420
   1.0000 1.0000 1.0000 1.0000 1.0000 1.0000 1.0000 1.0000 1.0000 1.0000
   1.0000 1.0000 1.0000 1.0000 1.0000 1.0000 1.0000 1.0000 1.0000 1.0000
   1.0000 1.0000 1.0000 1.0000 1.0000 1.0000 1.0000 1.0000 1.0000 1.0000
   q-Values
   0.0000 0.0000 0.0023 0.0033 0.0033 0.0068 0.0454 0.0454 0.1267 0.1861
   0.2509 0.2509 0.2509 0.3351 0.4198 0.4989 0.5681 0.6257 0.6257 0.6257
   0.6723 0.6723 0.6723 0.7090 0.7090 0.7375 0.7592 0.7592 0.7592 0.7592
   0.7592 0.7879 0.8032 0.8078 0.8078 0.8108 0.8108 0.8108 0.8129 0.8150
   0.8150 0.8150 0.8150 0.8155 0.8155 0.8159 0.8159 0.8160 0.8161 0.8161

(a) How many p-values are considered significant using BH adjusted p-values and thresholds of 0.05, 0.10, 0.15, 0.20, 0.25, and 0.30?

(b) How many p-values are considered significant using q-values and thresholds of 0.05, 0.10, 0.15, 0.20, 0.25, and 0.30?

(c) Compare the two sets of results.

Answer:

(a) From 0.05 to 0.30: 6, 7, 7, 9, 10, 13,

(b) From 0.05 to 0.30: 7, 7, 7, 10, 10, 13,

(c) The results from both approaches are very similar, but the q-values are always less than or equal to the corresponding adjusted p-values. Thus, the number of significant values for any threshold for the positive FDR approach is always greater than or equal to the number of significant values for the BH FDR approach.

10. An alternative to the definitions of the false discovery rate discussed so far is the local false discover rate, which is based on modeling the observed values of the test statistic as a mixture of two distributions, where most of the observations come from the null distribution and some observations (the interesting ones) come from the alternative distribution. (See Section 8.2.2 for more information on mixture models.) If Z is our test statistic, the density, $f(z)$ of Z, is given by the following:

$$f(z) = p_0 f_0(z) + p_1 f_1(z), \tag{10.2}$$

where p_0 is the probability an instance comes from the null distribution, $f_0(z)$ is the distribution of the p-values under the null hypothesis, p_1 is the probability an instance comes from the alternative distribution, and $f_1(z)$ is the distribution of p-values under the alternative hypothesis.

Using Bayes theorem, we can derive the probability of the null hypothesis for any value of z as follows.

$$p(H_0|z) = f(H_0 \text{ and } z)/f(z) = p_0 f_0(z)/f(z) \tag{10.3}$$

The quantity, $p(H_0|z)$, is the quantity that we would like to define as the local fdr. Since p_0 is often close to 1, the local false discovery rate, represented as fdr, all lowercase, is defined as the following:

$$\text{fdr(z)} = \frac{f_0(z)}{f(z)} \tag{10.4}$$

This is a point estimate, not an interval estimate as with the standard FDR, which is based on p-value, and as such, it will vary with the value of the test

statistic. Note that the local fdr has an easy interpretation, namely as the ratio of the density of observations from the null distribution to observations from both the null and alternative distributions. It also has the advantage of being interpretable directly as a real probability.

The challenge, of course, is in estimating the densities involved in Equation 10.4, which are usually estimated empirically. We consider the following simple case, where we specify the distributions by Gaussian distributions. The null distribution is given by, $f_0(z) = \mathcal{N}(0, 1)$, while the alternative distribution is given by $f_0(z) = \mathcal{N}(3, 1)$. $p_0 = 0.999$ and $p_1 = 0.001$.

(a) Compute $p(H_0|z)$ for the following values of z: 2, 2.5, 3, 3.5, 4, 4.5, 5.

