Statistics 202: Statistical Aspects of Data Mining

Professor David Mease

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

Lecture 8 = Finish chapter 6

Agenda:

1) Reminder about midterm exam (July 26)
2) Reminder about homework (due 9AM Tues)
3) Lecture over rest of Chapter 6 (sections 6.1and 6.7)
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) for 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

Announcement – Midterm Exam:

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

(see http://scpd.stanford.edu/scpd/enrollInfo/policy/proctors/monitor.asp)

I have heard from:

Catrina
Jack C
Steven V
Jeff N
Trent P
Duyen N
Jason E

If you are not one of these people, I will assume you will take the exam in the classroom unless you contact me and tell me otherwise

Homework Assignment:

Chapter 3 Homework Part 2 and Chapter 6 Homework is due 9AM Tuesday 7/24

Either email to me (dmease@stanford.edu), bring it to class, or put it under my office door.

SCPD students may use email or fax or mail.

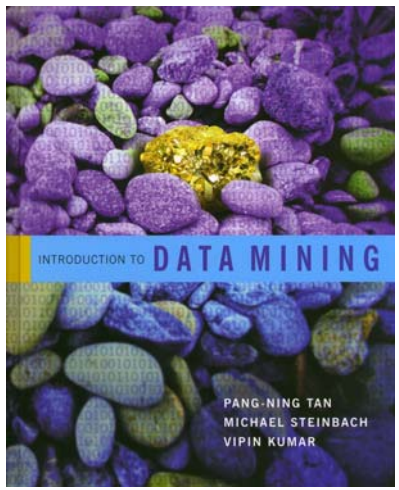
The assignment is posted at

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

Important: If using email, please submit only a single file (word or pdf) with your name and chapters in the file name. Also, include your name on the first page.

Introduction to Data Mining

by
Tan, Steinbach, Kumar



Chapter 6: Association Analysis

What is Association Analysis:

- 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

TID	*Items*
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

- Examples:

{Diaper} $\rightarrow$ {Beer},
{Milk, Bread} $\rightarrow$ {Eggs,Coke}
{Beer, Bread} $\rightarrow$ {Milk}

- Implication means co-occurrence, not causality!

Definitions:

TID	Items
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

- **Itemset**
  - A collection of one or more items
  - Example: {Milk, Bread, Diaper}
  - k-itemset = An itemset that contains k items
- **Support count (σ)**
  - Frequency of occurrence of an itemset
  - E.g. σ({Milk, Bread,Diaper}) = 2
- **Support**
  - Fraction of transactions that contain an itemset
  - E.g. s({Milk, Bread, Diaper}) = 2/5
- **Frequent Itemset**
  - An itemset whose support is greater than or equal to a *minsup* threshold

Another Definition:

- **Association Rule**
  - An implication expression of the form X → Y, where X and Y are itemsets
  - Example:
    {Milk, Diaper} → {Beer}

TID	*Items*
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

Even More Definitions:

- **Association Rule Evaluation Metrics**
  - **Support (s)**
    - **=Fraction of transactions that contain both X and Y**
  - **Confidence (c)**
    - **=Measures how often items in Y appear in transactions that contain X**

- **Example:**

$$\{\text{Milk}, \text{Diaper}\} \Rightarrow \text{Beer}$$

$$s = \frac{\sigma(\text{Milk}, \text{Diaper}, \text{Beer})}{|\text{T}|} = \frac{2}{5} = 0.4$$

$$c = \frac{\sigma(\text{Milk}, \text{Diaper}, \text{Beer})}{\sigma(\text{Milk}, \text{Diaper})} = \frac{2}{3} = 0.67$$

TID	*Items*
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

In class exercise #26:

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

Table 6.1. Example of market basket transactions.

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

In class exercise #27:

Use the results in the previous problem to compute the confidence for the association rules {b, d} → {a} and {a} → {b, d}. State what these values mean in plain English.

Table 6.1. Example of market basket transactions.

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

In class exercise #28:

Compute the support for itemsets {a}, {b, d}, and {a,b,d} by treating each customer ID as a market basket.

Table 6.1. Example of market basket transactions.

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

In class exercise #29:

Use the results in the previous problem to compute the confidence for the association rules {b, d} → {a} and {a} → {b, d}. State what these values mean in plain English.

Table 6.1. Example of market basket transactions.

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

In class exercise #30:

The data www.stats202.com/more_stats202_logs.txt contains access logs from May 7, 2007 to July 1, 2007. Treating each row as a "market basket" find the support and confidence for the rule

Mozilla/5.0 (compatible; Yahoo! Slurp; http://help.yahoo.com/help/us/ysearch/slurp)→ 74.6.19.105

An Association Rule Mining Task:

- Given a set of transactions T, find all rules having both
  - support ≥ minsup threshold
  - confidence ≥ minconf threshold

- Brute-force approach:
  - List all possible association rules
  - Compute the support and confidence for each rule
  - Prune rules that fail the minsup and minconf thresholds
  - Problem: this is computationally prohibitive!

The Support and Confidence Requirements can be Decoupled

TID	Items
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

{Milk,Diaper} → {Beer} (s=0.4, c=0.67)
{Milk,Beer} → {Diaper} (s=0.4, c=1.0)
{Diaper,Beer} → {Milk} (s=0.4, c=0.67)
{Beer} → {Milk,Diaper} (s=0.4, c=0.67)
{Diaper} → {Milk,Beer} (s=0.4, c=0.5)
{Milk} → {Diaper,Beer} (s=0.4, c=0.5)

- All the above rules are binary partitions of the same itemset: {Milk, Diaper, Beer}
- Rules originating from the same itemset have identical support but can have different confidence
- Thus, we may decouple the support and confidence requirements

Two Step Approach:

1) Frequent Itemset Generation

= Generate all itemsets whose support ≥ minsup

2) Rule Generation

= Generate high confidence (confidence ≥ minconf) rules from each frequent itemset, where each rule is a binary partitioning of a frequent itemset

- Note: Frequent itemset generation is still computationally expensive and your book discusses algorithms that can be used

In class exercise #31:

Use the two step approach to generate all rules having support ≥ .4 and confidence ≥ .6 for the transactions below.

