Exercise Tasks for the course on

Data Mining

in the degree program in Computer Science

Modified version (omitted and added items) of the tasks in
Tan/Steinbach/Kumar: *Introduction to Data Mining*. Pearson Education, 2006.

Ilmenau University of Technology
Faculty of Computer Science and Automation
Artificial Intelligence Group
apl. Prof. Dr.-Ing. habil. Rainer Knauf
Summer Term 2010
as developed so far by July 11, 2010

0 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.
   (b) Dividing the customers of a company according to their profitability.
   (c) Computing the total sales of a company.
   (d) Sorting a student database based on student identification numbers.
   (e) Predicting the outcomes of tossing a (fair) pair of dice.
   (f) Predicting the future stock price of a company using historical records.
   (g) Monitoring the heart rate of a patient for abnormalities.
   (h) Monitoring seismic waves for earthquake activities.
   (i) Extracting the frequencies of a sound wave.

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

3. For each of the following data sets, explain whether or not data privacy is an important issue.
   (a) Census data collected from 19001950.
   (b) IP addresses and visit times of Web users who visit your Website.
   (c) Images from Earth-orbiting satellites.
   (d) Names and addresses of people from the telephone book.
   (e) Names and email addresses collected from the Web.

1 Data

1. Each of the following Data Objects describes medical information of a patient as a 5-tuple of values:

012	232	33.5	0	10.7
020	121	16.9	2	210.1
027	156	24.0	0	427.6
...	...	...	...	...

A 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?

2. Classify the following attributes in Table 1 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.

3. You are approached by the marketing director of a local company, who believes that he has devised a foolproof way to measure customer satisfaction. He explains his scheme as follows:

> *Its so simple that I cant believe that no one has thought of it before. I just keep track of the number of customer complaints for each product. I read in a data mining book that counts are ratio attributes, and so, my measure of product satisfaction must be a ratio attribute. But when I rated the products based on my new customer satisfaction measure and showed them to my boss, he told me that I had overlooked the obvious, and that my measure was worthless. I think that he was just mad because our best-selling product had the worst satisfaction since it had the most complaints.*
>
> *Could you help me set him straight?*

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

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

4. A few months later, you are again approached by the same marketing director as in the 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.*

[1] An ISBN consists of 13 digits: 3 for a prefix that identifies the book branch (currently 978), followed by a 9 digit identification number and one digit that serves as a check number that checks the integrity of the complete ISBN.

Attribute	discrete		continuos	qualitative		quantitative	
	binary			nominal	ordinal	interval	ratio
Time in terms of AM or PM.							
Brightness as measured by a light meter.							
Brightness as measured by peoples judgments.							
Angles as measured in degrees between 0. and 360.							
Bronze, Silver, and Gold medals as awarded at the Olympics.							
Height above sea level.							
Number of patients in a hospital.							
ISBN numbers for books.[1]							
Ability to pass light in terms of the following values: opaque, translucent, transparent.							
Military rank.							
Distance from the center of campus.							
Density of a substance in g/cm^3.							
Coat check number in a theater.							
Licence plate numbers in Germany							

Table 1: Classification of Attributes

I suggested that we perform the comparisons in pairs and then use these comparisons to get the rankings. Thus, if we have three product variations, we have the customers compare variations 1 and 2, then 2 and 3, and finally 3 and 1.

Our testing time with my new procedure is a third of what it was for the old procedure, but the employees conducting the tests complain that they cannot come up with a consistent ranking from the results. And my boss wants the latest product evaluations, yesterday. I should also mention that he was the person who came up with the old product evaluation approach. Can you help me?

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

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

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

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

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

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

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

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

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

9. Consider the problem of finding the k nearest neighbors of a data object. A programmer designs the following algorithm for this task.

   Algorithm **kNN** for finding K nearest neighbors.
   1: **for** $i = 1$ to number of data objects **do**
   2: Find the distances of the i-th object to all other objects.
   3: Sort these distances in decreasing order.
   (Keep track of which object is associated with each distance.)
   4: **return** objects associated with the first K distances of the sorted list
   5: **end for**

   (a) Describe the potential problems with this algorithm if there are duplicate objects in the data set. Assume the distance function will only return a distance of 0 for objects that are the same.

(b) How would you fix this problem?

10. The following attributes are measured for members of a herd of Asian elephants: *weight*, *height*, *tusk length*, *trunk length*, and *ear area*. Based on these measurements, what sort of similarity measure would you use to compare or group these elephants? Justify your answer and explain any special circumstances.

11. This exercise compares and contrasts some similarity and distance measures.

    (a) Compute

        i. the Hamming Distance (L_1 Norm)

        ii. the Jaccard-Coeffizient

        for the two binary vectors $x = [0101010001]$ and $y = [0100011000]$.

