**Introduction to Data Mining**

**Instructor’s Solution Manual**

**Pang-Ning Tan Michael Steinbach Vipin Kumar**

**Copyright** Qc 2006 Pearson Addison-Wesley. All rights reserved.

**Contents**

**1 Introduction 1**

**2 Data 5**

**3 Exploring Data 19**

**4 Classiﬁcation: Basic Concepts, Decision Trees, and Model**

**Evaluation 25**

**5 Classiﬁcation: Alternative Techniques 45**

**6 Association Analysis: Basic Concepts and Algorithms 71**

**7 Association Analysis: Advanced Concepts 95**

**8 Cluster Analysis: Basic Concepts and Algorithms 125**

**9 Cluster Analysis: Additional Issues and Algorithms 147**

**10 Anomaly Detection 157**

iii

**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 prof- itability.

No. This is an accounting calculation, followed by the applica- tion 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 num- bers.

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, solu- tions 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

**2 Chapter 1** Introduction

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 de- tection. 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 In- ternet search engine company. Describe how data mining can help the company by giving speciﬁc examples of how techniques, such as clus- tering, 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**

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?

Field 3 ≈ 7 for the values displayed. While it can be dangerous to draw con- clusions from such a small sample, the two ﬁelds seem to contain essentially the same information.

Field 2

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, quan- titative, ratio

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

(f ) Height above sea level. Continuous, quantitative, interval/ratio (de- pends 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, translu- cent, transparent. Discrete, qualitative, ordinal

(j) Military rank. Discrete, qualitative, ordinal

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

(l) Density of a substance in grams per cubic centimeter. Discrete, quan- titative, 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 be- lieves 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 satisfac- tion 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

Satisfaction(product) =

number of complaints for the product 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 satis- faction 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 gener- ating an ordinal ranking of the product variations in terms of customer preference? Explain.

Yes, the marketing director is in trouble. A customer may give incon- sistent 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 order- ing the product, but still only allow pairwise comparisons. In general, creating an ordinal measurement scale based on pairwise comparison is diﬃcult because of possible inconsistencies.

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

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

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

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

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

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 | 0  1 | 0  0 | ...  ... | 1  0 | 0  1 | 0  0 | 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 autocorre- lation: 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 loca- tions 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 *ijth* 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 L2 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 ex- periments. Compare the data quality issues involved in observational science with those of experimental science and data mining.

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

ing satellites became available, measurements of sea surface temperature re- lied 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 experi- ment. 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 ex- plicity, precision is often expressed in terms of the number of signiﬁcant digits used to represent a value. Thus, a single precision number can only represent values with up to 32 bits, ≈ 9 decimal digits of precision. However, often the precision of a value represented using 32 bits (64 bits) is far less than 32 bits (64 bits).

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

(1) Text ﬁles can be easily inspected by typing the ﬁle or viewing it with a text editor.

(2) Text ﬁles are more portable than binary ﬁles, both across systems and programs.

(3) Text ﬁles can be more easily modiﬁed, for example, using a text editor or perl.

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

(a) Is noise ever interesting or desirable? Outliers?

No, by deﬁnition. Yes. (See Chapter 10.)

(b) Can noise objects be outliers?

Yes. Random distortion of the data is often responsible for outliers.

(c) Are noise objects always outliers?

No. Random distortion can result in an object or value much like a normal one.

(d) Are outliers always noise objects?

No. Often outliers merely represent a class of objects that are diﬀerent from normal objects.

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

Yes.

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 *ith* 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 dupli- cate 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 ele- phants: *weight, height, tusk length, trunk length,* and *ear area*. Based on these measurements, what sort of similarity measure from Section 2.4 would you use to compare or group these elephants? Justify your answer and ex- plain any special circumstances.

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. Eu- clidean 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 *ith* group is of size *mi* . 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* ∗ *mi /m* elements from each group.

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

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

16. Consider a document-term matrix, where *tfij* is the frequency of the *ith* word (term) in the *jth* document and *m* is the number of documents. Consider the variable transformation that is deﬁned by

*m*

t = *tfij* ∗ log *df*

*tf*

*ij*

*i*

*,* (2.1)

where *dfi* is the number of documents in which the *ith* 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 docu- ment? 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 vec- tors. The Jaccard similarity is a measure of the similarity between two binary vectors. Compute the Hamming distance and the Jaccard simi- larity between the following two binary vectors.

**x** = 0101010001

**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, **x** and **y**, calculate the indicated similarity or dis- tance measures.

(a) **x** = (1*,* 1*,* 1*,* 1)*,* **y** = (2*,* 2*,* 2*,* 2) cosine, correlation, Euclidean cos(**x***,* **y**) = 1, corr(**x***,* **y**) = 0/0 (undeﬁned), Euclidean(**x***,* **y**) = 2

(b) **x** = (0*,* 1*,* 0*,* 1)*,* **y** = (1*,* 0*,* 1*,* 0) cosine, correlation, Euclidean, Jaccard cos(**x***,* **y**) = 0, corr(**x***,* **y**) = −1*,* Euclidean(**x***,* **y**) = 2, Jaccard(**x***,* **y**) = 0

(c) **x** = (0*,* −1*,* 0*,* 1)*,* **y** = (1*,* 0*,* −1*,* 0) cosine, correlation, Euclidean

cos(**x***,* **y**) = 0, corr(**x***,* **y**)=0, Euclidean(**x***,* **y**) = 2

(d) **x** = (1*,* 1*,* 0*,* 1*,* 0*,* 1)*,* **y** = (1*,* 1*,* 1*,* 0*,* 0*,* 1) cosine, correlation, Jaccard cos(**x***,* **y**) = 0.75, corr(**x***,* **y**) = 0.25, Jaccard(**x***,* **y**) = 0.6

(e) **x** = (2*,* −1*,* 0*,* 2*,* 0*,* −3)*,* **y** = (−1*,* 1*,* −1*,* 0*,* 0*,* −1) cosine, correlation

cos(**x***,* **y**) = 0, corr(**x***,* **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 devia- tion in cases where cosine and correlation are the same and diﬀerent.)

For two vectors, **x** and **y** that have a mean of 0, corr(**x**, **y**) = cos(**x**, **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 normal- ized to have an L2 length of 1. What general observation can you make about the relationship between Euclidean distance and cosine similarity when vectors have an L2 norm of 1?

Since all the 100,000 points fall on the curve, there is a functional rela- tionship 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 Eu- clidean distance when each data object has an L2 length of 1.

Let **x** and **y** be two vectors where each vector has an L2 length of 1. For such vectors, the variance is just *n* times the sum of its squared attribute values and the correlation between the two vectors is their dot product divided by *n*.

*d*(**x***,* **y**) =

r I

*n*

I)(*xk* − *yk* )

\ 2

*k*=1

r I

*n*

= I) *x*2 − 2*xk yk* + *y*2

\ *k k*

*k*=1

= /1 − 2cos(**x***,* **y**)+1

= /2(1 − cos(**x***,* **y**))

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

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

*d*(**x***,* **y**) =

r I

*n*

I)(*xk* − *yk* )

\ 2

*k*=1

r I

*n*

= I) *x*2 − 2*xk yk* + *y*2

\ *k k*

*k*=1

= /*n* − 2*n*corr(**x***,* **y**)+ *n*

= /2*n*(1 − corr(**x***,* **y**))

21. Show that the set diﬀerence metric given by

*d*(*A, B*) = *size*(*A* − *B*)+ *size*(*B* − *A*)

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

1.4

1.2

1

Euclidean Distance

0.8

0.6

0.4

0.2

0

0 0.2 0.4 0.6 0.8 1

Cosine Similarity

1.4

1.2

1

Euclidean Distance

0.8

0.6

0.4

0.2

0

0 0.2 0.4 0.6 0.8 1

Correlation

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

(b) Relationship between Euclidean distance and correlation.

**Figure 2.1.** Figures for exercise 20.

1(a). Because the size of a set is greater than or equal to 0, *d*(**x***,* **y**) ≥ 0.

1(b). if *A* = *B*, then *A* − *B* = *B* − *A* = empty set and thus *d*(**x***,* **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*) − 2*size*(*A* ∩ *B*).

∴ *d*(*A, B*)+*d*(*B, C* ) = *size*(*A*)+*size*(*C* )+2*size*(*B*)−2*size*(*A*∩*B*)−2*size*(*B* ∩

*C* )

Since *size*(*A* ∩ *B*) ≤ *size*(*B*) and *size*(*B* ∩ *C* ) ≤ *size*(*B*),

*d*(*A, B*)+ *d*(*B, C* ) ≥ *size*(*A*)+ *size*(*C* )+ 2*size*(*B*) − 2*size*(*B*) = *size*(*A*)+

*size*(*C* ) ≥ *size*(*A*)+ *size*(*C* ) − 2*size*(*A* ∩ *C* ) = *d*(*A, C* )

∴ *d*(*A, C* ) ≤ *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* =

*corr* if *corr* ≥ 0

0 if *corr <* 0

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.

to transform this similarity value into a dissimilarity value in the interval

[0,∞].

*s* and *d* = − log *s*.

*d* = 1− *s*

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

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

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

(b) How might you deﬁne the distance between two sets of points in Eu- clidean 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 **x**. (It does not matter if **x** ∈ *S*.)

(a) If the goal is to ﬁnd all points within a speciﬁed distance *ε* of point **y**, **y** /= **x**, explain how you could use the triangle inequality and the al- ready calculated distances to **x** to potentially reduce the number of dis- tance calculations necessary? Hint: The triangle inequality, *d*(**x***,* **z**) ≤ *d*(**x***,* **y**)+ *d*(**y***,* **x**), can be rewritten as *d*(**x***,* **y**) ≥ *d*(**x***,* **z**) − *d*(**y***,* **z**).

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

Hint: If **z** is an arbitrary point of *S*, then the triangle inequality,

*d*(**x***,* **y**) ≤ *d*(**x***,* **z**)+*d*(**y***,* **z**), can be rewritten as *d*(**y***,* **z**) ≥ *d*(**x***,* **y**)−*d*(**x***,* **z**).

Another application of the triangle inequality starting with *d*(**x***,* **z**) ≤ *d*(**x***,* **y**) + *d*(**y***,* **z**), shows that *d*(**y***,* **z**) ≥ *d*(**x***,* **z**) − *d*(**x***,* **y**). If the lower bound of *d*(**y***,* **z**) obtained from either of these inequalities is greater

than *E*, then *d*(**y***,* **z**) does not need to be calculated. Also, if the upper bound of *d*(**y***,* **z**) obtained from the inequality *d*(**y***,* **z**) ≤ *d*(**y***,* **x**)+ *d*(**x***,* **z**) is less than or equal to *E*, then *d*(**x***,* **z**) does not need to be calculated.

(b) In general, how would the distance between **x** and **y** aﬀect the number of distance calculations?

If **x** = **y** then no calculations are necessary. As **x** becomes farther away, typically more distance calculations are needed.

(c) Suppose that you can ﬁnd a small subset of points *S*t, 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*t, and that you also have the pairwise distance matrix for *S*t. 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 **x** and **y** be the two points and let **x**∗ and **y**∗ be the points in *S*t that are closest to the two points, respectively. If *d*(**x**∗*,* **y**∗) + 2*E* ≤ *β*, then we can safely conclude *d*(**x***,* **y**) ≤ *β*. Likewise, if *d*(**x**∗*,* **y**∗) − 2*E* ≥ *β*, then we can safely conclude *d*(**x***,* **y**) ≥ *β*. These formulas are derived by considering the cases where **x** and **y** are as far from **x**∗ and **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, **x** and **y**, that satisﬁes the metric axioms given on page 70. Speciﬁcally, *d*(**x***,* **y**) = 1 − J(**x***,* **y**).

1(a). Because J(**x***,* **y**) ≤ 1*, d*(**x***,* **y**) ≥ 0.

1(b). Because J(**x***,* **x**) = 1*, d*(**x***,* **x**) = 0

2. Because J(**x***,* **y**) = J(**y***,* **x**)*, d*(**x***,* **y**) = *d*(**y***,* **x**)

3. (Proof due to Jeﬀrey Ullman)

minhash(**x**) is the index of ﬁrst nonzero entry of **x**

*prob*(minhash(**x**) = *k*) is the probability tha minhash(**x**) = *k* when **x** is ran- domly permuted.

Note that *prob*(minhash(**x**) = minhash(**y**)) = J(**x***,* **y**) (minhash lemma) Therefore, *d*(**x***,* **y**) = 1−*prob*(minhash(**x**) = minhash(**y**)) = *prob*(minhash(**x**) /= minhash(**y**))

We have to show that,

*prob*(minhash(**x**) /= minhash(**z**)) ≤ *prob*(minhash(**x**) /= minhash(**y**)) +

*prob*(minhash(**y**) /= minhash(**z**)

minhash(**x**) /= minhash(**y**) and minhash(**y**) /= minhash(**z**) must be true.

27. Show that the distance measure deﬁned as the angle between two data vec- tors, **x** and **y**, satisﬁes the metric axioms given on page 70. Speciﬁcally, *d*(**x***,* **y**) = arccos(cos(**x***,* **y**)).

Note that angles are in the range 0 to 180◦.

1(a). Because 0 ≤ cos(**x***,* **y**) ≤ 1*, d*(**x***,* **y**) ≥ 0.

1(b). Because cos(**x***,* **x**) = 1*, d*(**x***,* **x**) = arccos(1) = 0

2. Because cos(**x***,* **y**) = cos(**y***,* **x**)*, d*(**x***,* **y**) = *d*(**y***,* **x**)

3. If the three vectors lie in a plane then it is obvious that the angle between **x** and **z** must be less than or equal to the sum of the angles between **x** and **y** and **y** and **z**. If **y**t is the projection of **y** into the plane deﬁned by **x** and **z**, then note that the angles between **x** and **y** and **y** and **z** are greater than those between **x** and **y**t and **y**t and **z**.

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

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

**3**

**Exploring Data**

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

MATLAB and R have excellent facilities for visualization. Most of the ﬁg- ures in this chapter were created using MATLAB. R is freely available from [http://www.r-project.org/.](http://www.r-project.org/)

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

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

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

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

It would have been better to state this more generally as “What are the issues

... ,” since selection, as well as arrangement plays a key issue in displaying a three-dimensional plot.

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

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

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

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

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

Be sure to address the following issues:

• **Representation.** How will you map objects, attributes, and relation- ships to visual elements?

• **Arrangement.** Are there any special considerations that need to be taken into account with respect to how visual elements are displayed? Speciﬁc examples might be the choice of viewpoint, the use of trans-

parency, or the separation of certain groups of objects.

• **Selection.** How will you handle a large number of attributes and data objects?

The following solutions are intended for illustration.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

(a) If the line representing the median of the data is in the middle of the box, then the data is symmetrically distributed, at least in terms of the

75% of the data between the ﬁrst and third quartiles. For the remain- ing data, the length of the whiskers and outliers is also an indication, although, since these features do not involve as many points, they may be misleading.

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

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

Figure 3.12.

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

10. Comment on the use of a box plot to explore a data set with four attributes:

age, weight, height, and income.

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

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

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

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

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

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

The fact that the attribute values are ordered.

14. Describe the types of situations that produce sparse or dense data cubes.

Illustrate with examples other than those used in the book.

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

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

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

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

If it is sparse, identify the cells that are empty.

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

**Table 3.1.** Fact table for Exercise 16.

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

**Table 3.2.** Data cube for Exercise 16.

**Location ID**

**Product ID**

|  |  |  |
| --- | --- | --- |
|  | 1 2 3 | Total |
| 1  2 | 10 0 6  5 22 0 | 16  27 |
| Total | 15 22 6 | 43 |

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

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

**4**

**Classification: Basic Concepts, Decision Trees, and Model Evaluation**

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

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **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** |

T F

B **B**

T F T F

C C C C

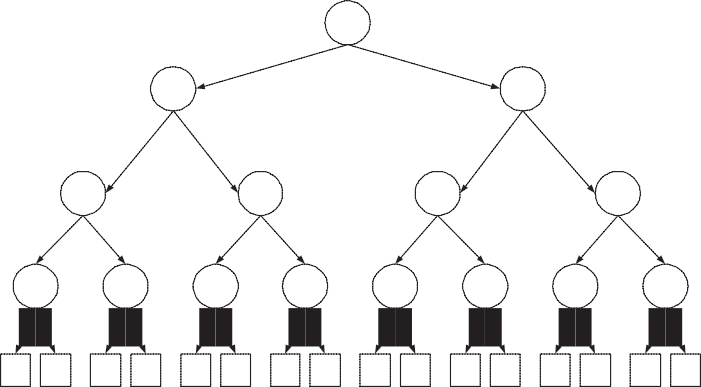
T F T F T F T F

D D D D D D D D

T F T F T F T F T F T F T F T F

T F F T F T T F F T T F T F F T

**Figure 4.1.** Decision tree for parity function of four Boolean attributes.



The preceding tree cannot be simpliﬁed.