Table 6.2. Market basket transactions.

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

Drawback of Confidence

	Coffee	$\overline{\text{Coffee}}$	
Tea	15	5	20
$\overline{\text{Tea}}$	75	5	80
	90	10	100

Association Rule: Tea → Coffee

Confidence(Tea → Coffee) = P(Coffee|Tea) = 0.75

Drawback of Confidence

	Coffee	$\overline{\text{Coffee}}$	
Tea	15	5	20
$\overline{\text{Tea}}$	75	5	80
	90	10	100

Association Rule: Tea → Coffee

Confidence(Tea → Coffee) = P(Coffee|Tea) = 0.75

but support(Coffee) = P(Coffee) = 0.9

Although confidence is high, rule is misleading

confidence($\overline{\text{Tea}}$ → Coffee) = P(Coffee|$\overline{\text{Tea}}$) = 0.9375

Other Proposed Metrics:

#	Measure	Formula
1	ϕ-coefficient	$\frac{P(A,B)-P(A)P(B)}{\sqrt{P(A)P(B)(1-P(A))(1-P(B))}}$
2	Goodman-Kruskal's (λ)	$\frac{\sum_j \max_k P(A_j,B_k)+\sum_k \max_j P(A_j,B_k)-\max_j P(A_j)-max_k P(B_k)}{2-\max_j P(A_j)-\max_k P(B_k)}$
3	Odds ratio (α)	$\frac{P(A,B)P(\overline{A},\overline{B})}{P(A,\overline{B})P(\overline{A},B)}$
4	Yule's Q	$\frac{P(A,B)P(\overline{AB})-P(A,\overline{B})P(\overline{A},B)}{P(A,B)P(\overline{AB})+P(A,\overline{B})P(\overline{A},B)} = \frac{\alpha-1}{\alpha+1}$
5	Yule's Y	$\frac{\sqrt{P(A,B)P(\overline{AB})}-\sqrt{P(A,\overline{B})P(\overline{A},B)}}{\sqrt{P(A,B)P(\overline{AB})}+\sqrt{P(A,\overline{B})P(\overline{A},B)}} = \frac{\sqrt{\alpha}-1}{\sqrt{\alpha}+1}$
6	Kappa (κ)	$\frac{P(A,B)+P(\overline{A},\overline{B})-P(A)P(B)-P(\overline{A})P(\overline{B})}{1-P(A)P(B)-P(\overline{A})P(\overline{B})}$
7	Mutual Information (M)	$\frac{\sum_i \sum_j P(A_i,B_j)\log\frac{P(A_i,B_j)}{P(A_i)P(B_j)}}{\min(-\sum_i P(A_i)\log P(A_i), -\sum_j P(B_j)\log P(B_j))}$
8	J-Measure (J)	$\max\left(P(A,B)\log(\frac{P(B\|A)}{P(B)})+P(A\overline{B})\log(\frac{P(\overline{B}\|A)}{P(\overline{B})}),\right.$ $\left.P(A,B)\log(\frac{P(A\|B)}{P(A)})+P(\overline{A}B)\log(\frac{P(\overline{A}\|B)}{P(\overline{A})})\right)$
9	Gini index (G)	$\max\Big(P(A)[P(B\|A)^2+P(\overline{B}\|A)^2]+P(\overline{A})[P(B\|\overline{A})^2+P(\overline{B}\|\overline{A})^2]$ $-P(B)^2-P(\overline{B})^2,$ $P(B)[P(A\|B)^2+P(\overline{A}\|B)^2]+P(\overline{B})[P(A\|\overline{B})^2+P(\overline{A}\|\overline{B})^2]$ $-P(A)^2-P(\overline{A})^2\Big)$
10	Support (s)	$P(A,B)$
11	Confidence (c)	$\max(P(B\|A),P(A\|B))$
12	Laplace (L)	$\max\left(\frac{NP(A,B)+1}{NP(A)+2},\frac{NP(A,B)+1}{NP(B)+2}\right)$
13	Conviction (V)	$\max\left(\frac{P(A)P(\overline{B})}{P(A\overline{B})},\frac{P(B)P(\overline{A})}{P(B\overline{A})}\right)$
14	Interest (I)	$\frac{P(A,B)}{P(A)P(B)}$
15	cosine (IS)	$\frac{P(A,B)}{\sqrt{P(A)P(B)}}$
16	Piatetsky-Shapiro's (PS)	$P(A,B)-P(A)P(B)$
17	Certainty factor (F)	$\max\left(\frac{P(B\|A)-P(B)}{1-P(B)},\frac{P(A\|B)-P(A)}{1-P(A)}\right)$
18	Added Value (AV)	$\max(P(B\|A)-P(B),P(A\|B)-P(A))$
19	Collective strength (S)	$\frac{P(A,B)+P(\overline{AB})}{P(A)P(B)+P(\overline{A})P(\overline{B})}\times\frac{1-P(A)P(B)-P(\overline{A})P(\overline{B})}{1-P(A,B)-P(\overline{AB})}$
20	Jaccard (ζ)	$\frac{P(A,B)}{P(A)+P(B)-P(A,B)}$
21	Klosgen (K)	$\sqrt{P(A,B)}\max(P(B\|A)-P(B),P(A\|B)-P(A))$

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

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