    (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[1]? Explain.

    (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 [2]. Explain.

    (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?[3]

12. For the following vectors, x and y, calculate the indicated similarity or distance measures.

    (a) cosine, correlation, Euclidean od $x = [1, 1, 1, 1]$ and $y = [2, 2, 2, 2]$.

    (b) cosine, correlation, Euclidean, Jaccard of $x = [0, 1, 0, 1]$ and $y = [1, 0, 1, 0]$.

    (c) cosine, correlation, Euclidean of $x = [0, -1, 0, 1]$ and $y = [1, 0, -1, 0]$.

    (d) cosine, correlation, Jaccard of $x = [1, 1, 0, 1, 0, 1]$ and $y = [1, 1, 1, 0, 0, 1]$.

    (e) cosine, correlation of $x = [2, -1, 0, 2, 0, -3]$ and $y = [-1, 1, -1, 0, 0, -1]$.

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

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

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

    (c) What is the relationship of the cosine measure to correlation, if any[4]?

[1] The Hamming measure is a distance, while the other three measures are similarities, but dont let this confuse you.

[2] 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.

[3] Note that two human beings share > 99.9 % of the same genes. This is similar with other species.

[4] Look at statistical measures such as mean and standard deviation in cases where cosine and correlation are the same and different.

(d) Figure 1(a) shows the relationship of the cosine measure to Euclidean distance for 100,000 randomly generated points (pairs of data objects with positive attribute values only) that have been normalized to have an L_2 norm of 1 (distance from the origin of the Euclidian space).

What general observation can you make about the relationship between Euclidean distance and cosine similarity when vectors have an L_2 norm of 1?

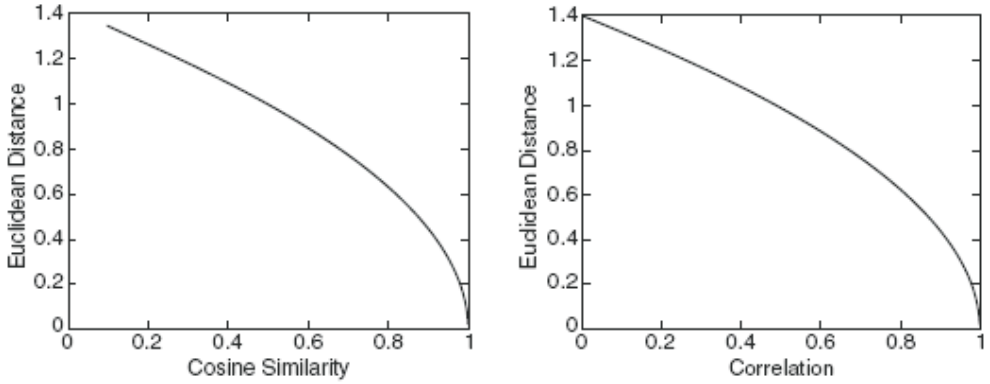


Figure 1: Euklidean Distance vs. Cosine Coefficient and Correlation

(e) Figure 1 (b) shows the relationship of correlation to Euclidean distance for 100,000 randomly generated points (pairs of data objects with positive attribute values only) 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?

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

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

14. Show that the set difference metric for data objects a and B given by

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

satisfies the following axioms:

(a) d is not negative: $\forall X, Y : d(X, Y) \geq 0$

(b) $\forall X, Y : d(X, Y) = 0$, iff $X = Y$

(c) d is symmetric: $\forall X, Y : d(X, Y) = d(Y, X)$

(d) $\forall X, Y, Z : d(X, Z) \leq d(X, Y) + d(Y, Z)$

15. Discuss how you might map correlation values from the interval $[-1, 1]$ to the interval $[0, 1]$. Note that the type of transformation that you use might depend on the application that you have in mind. Thus, consider two applications: (a) clustering time series and (b) predicting the behavior of one time series given another.

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

Table 2: Parity Function

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

17. 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.

    (b) How might you define the distance between two sets of points in Euclidean space?

2 Classification

2.1 Decision Trees

1. Draw the full decision tree for the parity function of four Boolean attributes A, B, C, and D (see Table 2). A, B, C, and D. Is it possible to simplify the tree?