2. Consider the training examples shown in Table 4.1 for a binary classiﬁcation problem.

**Table 4.1.** Data set for Exercise 2.

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

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

**Answer:**

Gini = 1 − 2 × 0*.*52 = 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 × 0*.*52 = 0*.*5. The gini for Female is also 0*.*5. Therefore, the overall gini for Gender is 0*.*5 × 0*.*5+ 0*.*5 × 0*.*5 = 0*.*5.

**Table 4.2.** Data set for Exercise 3.

|  |  |  |
| --- | --- | --- |
| Instance | *a*1 *a*2 *a*3 | Target Class |
| 1  2  3  4  5  6  7  8  9 | T T 1.0  T T 6.0  T F 5.0  F F 4.0  F T 7.0  F T 3.0  F F 8.0  T F 7.0  F T 5.0 | +  +  −  +  −  −  −  +  − |

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

**Answer:**

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 IDs.

3. Consider the training examples shown in Table 4.2 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 log2 (4*/*9) − 5*/*9 log2 (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  F | 3  1 | 1  4 |

The entropy for *a*1 is

4 1

9 − (3*/*4) log2 (3*/*4) − (1*/*4) log2 (1*/*4)

5 1

+ 9 − (1*/*5) log2 (1*/*5) − (4*/*5) log2 (4*/*5)

= 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  F | 2  2 | 3  2 |

The entropy for *a*2 is

5 1

9 − (2*/*5) log2 (2*/*5) − (3*/*5) log2 (3*/*5)

4 1

+ 9 − (2*/*4) log2 (2*/*4) − (2*/*4) log2 (2*/*4)

= 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 infor- mation 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

4 1 5

1 − (3*/*4)2 − (1*/*4)2 +

9

9

1

1 − (1*/*5)2 − (4*/*5)2

= 0*.*3444*.*

For attribute *a*2 , the gini index is

5 1 4

1 − (2*/*5)2 − (3*/*5)2 +

9

9

1

1 − (2*/*4)2 − (2*/*4)2

= 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 *,* ··· *, yc* } denote the *c* classes and *X* = {*x*1 *, x*2 *,* ··· *, xk* } denote the *k* attribute values of an attribute *X* . Before a node is split on *X* , the entropy is:

*c c k*

*E*(*Y* ) = − ) *P* (*yj* ) log2 *P* (*yj* ) = ) ) *P* (*xi , yj* ) log2 *P* (*yj* )*,* (4.1)

*j*=1

*j*=1 *i*=1

where we have used the fact that *P* (*yj* ) = L*k*

*i*=1

*P* (*xi , yj* ) from the law of

total probability.

After splitting on *X* , the entropy for each child node *X* = *xi* is:

*c*

*E*(*Y* |*xi* ) = − ) *P* (*yj* |*xi* ) log2 *P* (*yj* |*xi* ) (4.2)

*j*=1

where *P* (*yj* |*xi* ) is the fraction of examples with *X* = *xi* that belong to class *yj* . The entropy after splitting on *X* is given by the weighted entropy of the children nodes:

*E*(*Y* |*X* ) =

*k*

) *P* (*xi* )*E*(*Y* |*xi* )

*i*=1

*k c*

= − ) ) *P* (*xi* )*P* (*yj* |*xi* ) log2 *P* (*yj* |*xi* )

*i*=1 *j*=1

*k c*

= − ) ) *P* (*xi , yj* ) log2 *P* (*yj* |*xi* )*,* (4.3)

*i*=1 *j*=1

where we have used a known fact from probability theory that *P* (*xi , yj* ) = *P* (*yj* |*xi* ) × *P* (*xi* ). 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* ) ≤ *E*(*Y* ). Let us com- pute the diﬀerence between the entropies after splitting and before splitting, i.e., *E*(*Y* |*X* ) − *E*(*Y* ), using Equations 4.1 and 4.3:

*E*(*Y* |*X* ) − *E*(*Y* )

*k c k c*

= − ) ) *P* (*xi , yj* ) log2 *P* (*yj* |*xi* )+ ) ) *P* (*xi , yj* ) log2 *P* (*yj* )

*i*=1 *j*=1

*k c*

*P* (*y* )

*i*=1 *j*=1

= ) ) *P* (*xi , yj* ) log2

*i*=1 *j*=1

*j*

*P* (*yj* |*xi* )

*k c P* (*x* )*P* (*y* )

= ) ) *P* (*xi , yj* ) log2

*i*

*j*

(4.4)

*i*=1 *j*=1

*P* (*xi , yj* )

To prove that Equation 4.4 is non-positive, we use the following property of a logarithmic function:

*d d*

) *ak* log(*zk* ) ≤ log ( ) *ak zk* )*,* (4.5)

*k*=1

subject to the condition that L*d*

*k*=1

*k*=1

*ak* = 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 4.4 can be bounded as follows:

*k c P* (*x* )*P* (*y* ) 1

*E*(*Y* |*X* ) − *E*(*Y* ) ≤ log2

) ) *P* (*xi , yj* ) *i j*

*i*=1 *j*=1

*P* (*xi , yj* )

*k c* 1

= log2

) *P* (*xi* ) ) *P* (*yj* )

*i*=1

= log2 (1)

= 0

*j*=1

Because *E*(*Y* |*X* ) − *E*(*Y* ) ≤ 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  T T T T F F F T T | F  T T F T F F F 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:

*Eorig* = −0*.*4 log 0*.*4 − 0*.*6 log 0*.*6 = 0*.*9710

The information gain after splitting on A is:

4 4 3 3

*EA*=*T* = − 7 log 7 − 7 log 7 = 0*.*9852

3 3 0 0

*EA*=*F* = − 3 log 3 − 3 log 3 = 0

∆ = *Eorig* − 7*/*10*EA*=*T* − 3*/*10*EA*=*F* = 0*.*2813

The information gain after splitting on B is:

3 3 1 1

*EB*=*T* = − 4 log 4 − 4 log 4 = 0*.*8113

1 1 5 5

*EB*=*F* = − 6 log 6 − 6 log 6 = 0*.*6500

∆ = *Eorig* − 4*/*10*EB*=*T* − 6*/*10*EB*=*F* = 0*.*2565

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:

*Gorig* = 1 − 0*.*42 − 0*.*62 = 0*.*48

The gain in gini after splitting on A is:

3 − 3

|  |  |  |
| --- | --- | --- |
| *GA*=*T* | = | 1 − |
| *GA*=*F* | = | 1 = |

|  |  |
| --- | --- |
| 4 2 | 3 2 |
| 7 | − 7 |
| 3 2 | 0 2 |

= 0*.*4898

= 0

∆ = *Gorig* − 7*/*10*GA*=*T* − 3*/*10*GA*=*F* = 0*.*1371

The gain in gini after splitting on B is:

6 − 6

|  |  |  |
| --- | --- | --- |
| *GB*=*T* | = | 1 − |
| *GB*=*F* | = | 1 = |

|  |  |
| --- | --- |
| 1 2 | 3 2 |
| 4 | − 4 |
| 1 2 | 5 2 |

= 0*.*3750

= 0*.*2778

∆ = *Gorig* − 4*/*10*GB*=*T* − 6*/*10*GB*=*F* = 0*.*1633

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 decreas- ing 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, ∆, 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 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 com- pute the error rates for attributes *X* , *Y* , and *Z* . For attribute *X* , the corresponding counts are:

|  |  |  |
| --- | --- | --- |
| *X* | C1 | C2 |
| 0  1 | 60  40 | 60  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  1 | 40  60 | 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  1 | 30  70 | 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 in- volve 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  1 | 15  15 | 45  25 |

The error rate in both cases (*X* and *Y* ) are (15 + 15)*/*100 = 0*.*3.

|  |  |  |
| --- | --- | --- |
| *Y* | C1 | C2 |
| 0  1 | 15  15 | 45  25 |

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

|  |  |  |
| --- | --- | --- |
| *Y* | C1 | C2 |
| 0  1 | 25  45 | 15  15 |

|  |  |  |
| --- | --- | --- |
| *X* | C1 | C2 |
| 0  1 | 45  25 | 15  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.

Z

0 1

X X or or Y Y

0 1 0 1

C2 C2 C1 C1

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 sub- sequent 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.

|  |  |  |
| --- | --- | --- |
| *Z* | C1 | C2 |
| 0  1 | 15  45 | 45  15 |

|  |  |  |
| --- | --- | --- |
| *Y* | C1 | C2 |
| 0  1 | 5  55 | 55  5 |

The error rate using attributes *Y* and *Z* are 10*/*120 and 30*/*120, re- spectively. 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  1 | 35  5 | 5  35 |

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

|  |  |  |
| --- | --- | --- |
| *Z* | C1 | C2 |
| 0  1 | 15  25 | 25  15 |

The corresponding two-level decision tree is shown below.

X

0 1

Y Y

0 1 0 1

C2 C1 C1 C2

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 heuris- tic does not always produce an optimal solution.

7. 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  F T F T F T F | T  T F F T T F F | T  T T T F F F F | 5  0  20  0  0  25  0  0 | 0  20  0  5  0  0  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

50 50 50

*Eorig* = 1 − max( 100 *,* 100 ) = 100 *.*

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

*A* = *T A* = *F*

+

|  |  |
| --- | --- |
| 25 | 25 |
| 0 | 50 |

−

*EA*=*T* = 1 − max(

*EA*=*F* = 1 − max(

25

25 *,*

25

75 *,*

0 0

) = = 0

25 25

50 25

) =

75 75

25 75 25

∆*A* = *Eorig* − 100 *EA*=*T* − 100 *EA*=*F* = 100

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

*B* = *T B* = *F*

+

|  |  |
| --- | --- |
| 30 | 20 |
| 20 | 30 |

−

20

*EB*=*T* = 50

20

*EB*=*F* = 50

50 50 10

∆*B* = *Eorig* − 100 *EB*=*T* − 100 *EB*=*F* = 100

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

*C* = *T C* = *F*

+

|  |  |
| --- | --- |
| 25 | 25 |
| 25 | 25 |

−

25

*EC* =*T* = 50

25

*EC* =*F* = 50

50 50 0

∆*C* = *Eorig* − 100 *EC* =*T* − 100 *EC* =*F* = 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  F T F | T  T F F | 0  0  25  0 | 20  5  0  25 |

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

25

*Eorig* = 75

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

*B* = *T B* = *F*

+

|  |  |
| --- | --- |
| 25 | 0 |
| 20 | 30 |

20

*EB*=*T* = 45

*EB*=*F* = 0

− ∆ = *E* 45 20 5

*B orig* − 75 *EB*=*T* − 75 *EB*=*F* = 75

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

*C* = *T C* = *F*

+

|  |  |
| --- | --- |
| 0 | 25 |
| 25 | 25 |

−

0

*EC* =*T* = 25

25

*EC* =*F* = 50

25 50

∆*C* = *Eorig* − 75 *EC* =*T* − 75 *EC* =*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 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:

*Eorig* = 25

50 *.*

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

*A* = *T A* = *F*

+

|  |  |
| --- | --- |
| 25 | 0 |
| 0 | 25 |

*EA*=*T* = 0

*EA*=*F* = 0

25

∆

−

*A* =

50

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

*B* = *T B* = *F*

+

|  |  |
| --- | --- |
| 5 | 20 |
| 20 | 5 |

−

5

*EB*=*T* = 25

5

*EB*=*F* = 25

15

∆*B* = 50

Therefore, *A* is chosen as the splitting attribute.

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 | – |

A

0 1

B C

0 1 0 1

+ \_ + \_

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 4.2.** Decision tree and data sets for Exercise 8.

For the *C* = *F* child, the error rate before splitting is: *Eorig* = 25 *.*

50

After splitting on attribute *A*, the error rate is:

*A* = *T A* = *F*

*EA*=*T* = 0

25

|  |  |
| --- | --- |
| 0 | 25 |
| 0 | 25 |

+ *EA*=*F* = 50

− ∆*A* = 0

After splitting on attribute *B*, the error rate is:

*B* = *T B* = *F*

+

|  |  |
| --- | --- |
| 25 | 0 |
| 0 | 25 |

*EB*=*T* = 0

*EB*=*F* = 0

25

∆

−

*B* =

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.

8. Consider the decision tree shown in Figure 4.2.

(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 × 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.

9. Consider the decision trees shown in Figure 4.3. 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.

C1

C1 C2 C3

C2 C3

C1 C2

(a) Decision tree with 7 errors (b) Decision tree with 4 errors

**Figure 4.3.** Decision trees for Exercise 9.

• 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 log2 *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 log2 *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 com- mits on the training set. Each error is encoded by log2 *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:

log2 (*m*) = log2 (16) = 4

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

plog2 (*k*)l = plog2 (3)l = 2

The cost for each misclassiﬁcation error is log2 (*n*).

The overall cost for the decision tree (a) is 2×4+3×2+7×log2 *n* = 14+7 log2 *n* and the overall cost for the decision tree (b) is 4×4+5×2+4×5 = 26+4 log2 *n*. According to the MDL principle, tree (a) is better than (b) if *n <* 16 and is

worse than (b) if *n>* 16.

10. While the .632 bootstrap approach is useful for obtaining a reliable estimate of model accuracy, it has a known limitation. Consider a two-class problem, where there are equal number of positive and negative examples in the data. Suppose the class labels for the examples are generated randomly. The clas- siﬁer used is an unpruned decision tree (i.e., a perfect memorizer). Determine the accuracy of the classiﬁer using each of the following methods.

(a) The holdout method, where two-thirds of the data are used for training and the remaining one-third are used for testing.

**Answer:**

Assuming that the training and test samples are equally representative, the test error rate will be close to 50%.

(b) Ten-fold cross-validation.

**Answer:**

Assuming that the training and test samples for each fold are equally representative, the test error rate will be close to 50%.

(c) The .632 bootstrap method.

**Answer:**

The training error for a perfect memorizer is 100% while the error rate for each bootstrap sample is close to 50%. Substituting this information into the formula for .632 bootstrap method, the error estimate is:

1 )

*b*

*b*

*i*=1

1

0*.*632 × 0*.*5+ 0*.*368 × 1

= 0*.*684*.*

(d) From the results in parts (a), (b), and (c), which method provides a more reliable evaluation of the classiﬁer’s accuracy?

**Answer:**

The ten-fold cross-validation and holdout method provides a better error estimate than the .632 boostrap method.

11. Consider the following approach for testing whether a classiﬁer A beats an- other classiﬁer B. Let *N* be the size of a given data set, *pA* be the accuracy of classiﬁer A, *pB* be the accuracy of classiﬁer *B*, and *p* = (*pA* + *pB* )*/*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* = *pA* − *pB .*

2*p*(1−*p*)

*N*

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

Table 4.3 compares the accuracies of three diﬀerent classiﬁers, decision tree classiﬁers, na¨ıve Bayes classiﬁers, and support vector machines, on various data sets. (The latter two classiﬁers are described in Chapter 5.)

