Statistics 202: Statistical Aspects of Data Mining

Professor David Mease

Tuesday, Thursday 9:00-10:15 AM Terman 156

Lecture 9 = Review for midterm exam

Agenda:

1) Reminder about midterm exam (July 26)
2) Review Simpson's Paradox
3) Go over homework solutions
4) A few sample midterm questions

Announcement – Midterm Exam:

The midterm exam will be Thursday, July 26

The best thing will be to take it in the classroom (9:00-10:15 AM in Terman 156)

For remote students who absolutely can not come to the classroom that day please email me to confirm arrangements with SCPD

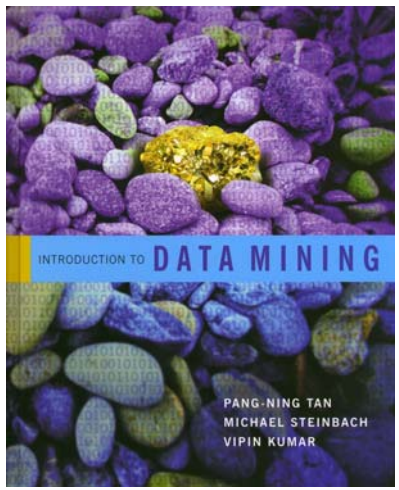
You are allowed one 8.5 x 11 inch sheet (front and back) containing notes

No books or computers are allowed, but please bring a hand held calculator

The exam will cover the material that we covered in class from Chapters 1,2,3 and 6

Introduction to Data Mining

by
Tan, Steinbach, Kumar



Chapter 6: Association Analysis

Simpson's "Paradox" (page 384)

- Occurs when a 3rd (possibly hidden) variable causes the observed relationship between a pair of variables to disappear or reverse directions

- Example: My friend and I play a basketball game and each shoot 20 shots. Who is the better shooter?

	me
make	10
miss	10
total	20

	my friend
make	8
miss	12
total	20

Simpson's "Paradox" (page 384)

- Occurs when a 3rd (possibly hidden) variable causes the observed relationship between a pair of variables to disappear or reverse directions

- Example: My friend and I play a basketball game and each shoot 20 shots. Who is the better shooter?

	me
make	10
miss	10
total	20

	my friend
make	8
miss	12
total	20

- But, who is the better shooter if you *control for* the distance of the shot? Who would you rather have on your team?

	me		
	far	close	total
make	1	9	10
miss	3	7	10
total	4	16	20

	my friend		
	far	close	total
make	5	3	8
miss	10	2	12
total	15	5	20

Another example of Simpson's "Paradox"

- A search engine labels web pages as good and bad. A researcher is interested in studying the relationship between the duration of time a user spends on the web page (long/short) and the good/bad attribute.

	good
long	10
short	10
total	20

	bad
long	8
short	12
total	20

Another example of Simpson's "Paradox"

- A search engine labels web pages as good and bad. A researcher is interested in studying the relationship between the duration of time a user spends on the web page (long/short) and the good/bad attribute.

	good
long	10
short	10
total	20

	bad
long	8
short	12
total	20

- It is possible that this relationship reverses direction when you *control for* the type of query (adult/non-adult). Which relationship is more relevant?

	good		
	adult	non-adult	total
long	1	9	10
short	3	7	10
total	4	16	20

	bad		
	adult	non-adult	total
long	5	3	8
short	10	2	12
total	15	5	20

Yet another example of Simpson's "Paradox"

- Height and reading ability are strongly correlated in grade schools. Why?

Homework Solutions

- As of 9AM Tuesday, July 24, solutions to all three homework assignments will be posted at

http://www.stats202.com/solutions.html

- Review these for the exam

- Note that even if you had a prefect score, you may still have missed some parts, so check your answers against these solutions carefully

Sample Midterm Question #1:

What is the definition of data mining used in your textbook?

A) the process of automatically discovering useful information in large data repositories

B) the computer-assisted process of digging through and analyzing enormous sets of data and then extracting the meaning of the data

C) an analytic process designed to explore data in search of consistent patterns and/or systematic relationships between variables, and then to validate the findings by applying the detected patterns to new subsets of data

Sample Midterm Question #2:

If height is measured as short, medium or tall then it is what kind of attribute?

A) Nominal

B) Ordinal

C) Interval

D) Ratio

Sample Midterm Question #3:

If my data frame in R is called "data", which of the following will give me the third column?

A) data[2,]

B) data[3,]

C) data[,2]

D) data[,3]

E) data(2,)

F) data(3,)

G) data(,2)

H) data(,3)

Sample Midterm Question #4:

Compute the confidence for the association rule {b, d} → {a} by treating each row as a market basket. Also, state what this value means in plain English.

Items Bought
$\{a, d, e\}$
$\{a, b, c, e\}$
$\{a, b, d, e\}$
$\{a, c, d, e\}$
$\{b, c, e\}$
$\{b, d, e\}$
$\{c, d\}$
$\{a, b, c\}$
$\{a, d, e\}$
$\{a, b, e\}$

Sample Midterm Question #5:

If a data set is space delimited, what should be done to allow a text string that includes a space so that R or Excel will not split the string into 2 columns?

A) Escape it
B) Remove the space
C) Use all capitals in the string
D) Select “Fix the spaces” from the menu bar

Sample Midterm Question #6:

Compute the standard deviation for the numbers 23, 25, 30. Show your work below.