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

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

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

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

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

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

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

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

3. Consider the training examples shown in Table 4 for a binary classification problem.

Customer ID	*Gender*	*Car Type*	*Shirt Size*	*Class*
1	m	Van	S	C0
2	m	Sport	M	C0
3	m	Sport	M	C0
4	m	Sport	L	C0
5	m	Sport	XL	C0
6	m	Sport	XL	C0
7	f	Sport	S	C0
8	f	Sport	S	C0
9	f	Sport	M	C0
10	f	Luxus	L	C0
11	m	Van	L	C1
12	m	Van	XL	C1
13	m	Van	M	C1
14	m	Luxus	XL	C1
15	f	Luxus	S	C1
16	f	Luxus	S	C1
17	f	Luxus	M	C1
18	f	Luxus	M	C1
19	f	Luxus	M	C1
20	f	Luxus	L	C1

Table 3: Training Data for a binary classification

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

Table 4: Training Data for a binary classification

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

Table 5: Trainingsdaten für eine binäre Klassifikation

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

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

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

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

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

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

4. Consider the data set in Table 5 for a binary class problem.

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

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

(c) Figure 2 shows that entropy and the Gini index are both monotonously increasing on the range $[0, 0.5]$ and they are both monotonously decreasing on the range $[0.5, 1]$. Is it possible that information gain and the gain in the Gini index favor different attributes? Explain.

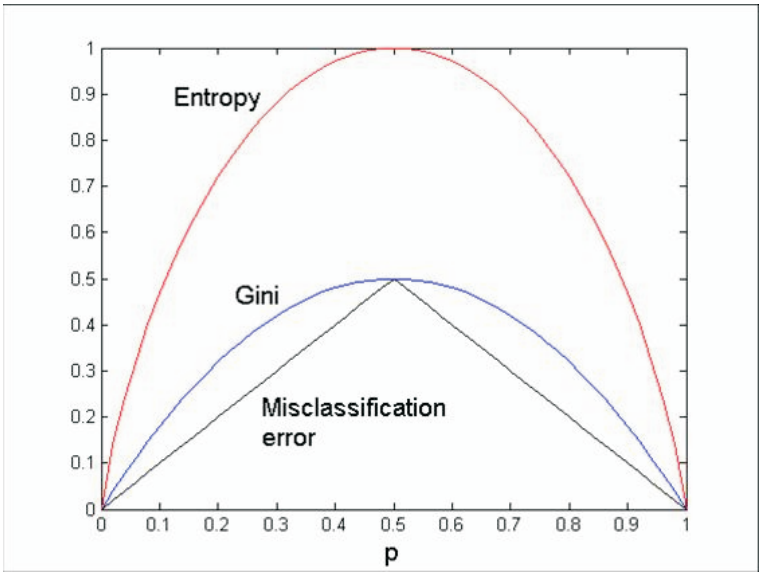


Figure 2: Gini Index, Entropy and Classification Error

X	Y	Z	Number of DO	
			in class +	in class –
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

Table 6: Training Data for a binary classification

5. Consider the set of training examples for a binary classification shown in Table 6.

   (a) Compute a decision tree using the greedy approach. Use the classification error rate as the criterion for splitting. What is the overall error rate of the induced tree?

   (b) Repeat the tree development, but use 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?

   (c) Compare and discuss the results of both sub–exercises. Comment on the suitability of the greedy heuristic used for splitting attribute selection.

6. Table 7 summarizes a data set with three attributes A, B, C and two class labels + and –. Build a two-level decision tree.

A	B	C	Number of DO in class	
			+	–
F	F	F	0	25
F	F	T	0	5
F	T	F	25	0
F	T	T	0	20
T	F	F	0	0
T	F	T	20	0
T	T	F	0	0
T	T	T	5	0

Table 7: Trainings Data for a binary classification

   (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.

   (b) Continue the tree construction with the two children of the root node.

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

   (d) Repeat parts (a), (b), and (c) using C as the first level splitting attribute. What is the classification error in the resulting decision tree?

(e) Use the results in parts (c) and (d) to conclude about the greedy nature of the decision tree induction algorithm.

7. By using the training set shown in Table 8 one developed a decision tree as shown in Figure 3.

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	−

Table 8: Training Data for a binary classification

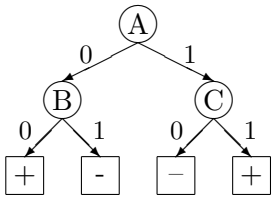


Figure 3: Decision Tree T, developed from the data in Table 8

(a) Compute the generalization error rate of the tree using the optimistic approach, i.e. with the Training Set.

(b) Compute the generalization error rate of the tree using the pessimistic approach. Use a malus $\Omega(t_i) = 0.5$ for each leaf node t_i.

(c) Compute the generalization error rate of the tree using the validation set shown 9. This approach is known as **reduced error pruning**.

lfd. Nr.	A	B	C	Klasse
11	0	0	0	+
12	0	1	1	+
13	1	1	0	+
14	1	0	1	−
15	1	0	0	+

Table 9: Validation Set for decision tree T (Figure 3)

8. *This exercise refers to decision trees on Regular Patterns, which was not part of the lecture in the Doctorate Programm on Mobile Communication, but is part of the lecture*

in the regular Data Mining course for Computer Science students. However, Ph.D. students in the Doctorate Program may be attracted to study this topic as well and solve this exercise!