**Table 4.3.** Comparing the accuracy of various classification methods.

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

**Answer:**

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

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

12. Let *X* be a binomial random variable with mean *Np* and variance *N p*(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:

Mean*, E*[*r*] = *E*[*X/N* ] = *E*[*X* ]*/N* = (*N p*)*/N* = *p*;

Variance*, E*[(*r* − *E*[*r*])2 ] = *E*[(*X/N* − *E*[*X/N* ])2 ]

= *E*[(*X* − *E*[*X* ])2 ]*/N* 2

= *N p*(1 − *p*)*/N* 2

= *p*(1 − *p*)*/N*

**5**

**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 −→ Value = Low

Mileage = Low −→ Value = High

Air Conditioner = Working, Engine = Good −→ Value = High

Air Conditioner = Working, Engine = Bad −→ Value = Low

Air Conditioner = Broken −→ 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 Fu¨rnkranz 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 : *A* −→ *C*

*R*2 : *A* ∧ *B* −→ *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:

*p* + (*N* − *n*)

*vIREP* =

*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. *vIREP* is actually similar to classiﬁcation accuracy for the validation set. IREP favors rules that have higher values of *vIREP* . On the other hand, RIPPER applies the following measure to determine whether a rule should be pruned:

*v* = *p* − *n .*

*RI PPER p* + *n*

(a) Suppose *R*1 is covered by 350 positive examples and 150 negative ex- amples, 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. There- fore, the FOIL’s information gain for *R*2 with respect to *R*1 is:

300

350 1

*Gain* = 300 × log2 350 − log2 500

= 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 *vIREP* for both rules. Which rule does IREP prefer?

**Answer:**

For this problem, *P* = 500, and *N* = 500.

For rule *R*1, *p* = 200 and *n* = 50. Therefore,

*VIREP* (*R*1) =

*p* + (*N* − *n*) =

*P* + *N*

200 + (500 − 50) = 0*.*65

1000

For rule *R*2, *p* = 100 and *n* = 5.

*p* + (*N* − *n*)

100 + (500 − 5)

*VIREP* (*R*2) =

=

*P* + *N*

= 0*.*595

1000

Thus, IREP prefers rule R1.

(c) Compute *vRI PPER* for the previous problem. Which rule does RIPPER

prefer?

**Answer:**

*p* − *n*

150

*VRI PPER* (*R*1) = *p* + *n* = 250 = 0*.*6

*p* − *n* 95

*VRI PPER* (*R*2) = *p* + *n* = 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* −→ + (covers 4 positive and 1 negative examples), *R*2 : *B* −→ + (covers 30 positive and 10 negative examples), *R*3 : *C* −→ + (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 ∅ −→ +. 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

4 100

4 × log2 5 − log2 500 = 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

30 100

30 × log2 40 − log2 500 = 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

100

100 × log2 190 − log2 500 = 139*.*6*.*

Therefore, *R*3 is the best candidate and *R*1 is the worst candidate ac- cording to FOIL’s information gain.

(c) The likelihood ratio statistic.

**Answer:**

For *R*1 , the expected frequency for the positive class is 5 × 100*/*500 = 1 and the expected frequency for the negative class is 5 × 400*/*500 = 4. Therefore, the likelihood ratio for *R*1 is

1

2 × 4 × log2 (4*/*1) + 1 × log2 (1*/*4)

= 12*.*

For *R*2 , the expected frequency for the positive class is 40 ×100*/*500 = 8 and the expected frequency for the negative class is 40 × 400*/*500 = 32. Therefore, the likelihood ratio for *R*2 is

1

2 × 30 × log2 (30*/*8) + 10 × log2 (10*/*32)

= 80*.*85

For *R*3 , the expected frequency for the positive class is 190 × 100*/*500 =

38 and the expected frequency for the negative class is 190 × 400*/*500 =

152. Therefore, the likelihood ratio for *R*3 is

1

2 × 100 × log2 (100*/*38) + 90 × log2 (90*/*152)

= 143*.*09

Therefore, *R*3 is the best candidate and *R*1 is the worst candidate ac- cording 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 candi- date and *R*3 is the worst candidate according to the m-estimate mea- sure.

5. Figure 5.1 illustrates the coverage of the classiﬁcation rules *R*1, *R*2, and *R*3.

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. *R*1 covers 12 positive examples and 3 negative examples. The expected frequency for the positive class is 15 × 29*/*50 = 8*.*7 and the

expected frequency for the negative class is 15×21*/*50 = 6*.*3. Therefore,

the likelihood ratio for *R*1 is

1

2 × 12 × log2 (12*/*8*.*7) + 3 × log2 (3*/*6*.*3)

= 4*.*71*.*

*R*2 covers 7 positive examples and 3 negative examples. The expected frequency for the positive class is 10 × 29*/*50 = 5*.*8 and the expected

**R1**

+

class = +

**R3 R2**

+ + +

|  |  |  |
| --- | --- | --- |
|  | + + | |
| ++  +  + +  + +  + + +  - | +  +  - - | + +  +  +  -  - |
|  |  | |

+

+ +

+ +

class = -

-

- - -

-

-

- -

+ +

- - -

-

-

- -

-

**Figure 5.1.** Elimination of training records by the sequential covering algorithm. *R*1, *R*2, and *R*3

represent regions covered by three different rules.

frequency for the negative class is 10 × 21*/*50 = 4*.*2. Therefore, the likelihood ratio for *R*2 is

1

2 × 7 × log2 (7*/*5*.*8) + 3 × log2 (3*/*4*.*2)

= 0*.*89*.*

*R*3 covers 8 positive examples and 4 negative examples. The expected frequency for the positive class is 12 × 29*/*50 = 6*.*96 and the expected frequency for the negative class is 12 × 21*/*50 = 5*.*04. Therefore, the likelihood ratio for *R*3 is

1

2 × 8 × log2 (8*/*6*.*96) + 4 × log2 (4*/*5*.*04)

= 0*.*5472*.*

*R*1 is the best rule and *R*3 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 *R*1), 66.67% (for *R*2), and 64.29% (for *R*3), respectively. Therefore *R*1 is the best rule and *R*3 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 *R*1), 68.0% (for *R*2), and 65.43% (for *R*3), respectively. Therefore *R*1 is the best rule and *R*3 is the worst rule according to the m-estimate measure.

(d) The rule accuracy after *R*1 has been discovered, where none of the examples covered by *R*1 are discarded).

**Answer:**

If the examples for *R*1 are not discarded, then *R*2 will be chosen because it has a higher accuracy (70%) than *R*3 (66.7%).

(e) The rule accuracy after *R*1 has been discovered, where only the positive examples covered by *R*1 are discarded).

**Answer:**

If the positive examples covered by *R*1 are discarded, the new accuracies for *R*2 and *R*3 are 70% and 60%, respectively. Therefore *R*2 is preferred over *R*3.

(f ) The rule accuracy after *R*1 has been discovered, where both positive and negative examples covered by *R*1 are discarded.

**Answer:**

If the positive and negative examples covered by *R*1 are discarded, the new accuracies for *R*2 and *R*3 are 70% and 75%, respectively. In this case, *R*3 is preferred over *R*2.

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*|*U G*) = 0*.*15, *P* (*S*|*G*) = 0*.*23, *P* (*G*) = 0*.*2, *P* (*U G*) = 0*.*8. We want to compute *P* (*G*|*S*).

According to Bayesian Theorem,

0*.*23 × 0*.*2

*P* (*G*|*S*) = 0*.*15

× 0*.*8+ 0*.*23

= 0*.*277*.* (5.1)

× 0*.*2

(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* (*U G*) *> P* (*G*).

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

**Answer:**

An undergraduate student because *P* (*U G*|*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*|*U G*) = 0*.*1, *P* (*D*|*G*) = 0*.*3.

*P* (*D*) = *P* (*U G*)*.P* (*D*|*U G*)+ *P* (*G*)*.P* (*D*|*G*) = 0*.*8 ∗ 0*.*1+ 0*.*2 ∗ 0*.*3 = 0*.*14.

*P* (*S*) = *P* (*S*|*U G*)*P* (*U G*)+ *P* (*S*|*G*)*P* (*G*) = 0*.*15 ∗0*.*8+0*.*23 ∗0*.*2 = 0*.*166.

*P* (*DS*|*G*) = *P* (*D*|*G*) × *P* (*S*|*G*) = 0*.*3 × 0*.*23 = 0*.*069 (using conditional

independent assumption)

*P* (*DS*|*U G*) = *P* (*D*|*U G*) × *P* (*S*|*U G*) = 0*.*1 × 0*.*15 = 0*.*015. We need to compute *P* (*G*|*DS*) and *P* (*U G*|*DS*).

0*.*069 × 0*.*2

0*.*0138

*P* (*G*|*DS*) =

=

*P* (*DS*) *P* (*DS*)

0*.*015 × 0*.*8

0*.*012

*P* (*U G*|*DS*) =

=

*P* (*DS*) *P* (*DS*)

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

7. Consider the data set shown in Table 5.1

**Table 5.1.** Data set for Exercise 7.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Record | *A* | *B* | *C* | Class |
| 1  2  3  4  5  6  7  8  9  10 | 0  0  0  0  0  1  1  1  1  1 | 0  0  1  1  0  0  0  0  1  0 | 0  1  1  1  1  1  1  1  1  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 ques- tion to predict the class label for a test sample (*A* = 0*,B* = 1*,C* = 0) using the na¨ıve Bayes approach.

**Answer:**

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

*P* (+|*A* = 0*,B* = 1*,C* = 0)

= *P* (*A* = 0*,B* = 1*,C* = 0|+) × *P* (+)

*P* (*A* = 0*,B* = 1*,C* = 0)

= *P* (*A* = 0|+)*P* (*B* = 1|+)*P* (*C* = 0|+) × *P* (+)

*K*

= 0*.*4 × 0*.*2 × 0*.*6 × 0*.*5*/K*

= 0*.*024*/K.*

*P* (−|*A* = 0*,B* = 1*,C* = 0)

= *P* (*A* = 0*,B* = 1*,C* = 0|−) × *P* (−)

*P* (*A* = 0*,B* = 1*,C* = 0)

= *P* (*A* = 0|−) × *P* (*B* = 1|−) × *P* (*C* = 0|−) × *P* (−)

*K*

= 0*/K*

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*

*P* (+|*A* = 0*,B* = 1*,C* = 0)

= *P* (*A* = 0*,B* = 1*,C* = 0|+) × *P* (+)

*P* (*A* = 0*,B* = 1*,C* = 0)

= *P* (*A* = 0|+)*P* (*B* = 1|+)*P* (*C* = 0|+) × *P* (+)

*K*

(4*/*9) × (3*/*9) × (5*/*9) × 0*.*5

=

*K*

= 0*.*0412*/K*

*P* (−|*A* = 0*,B* = 1*,C* = 0)

= *P* (*A* = 0*,B* = 1*,C* = 0|−) × *P* (−)

*P* (*A* = 0*,B* = 1*,C* = 0)

= *P* (*A* = 0|−) × *P* (*B* = 1|−) × *P* (*C* = 0|−) × *P* (−)

*K*

(5*/*9) × (4*/*9) × (2*/*9) × 0*.*5

=

*K*

= 0*.*0274*/K*

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 condi- tional 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 5.2.

(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 na¨ıve 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,

**Table 5.2.** Data set for Exercise 8.

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

*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,

*P* (*R*|+) = *P* (*A* = 1|+) × *P* (*B* = 1|+) × *P* (*C* = 1|+) = 0*.*192

*P* (*R*|−) = *P* (*A* = 1|−) × *P* (*B* = 1|−) × *P* (*C* = 1|−) = 0*.*032

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*) ×

*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) × *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.

**Attributes**

Distinguishing Attributes Noise Attributes

A1

Class A

A2

**Records**

B1

Class B

B2

**Figure 5.2.** Data set for Exercise 9.

9. (a) Explain how na¨ıve Bayes performs on the data set shown in Figure 5.2.

**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 (*A*1, *A*2,

*B*1, and *B*2), will na¨ıve 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.

(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 distin- guishing attributes. If there are four classes, then decision tree will improve considerably.

10. Repeat the analysis shown in Example 5.3 for ﬁnding the location of a decision boundary using the following information:

(a) The prior probabilities are *P* (Crocodile) = 2 × *P* (Alligator).

**Answer:** *x*ˆ = 13*.*0379.

(b) The prior probabilities are *P* (Alligator) = 2 × *P* (Crocodile).

**Answer:** *x*ˆ = 13*.*9621.

(c) The prior probabilities are the same, but their standard deviations are diﬀerent; i.e., *σ*(Crocodile) = 4 and *σ*(Alligator) = 2.

**Answer:** *x*ˆ = 22*.*1668.

11. Figure 5.3 illustrates the Bayesian belief network for the data set shown in

Table 5.3. (Assume that all the attributes are binary).

Mileage

Engine

Air

Conditioner

Car

Value

**Figure 5.3.** Bayesian belief network.

**Table 5.3.** Data set for Exercise 11.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Mileage | Engine | Air Conditioner | Number of Records  with Car Value=Hi | Number of Records  with Car Value=Lo |
| Hi  Hi Hi Hi Lo Lo Lo Lo | Good  Good Bad Bad Good Good Bad Bad | Working  Broken Working Broken Working Broken Working Broken | 3  1  1  0  9  5  1  0 | 4  2  5  4  0  1  2  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(B = bad) = 0.1 P(F = empty) = 0.2

Battery

Fuel

Gauge

P(G = empty | B = good, F = not empty) = 0.1

P(G = empty | B = good, F = empty) = 0.8

P(G = empty | B = bad, F = not empty) = 0.2

P(G = empty | B = bad, F = empty) = 0.9

Start

P(S = no | B = good, F = not empty) = 0.1

P(S = no | B = good, F = empty) = 0.8

P(S = no | B = bad, F = not empty) = 0.9

P(S = no | B = bad, F = empty) = 1.0

**Figure 5.4.** Bayesian belief network for Exercise 12.

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).

*P* (*Engine* = *Bad, Air C ond* = *Broken*)

= ) *P* (*Engine* = *Bad, Air C ond* = *Broken, M ileage* = *α, V alue* = *β*)

*αβ*

= ) *P* (*V alue* = *β*|*Engine* = *Bad, Air C ond* = *Broken*)

*αβ*

× *P* (*Engine* = *Bad*|*M ileage* = *α*)*P* (*M ileage* = *α*)*P* (*Air C ond* = *Broken*)

= 0*.*1453*.*

12. Given the Bayesian network shown in Figure 5.4, compute the following prob- abilities:

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

**Answer:**

*P* (*B* = *good, F* = *empty, G* = *empty, S* = *yes*)

= *P* (*B* = *good*) × *P* (*F* = *empty*) × *P* (*G* = *empty*|*B* = *good, F* = *empty*)

×*P* (*S* = *yes*|*B* = *good, F* = *empty*)

= 0*.*9 × 0*.*2 × 0*.*8 × 0*.*2 = 0*.*0288*.*

(b) *P* (B = bad, F = empty, G = not empty, S = no).

**Answer:**

*P* (*B* = *bad, F* = *empty, G* = *not empty, S* = *no*)

= *P* (*B* = *bad*) × *P* (*F* = *empty*) × *P* (*G* = *not empty*|*B* = *bad, F* = *empty*)

×*P* (*S* = *no*|*B* = *bad, F* = *empty*)

= 0*.*1 × 0*.*2 × 0*.*1 × 1*.*0 = 0*.*002*.*

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

**Answer:**

*P* (*S* = *yes*|*B* = *bad*)

= ) *P* (*S* = *yes*|*B* = *bad, F* = *α*)*P* (*B* = *bad*)*P* (*F* = *α*)

*α*

= 0*.*1 × 0*.*1 × 0*.*8

= 0*.*008

13. Consider the one-dimensional data set shown in Table 5.4.

**Table 5.4.** Data set for Exercise 13.

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

(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: +,

3-nearest neighbor: −,

5-nearest neighbor: +,

9-nearest neighbor: −.

(b) Repeat the previous analysis using the distance-weighted voting ap- proach described in Section 5.2.1.

**Answer:**

1-nearest neighbor: +,

3-nearest neighbor: +,

5-nearest neighbor: +,

9-nearest neighbor: +.

14. The nearest-neighbor algorithm described in Section 5.2 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:

*k*  *n n*

*d*(*V*1 *, V*2 ) = ) *i*1 −

*i*2 *,* (5.2)

*i*=1 *n*1

*n*2

where *nij* is the number of examples from class *i* with attribute value *Vj* and

*nj* is the number of examples with attribute value *Vj* .

Consider the training set for the loan classiﬁcation problem shown in Figure

5.9. 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 5.9 can be summarized for the Home Owner

and Marital Status attributes as follows.

