The Right Stuff: Appropriate Mathematics for All Students

Promoting the use of materials that engage students in meaningful activities that promote the effective use of technology to support mathematics, further equip students with stronger problem solving and critical thinking skills, and enhance numeracy.

Overview

Students compute support and confidence for item set by applying the concepts of

- Numeracy—Students will determine conditional probability through methods of Association Analysis.
- Quantitative Literacy – Students will explore discrete data.
- Connections to Other Disciplines—Students will explore data mining techniques using mathematics.

Supplies and Materials

- Student Worksheet 4.1

Prerequisite Knowledge

Students must be able to reduce fractions, convert fractions to decimals, and decimals to percents.

Instructional Suggestions

1. Ask students to bring in a grocery receipt that lists the items purchased. This could be used in addition to or instead of the list used with items #6 - #14.

Assessment Ideas

The student will demonstrate numeracy, quantitative literacy and connections to other disciplines by computing support and confidence for the item set found in # 6 - #14 as a cumulative assessment activity. As an extension, the student may be given an edited item list based on personal grocery receipts and asked to make additional suggestions for the retail store.

This material is based upon work supported by the National Science Foundation under Grant No. DUE 0632883
Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Introduction

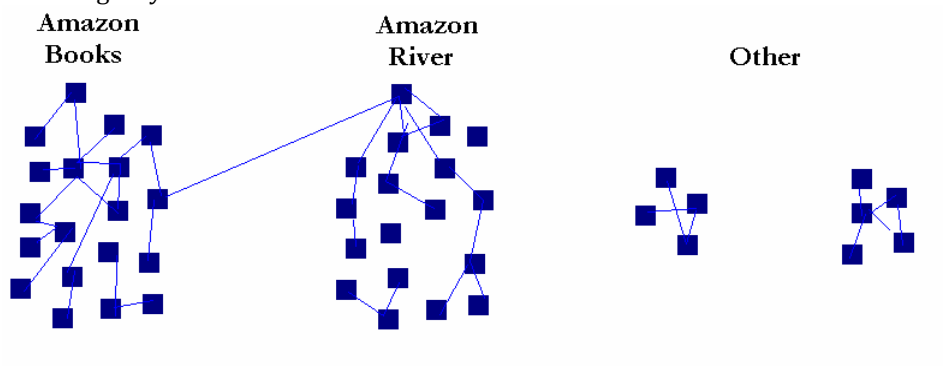
Businesses have found that if they collect a lot of data they need that data to work for them.

This is opposite of the traditional statistical process of posing a question and then performing an experiment to find the answer.

Businesses and the government collect data and mine it to find relationships and patterns in order to make decisions about marketing, inventory, product placement, etc. Government may make decisions about who or what to examine more closely because of the data and trends they find.

A simple example of data mining is the classification of a new web page. How does a search engine classify a new web page? A search engine classifies a new web page by evaluating the text that is on the page and classifying that page with other pages that have similar text. For example, suppose we have a classification: "AMAZON."

The web pages that are classified as "AMAZON" are linked by user routing in the following way:



What "trends" in the words that are found on these pages would lead a search engine to classify these pages correctly?

In other words, what words would most likely strongly influence the classification?

Data mining, in this case, is the search for words that would lead to a correct classification. The mathematics comes in when we want to quantify the degree to which the algorithm produces the correct classification with the least amount of time, work, and money.

This material is based upon work supported by the National Science Foundation under Grant No. DUE 0632883
Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Random sampling is often said to be the fairest way to find an unbiased sample in order to make an accurate prediction about the population.

However, in order to classify a web page correctly, randomly selecting words from that page may not provide the most useful information.

Consider the random selection of words taken from a well-known song (top right).

Then consider the random selection of larger "chunks" (bottom right).

Which collection of words provides you with the best information on which to make a decision?

The missing words are hidden in a white font.

The song is *Yesterday* by the Beatles.