For the treatment of a particular heavy disease, there are three therapies a, b, and c, which are typically applied. They have to be applied sequentially. Class P contains those therapy sequences, which ended up with a complete recovery of the patient; class N contains those therapy sequences, which did not lead to a complete recovery.

Construct a decision tree that estimates the recovery chances of therapy sequences for $P = \{bc, ac\}$ and $N = \{cb, acb\}$.

2.2 Rule Based Classification

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

   - $AirConditioner : \{working, broken\}$
   - $Engine : \{good, bad\}$
   - $Milage : \{high, medium, low\}$
   - $Rust : \{yes, no\}$

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

   $Milage = high \longrightarrow Value = low$
   $Milage = low \longrightarrow Value = high$
   $AirConditioner = working \wedge Engine = good \longrightarrow Value = high$
   $AirConditioner = working \wedge Engine = bad \longrightarrow Value = low$
   $AirConditioner = broken \longrightarrow Value = low$

   (a) Are the rules mutually exclusive?

   (b) Is the rule set exhaustive?

   (c) Is ordering needed for this set of rules?

   (d) Do you need a default class for the rule set?

2. For a top down rule growing technology, the value of a conjunct B when specifying a rule

   - $r_1 : A \longrightarrow C$

   towards

   - $r_2 : A \wedge B \longrightarrow C$

   needs to be determined.

   (a) In the RIPPER Algorithm, FOIL's information gain decides about the best conjunct to be added.

   Suppose

   - r_1 covers 350 positive and 150 negative DO and
   - r_2 covers 300 positive and 50 negative DO.

Compute FOIL's information gain for the rule r_2 with respect to r_1.

(b) A former version of RIPPER called IREP favors those rule refinements, which increase the value of

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

over a validation set most respectively stops rule growing, if no refinement increases this value. Here,

- P is the number of positive DO in the validation set,
- N is the number of negative DO in the validation set,
- p is the number of positive DO covered by the rule, and
- n is the number of negative DO covered by the rule.

Suppose a validation set containing 500 positive and 500 negative DO.

- r_1 covers 200 positive and 50 negative DO.
- r_2 covers 100 positive and 5 negative DO.

Which one of the rules would be preferred by IREP?

(c) Compute v_{RIPPER} of r_1 und r_2 for previous problem. Which one of the rules does RIPPER prefer?

3. C4.5 rules 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. Discuss the strengths and weaknesses of both methods.

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

r_1: $A \longrightarrow +$ (covers 4 positive and 1 negative DO)

r_2: $B \longrightarrow +$ (covers 30 positive and 10 negative DO)

r_3: $C \longrightarrow +$ (covers 100 positive and 90 negative DO)

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

(a) Accuracy

(b) FOIL's Information Gain

(c) likelihood ratio statistic

(d) Laplace mesure

(e) m-estimate mesure

5. Figure 4 illustrates the coverage of the classification rules r_1, r_2, and r_3 in a data set that contains 29 positive and 21 negative DO. Determine which is the best and worst rule according to:

(a) the likelihood ratio statistic

(b) the Laplace measure, and

(c) m-estimate measure ($p_+ = \frac{29}{29+21} = 0.58$)

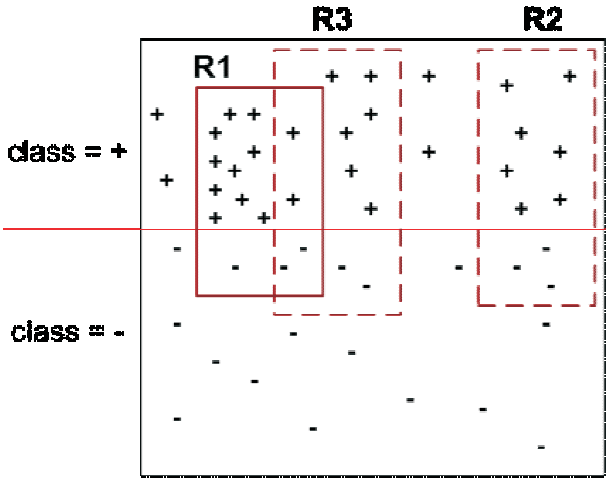


Figure 4: Data set and coverage of three rules

Determine the better one of the rules r_2 or r_3 based on the accuracy after r_1 was already included into the rule base and

(d) non of the DO covered by r_1 has been removed from the data set.

(e) only the positive DO covered by r_1 have been removed from the data set.

(f) all (positive and negative) DO covered by r_1 have been removed from the data set.