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

|  |  |  |
| --- | --- | --- |
| Class | Home Owner | |
| Yes | No |
| Yes  No | 0  3 | 3  4 |

*d*(Single, Married) = 1 *d*(Single, Divorced) = 0 *d*(Married, Divorced) = 1

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

15. 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

16. (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:

1

*y* = *sgn x*1 + *x*2 − 1*.*5 *.*

For OR function, a possible perceptron model is:

1

*y* = *sgn x*1 + *x*2 − 0*.*5 *.*

(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 relation- ships between the input and output attributes. However, if linear func- tions 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.

17. 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 5.5 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,...,Z* ) = 1 −

*P* (+|*A,...,Z* ). Assume that we are mostly interested in detecting instances

from the positive class.

**Table 5.5.** Posterior probabilities for Exercise 17.

|  |  |  |  |
| --- | --- | --- | --- |
| Instance | True Class | *P* (+|*A,..., Z, M*1 ) | *P* (+|*A,..., Z, M*2 ) |
| 1  2  3  4  5  6  7  8  9  10 | +  +  −  −  +  +  −  −  +  − | 0.73  0.69  0.44  0.55  0.67  0.47  0.08  0.15  0.45  0.35 | 0.61  0.03  0.68  0.31  0.45  0.09  0.38  0.05  0.01  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 *M* 1 and *M* 2 are shown in the Figure 5.5.

1

0.9

0.8

0.7

M1 M2

TPR

0.6

0.5

0.4

0.3

0.2

0.1

0

0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1

FPR

**Figure 5.5.** ROC curve.

*M* 1 is better, since its area under the ROC curve is larger than the area under ROC curve for *M* 2.

(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 *M* 1 is shown below.

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | + | - |
| Actual | + | 3 | 2 |
| - | 1 | 4 |

Precision = 3*/*4 = 75%. Recall = 3*/*5 = 60%.

F-measure = (2 × *.*75 × *.*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 *M* 2 is shown below.

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | + | - |
| Actual | + | 1 | 4 |
| - | 1 | 4 |

Precision = 1*/*2 = 50%. Recall = 1*/*5 = 20%.

F-measure = (2 × *.*5 × *.*2)*/*(*.*5+ *.*2) = 0*.*2857.

Based on F-measure, *M* 1 is still better than *M* 2. This result is consis- tent 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 *M* 1 is shown below.

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | + | - |
| Actual | + | 5 | 0 |
| - | 4 | 1 |

Precision = 5*/*9 = 55*.*6%. Recall = 5*/*5 = 100%.

F-measure = (2 × *.*556 × 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 *T RP* = 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 × (1 − 0*.*2) = 0*.*6 × 0*.*8 = 0*.*48. For *t* = 0*.*1, area = 1 × (1 − 0*.*8) = 1 × 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.

18. 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  1  2  0  1  2  0  1  2 | 0  0  0  1  1  1  2  2  2 | 0  0  0  10  10  10  0  0  0 | 100  0  100  100  0  100  100  0  100 |

The concept for the “+” class is *Y* = 1 and the concept for the “−” class is

*X* = 0 ∨ *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

*Eorig* = 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

*EX* =0 = 10*/*310

*EX* =1 = 0

*EX* =2 = 10*/*310

310 10 10

310 10

∆*X* = *Eorig* − 630 310 − 630 0 − 630 310 = 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

*EY* =0 = 0

*EY* =1 = 30*/*230

*EY* =2 = 0

230 30

∆*Y* = *Eorig* − 630 230 = 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

class = +*,* X = 1;

−*,* otherwise.

(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

accuracy :

610

630

10

= 0*.*9683

+ − precision :

= 1*.*0

10

Actual + 10 20

− 0 600

10

recall : 30 = 0*.*3333

2 ∗ 0*.*3333 ∗ 1*.*0

F − measure :

= 0*.*5

1*.*0+ 0*.*3333

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

⎧ 0*,* if *i* = *j*;

⎨

*C* (*i, j*) =

⎩

1*,* if *i* = +, *j* = −;

Number of − instances

*,* if *i* = , *j* = +.

Number of + instances −

(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

+ −

Actual + 0 600/30=20

− 1 0

The decision tree in part (a) has 7 leaf nodes, *X* = 1, *X* = 0 ∧ *Y* = 0, *X* = 0 ∧ *Y* = 1, *X* = 0 ∧ *Y* = 2, *X* = 2 ∧ *Y* = 0, *X* = 2 ∧ *Y* = 1, and *X* = 2 ∧ *Y* = 2. Only *X* = 0 ∧ *Y* = 1 and *X* = 2 ∧ *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

class = +*, X* = 1 ∨ (*X* = 0 ∧ *Y* = 1) ∨ (*X* = 2 ∧ *Y* = 2);

−*,* otherwise.

(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

accuracy :

430

630

30

= 0*.*6825

+ − precision :

= 0*.*1304

230

Actual + 30 0

− 200 400

30

recall : 30 = 1*.*0

2 ∗ 0*.*1304 ∗ 1*.*0

F − measure :

= 0*.*2307

1*.*0+ 0*.*1304

19. (a) Consider the cost matrix for a two-class problem. Let *C* (+*,* +) = *C* (−*,* −) = *p*, *C* (+*,* −) = *C* (−*,* +) = *q*, and *q > p*. Show that min- imizing the cost function is equivalent to maximizing the classiﬁer’s accuracy.

**Answer:**

Confusion Matrix + −

|  |  |
| --- | --- |
| a | b |
| c | d |

+

−

Cost Matrix + -

+

|  |  |
| --- | --- |
| p | q |
| q | p |

−

The total cost is *F* = *p*(*a* + *d*)+ *q*(*b* + *c*).

Since *acc* = *a*+*d* , where *N* = *a* + *b* + *c* + *d*, we may write

*N*

*F* = *N* f*acc*(*p* − *q*)+ *q*l*.*

Because *p* − *q* is negative, minimizing the total cost is equivalent to maximizing accuracy.

(b) Show that a cost matrix is scale-invariant. For example, if the cost matrix is rescaled from *C* (*i, j*) −→ *βC* (*i, j*), where *β* is the scaling factor, the decision threshold (Equation 5.82) will remain unchanged.

**Answer:**

The cost matrix is:

Cost Matrix + −

|  |  |
| --- | --- |
| *c*(+*,* +) | *c*(+*,* −) |
| *c*(−*,* +) | *c*(−*,* −) |

+

−

A node *t* is classiﬁed as positive if:

*c*(+*,* −)*p*(+|*t*)+ *c*(−*,* −)*p*(−|*t*) *> c*(−*,* +)*p*(−|*t*)+ *c*(+*,* +)*p*(+|*t*)

=⇒ *c*(+*,* −)*p*(+|*t*)+ *c*(−*,* −)[1 − *p*(+|*t*)] *> c*(−*,* +)[1 − *p*(+|*t*)] + *c*(+*,* +)*p*(+|*t*)

*c*(−*,* +) − *c*(−*,* −)

=⇒ *p*(+|*t*) *>* [*c*(

−

*,* +)

− *c*(

−*,* −

)] + [*c*(+*,* −)

− *c*(+*,* +)]

The transformed cost matrix is:

Cost Matrix + −

|  |  |
| --- | --- |
| *βc*(+*,* +) | *βc*(+*,* −) |
| *βc*(−*,* +) | *βc*(−*,* −) |

+

−

Therefore, the decision rule is:

*βc*(−*,* +) − *βc*(−*,* −)

*p*(+|*t*) *>*

[*βc*(−

*,* +)

− *βc*(

−*,* −

)] + [*βc*(+*,* −)

− *βc*(+*,* +)]

= *c*(−*,* +) − *c*(−*,* −)

[*c*(−*,* +) − *c*(−*,* −)] + [*c*(+*,* −) − *c*(+*,* +)]

which is the same as the original decision rule.

(c) Show that a cost matrix is translation-invariant. In other words, adding a constant factor to all entries in the cost matrix will not aﬀect the decision threshold (Equation 5.82).

**Answer:**

The transformed cost matrix is:

Cost Matrix + −

|  |  |
| --- | --- |
| *c*(+*,* +) + *β* | *c*(+*,* −)+ *β* |
| *c*(−*,* +) + *β* | *c*(−*,* −)+ *β* |

+

−

Therefore, the decision rule is:

*β* + *c*(−*,* +) − *β* − *c*(−*,* −)

*p*(+|*t*) *>*

[*β* + *c*(−

*,* +) − *β*

− *c*(

−*,* −

)] + [*β* + *c*(+*,* −) − *β*

− *c*(+*,* +)]

= *c*(−*,* +) − *c*(−*,* −)

[*c*(−*,* +) − *c*(−*,* −)] + [*c*(+*,* −) − *c*(+*,* +)]

which is the same as the original decision rule.

20. Consider the task of building a classiﬁer from random data, where the at- tribute 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%.

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

**Answer:**

*f* (**w**) = **w**

2

2

*N* 1

2

*N*

+ *C* ( ) *ξ* )2 *.*

*i*

*i*=1

2

*LD* = ) *λi* −

) *λi λj yi yj* **xi** · **xj** − *C* ) *ξi .*

*i*=1

*i,j i*

Notice that the dual Lagrangian depends on the slack variables *ξi* ’s.

22. 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:

Φ = (1*,* √2 *x*1 *,* √2 *x*2 *,* √2*x*1 *x*2 *, x*2 *, x*2 )*.*

1 2

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

**Answer:**

The decision boundary is *f* (*x*1 *, x*2 ) = *x*1 *x*2 .

23. Given the data sets shown in Figures 5.6, explain how the decision tree, na¨ıve

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 at- tribute 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.

**Attributes**

**Attributes**

Distinguishing

Attributes Noise Attributes

Distinguishing Attributes Noise Attributes

Class A

Class A

**Records**

**Records**

Class B

Class B

(a) Synthetic data set 1.

(b) Synthetic data set 2.

**Attributes**

A

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Class A | Class B | Class A | Class B | Class A |
| Class B | Class A | Class B | Class A | Class B |
| Class A | Class B | Class A | Class B | Class A |
| Class B | Class A | Class B | Class A | Class B |

|  |  |  |  |
| --- | --- | --- | --- |
| Distinguishing  Attribute set 1 | Distinguishing  Attribute set 2 | Noise Attributes | |
| 60% filled with 1 | 40% filled with 1 |  | Class |
| 40% filled with 1 | 60% filled with 1 |  | Class |

**Attribute Y**

**Records**

B

**Attribute X**

(c) Synthetic data set 3.

(d) Synthetic data set 4

Class A

Class B

Class A

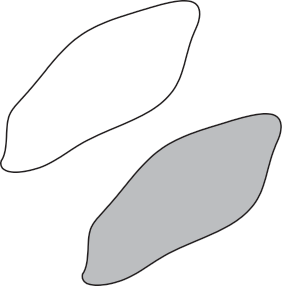
**Attribute Y**

**Attribute Y**

Class B

Class B

**Attribute X**



**Attribute X**

(e) Synthetic data set 5.

(f ) Synthetic data set 6.

**Figure 5.6.** Data set for Exercise 23.

**6**

**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 −→ Bread. Such obvious rule tends to be uninteresting. (b) A rule that has reasonably high support but low conﬁdence.

**Answer:** Milk −→ 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 −→ Laundry detergent. Such low conﬁdence rule

tends to be uninteresting.

(d) A rule that has low support and high conﬁdence.

**Answer:** Vodka −→ Caviar. Such rule tends to be interesting.

2. Consider the data set shown in Table 6.1.

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

**Answer:**

**Table 6.1.** Example of market basket transactions.

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

*s*({*e*}) = *s*({*b, d*}) = *s*({*b, d, e*}) =

8

= 0*.*8

10

2

= 0*.*2

10

2

= 0*.*2

10

(6.1)

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

**Answer:**

*c*(*bd* −→ *e*) =

*c*(*e* −→ *bd*) =

0*.*2 = 100%

0*.*2

0*.*2 = 25%

0*.*8

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:**

*s*({*e*}) = *s*({*b, d*}) = *s*({*b, d, e*}) =

4

= 0*.*8

5

5

= 1

5

4

= 0*.*8

5

(d) Use the results in part (c) to compute the conﬁdence for the association rules {*b, d*} −→ {*e*} and {*e*} −→ {*b, d*}.

**Answer:**

*c*(*bd* −→ *e*) =

*c*(*e* −→ *bd*) =

0*.*8 = 80%

1

0*.*8 = 100%

0*.*8

(e) Suppose *s*1 and *c*1 are the support and conﬁdence values of an associa- tion 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 ∅ −→ *A* and *A* −→ ∅?

**Answer:**

*c*(∅ −→ *A*) = *s*(∅ −→ *A*).

*c*(*A* −→ ∅) = 100%.

(b) Let *c*1 , *c*2 , and *c*3 be the conﬁdence values of the rules {*p*} −→ {*q*},

{*p*} −→ {*q, r*}, and {*p, r*} −→ {*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:**

*s*(*p*∪*q*) *s*(*p*) *s*(*p*∪*q*∪*r*)

*c*1 =

*c*2 =

*s*(*p*)

*c* = *s*(*p*∪*q*∪*r*)

3

*s*(*p*∪*r*)

Considering *s*(*p*) ≥ *s*(*p* ∪ *q*) ≥ *s*(*p* ∪ *q* ∪ *r*)

Thus: *c*1 ≥ *c*2 & *c*3 ≥ *c*2.

Therefore *c*2 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* ∪ *q*) = *s*(*p* ∪ *q* ∪ *r*)

but *s*(*p*) ≥ *s*(*p* ∪ *r*)

Thus: *c*3 ≥ (*c*1 = *c*2)

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

(d) Transitivity: Suppose the conﬁdence of the rules *A* −→ *B* and *B* −→ *C* are larger than some threshold, *minconf* . Is it possible that *A* −→ *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* → *B*) = 66% *> minconf c*(*B* → *C* ) = 71% *> minconf*

But *c*(*A* → *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* = *σ*(*X* ) is anti-monotone because *s*(*X* )

≥

|*T* |

*s*(*Y* ) whenever *X* ⊂ *Y* .

(a) A characteristic rule is a rule of the form {*p*} −→ {*q*1 *, q*2 *,..., qn* }, 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:

{ )

*ζ* ({*p*1 *, p*2 *,..., pk* }) = min f *c*( *p*1 } −→ {*p*2 *, p*3 *,..., pk* } *,...*

*c*( ) l

{*pk* } −→ {*p*1 *, p*3 *..., pk*−1 }

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

**Answer:**

*ζ* is an anti-monotone measure because

*ζ* ({*A*1 *, A*2 *,* ··· *, Ak* }) ≥ *ζ* ({*A*1 *, A*2 *,* ··· *, Ak , Ak*+1 }) (6.2) For example, we can compare the values of *ζ* for {*A, B*} and {*A, B, C* }.

*ζ* ({*A, B*}) = min (*c*(*A* −→ *B*)*, c*(*B* −→ *A*))

= min ( *s*(*A, B*)

*,*

*s*(*A*)

= *s*(*A, B*)

*s*(*A, B*) )

*s*(*B*)

(6.3)

max(*s*(*A*)*, s*(*B*))

*ζ* ({*A, B, C* }) = min (*c*(*A* −→ *BC* )*, c*(*B* −→ *AC* )*, c*(*C* −→ *AB*))

= min ( *s*(*A, B, C* )

*,*

*s*(*A*)

= *s*(*A, B, C* )

*s*(*A, B, C* )

*s*(*B*) *,*

*s*(*A, B, C* ) )

*s*(*C* )

(6.4)

max(*s*(*A*)*, s*(*B*)*, s*(*C* ))

Since *s*(*A, B, C* ) ≤ *s*(*A, B*) and max(*s*(*A*)*, s*(*B*)*, s*(*C* )) ≥ max(*s*(*A*)*, s*(*B*)), therefore *ζ* ({*A, B*}) ≥ *ζ* ({*A, B, C* }).

(b) A discriminant rule is a rule of the form {*p*1 *, p*2 *,..., pn* } −→ {*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:

{ )*,...*

*c*( ) l

{*p*1 *, p*2 *,... pk*−1 } −→ {*pk* }

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

**Answer:**

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

*η*({*A, B, C* }).