Most people will recognize the complete lines rather than the random selection of words.

				seemed				
Now								stay
	I							
	I'm				man			
		shadow						
oh,								
Why					I			she
	said						for	
					easy		to	
			a				away	
oh,				yesterday				
Why					I		know,	
			wrong,			long		
		was				game		
Now					to			
			in					

Now	it	looks	as	though	they're	here	to	stay
oh,	yesterday	came	suddenly					
Now	I	need	a	place	to	hide	away	
oh,	I	believe	in	yesterday,	Mm			

This material is based upon work supported by the National Science Foundation under Grant No. DUE 0632883
Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Data from web pages provides search engines with information on how to classify the pages. Data about you is also being collected in order to classify you!

1. In what ways are data being collected on you?

Data is gathered through credit card usage, EZ Pass, computer usage, Internet cookies, cell phone usage, frequent buyer cards, …

The classic example of data mining comes from the grocery store – the frequent buyer cards provide the store with all kinds of information.

The question is, with all the data that is accessible to the grocery chain, how do they determine which trends occur most often and then how do they use that information?

Association Analysis uses a set of transactions to discover rules that indicate the likely occurrence of an item based on the occurrences of other items in the transaction.

The implication symbol, →, means co-occurrence, not causality.

An "**Item Set**" is a list of items purchased by a customer during one transaction.

The "**Support Count**" is the frequency of occurrence of an item set.

Example: The support count for the item set { Diapers, Milk } is 2 since customers, 3 and 5, bought diapers and milk.

1. Find the support count for each item set.
   a. { Bread, Milk, OJ }
   b. { Eggs, Milk }
   c. { Bread, Eggs, Milk }

Customer	Items
1	Bread, Milk, OJ
2	Bread, Diapers, Eggs, OJ
3	Coke, Diapers, Milk, OJ
4	Bread, Coke, Diapers, OJ
5	Bread, Coke, Diapers, Milk
6	Bread, Eggs, Milk
7	Bread, Coke, OJ
8	Bread, Eggs, Milk, OJ
9	Eggs, Milk, Napkins
10	Bread, Eggs, Milk

Item Set #1

a. { Bread, Milk, OJ } = 2

b. { Eggs, Milk } = 4

c. { Bread, Eggs, Milk } = 3

The "**Support**" for an item set X is the fraction of all transactions that contain an item set. The support for a set X is denoted by supp(X).

Example: The support for the item set { Diapers, Milk } is supp({Diapers, Milk}) = 2/10 = 1/5 since 2 of the total 10 transaction contain diapers and milk.

2. Find the support for each item set.
   a. supp({ Bread, Milk, OJ })
   b. supp({ Eggs, Milk })
   c. supp({ Bread, Eggs, Milk })

a. supp({ Bread, Milk, OJ }) = 2/10 = 1/5

b. supp({ Eggs, Milk }) = 4/10 = 2/5

c. supp({ Bread, Eggs, Milk }) = 3/10

This material is based upon work supported by the National Science Foundation under Grant No. DUE 0632883
Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

The **"Confidence for an Association Rule"** measures how often items in Y appear in transactions that contain X and is defined as support for all the items in X or Y divided by the support for the items in X. Confidence is denoted c ({ X } → { Y }).

Example: c ({Eggs, Milk} → {Bread}) =

$$\frac{\text{supp } \{\text{Bread, Eggs, Milk}\}}{\text{supp } \{\text{Eggs, Milk}\}} = \frac{3/10}{4/10} = \frac{3}{4}$$

3. Find the confidence for each association rule.
   a. c ({ Bread, Milk } → { Eggs })
   b. c ({ Bread, Milk } → { Diapers })
   c. c ({ Bread, Diapers } → { Coke })

Confidence provides an estimate of the conditional probability of Y given X. Probabilities are often represented as percents. Convert each confidence to a percentage.

a. c ({ Bread, Milk } → { Eggs })
= (3/10)/(5/10) = 3/5 = 60%

b. c ({ Bread, Milk } → { Diapers })
= (1/10)/(5/10) = 1/5 = 20%

c. c ({ Bread, Diapers } → { Coke })
= (2/10)/(3/10) = 2/3 = 67%

This material is based upon work supported by the National Science Foundation under Grant No. DUE 0632883
Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Consider the transactions shown here. Find the value of the support and confidence in #6 – 14.