2.3 kNN Classification

1. Consider the one-dimensional data set shown in Table 10.

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

Table 10: Data Set for exercise 1

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

(b) Repeat the previous analysis using the distance-weighted voting approach

$$y' = \underset{v}{\operatorname{argmax}} \sum_{[x_i,y_i]\in kNN(x')} \frac{1}{d(x',x_i)^2} * ord(v = y_i)$$

$$\text{with } ord(prop) = \begin{cases} 1 & \text{, if } \quad prop = true \\ 0 & \text{, if } \quad prop = false \end{cases} .$$

2. The nearest-neighbor algorithm described can be extended to handle nominal attributes. A variant of the algorithm called PEBLS (Parallel Exemplar-Based Learning System) measures the distance between two values of a nominal attribute using the

modified value difference metric (MVDM). Given a pair of nominal attribute values, v_1 and v_2, the distance between them is defined as follows:

$$d(v_1, v_2) = \sum_{i=1}^{k} \left| \frac{n_{i1}}{n_1} - \frac{n_{i2}}{n_2} \right|$$

where

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

home_owner	*martial_status*	*annual_income*	*defaulted_borrower*
yes	*single*	124	*no*
no	*married*	100	*no*
no	*single*	70	*no*
yes	*married*	120	*no*
no	*divorced*	95	*yes*
no	*married*	60	*no*
yes	*divorced*	220	*no*
no	*single*	85	*yes*
no	*married*	75	*no*
no	*single*	90	*yes*

Table 11: Data Set for exercise 2

Consider the training set for the loan classification problem shown in Table 11. Use the MVDM measure to compute the distance between every pair of attribute values for the nominal attributes *home_owner* and *martial_status*.

2.4 Bayesian Classification

1. 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.

   (a) What is the probability that a student who smokes is a graduate student?

   (b) Is a randomly chosen college student more likely to be a graduate or undergraduate student?

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

   (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.

2. Consider the data set shown in Table 12.

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

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

Table 12: Data Set for exercise 2

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

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

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

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

3. Consider the data set shown in Table 13.

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

Table 13: Data Set for exercise 3

(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.

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

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

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

(e) Compare $p((A = 1) \wedge (B = 1)|+)$ against $p(\mathrm{A} = 1|+)$ and $p(B = 1|+)$. Are the variables conditionally independent given the class?

4. Figure 5 illustrates the Bayesian Belief Network for the Data Set shown in Table 14 (Assume that all the attributes are binary).

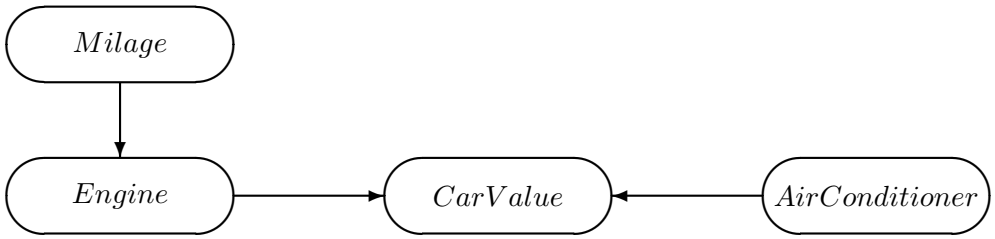


Figure 5: Bayesian Belief Network to exercise 4

Milage	*Engine*	*AirConditioner*	Number of Records with	
			$CarValue = high$	$CarValue = low$
high	*good*	*working*	3	4
high	*good*	*broken*	1	2
high	*bad*	*working*	1	5
high	*bad*	*broken*	0	4
low	*good*	*working*	9	0
low	*good*	*broken*	5	1
low	*bad*	*working*	1	2
low	*bad*	*broken*	0	2

Table 14: Data Set to exercise 4

(a) Draw the probability table for each node in the network.

(b) Use the Bayesian Network to compute $p(Engine = Bad \wedge AirConditioner = Broken)$.