*η*({*A, B*}) = min (*c*(*A* −→ *B*)*, c*(*B* −→ *A*))

= min ( *s*(*A, B*)

*,*

*s*(*A*)

= *s*(*A, B*)

*s*(*A, B*) )

*s*(*B*)

(6.5)

max(*s*(*A*)*, s*(*B*))

*η*({*A, B, C* }) = min (*c*(*AB* −→ *C* )*, c*(*AC* −→ *B*)*, c*(*BC* −→ *A*))

= min ( *s*(*A, B, C* )

*,*

*s*(*A, B*)

*s*(*A, B, C* )

*s*(*A, C* ) *,*

*s*(*A, B, C* ) )

*s*(*B, C* )

= *s*(*A, B, C* )

max(*s*(*A, B*)*, s*(*A, C* )*, s*(*B, C* ))

(6.6)

Since *s*(*A, B, C* ) ≤ *s*(*A, B*) and max(*s*(*A, B*)*, s*(*A, C* )*, s*(*B, C* )) ≤ max(*s*(*A*)*, s*(*B*)), therefore *η*({*A, B, C* }) can be greater than or less than *η*({*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

*ζ* t({*A*1 *, A*2 *,* ··· *, Ak* }) = max( *c*(*A*1 −→ *A*2 *, A*3 *,* ··· *, Ak* )*,* ···

*c*(*Ak* −→ *A*1 *, A*3 ··· *, Ak*−1 ))

*ζ* t({*A, B*}) = max (*c*(*A* −→ *B*)*, c*(*B* −→ *A*))

= max ( *s*(*A, B*)

*,*

*s*(*A*)

= *s*(*A, B*)

*s*(*A, B*) )

*s*(*B*)

(6.7)

min(*s*(*A*)*, s*(*B*))

*ζ* t({*A, B, C* }) = max (*c*(*A* −→ *BC* )*, c*(*B* −→ *AC* )*, c*(*C* −→ *AB*))

= max ( *s*(*A, B, C* )

*,*

*s*(*A*)

= *s*(*A, B, C* )

*s*(*A, B, C* )

*s*(*B*) *,*

*s*(*A, B, C* ) )

*s*(*C* )

(6.8)

min(*s*(*A*)*, s*(*B*)*, s*(*C* ))

Since *s*(*A, B, C* ) ≤ *s*(*A, B*) and min(*s*(*A*)*, s*(*B*)*, s*(*C* )) ≤ min(*s*(*A*)*, s*(*B*)), *ζ* t({*A, B, C* }) can be greater than or less than *ζ* t({*A, B*}). Therefore, the measure is non-monotone.

Let

*η*t({*A*1 *, A*2 *,* ··· *, Ak* }) = max( *c*(*A*2 *, A*3 *,* ··· *, Ak* −→ *A*1 )*,* ···

*c*(*A*1 *, A*2 *,* ··· *Ak*−1 −→ *Ak* ))

*η*t({*A, B*}) = max (*c*(*A* −→ *B*)*, c*(*B* −→ *A*))

= max ( *s*(*A, B*)

*,*

*s*(*A*)

= *s*(*A, B*)

*s*(*A, B*) )

*s*(*B*)

(6.9)

min(*s*(*A*)*, s*(*B*))

*η*({*A, B, C* }) = max (*c*(*AB* −→ *C* )*, c*(*AC* −→ *B*)*, c*(*BC* −→ *A*))

= max ( *s*(*A, B, C* )

*,*

*s*(*A, B*)

*s*(*A, B, C* )

*s*(*A, C* ) *,*

*s*(*A, B, C* ) )

*s*(*B, C* )

= *s*(*A, B, C* )

min(*s*(*A, B*)*, s*(*A, C* )*, s*(*B, C* ))

(6.10)

Since *s*(*A, B, C* ) ≤ *s*(*A, B*) and min(*s*(*A, B*)*, s*(*A, C* )*, s*(*B, C* )) ≤ min(*s*(*A*)*, s*(*B*)*, s*(*C* ))

≤ min(*s*(*A*)*, s*(*B*)), *η*t({*A, B, C* }) can be greater than or less than *η*t({*A, B*}).

Hence, the measure is non-monotone.

5. Prove Equation 6.3. (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 (*d*) ways for doing this. After selecting the

*k*

items for the left-hand side, there are (*d*−*k*) ways to choose the remaining

*i*

items to form the right hand side of the rule, where 1 ≤ *i* ≤ *d* − *k*. Therefore the total number of rules (*R*) is:

*R* = ) *d k*

*d*

*k*=1

*d*

*d*−*k*

)

*i*=1

*d* − *k i*

= )

*k*=1

*d*

= )

*k*=1

*d*

*d* 2*d*−*k* 1

*k*

−

*d* 2*d*−*k* ) *d k k*

*d*

−

*k*=1

1

where

Since

= ) *d*

*k*

*k*=1

) *n i*

*n*

*i*=1

2*d*−*k* − 2*d* +1 *,*

= 2*n* − 1*.*

*d*

(1 + *x*)*d* = )

*i*=1

*d xd*−*i* + *xd , i*

substituting *x* = 2 leads to:

*d*

3*d* = )

*i*=1

*d* 2*d*−*i* + 2*d . i*

Therefore, the total number of rules is:

1

*R* = 3*d* − 2*d* − 2*d* +1

= 3*d* − 2*d*+1 + 1*.*

**Table 6.2.** Market basket transactions.

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

6. Consider the market basket transactions shown in Table 6.2.

(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 maxi- mum 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:** (6) = 20*.*

3

(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*} −→ {*b*} and

{*b*} −→ {*a*} have the same conﬁdence.

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

7. 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 proce- dure using the *Fk*−1 × *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 pro- cedure 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}

8. The *Apriori* algorithm uses a generate-and-count strategy for deriving fre- quent 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 6.3 with *minsup* = 30%, i.e., any itemset occurring in less than 3 transactions is considered to be infrequent.

**Table 6.3.** Example of market basket transactions.

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

(a) Draw an itemset lattice representing the data set given in Table 6.3.

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

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.

null F

F F F F F

A B C D E

F I F F F F F F

AB AC AD AE BC BD BE CD

F I CE DE

N I I N N F I N F N

ABC ABD ABE ACD ACE ADE BCD BCE BDE CDE

N ABCD

N ABCE

N ABDE

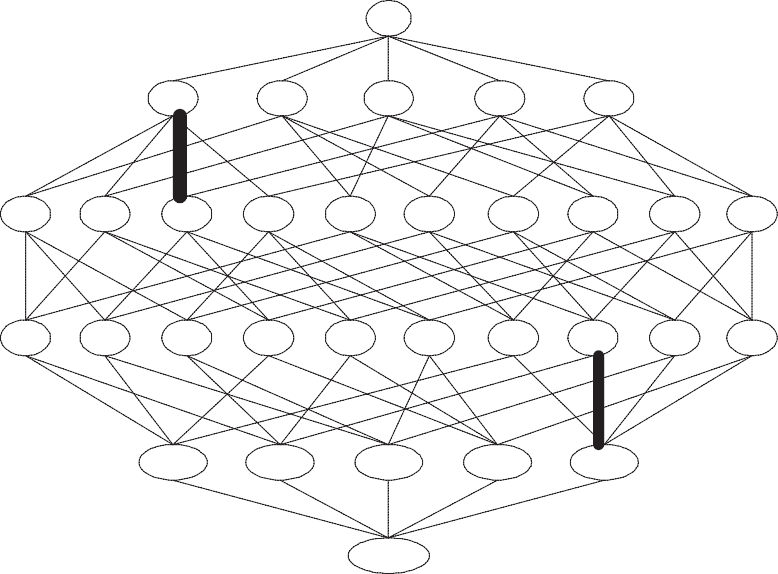
N ACDE

N BCDE

N

ABCDE

**Figure 6.1.** 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:**

1,4,7

2,5,8

3,6,9

1,4,7

2,5,8

3,6,9

1,4,7

3,6,9

2,5,8 1,4,7 2,5,8

3,6,9

L1 L5 L6 L7 L8 L9 L11 L12

{145}

{178}

1,4,7

2,5,8

{168}

3,6,9

{246}

{278}

{258}

{289}

{568}

{346}

{379}

{678}

{356}

{689}

{367}

L2 L3 L4

{127}

{457}

{125}

{158}

{458}

{459}

{456}

{789}

**Figure 6.2.** An example of a hash tree structure.

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%.

9. 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 6.2.

(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 trans- action?

**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 item- sets 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}.

10. 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:**

1 3 4

**L1**

2 3 4 2 4 5

4 5 6 **L5**

1 2 3 1 2 6 **L4**

**L2** 3 4 6

**L3**

**Figure 6.3.** Hash tree for Exercise 10.

null

a b c d e

ab ac

ad ae

bc bd

be cd

ce de

abc abd abe

acd

ace ade bcd bce bde cde

abcd

abce abde acde bcde

abcde

**Figure 6.4.** An itemset lattice

(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}.

11. Given the lattice structure shown in Figure 6.4 and the transactions given in

Table 6.3, label each node with the following letter(s):

• *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.

null C

C C C C F

A B C D E

MC I

F F M

C C F M I

AB AC AD AE

C BC BD BE

C CD CE DE

I I I I I M

I I M I

ABC ABD ABE ACD ACE

C ADE BCD BCE

C BDE CDE

I ABCD

I ABCE

I ABDE

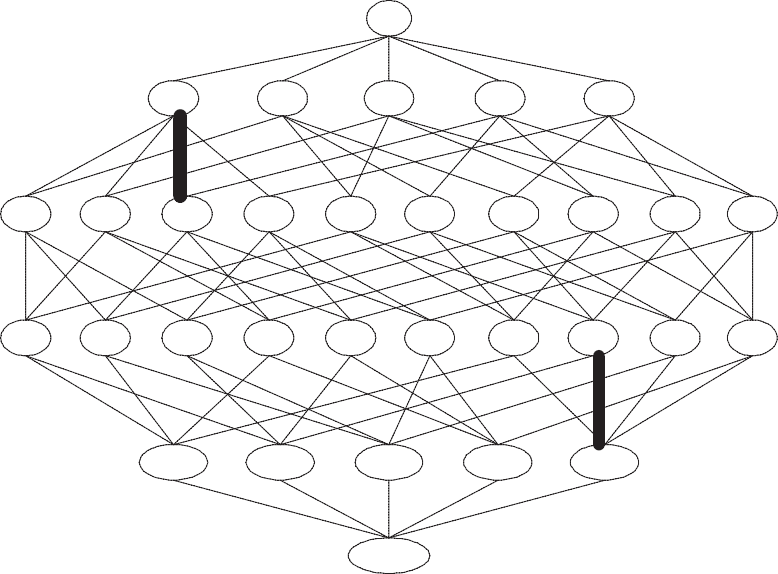
I ACDE

I BCDE

I

ABCDE

**Figure 6.5.** Solution for Exercise 11.



12. The original association rule mining formulation uses the support and conﬁ- dence measures to prune uninteresting rules.

(a) Draw a contingency table for each of the following rules using the trans- actions shown in Table 6.4.

Rules: {*b*} −→ {*c*}, {*a*} −→ {*d*}, {*b*} −→ {*d*}, {*e*} −→ {*c*},

{*c*} −→ {*a*}.

**Answer:**

|  |  |  |
| --- | --- | --- |
|  | *d* | *d* |
| *a*  *a* | 4  5 | 1  0 |

|  |  |  |
| --- | --- | --- |
|  | *d* | *d* |
| *b*  *b* | 6  3 | 1  0 |

|  |  |  |
| --- | --- | --- |
|  | *c* | *c* |
| *b*  *b* | 3  2 | 4  1 |

|  |  |  |
| --- | --- | --- |
|  | *c* | *c* |
| *e*  *e* | 2  3 | 4  1 |

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

|  |  |  |
| --- | --- | --- |
|  | *a* | *a* |
| *c*  *c* | 2  3 | 3  2 |

**Table 6.4.** Example of market basket transactions.

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

i. Support.

**Answer:**

|  |  |  |
| --- | --- | --- |
| Rules | Support | Rank |
| *b* −→ *c*  *a* −→ *d*  *b* −→ *d*  *e* −→ *c*  *c* −→ *a* | 0.3  0.4  0.6  0.2  0.2 | 3  2  1  4  4 |

ii. Conﬁdence.

**Answer:**

|  |  |  |
| --- | --- | --- |
| Rules | Conﬁdence | Rank |
| *b* −→ *c*  *a* −→ *d*  *b* −→ *d*  *e* −→ *c*  *c* −→ *a* | 3/7  4/5  6/7  2/6  2/5 | 3  2  1  5  4 |

iii. Interest(*X* −→ *Y* ) = *P* (*X,Y* ) *P* (*Y* ).

*P* (*X* )

**Answer:**

|  |  |  |
| --- | --- | --- |
| Rules | Interest | Rank |
| *b* −→ *c*  *a* −→ *d*  *b* −→ *d*  *e* −→ *c*  *c* −→ *a* | 0.214  0.72  0.771  0.167  0.2 | 3  2  1  5  4 |

iv. IS(*X Y* ) = *P* (*X,Y* ) .

−→  ~~√~~

*P* (*X* )*P* (*Y* )

**Answer:**

|  |  |  |
| --- | --- | --- |
| Rules | IS | Rank |
| *b* −→ *c*  *a* −→ *d*  *b* −→ *d*  *e* −→ *c*  *c* −→ *a* | 0.507  0.596  0.756  0.365  0.4 | 3  2  1  5  4 |

v. Klosgen(*X* −→ *Y* )= /*P* (*X, Y* )×(*P* (*Y* |*X* )−*P* (*Y* )), where *P* (*Y* |*X* ) =

*P* (*X,Y* )

.

*P* (*X* )

**Answer:**

|  |  |  |
| --- | --- | --- |
| Rules | Klosgen | Rank |
| *b* −→ *c*  *a* −→ *d*  *b* −→ *d*  *e* −→ *c*  *c* −→ *a* | -0.039  -0.063  -0.033  -0.075  -0.045 | 2  4  1  5  3 |

vi. Odds ratio(*X Y* ) = *P* (*X,Y* )*P* (*X,Y* ) .

−→

*P* (*X,Y* )*P* (*X,Y* )

**Answer:**

|  |  |  |
| --- | --- | --- |
| Rules | Odds Ratio | Rank |
| *b* −→ *c*  *a* −→ *d*  *b* −→ *d*  *e* −→ *c*  *c* −→ *a* | 0.375  0  0  0.167  0.444 | 2  4  4  3  1 |

13. Given the rankings you had obtained in Exercise 12, compute the correla- tion 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.

14. Answer the following questions using the data sets shown in Figure 6.6.

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

tions)?

(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).

15. (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:

*φ* = *P* (*A, B*) − *P* (*A*)*P* (*B*) = 1

*P* (*A*)*P* (*B*)f1 − *P* (*A*)lf1 − *P* (*B*)l

The preceding equation can be simpliﬁed as follows:

12

*P* (*A, B*) − *P* (*A*)*P* (*B*)

= *P* (*A*)*P* (*B*)f1 − *P* (*A*)lf1 − *P* (*B*)l

*P* (*A, B*)2 − 2*P* (*A, B*)*P* (*A*)*P* (*B*) = *P* (*A*)*P* (*B*)f1 − *P* (*A*) − *P* (*B*)l

*P* (*A, B*)2 = *P* (*A*)*P* (*B*)f1 − *P* (*A*) − *P* (*B*)+ 2*P* (*A, B*)l

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

*P* (*A*)*P* (*B*)2 − *P* (*A*)f1 − *P* (*A*)+ 2*P* (*A, B*)l*P* (*B*)+ *P* (*A, B*)2 = 0

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

*P* (*B*) = *P* (*A*)*β* −

/*P* (*A*)2 *β*2

− 4*P* (*A*)*P* (*A, B*)2

2*P* (*A*) *,*

Items

Items

2000

Transactions

4000

6000

2000

4000

6000

8000

Transactions

8000

200

400

600 800

200

400

600 800

(a) (b)

Items

Items

2000

Transactions

4000

6000

2000

4000

6000

8000

Transactions

8000

200

400

600 800

200

400

600 800

(c)

(d)

Items

Items

2000

2000

4000

Transactions

6000

10% are 1s

90% are 0s

(uniformly distributed)

4000

6000

8000

Transactions

8000

200

400

600 800

200

400

600 800

(e)

(f)

**Figure 6.6.** Figures for Exercise 14.

solution because it corresponds to *φ* = −1. Furthermore, the solution for *P* (*B*) must satisfy the following constraint: *P* (*B*) ≥ *P* (*A, B*). It can be shown that:

*P* (*B*) − *P* (*A, B*)

= 1 − *P* (*A*)

2

/(1 − *P* (*A*))2 + 4*P* (*A, B*)(1 − *P* (*A*))(1 − *P* (*A, B*)*/P* (*A*))

− 2

≤ 0

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*) × *P* (*A, B*) =

*P* (*A, B*) × *P* (*A, B*).