Customer ID	Transaction ID	Items Bought
1	0001	{a, b, c}
2	0024	{a, b, c, e}
2	0038	{a, b, e}
3	0012	{a, b, d, e}
3	0044	{a, c, e}
3	0088	{b, d, e}
4	0015	{a, b, c}
4	0055	{a, b, c, d, e}
5	0022	{a, b, c}
6	0035	{a, c, e}
6	0042	{a, d, e}
6	0080	{b, c, d, e}

Item Set #2

Find:

6. s(a) = **10/12** = **.83**
7. s(e) = **9/12** = **.75**
8. s(b, d) = **4/12** = **.33**
9. s(b, c) = **6/12** = **.50**
10. s(a, b, d) = **2/12** = **.17**
11. s(b, c, e) = **3/12** = **.25**
12. c({b, d} → {a}) = $^{2/12}/_{4/12}$ = **.50**
13. c({a} → {b, d}) = $^{2/12}/_{10/12}$ = **.20**
14. c({b, e} → {c}) = $^{3/12}/_{6/12}$ = **.50**
15. c({e} → {b, c}) = $^{3/12}/_{9/12}$ = **.33**
16. c({b, e} → {c}) = $^{3/12}/_{6/12}$ = **.50**
17. c({a, b} → {e}) = $^{4/12}/_{7/12}$ = **.57**
18. c({a, b} → {c}) = $^{5/12}/_{7/12}$ = **.71**

Which association rule has the highest confidence? What might that mean to the chain?

The highest confidence is that of #18. Therefore, the chain might place some items with a large profit margin between the two sets of items in order to make the customer walk by them on their way from items a and b to item c.

This material is based upon work supported by the National Science Foundation under Grant No. DUE 0632883
Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Many of the calculations made in the previous set of exercises can be done with the aid of a spreadsheet.

The picture (below) is taken from 4.3 Excel.

Examine how the calculations are made in the G – J columns and then in rows 20, 21, and 23.

19. How is column G calculated?

By multiplying the number in the column C (b) by the number in column E (d).

Colulmn J

By multiplying the numbers in the columns C (b), D (c) and F (e).

20. How is c ({b, d} --> {a}) (cell B20) calculated?

By dividing the number in cell I18 by the number in cell G18.

3										
4	Customer	Items Bought								
5	ID	a	b	c	d	e	b & d	b & c	a&b&d	b&c&e
6	1	1	1	1			0	1	0	0
7	2	1	1	1		1	0	1	0	1
8	2	1	1			1	0	0	0	0
9	3	1	1		1	1	1	0	1	0
10	3	1		1		1	0	0	0	0
11	3		1		1	1	1	0	0	0
12	4	1	1	1			0	1	0	0
13	4	1	1	1	1	1	1	1	1	1
14	5	1	1	1			0	1	0	0
15	6	1		1		1	0	0	0	0
16	6	1			1	1	0	0	0	0
17	6		1	1	1	1	1	1	0	1
18	Support for each item	0.83	0.75	0.67	0.42	0.75	0.33	0.50	0.17	0.25
19										
20	c ({b, d} --> {a}) =	0.5								
21	c({a} --> {b, d}) =	0.2								
22	c({b, e} --> {c}) =									
23	c({e} --> {b, c}) =	0.33								
24	c({b, e} --> {c}) =									
25	c({a, b} --> {e}) =									
26	c({a, b} --> {c}) =									

Add additional columns to the matrix in the spreadsheet and make the computations to verify the calculations done in #12-18.

This material is based upon work supported by the National Science Foundation under Grant No. DUE 0632883
Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.