5. Given the Bayesian Network shown in Figure 6, compute the following probabilities:

(a) $p(B = good \wedge F = empty \wedge G = empty \wedge S = yes)$

(b) $p(B = bad \wedge F = empty \wedge G = not\_empty \wedge S = no)$

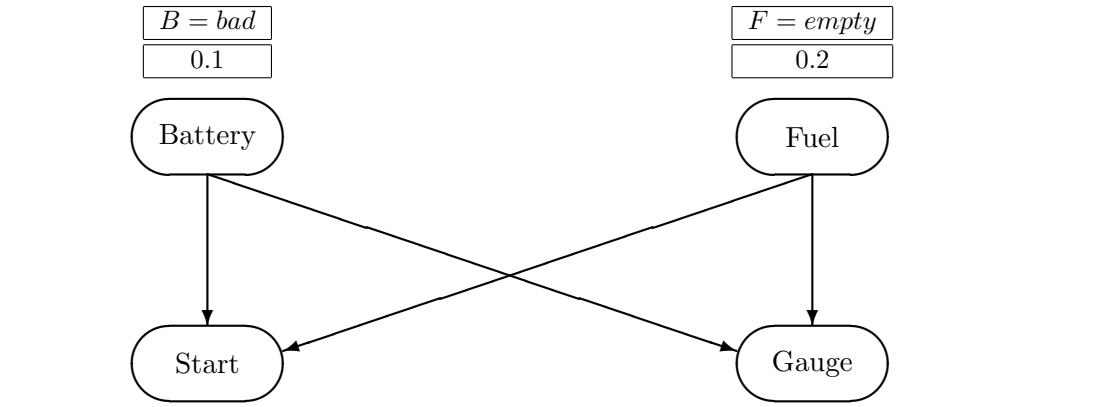
(c) Given that the battery is bad, compute the probability that the car will start.

3 Association Analysis

1. Consider the data in Table 15.

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

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



	$S = no$
$B = good \wedge F = not\_empty$	0.1
$B = good \wedge F = empty$	0.8
$B = bad \wedge F = not\_empty$	0.9
$B = bad \wedge F = empty$	1.0

	$G = empty$
$B = good \wedge F = not\_empty$	0.1
$B = good \wedge F = empty$	0.8
$B = bad \wedge F = not\_empty$	0.2
$B = bad \wedge F = empty$	0.9

Figure 6: Bayesian Belief Network to exercise 5

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

Table 15: Example of market basket transactions for exercise 1

(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.)

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

(e) Let

- s_1 and c_1 are the support and confidence values of an association rule r when treating each transaction ID as a market basket and
- 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!

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

(b) Let c_1, c_2, and c_3 be the confidence values of the rules

- $r_1 : \{p\} \rightarrow \{q\}$
- $r_2 : \{p\} \rightarrow \{q, r\}$
- $r_3 : \{p, r\} \rightarrow \{q\}$

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?

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

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

3. Consider the market basket transactions shown in Table 16.

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

Table 16: Market basket transactions for exercise 3

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

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

(c) What is the maximum number of size-3 itemsets that can be derived from this data set? Write a general expression for the maximum number of k-size itemsets drived from n ($n \geq k$) different items.

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

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

4. Consider the following set of frequent 3-itemsets, which have been developed out of the frequent items $F_1 = \{a, b, c, d, e, f\}$:

$F_3 = \{\{a, b, c\}, \{a, b, d\}, \{a, b, e\}, \{a, c, d\}, \{a, c, e\}, \{b, c, d\}, \{b, c, e\}, \{c, d, e\}\}$.

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

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

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

5. The *a priori* algorithm uses a generate-and-count strategy for deriving frequent itemsets. Candidate itemsets of size k are created by joining a pair of frequent itemsets of size $k-1$ (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.

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

Table 17: Example of market basket transactions for exercise 5

Suppose the *a priori* algorithm is applied to the data set shown in Table 17 with $minsup = 0.3$, i.e., any itemset occurring in less than 3 transactions is considered to be infrequent.

(a) Draw an itemset lattice representing the relation $\subseteq$ in up to 4-Intemsets $\{a, b, c, d, e\}^*$ of $\{a, b, c, d, e\}$!
Label (or color) each node in the lattice with the following letter(s):

N: If the itemset is not considered to be a candidate itemset by the *a priori* algorithm.
There are two reasons for an itemset not to be considered as a candidate itemset:

i. it is not generated at all during the candidate generation step, or

ii. 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.

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

(c) What is the pruning ratio of the *a priori* 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.