**Answer:**

When *A* and *B* are independent, *P* (*A, B*) = *P* (*A*) × *P* (*B*) or equiva- lently:

*P* (*A, B*) − *P* (*A*)*P* (*B*) = 0

*P* (*A, B*) − [*P* (*A, B*)+ *P* (*A, B*)][*P* (*A, B*)+ *P* (*A, B*)] = 0

*P* (*A, B*)[1 − *P* (*A, B*) − *P* (*A, B*) − *P* (*A, B*)] − *P* (*A, B*)*P* (*A, B*) = 0

*P* (*A, B*)*P* (*A, B*) − *P* (*A, B*)*P* (*A, B*) = 0*.*

(c) Show that Yule’s *Q* and *Y* coeﬃcients

*Q* = *f*11 *f*00 − *f*10 *f*01 1

*f*11 *f*00 + *f*10 *f*01

√*f* 11 *f*00 − √*f* 10 *f*01 1

*Y* = √

*f*

11

*f*00

+ √*f*10

*f*01

are normalized versions of the odds ratio.

**Answer:**

Odds ratio can be written as:

*α* = *f*11 *f*00 *. f*10 *f*01

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

*Q* = *α* − 1

*α* +1

√*α* − 1

*Y* = √*α* +1

In both cases, *Q* and *Y* increase monotonically with *α*. Furthermore, when *α* = 0, *Q* = *Y* = −1 to represent perfect negative correlation. When *α* = 1, which is the condition for attribute independence, *Q* = *Y* = 1. Finally, when *α* = ∞, *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  Odds ratio Kappa *κ* Interest Cosine, *IS*  Piatetsky-Shapiro’s  Collective strength Jaccard Conviction Certainty factor Added value | 0  1  0  1  /*P* (*A, B*)  0  1  0 ··· 1  1  0  0 |

16. Consider the interestingness measure, *M* = *P* (*B*|*A*)−*P* (*B*) , for an association

1−*P* (*B*)

rule *A* −→ *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 max- imum 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:

*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 inde- pendent?

**Answer:** 0.

(g) Is the measure null-invariant?

**Answer:** No.

(h) Does the measure remain invariant under row or column scaling oper- ations?

**Answer:** No.

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

**Answer:** Asymmetric.

17. 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*}→ {*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*}. De- scribe 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 × 0*.*9) = 0*.*889. The items are nega- tively 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*} −→ {*b*} is less than the support of {*b*}, then:

i. *c*({*a*} −→ {*b*}) *> c*({*a*} −→ {*b*}), ii. *c*({*a*} −→ {*b*}) *> s*({*b*}),

where *c*(·) denote the rule conﬁdence and *s*(·) denote the support of an itemset.

**Answer:**

Let

*c*({*a*} −→ {*b*}) =

*P* ({*a, b*})

*P* ( *a* ) *< P* ({*b*})*,*

{ }

which implies that

Furthermore,

*P* ({*a*})*P* ({*b*}) *> P* ({*a, b*})*.*

*c*({*a*} −→ {*b*}) =

*P* ({*a, b*}) =

*P* ( *a* )

*P* ({*b*}) − *P* ({*a, b*})

{ }

i. Therefore, we may write

1 − *P* ({*a*})

*P* ({*b*}) − *P* ({*a, b*})

*P* ({*a, b*})

*c*({*a*} −→ {*b*}) − *c*({*a*} −→ {*b*}) =

1 *P* ( *a* ) −

− { }

*P* ({*a*})

= *P* ({*a*})*P* ({*b*}) − *P* ({*a, b*})

*P* ({*a*})(1 − *P* ({*a*}))

which is positive because *P* ({*a*})*P* ({*b*}) *> P* ({*a, b*}). ii. We can also show that

*P* ({*b*}) − *P* ({*a, b*})

*c*({*a*} −→ {*b*}) − *s*({*b*}) =

1 *P* ( *a* ) − *P* ({*b*})

− { }

= *P* ({*a*})*P* ({*b*}) − *P* ({*a, b*})

1 − *P* ({*a*})

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

18. Table 6.5 shows a 2 × 2 × 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 *φ*( *A, B* ) =  ~~√~~ *P* (*A,B*)− *P* (*A*)*P* (*B*) .

{ }

*P* (*A*)*P* (*B*)(1−*P* (*A*))(1−*P* (*B*))

**Answer:**

i. When *C* = 0, *φ*(*A, B*) = −1*/*3. ii. When *C* = 1, *φ*(*A, B*) = 1.

iii. When *C* = 0 or *C* = 1, *φ* = 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 6.5.** 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 6.6.** Contingency tables for Exercise 19.

*B B B B A A*

|  |  |
| --- | --- |
| 9 | 1 |
| 1 | 89 |

|  |  |
| --- | --- |
| 89 | 1 |
| 1 | 9 |

*A A*

(a) Table I. (b) Table II.

19. Consider the contingency tables shown in Table 6.6.

(a) For table I, compute support, the interest measure, and the *φ* correla- tion coeﬃcient for the association pattern {A, B}. Also, compute the conﬁdence of rules *A* → *B* and *B* → *A*.

**Answer:**

*s*(*A*) = 0*.*1, *s*(*B*) = 0*.*9, *s*(*A, B*) = 0*.*09.

*I* (*A, B*) = 9, *φ*(*A, B*) = 0*.*89.

*c*(*A* −→ *B*) = 0*.*9, *c*(*B* −→ *A*) = 0*.*9.

(b) For table II, compute support, the interest measure, and the *φ* correla- tion coeﬃcient for the association pattern {A, B}. Also, compute the conﬁdence of rules *A* → *B* and *B* → *A*.

**Answer:**

*s*(*A*) = 0*.*9, *s*(*B*) = 0*.*9, *s*(*A, B*) = 0*.*89.

*I* (*A, B*) = 1*.*09, *φ*(*A, B*) = 0*.*89.

*c*(*A* −→ *B*) = 0*.*98, *c*(*B* −→ *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.

20. Consider the relationship between customers who buy high-deﬁnition televi- sions and exercise machines as shown in Tables 6.19 and 6.20.

(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, *φ* = 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 associa- tion changes when data is pooled together instead of being stratiﬁed.

**Answer:**

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

**7**

**Association Analysis: Advanced Concepts**

1. Consider the traﬃc accident data set shown in Table 7.1.

**Table 7.1.** Traffic accident data set.

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

(a) Show a binarized version of the data set.

**Answer:** See Table 7.2.

(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 fre- quent itemsets will be generated?

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Good | Bad | Alcohol | Sober | Exceed  speed | None | Disobey  stop | Disobey  traﬃc | Belt  = No | Belt  = Yes | Major | Minor |
| 1  0  1  1  0  1  0  1  1  0  1  0 | 0  1  0  0  1  0  1  0  0  1  0  1 | 1  0  0  0  0  1  1  0  1  0  1  0 | 0  1  1  1  1  0  0  1  0  1  0  1 | 1  0  0  1  0  0  0  0  0  0  1  0 | 0  1  0  0  0  0  1  0  1  0  0  0 | 0  0  1  0  0  1  0  0  0  0  0  1 | 0  0  0  0  1  0  0  1  0  1  0  0 | 1  0  0  0  1  0  0  0  1  1  0  0 | 0  1  1  1  0  1  1  1  0  0  1  1 | 1  0  0  1  1  0  1  1  1  1  1  0 | 0  1  1  0  0  1  0  0  0  0  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 at- tributes: (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 7.3.

**Table 7.3.** Traffic accident data set.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Bad | Alcohol  Impaired | Traﬃc  violation | Belt  = No | Major |
| 0  1  0  0  1  0  1  0  0  1  0  1 | 1  0  0  0  0  1  1  0  1  0  1  0 | 1  0  1  1  1  1  0  1  0  1  1  1 | 1  0  0  0  1  0  0  0  1  1  0  0 | 1  0  0  1  1  0  1  1  1  1  1  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 7.4. Suppose we apply the fol- lowing 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 ≥ 30%.

**Table 7.4.** Data set for Exercise 2.

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

**Answer:**

Table 7.5 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 7.5.** Discretized data using D1.

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

**Table 7.6.** Discretized data using D2.

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

Table 7.6 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 ap- proach.

i. Plot a graph of temperature versus pressure for the data points shown in Table 7.4.

**Answer:**

The graph of Temperature and Pressure is shown below.

**Pressure vs Temperature**

1110 C2

1100

1090

1080

**Pressure**

1070

1060

1050

C1

1040

1030

1020

75 80 85 90 95 100 105

**Temperature**

**Figure 7.1.** 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 7.4 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 7.7.** Example of numeric data set.

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

v. Derive all the frequent itemsets having support ≥ 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 7.8. The ﬁrst attribute is continuous, while the remaining two attributes are asymmetric binary. A rule is consid- ered to be strong if its support exceeds 15% and its conﬁdence exceeds 60%. The data given in Table 7.8 supports the following two strong rules:

(i) {(1 ≤ *A* ≤ 2)*,B* = 1}→ {*C* = 1}

(ii) {(5 ≤ *A* ≤ 8)*,B* = 1}→ {*C* = 1}

**Table 7.8.** Data set for Exercise 3.

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

(a) Compute the support and conﬁdence for both rules.

**Answer:**

*s*({(1 ≤ *A* ≤ 2)*,B* = 1}→ {*C* = 1}) = 1*/*6 *c*({(1 ≤ *A* ≤ 2)*,B* = 1}→ {*C* = 1}) = 1 *s*({(5 ≤ *A* ≤ 8)*,B* = 1}→ {*C* = 1}) = 1*/*6 *c*({(5 ≤ *A* ≤ 8)*,B* = 1}→ {*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 7.9.** A Synthetic Data set

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

Where

*A*1 = 1 ≤ *A* ≤ 2; *A*2 = 3 ≤ *A* ≤ 4;

*A*3 = 5 ≤ *A* ≤ 6; *A*4 = 7 ≤ *A* ≤ 8;

*A*5 = 9 ≤ *A* ≤ 10; *A*6 = 11 ≤ *A* ≤ 12;

For the ﬁrst rule, there is one corresponding rule:

{*A*1 = 1*,B* = 1}→ {*C* = 1}

*s*(*A*1 = 1*,B* = 1}→ {*C* = 1}) = 1*/*6

*c*(*A*1 = 1*,B* = 1}→ {*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:

{*A*3 = 1*,B* = 1}→ {*C* = 1}

{*A*4 = 1*,B* = 1}→ {*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 7.10.** A Synthetic Data set

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

Where

*A*1 = 1 ≤ *A* ≤ 3; *A*2 = 4 ≤ *A* ≤ 6;

*A*3 = 7 ≤ *A* ≤ 9; *A*4 = 10 ≤ *A* ≤ 12;

For the ﬁrst rule, there is one corresponding rule:

{*A*1 = 1*,B* = 1}→ {*C* = 1}

*s*(*A*1 = 1*,B* = 1}→ {*C* = 1}) = 1*/*6

*c*(*A*1 = 1*,B* = 1}→ {*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:

{*A*2 = 1*,B* = 1}→ {*C* = 1}

{*A*3 = 1*,B* = 1}→ {*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 7.11.** A Synthetic Data set

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| A1 | A2 | A3 | B | C |
| 1  1  1  1  0  0  0  0  0  0  0  0 | 0  0  0  0  1  1  1  1  0  0  0  0 | 0  0  0  0  0  0  0  0  1  1  1  1 | 1  1  1  1  1  0  0  1  0  0  0  0 | 1  1  0  0  1  1  0  1  0  0  0  1 |

Where

*A*1 = 1 ≤ *A* ≤ 4; *A*2 = 5 ≤ *A* ≤ 8;

*A*3 = 9 ≤ *A* ≤ 12;

For the ﬁrst rule, there is one correspomding rule:

{*A*1 = 1*,B* = 1}→ {*C* = 1}

*s*(*A*1 = 1*,B* = 1}→ {*C* = 1}) = 1*/*6

*c*(*A*1 = 1*,B* = 1}→ {*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:

{*A*2 = 1*,B* = 1}→ {*C* = 1}

*s*(*A*2 = 1*,B* = 1}→ {*C* = 1}) = 1*/*6

*c*(*A*2 = 1*,B* = 1}→ {*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, ·, 11 *<*= *A <*= 12

1 *<*= *A <*= 4, 5 *<*= *A <*= 8 , 9 *<*= *A <*= 12

4. Consider the data set shown in Table 7.12.

**Table 7.12.** 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  15 – 25  25 – 35  35 – 50 | 2  2  10  4 | 3  5  15  6 | 5  10  5  5 | 3  10  3  3 | 2  3  2  2 |

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

i. 15 *< A <* 25 −→ 10 *< B <* 20, 10 *< A <* 25 −→ 10 *< B <* 20, and 15 *<A<* 35 −→ 10 *<B <* 20.

**Answer:**

Both 15 *< A <* 25 −→ 10 *< B <* 20 and 10 *< A <* 25 −→

10 *<B <* 20 have conﬁdence 33*.*3%.

ii. 15 *< A <* 25 −→ 10 *< B <* 20, 15 *< A <* 25 −→ 5 *< B <* 20, and

15 *<A<* 25 −→ 5 *<B <* 30.

**Answer:**

The rule 15 *< A <* 25 −→ 5 *< B <* 30 has the highest conﬁdence

(83*.*3%).

iii. 15 *<A<* 25 −→ 10 *<B <* 20 and 10 *<A<* 35 −→ 5 *<B <* 30.

**Answer:**

The rule 10 *< A <* 35 −→ 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 × 12*/*65 + 7*.*5 × 20*/*65 + 15 × 15*/*65 + 25 × 13*/*65 + 35 × 5*/*65 = 13*.*92*.*

Therefore the statistics-based association rule is:

15 ≤ *A<* 35 −→ *B* : *µ* = 13*.*82*.*

(c) Test whether the quantitative association rule given in part (b) is sta- tistically 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 × 6*/*35 + 7*.*5 × 9*/*35 + 15 × 10*/*35 + 25 × 6*/*65 + 35 × 4*/*65 = 14*.*93*.*

The standard deviations for the two groups are 9.786 (15 ≤ Age *<* 35)

and 10.203 (Age *<* 15 or Age ≥ 35), respectively.

14*.*93 − 13*.*82

*Z* =

9*.*7862

+

65

10*.*2032

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 con- vert 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, Grad- uate:PhD, Professional

**Answer:**

(a) Each attribute value can be represented using an asymmetric bi- nary 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 at- tribute. 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 at- tribute. 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 bi- nary 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 7.13. Suppose we are interested in extracting the following association rule:

{*α*1 ≤ Age ≤ *α*2 *,* Play Piano = Yes} −→ {Enjoy Classical Music = Yes}

**Table 7.13.** Data set for Exercise 6.

|  |  |  |
| --- | --- | --- |
| Age | Play Piano | Enjoy Classical Music |
| 9  11  14  17  19  21  25  29  33  39  41  47 | Yes  Yes Yes Yes Yes No No Yes No No No No | Yes  Yes No No Yes No No Yes No Yes 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 val- ues. 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:**

(*α*1 = 19*, α*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 ob- tained 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 ob- tained in part (a).

**Answer:**

(*α*1 = 9*, α*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 7.14, with an item taxonomy given in Figure 7.25.

**Table 7.14.** Example of market basket transactions.

|  |  |
| --- | --- |
| Transaction ID | Items Bought |
| 1  2  3  4  5  6  7 | Chips, Cookies, Regular Soda, Ham  Chips, Ham, Boneless Chicken, Diet Soda Ham, Bacon, Whole Chicken, Regular Soda Chips, Ham, Boneless Chicken, Diet Soda Chips, Bacon, Boneless Chicken  Chips, Ham, Bacon, Whole Chicken, Regular Soda  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 ex- tended transaction *t*t that contains all the items in *t* as well as their re- spective ancestors. For example, the transaction *t* = { Chips, Cookies} will be replaced by *t*t = {Chips, Cookies, Snack Food, Food}. Use this approach to derive all frequent itemsets (up to size 4) with support ≥

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 gen- erated 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 hi-

erarchy. 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 ≥ 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 asso- ciation rule may vary in the presence of a concept hierarchy.