(b) Compute the local fdr for the following values of z: 2, 2.5, 3, 3.5, 4, 4.5, 5.

(c) How close are these two sets of values?

Answer:

(a) From the smallest value of z to the largest, the results are as follows: 0.9955, 0.9803, 0.9173, 0.7123, 0.3559, 0.1098, 0.0268

(b) From the smallest value of z to the largest, the results are as follows: 0.9965, 0.9813, 0.9183, 0.7130, 0.3562, 0.1099, 0.0268

(c) The approximation is actually quite good in this case, but admittedly, we did not have to estimate the distributions as is typically the case.

11. The following are two alternatives to swap randomization—presented in Section 10..4.2—for randomizing a binary matrix so that the number of 1's in any row and column are preserved. Examine each method and (i) verify that it does indeed preserve the number of 1's in any row and column, and (ii) identify the problem with the alternative approach.

(a) Randomly permute the order of the columns and rows. An example is shown in Figure 10.1.

	i_1	i_2	i_3
t_1	1	1	0
t_2	1	0	1
t_3	1	1	1

	i_3	i_2	i_1
t_2	0	0	1
t_3	1	1	1
t_1	1	1	1

Figure 10.1. A 3 × 3 matrix before and after randomizing the order of the rows and columns. The leftmost matrix is the original.

(b) Figure 10.2 shows another approach to randomizing a binary matrix. This approach converts the binary matrix to a row-column representation, then randomly reassigns the columns to various entries, and finally converts the data back into the original binary matrix format.

	i_1	i_2	i_3	i_4
t_1	1	1	0	0
t_2	0	1	0	1
t_3	1	1	1	0
t_4	0	0	1	1

row	*col*
1	1
1	2
2	2
2	4
3	1
3	2
3	3
4	3
4	4

row	*col*
1	4
1	2
2	2
2	1
3	1
3	2
3	3
4	3
4	1

	i_1	i_2	i_3	i_4
t_1	0	1	0	1
t_2	1	1	0	0
t_3	0	1	1	1
t_4	1	0	1	0

Figure 10.2. A 4 $\times$ 4 matrix before and after randomizing the entries. From right to left, the tables represent the following: The original binary data matrix, the matrix in row-column format, the row-column format after randomly permuting the entries in the *col* column, and the matrix reconstructed from the randomized row-column representation.

Answer:

(a) The method works for the given example, but it doesn't achieve the goal which is to produce a data matrix that is different from the original because in association analysis the order of rows and columns does not matter to the results produced. In other words, the itemsets and association rules found in the original matrix will be the same as the old.

(b) The method words for the given example, but does not work in general. The reason is that randomly permuting the columns assigned to the row numbers can result in collision, i.e., the same column number can be assigned to identical row numbers. Depending on how this scheme is implemented this can result in a non-binary matrix or a missing entry both of which will violate the desired outcome of preserving the number of 1's in each row and column.

Bibliography

[1] W. W. Cohen. Fast effective rule induction. In *Proc. of the 12th Intl. Conf. on Machine Learning*, pages 115–123, Tahoe City, CA, July 1995.

[2] S. Cost and S. Salzberg. A weighted nearest neighbor algorithm for learning with symbolic features. *Machine Learning*, 10:57–78, 1993.

[3] J. Fürnkranz and G. Widmer. Incremental reduced error pruning. In *Proc. of the 11th Intl. Conf. on Machine Learning*, pages 70–77, New Brunswick, NJ, July 1994.

[4] J. Hartigan. *Clustering Algorithms*. Wiley, New York, 1975.

[5] R. A. Jarvis and E. A. Patrick. Clustering using a similarity measure based on shared nearest neighbors. *IEEE Transactions on Computers*, C-22(11):1025–1034, 1973.

[6] R. Kohavi. A study on cross-validation and bootstrap for accuracy estimation and model selection. In *Proc. of the 15th Intl. Joint Conf. on Artificial Intelligence*, pages 1137–1145, Montreal, Canada, August 1995.