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

6. The *a priori* 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 7.

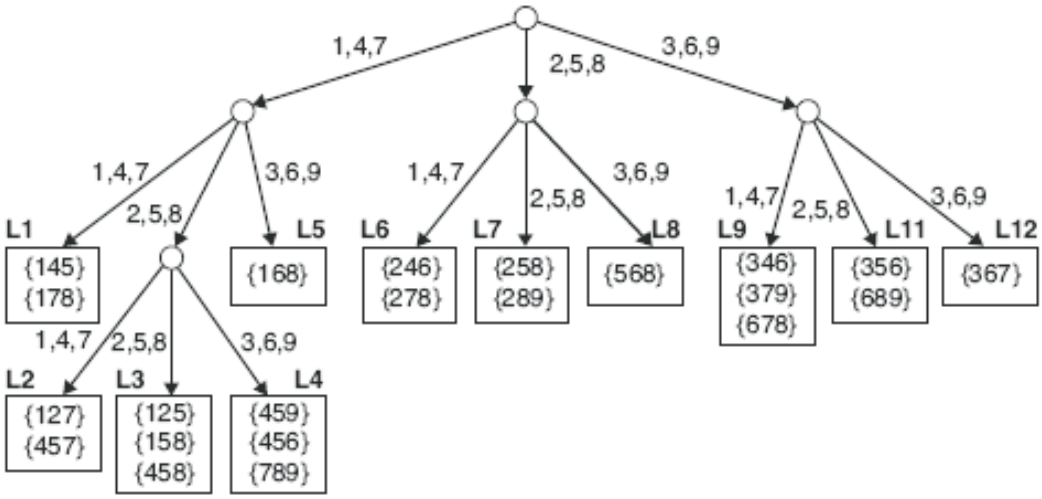


Figure 7: Hash Tree structure for exercise 6

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

(b) Use the visited leaf nodes in part (a) to determine the candidate itemsets that are contained in the transaction $t = \{1, 3, 4, 5, 8\}$. Which of the candidate itemsets in these visited leaf nodes are in t and thus, increase the support?

7. 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 binary 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.

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

Table 18: Transaction data for exercise 8

- **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.

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

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

8. Given the lattice structure shown in Figure 8 and the transactions given in Table 18, 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,
   - F if it is frequent but neither maximal nor closed, and
   - I if it is infrequent.

   Assume that the support threshold is $minsup = 0.3$.

9. Consider the transactions given in table 19.

   (a) Draw the lattice of the rules that can be formed from the frequent itemset $\{a, c, e, g\}$! Draw edges between the nodes only, it their origins represent the rules, which are merged for building the edges's destinations when using a redundancy free technique!

   (b) Label the nodes with the confidence of the related rule!

   (c) Let $minconf = 0.75$ be the minimum confidence. Label each node

   - with C (confident), if this rule is built when merging two rules of the upper level with a redundancy free technique and will be subject of further merging for the next level rules.

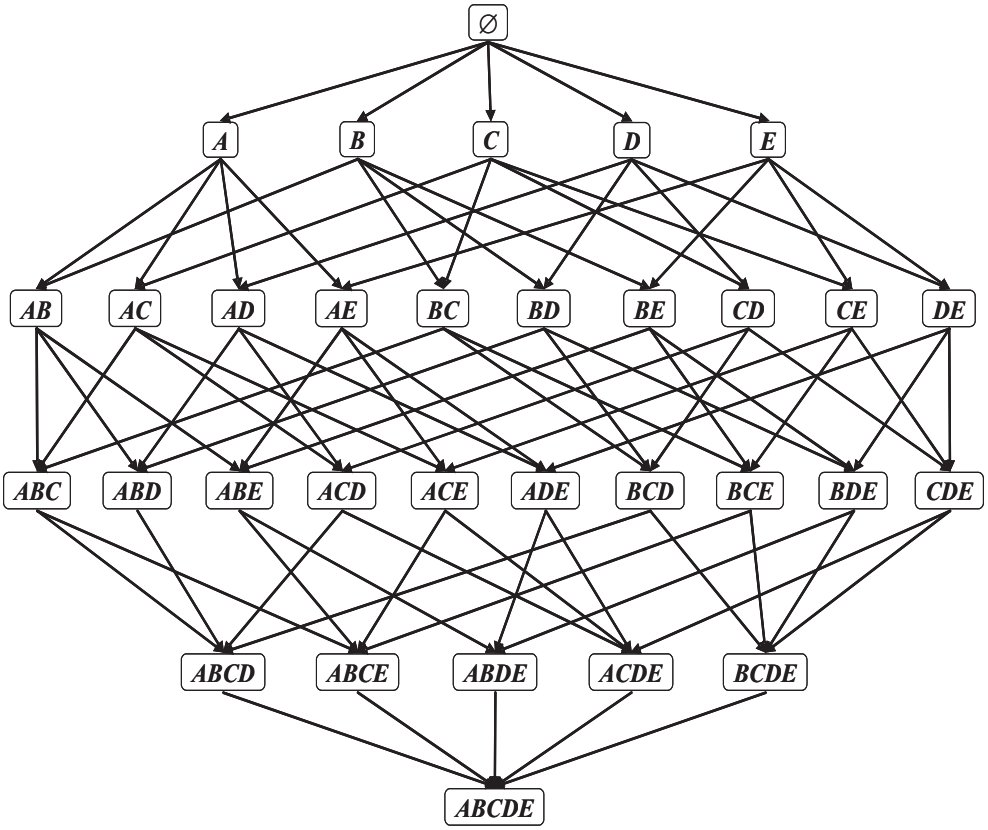


Figure 8: Itemset lattice for exercise 8

- with I (inconfident), if this rule is built when merging two rules of the upper level with a redundancy free technique, but will not be subject of further merging for the next level rules because of insufficient confidence.
- with N (not formed), if this rule is not built when merging two rules of the upper level with a redundancy free technique.