(a) Consider an item *x* in a given concept hierarchy. Let *x*1 , *x*2 , *.. .*, *xk* denote the *k* children of *x* in the concept hierarchy. Show that *s*(*x*) ≤ *i*=1 *s*(*xi* ), where *s*(·) is the support of an item. Under what conditions

L*k*

will the inequality become an equality?

**Answer:**

If no transaction contains more than one child of *x*, then *s*(*x*) = L*k*

*i*=1

*s*(*xi* ).

(b) Let *p* and *q* denote a pair of items, while *p*ˆ and *q*ˆ 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*({*p*ˆ*, q*}), (ii) *s*({*p, q*ˆ}), and (iii) *s*({*p*ˆ*, q*ˆ}).

**Answer:**

All three itemsets are guaranteed to be frequent.

(c) Consider the association rule {*p*} −→ {*q*}. Suppose the conﬁdence of the rule exceeds *minconf* . Which of the following rules are guaranteed to have conﬁdence higher than *minconf* ? (i) {*p*} −→ {*q*ˆ}, (ii) {*p*ˆ} −→

{*q*}, and (iii) {*p*ˆ} −→ {*q*ˆ}.

**Answer:**

Only {*p*} −→ {*q*ˆ} 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 ≥ 50% given the sequence database shown in Table 7.15. 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 *... ei ... ei*+1 *... elast >* given be- low, determine whether they are subsequences of the sequence

*<* {1*,* 2*,* 3}{2*,* 4}{2*,* 4*,* 5}{3*,* 5}{6} *>*

**Table 7.15.** Example of event sequences generated by various sensors.

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

subjected to the following timing constraints:

mingap = 0 (interval between last event in *ei* and ﬁrst event in *ei*+1 is *>* 0)

maxgap = 3 (interval between ﬁrst event in *ei* and last event in *ei*+1 is ≤ 3)

maxspan = 5 (interval between ﬁrst event in *e*1 and last event in *elast* is ≤ 5)

*ws* = 1 (time between ﬁrst and last events in *ei* 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* = (*e*1 *,..., elast* ) below, determine whether they are subsequences of the following data sequence:

({*A, B*}{*C, D*}{*A, B*}{*C, D*}{*A, B*}{*C, D*})

subjected to the following timing constraints:

mingap = 0 (interval between last event in *ei* and ﬁrst event in *ei*+1 is *>* 0)

maxgap = 2 (interval between ﬁrst event in *ei* and last event in *ei*+1 is ≤ 2)

maxspan = 6 (interval between ﬁrst event in *e*1 and last event in *elast* is ≤ 6)

*ws* = 1 (time between ﬁrst and last events in *ei* is ≤ 1)

(a) *w* = ({*A*}{*B*}{*C* }{*D*})

**Answer:** Yes.

(b) *w* = ({*A*}{*B, C, D*}{*A*})

**Answer:** No.

(c) *w* = ({*A*}{*B, C, D*}{*A*})

**Answer:** No.

(d) *w* = ({*B, C* }{*A, D*}{*B, C* })

**Answer:** No.

(e) *w* = ({*A, B, C, D*}{*A, B, C, D*})

**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 7.16 for a given object. Count the number of occurrences for the sequence ({*p*}{*q*}{*r*}) according to the following counting methods:

Assume that *ws* = 0, *mingap* = 0, *maxgap* = 3, *maxspan* = 5).

**Table 7.16.** Example of event sequence data for Exercise 14.

|  |  |
| --- | --- |
| Timestamp | Events |
| 1  2  3  4  5  6  7  8  9  10 | p, q  r s  p, q r, s p  q, r q, s p  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 over- lap).

**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 gen- eration using vertex growing approach.

(b) An unlabeled graph is equivalent to a labeled graph where all the ver- tices 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 7.2. Assume the edge-growing method is used to expand the sub- graphs.

**Answer:** See Figure 7.3.

17. Draw all the candidate subgraphs obtained by joining the pair of graphs shown in Figure 7.4. Assume the edge-growing method is used to expand the subgraphs.

**Answer:** See Figure 7.5.

18. (a) If support is deﬁned in terms of induced subgraph relationship, show that the conﬁdence of the rule *g*1 −→ *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 ,

··· , *G*5 , shown in Figure 7.6. The graph *g*1 shown on the top-right hand

b b

a a a a

a a a a

(a)

b a a

b a a

b a a

c a a

(b)

**Figure 7.2.** Graphs for Exercise 16.

b b b b b

a a a a a a

a a a

a

(a)

a a b

b a a

b a a

c

(b)

**Figure 7.3.** Solution for Exercise 16.

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 −→ *g*1 . Using the standard deﬁnition of conﬁdence as the ratio between the support of *g*2 ∪ *g*1 ≡ *g*3 to the

b a b

b a b

b b b b

(a)

b a b

b a a

b a a

c a a

(b)

**Figure 7.4.** Graphs for Exercise 17.

b a b

b a b

b b b b

b a b

b a b

b b b

b b

b

b a b b b a b

b b b b

b a a b a a

b a a c a a

b a a

b a a

C

**Figure 7.5.** Solution for Exercise 17.

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 1 a 1

e

Subgraph g1

d a 1 e

1 1 1

b 1 d c

G1

1 1 1 c 1 e

b

G2

Subgraph support = 80%

Induced subgraph support = 80%

Subgraph g2

a 1 e

b 1 d

a 1 d

1 1 1

1 a 1 e

1

d

G3 G4 c

1

e

Subgraph support = 60%

Induced subgraph support = 20%

b d

1 1 1 G

a e 5

1 c

Subgraph g3

a 1 d

1 1 e

Graph Data Set

Subgraph support = 40%

Induced subgraph support = 40%

**Figure 7.6.** Computing the support of a subgraph from a set of graphs.

A na¨ıve 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*.

**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.

19. (a) Consider a graph mining algorithm that uses the edge-growing method to join the two undirected and unweighted subgraphs shown in Figure

19a.

i. Draw all the distinct cores obtained when merging the two sub- graphs.

**Answer:** See Figure 7.7.

B B

A A A A

A A A A A A

B B B

A A A A A

**Figure 7.7.** Solution to Exercise 19.

ii. How many candidates are generated using the following core?

A A

B

A A

**Answer:** No candidate *k* + 1-subgraph can be generated from the core.

20. 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 na¨ıve way for deriving negative itemsets is to extend each transaction to include absence of items as shown in Table 7.17.

i. Suppose the transaction database contains 1000 distinct items.

What is the total number of positive itemsets that can be gener-

**Table 7.17.** Example of numeric data set.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| TID | TV | ¬TV | DVD | ¬DVD | VCR | ¬VCR | *...* |
| 1  2 | 1  1 | 0  0 | 0  0 | 1  1 | 0  0 | 1  1 | *...*  *...* |

ated from these items? (Note: A positive itemset does not contain any negated items).

**Answer:** 21000 − 1.

ii. What is the maximum number of frequent itemsets that can be

generated from these transactions? (Assume that a frequent item- set may contain positive, negative, or both types of items) **Answer:** 22000 − 1.

iii. Explain why such a na¨ıve 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 *x*).

(b) Consider the database shown in Table 7.14. What are the support and conﬁdence values for the following negative association rules involving regular and diet soda?

i. ¬Regular −→ Diet.

**Answer:** *s* = 42*.*9%, *c* = 100%.

ii. Regular −→ ¬Diet.

**Answer:** *s* = 42*.*9%, *c* = 100%.

iii. ¬Diet −→ Regular.

**Answer:** *s* = 42*.*9%, *c* = 100%.

iv. Diet −→ ¬Regular.

**Answer:** *s* = 42*.*9%, *c* = 100%.

21. 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 2*d*. 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:** 22*d* .

(b) Assume that an itemset must contain positive or negative items of dif- ferent variables. For example, the itemset {*a, a, b, c*} is invalid because it contains both positive and negative items for variable *a*. What is the total size of the itemset lattice?

**Answer:** L*d*

(*d*) L*k*

(*k*) = L*d*

(*d*)2*k* = 3*d* − 1.

*k*=1 *k*

*i*=0 *i*

*k*=1 *k*

sure 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, c, 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 ∨ b ∨ 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 pat-

terns?

**Answer:** Non-monotone.

23. Many association analysis algorithms rely on an *Apriori* -like approach for ﬁnding frequent patterns. The overall structure of the algorithm is given below.

**Algorithm 7.1** *Apriori* -like algorithm.

1: *k* = 1.

2: *Fk* = { *i* | *i* ∈ *I* ∧ *σ*({*i*}) ≥ *minsup*}. {Find frequent 1-patterns.}

*N*

3: **repeat**

4: *k* = *k* + 1.

5: *Ck* = genCandidate(*Fk*−1 ). {Candidate Generation}

6: *Ck* = pruneCandidate(*Ck* , *Fk*−1 ). {Candidate Pruning}

7: *Ck* = count(*Ck* , *D*). {Support Counting}

8: *Fk* = { *c* | *c* ∈ *Ck* ∧ *σ*(*c*) ≥ *minsup*}. {Extract frequent patterns}

*N*

9: **until** *Fk* = ∅

10: Answer = U *Fk* .

Suppose we are interested in ﬁnding boolean logical rules such as

{*a* ∨ *b*} −→ {*c, d*}*,*

which may contain both disjunctions and conjunctions of items. The corre- sponding itemset can be written as {(*a* ∨ *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 pat- terns?

(d) How should the support counting step be modiﬁed to ﬁnd such pat- terns?

**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.

**8**

**Cluster Analysis: Basic Concepts and Algorithms**

1. Consider a data set consisting of 220 data vectors, where each vector has

32 components and each component is a 4-byte value. Suppose that vec- tor quantization is used for compression and that 216 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 × 32 × 220 = 134*,* 217*,* 728 bytes. After compression, the data set requires 4×32×216 = 8*,* 388*,* 608 bytes for the prototype vectors and 2 × 220 = 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 8.1.

The solutions are also indicated in Figure 8.1.

**Figure 8.1.** Points for Exercise 2.

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 8.1. (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!*/*44 = 0*.*0938.

number of ways to select one centroid from each cluster

*p* =

number of ways to select *K* centroids

*K* !*nK K* !

= =

(*Kn*)*K K K*

(8.1)

(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 essen- tially 0 by the time *K* = 10.

(b) For *K* clusters, *K* = 10*,* 100*,* and 1000, ﬁnd the probability that a sample of size 2*K* 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 proba- bilities are 0.21, *<* 10−6 , and *<* 10−6 .

Proceeding analytically, the probability that a point doesn’t come from a particular cluster is, 1 − 1 , and thus, the probability that all 2*K* points don’t come from a particular cluster is (1 − 1 )2*K* . Hence, the probability that at least one of the 200 points comes from a particular cluster is 1 − (1 − 1 )2*K* . If we assume independence (which is too optimistic, but becomes approximately true for larger values of *K* ), then

*K*

*K*

*K*

an upper bound for the probability that all clusters are represented in the ﬁnal sample is given by (1 − (1 − 1 )2*K* )*K* . The values given by this bound are 0.27, 5.7e-07, and 8.2e-64, respectively.

*K*

0.5

0.45

0.4

0.35

0.3

Probability

0.25

0.2

0.15

0.1

0.05

0

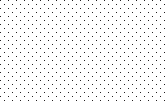
0 10 20 30 40 50 60 70 80 90 100

K

**Figure 8.2.** Probability of at least one point from each cluster. Exercise 4.

5. Identify the clusters in Figure 8.3 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 8.3.** 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.

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 be- comes 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. As- sume 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 8.4 matches the corresponding part of this question, e.g., Figure

8.4(a) goes with part (a).

Global minimum



Local minimum Global minimum



Local minimum

(a) (b) (c)

(d)

(e)

**Figure 8.4.** 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◦ ≤ *θ* ≤ 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 necessar- ily 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 en- closed 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 lo- cal 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 cir- cle 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 mini- mize 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 signiﬁ- cant 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?

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 8.5.) 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 per- pendicular line splits the plane into two regions, each containing points that are closest to the centroid the region contains.

**Figure 8.5.** Voronoi diagram for Exercise 13.

You use K-means to cluster the data, but for all values of *K* , 1 ≤ *K* ≤ 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 prox- imity 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 struc- ture of the data. For example, consider a set of three clusters, each of which has two, three, and four subclusters, respectively. An ideal hi- erarchical 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 hier- archical clustering by moving clusters up the tree.

16. Use the similarity matrix in Table 8.1 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 8.6(a) and 8.6(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 8.1.** 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 |

0.55

0.1

0.6

0.2

0.65

0.3

0.7

0.4

0.75

Similarity

0.5

Similarity

0.8

0.6

0.85

0.7

0.9

0.8

0.95

0.9

1

2 5 3 4 1

1

2 5 3 1 4

(a) Single link.

(b) Complete link.

**Figure 8.6.** 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 accept- able 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 mini- mizing SSE, there is no reﬁnement step as in regular K-means. Likewise, bisecting K-means has no overall reﬁnement step. Thus, unless such a reﬁne- ment 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 conse- quently, are impractical to use directly on larger data sets. One possible technique for reducing the time required is to sample the data set. For ex-

ample, if *K* clusters are desired and √*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 hier- archical 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* − √*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) sam- pling will cause problems for this approach and (2) what those problems are. Assume that the sampling technique randomly chooses points from the to- tal 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 typi- cally 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.

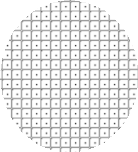
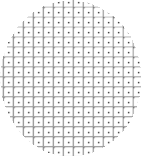
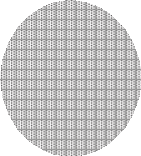
This has no particular impact.

(j) Data with many and mixed attribute types.

Many attributes was discussed under high-dimensionality. Mixed at- tributes have no particular impact.

20. Consider the following four faces shown in Figure 8.7. 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 8.7.** 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 8.7(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 8.2.

**Table 8.2.** 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 8.3.** 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 8.4 and

8.5, respectively, compute the correlation between the similarity matrix and the ideal similarity matrix, i.e., the matrix whose *ijth* entry is 1 if two objects belong to the same cluster, and 0 otherwise.

**Table 8.4.** Table of cluster labels 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 |

**Table 8.5.** Similarity matrix for Exercise 24.

|  |  |
| --- | --- |
| Point | Cluster Label |
| P1 | 1 |
| P2 | 1 |
| P3 | 2 |
| P4 | 2 |

We need to compute the correlation between the vector **x** =*<* 1*,* 0*,* 0*,* 0*,* 0*,* 1 *>* and the vector **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 simi- larity matrix.

We get:

Standard deviation of the vector **x** : *σx* = 0*.*5164

Standard deviation of the vector **y** : *σy* = 0*.*1703

Covariance of **x** and **y**: cov(**x***,* **y**) = −0*.*200

Therefore, corr(**x***,* **y**) = cov(**x***,* **y**)*/σx σy* = −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 8.8. Class A contains points p1, p2, and p3, while p4, p5, p6, p7, and p8 belong to class B.

{p1, p2, p3, p4, p5, p6, p7, p8}

{p1, p2, p4, p5,}

{p3, p6, p7, p8}

{p1, p2} {p4, p5} {p3, p6} {p7, p8}

**Figure 8.8.** Hierarchical clustering for Exercise 25.

Let *R*(*i, j*) = *nij /ni* indicate the recall of class *i* with respect to cluster *j*. Let *P* (*i, j*) = *nij /nj* indicate the precision of class *i* with respect to cluster *j*.

*F* (*i, j*) = 2*R*(*i, j*) × *P* (*i, j*)*/*(*R*(*i, j*)+ *P* (*i, j*)) is the F-measure for class *i* and

cluster *j*.