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

    (a) Draw a contingency table for each of the following rules using the transactions shown in Table 20:

        i. $b \rightarrow c$
        ii. $a \rightarrow d$
        iii. $b \rightarrow d$
        iv. $e \rightarrow c$
        v. $c \rightarrow a$

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

        i. Support
        ii. Confidence

TID	items
1	$\{a, b, c, e, g\}$
2	$\{a, b, e, g\}$
3	$\{a, c, e, f, g\}$
4	$\{c, d, e, g\}$
5	$\{a, c, d, e, g\}$
6	$\{c, e, f, g\}$
7	$\{a, b, c, d, e, g\}$
8	$\{a, b, c, e, f\}$
9	$\{a, c, d, e, f, g\}$
10	$\{a, b, c, f, g\}$
11	$\{a, c, d, e, g\}$
12	$\{a, b, e, f\}$
13	$\{a, b, c, e, f, g\}$
14	$\{c, d, f, g\}$
15	$\{a, c, d, e, f, g\}$
16	$\{b, c, d, e, f\}$
17	$\{a, b, c, e, g\}$
18	$\{b, d, e, g\}$
19	$\{a, c, d, e, g\}$
20	$\{b, c, e, f\}$

Table 19: Transaction Data for exercise 9

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

Table 20: Transaction data for exercise 10

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

Table 21: Contingency Table for exercise 12

buy HDTV	buy EM		
	yes	no	
yes	99	81	180
no	54	66	120
	153	147	300

Table 22: A two-way contingency table between the sale of HDTV and EM for exercise 13

iii. Interest Factor

iv. Correlation (ϕ- Coefficient)

v. Interest Support (IS)

vi. Odds Ratio (α - Coefficient)

(c) Given the rankings you had obtained (b), compute the correlation between the rankings of confidence and the other five measures.

Which measure is most highly correlated with confidence?

11. Suppose we have market basket data consisting of 100 transactions and 20 items. Assume the support for item a is 0.25, the support for item b is 0.9 and the support for itemset $\{a, b\}$ is 0.2. Let the support and confidence thresholds be 0.1 and 0.6, respectively.

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

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

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

12. Table 21 shows a $2 \times 2 \times 2$ contingency table for the binary variables A and B at different values of the control variable C.

(a) Compute the φ coefficient for A and B when $C = 0$, $C = 1$, and $C = 0$ *or* 1.

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

13. Consider the relationship between customers who buy high-definition televisions (HDTV) and exercise machines (EM) as shown in Tables 22 and 23.

(a) Compute the odds ratios (α coefficients) for both tables.

(b) Compute the correlations (φ-coefficients) for both tables.

costumer group	buy HDTV	buy EM		total
		yes	no	
college students	yes	1	9	10
	no	4	30	34
working adult	yes	98	72	170
	no	50	36	86

Table 23: A three-way contingency table for exercise 13

(c) Compute the Lift measure for both tables.

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

4 Cluster Analysis

1. Suppose that for a data set

   - there are m points and K clusters,
   - half the points and clusters are in "more dense" regions,
   - half the points and clusters are in "less dense" regions, and
   - the two regions are well-separated from each other.

   For the given data set, which of the following should occur in order to minimize the squared error when finding K clusters:

   (a) Centroids should be equally distributed between more dense and less dense regions.
   (b) More centroids should be allocated to the less dense region.
   (c) More centroids should be allocated to the denser region.

2. Consider the mean of a cluster of objects from a binary transaction data set.

   (a) What are the minimum and maximum values of the components of the mean?
   (b) What is the interpretation of components of the cluster mean?
   (c) Which components most accurately characterize the objects in the cluster?

3. 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.

4. Total SSE is the sum of the SSE for each separate attribute.

   (a) What does it mean if the SSE for one variable is low for all clusters?
   (b) What does it mean if the SSE for one variable is low for just one cluster?
   (c) What does it mean if the SSE for one variable is high for all clusters?
   (d) What does it mean if the SSE for one variable is high for just one cluster?
   (e) How could you use the per variable SSE information to improve your clustering?

5. The leader algorithm of HARTIGAN 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.[5]

   What are the advantages and disadvantages of the leader algorithm as compared to K-means?

6. 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 9)

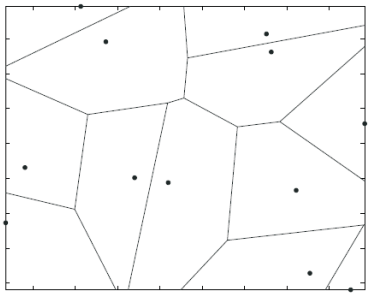


Figure 9: Voronoi Diagram for Exercise 6

   (a) What is the relationship between Voronoi diagrams and K-means clusters?

   (b) What do Voronoi diagrams tell us about the possible shapes of K-means clusters?

[5]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.