For cluster #1 = {*p*1*, p*2*, p*3*, p*4*, p*5*, p*6*, p*7*, p*8}: Class = A:

*R*(*A,* 1) = 3*/*3 = 1*, P* (*A,* 1) = 3*/*8 = 0*.*375

*F* (*A,* 1) = 2 × 1 × 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* = L2 *ni* max*i F* (*i, j*) = 3*/*8 ∗ *F* (*A*)+5*/*8 ∗ *F* (*B*) = 0*.*78

1 *n*

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 8.14.

1 ) ) (*x − y*)2 = 1

) ) ((*x − c* ) *−* (*y − c* ))2

2*|Ci | x*∈*Ci y*∈*Ci*

*i i*

2*|Ci | x*∈*Ci y*∈*Ci*

= 1 )

) (*x − ci* )2 *−* 2 )

) (*x − ci* )(*y − ci* )

2*|Ci |*

+ )

*x*∈*Ci y*∈*Ci*

) (*y − ci* )2

*x*∈*Ci y*∈*Ci*

*x*∈*Ci y*∈*Ci*

= 1 )

) (*x − ci* )2 + )

) (*y − ci* )2

2*|Ci |*

1

=

*x*∈*Ci y*∈*Ci*

*|Ci |*(*x − ci* )2

)

*x*∈*Ci y*∈*Ci*

*|Ci | x*∈*Ci*

= SSE

The cross term L*x*∈*Ci* L*y*∈*Ci* (*x* − *ci* )(*y* − *ci* ) is 0.

28. Prove Equation 8.15.

1 *K K*

1 *K K*

) )

*K i*=1 *j*=1

*|Ci |*(*cj − ci* )2 =

) )

2*K i*=1 *j*=1

*|Ci |*((*m − ci* ) *−* (*m − cj* ))2

1 *K K K K*

) ) ) )

= *|Ci |*(*m − ci* )2 *−* 2 *|Ci |*(*m − ci* )(*m − cj* )

2*K i*=1 *j*=1

*K K*

) )

*i*=1 *j*=1

+ *|Ci |*(*m − cj* )2

*i*=1 *j*=1

1 *K K K K*

) ) ) )

= *|Ci |*(*m − ci* )2 + *|Ci |*(*m − cj* )2

2*K i*=1 *j*=1

1 *K*

*i*=1 *j*=1

= ) *K |Ci |*(*m − ci* )2

*K i*=1

= SSB

Again, the cross term cancels.

29. Prove that L*K*

*i*=1

L

*x*∈*Ci*

(*x* − *mi* )(*m* − *mi* ) = 0. This fact was used in the

proof that TSS = SSE + SSB on page 557.

*K*

) ) (*x − ci* )(*c − ci* ) =

*K*

) ) (*xc − ci c − xci* + *c*2 )

*i*

*i*=1 *x*∈*Ci*

*i*=1 *x*∈*Ci*

*K K K K*

= ) ) *xc −* ) ) *ci c −* ) ) *xci* + ) ) *c*2

*i*

*i*=1 *x*∈*Ci*

*i*=1 *x*∈*Ci*

*i*=1 *x*∈*Ci*

*i*=1 *x*∈*Ci*

= *mi ci c − mi ci c − mi c*2 + *mi c*2

*i i*

= 0

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 fre- quently 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 at- tribute, 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 ob- jects or attributes, is applied. In co-clustering both sets of constraints

are applied simultaneously. Thus, partitioning the objects and at- tributes 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 8.9, 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

**1**

1

0.9

0.8

0.7

0.6

0.5

**y**

0.4

0.3

0.2

0.1

0

**2**

1

0.9

0.8

0.7

0.6

0.5

**y**

0.4

0.3

0.2

0.1

0

0 0.2 0.4 0.6 0.8 1

**x**

**3**

1

0.9

0.8

0.7

0.6

0.5

**y**

0.4

0.3

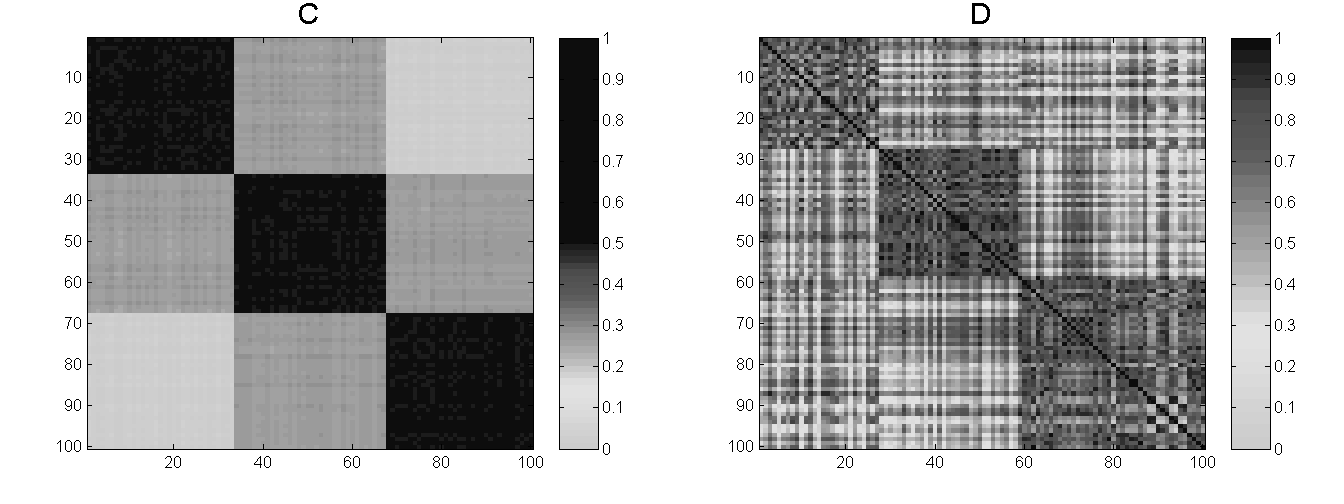
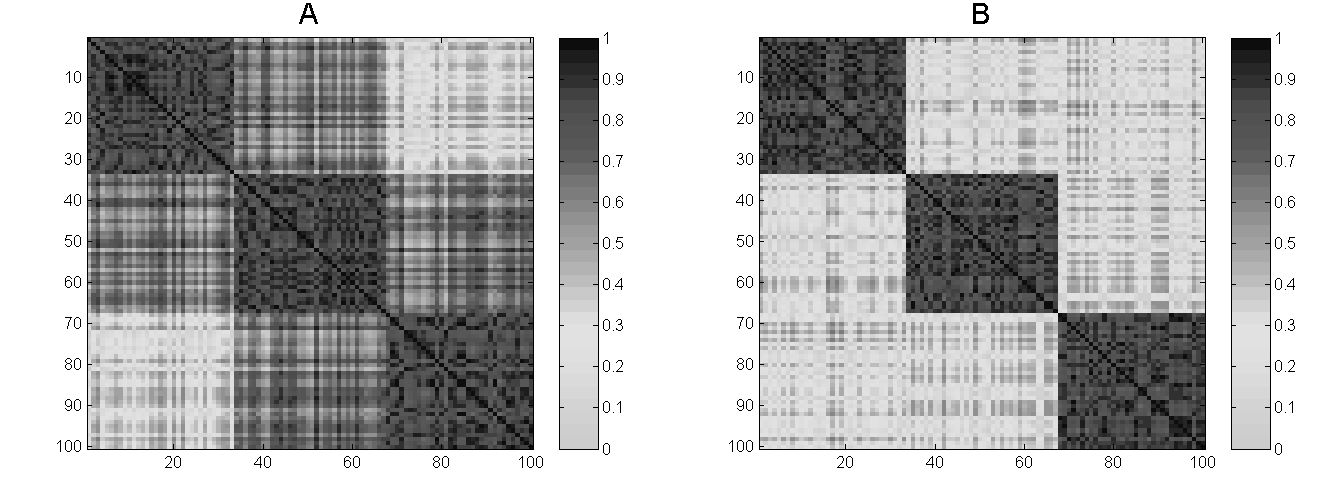
0.2

0.1

0

0 0.2 0.4 0.6 0.8 1

**x**



0 0.2 0.4 0.6 0.8 1

**x**

**4**

1

0.9

0.8

0.7

0.6

0.5

**y**

0.4

0.3

0.2

0.1

0

0 0.2 0.4 0.6 0.8 1

**x**

**Figure 8.9.** Points and similarity matrices for Exercise 32.

**9**

**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 con- tain many of the same words. Although we can also include the frequency with which those words occur in the similarity computation, this can some- times 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 normal- ized 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*) ≥ *δ*, where 0 *<δ* ≤ 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 opti- mization 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 do- ing, and (2) it allows the use of powerful optimization techniques that have been developed in a wide variety of ﬁelds. Unfortunately, most of these op- timization 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 out- liers 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 mem- bership 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 9.12 gives the likelihood for a set of points from a Gaussian dis- tribution 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*.

*∂1*((*µ, σ*)|X ) = *∂*

) (*xi* − *µ*)2 \

*∂µ ∂µ*

*m*

−

*i*=1

2*σ*2 − 0*.*5*m* log 2*π* − *m* log *σ*

*m*

= − )

2(*xi*

− *µ*)

*i*=1

2*σ*2

Setting this equal to 0 and solving, we get *µ* = 1 L*m*

*xi* .

Likewise, we can solve for *σ*.

*m*

*m i*=1

2 \

*∂1*((*µ, σ*)|X )

*∂σ*

*∂*

= *∂σ* −

*m*

) (*xi* − *µ*)

2*σ*2 − 0*.*5*m* log 2*π* − *m* log *σ*

*i*=1

2

= ) 2(*xi* − *µ*) *m*

*i*=1

2*σ*3 − *σ*

Setting this equal to 0 and solving, we get *σ*2 = 1 L*m*

(*xi* − *µ*)2 .

*m i*=1

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 informa- tion 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 9.1 and 9.4, 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 9.1 shows a clustering of a two-dimensional point data set with two clusters: The leftmost cluster, whose points are marked by asterisks, is some- what 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.

6

4



2

0

y

–2

–4

–6

–8

–10 –8 –6 –4 –2 0 2 4

x

**Figure 9.1.** 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 9.4.2 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 succes- sively 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 con- nected 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 sparsi- ﬁed proximity matrix symmetric.

One approach is to set the *ijth* entry to 0 if the *jith* entry is 0, or vice versa. Another approach is to set the *ijth* entry to 1 if the *jith* 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 9.1 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 9.10.

The following is the SNN similarity matrix.

17. For the deﬁnition of SNN similarity provided by Algorithm 9.10, the cal- culation of SNN distance does not take into account the position of shared

**Table 9.1.** Two nearest neighbors of four points.

|  |  |  |
| --- | --- | --- |
| Point | First Neighbor | Second Neighbor |
| 1 | 4 | 3 |
| 2 | 3 | 4 |
| 3 | 4 | 2 |
| 4 | 3 | 1 |

**Table 9.2.** Two nearest neighbors of four points.

|  |  |
| --- | --- |
| Point | 1 2 3 4 |
| 1  2  3  4 | 2 0 0 1  0 2 1 0  0 1 2 0  1 0 0 2 |

neighbors in the two nearest neighbor lists. In other words, it might be de- sirable 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 *ith* 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 collec- tions 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 8, page 564.)

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 diﬃ- cult 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*) ≥ 2*d*(*x, b*) then *d*(*x, c*) ≥ *d*(*x, b*).

**Proof:**

We know *d*(*b, c*) ≤ *d*(*b, x*)+ *d*(*x, c*). So *d*(*b, c*) − *d*(*x, b*) ≤ *d*(*x, c*).

Now *d*(*b, c*) − *d*(*x, b*) ≥ 2*d*(*x, b*) − *d*(*x, b*) = *d*(*x, b*). So *d*(*x, b*) ≤ *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 9.19—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.

**10**

**Anomaly Detection**

1. Compare and contrast the diﬀerent techniques for anomaly detection that were presented in Section 10.1.2. 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 tech- niques 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 tech- nique that ﬁts a model to the data. However, note that a particular model, statistical or otherwise, must be assumed. Consequently, model-based ap- proaches 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 pro- duce 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 unusu- ally 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 Section 10.1.2 is this deﬁ- nition 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 con- vex 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 numer- ical anomaly scores.

4. Association analysis can be used to ﬁnd anomalies as follows. Find strong as- sociation patterns, which involve some minimum number of objects. Anoma- lies 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 6.8 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 similar- ity 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 ap- proach.

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 unsuper- vised 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 10.1, is a more statistically sophisticated procedure for detecting outliers than that of Deﬁnition 10.3. 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 *gc* , 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 *gc* are updated at each iteration.

**Algorithm 10.1** Grubbs’ approach for outlier elimination.

1: Input the values and *α*

{*m* is the number of values, *α* is a parameter, and *tc* is a value chosen so that

*α* = *prob*(*x* ≥ *tc* ) for a *t* distribution with *m* − 2 degrees of freedom.}

2: **repeat**

3: Compute the sample mean (*x*) and standard deviation (*sx* ).

4: Compute a value *gc* so that *prob*(|*z*|≥ *gc* ) = *α*.

2

(In terms of *tc* and *m*, *gc* = *m*− 1

√*m*

*tc* .)

*m*−2+*t*2

*c*

5: Compute the z-score of each value, i.e.,

*z* = (*x* − *x*)*/sx* .

6: Let

*g* = max |*z*|, i.e., ﬁnd the z-score of largest magnitude and call it *g*.

7: **if** *g > gc* **then**

8: Eliminate the value corresponding to *g*.

9: *m* ← *m* − 1

10: **end if**

11: **until** No objects are eliminated.

(a) What is the limit of the value *m*− 1

√

*m*

2 *c m*−2+*t*2

used for Grubbs’ test as

*m* approaches inﬁnity? Use a signiﬁcance level of 0.05.

*t*

*c*

Note that this could have been phrased better. The value of this ex- pression approaches *tc* , but strictly speaking this is not a limit as *tc* depends on *m*.

lim

*m* − 1

2

*c* = lim

*c*

*m* − 1

*tc* = 1 × *tc* = *tc .*

*t*

*m*→∞ √*m*

*c*

*m* − 2+ *t*2

*m*→∞ /*m*(*m* − 2+ *t*2 ) ×

Also, the value of *tc* will continue to increase with *m*, although slowly. For *m* = 1020 , *tc* = 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 stan- dard deviations or 0.0027 for a double-sided deviation. Thus, the num- ber 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 **x** with respect to a multivariate normal distribution having a mean ***µ*** and covariance matrix **Σ** is given by the equa- tion

1

*prob*(**x**) = √

*e*−

1*/*2

(**x***−****µ***)Σ*−*1 (**x***−****µ***)*T*

2

*.* (10.1)

( 2*π*)*m* |Σ|

Using the sample mean **x** and covariance matrix **S** as estimates of the mean

***µ*** and covariance matrix **Σ**, respectively, show that the log(*prob*(**x**)) is equal to the Mahalanobis distance between a data point **x** and the sample mean **x** plus a constant that does not depend on **x**.

log *prob*(**x**) = − log ((√2*π*)*m* |Σ 1*/*2 ) −

|

1 1

2 (**x** − ***µ***)Σ−

(**x** − ***µ***)*T .*

If we use the sample mean and covariance as estimates of ***µ*** and Σ, respec- tively, then

log *prob*(**x**) = − log ((√2*π*)*m* |*S*|1*/*2 ) −

1

(**x** − **x**)**S**−1 (**x** − **x**)*T*

2

tity, 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 dis- tance.

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 8.5.2.

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 de- scribed in Section 10.5 and the accompanying ﬁgure, Figure 10.10.

(a) The points at the bottom of the compact cluster shown in Figure 10.10 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.)

number of anomalies detected

detection rate =

false alarm rate =

total number of anomalies

number of false anomalies

number of objects classiﬁed as anomalies

The detection rate is simply 99%.

The false alarm rate = 0*.*99*m* × 0*.*01*/*(0*.*99*m* × 0*.*01 + 0*.*01*m* × 0*.*99) = 0*.*50 =

50%.

13. When a comprehensive training set is available, a supervised anomaly detec- tion 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 de- termine, 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 super- vised and unsupervised anomaly detection methods should be used. A good example of such a situation is network intrusion detection. Proﬁles or sig- natures 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.

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 mean- ingful 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 de- scribed 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 assump- tions 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.

**Bibliography**

BIBLIOGRAPHY **165**

[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. Fu¨rnkranz 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.