Data Mining (W4240 Section 001)
A Priori

Giovanni Motta

Columbia University, Department of Statistics

December 7, 2015

Outline

Outline

Shopping Histories

Let's look at some data sets. What questions can these answer?

Order Date	Item	Price
August 31, 2012	Girl Genius Omnibus Volume One: Agatha Awakens	$8.97
August 25, 2012	Ito En Oi Ocha Japanese Green Tea, 16.9-Ounce Bottles (12 Pack)	$19.92
August 24, 2012	R in a Nutshell: A Quick Desktop Reference	$33.90
August 21, 2012	Creative HS-800 Fatal1ty Gaming Headset	$3225
August 21, 2012	Ensign Peak Everyday Duffel Bag, Gray	$10.59
August 16, 2012	Lucky Peach Issue 4	$7.79
July 25, 2012	Ito En Oi Ocha Japanese Green Tea, 16.9-Ounce Bottles (12 Pack)	$19.92
July 25, 2012	Feliway – Refil, 48 ml	$19.68
July 23, 2012	DRINKWELL Original Filters 12 pack	$16.95

User Ratings

VeggieTales: Lions, Shepherds and Queens	★☆☆☆☆ Not Interested
Bio-Dome	★☆☆☆☆ Not Interested
Revenge of the Pink Panther	★☆☆☆☆ Not Interested
Joe Versus the Volcano	★☆☆☆☆ Not Interested
Sky Captain and the World of Tomorrow	★☆☆☆☆ Not Interested
Roman Holiday	★☆☆☆☆ Not Interested
Porky's	★★☆☆☆ Not Interested
Signs	★★☆☆☆ Not Interested
Men in Black II	☆☆☆☆☆ Not Interested

Document Collections

Images




listverse.com



REGRETS
Those were the droids you were looking for.



I CAN HAS
CHEEZBURGER?


Supervised vs. Unsupervised Learning

So far, we have focused (mostly) on *supervised learning*

- data have labels (category, response value)
- use training data to predict new labels

We also considered unsupervised learning:

- want to learn a feature of the data that is not labeled
- examples: clustering, document topics, associations, low dimensional structure

Supervised vs. Unsupervised Learning

Data	**Supervised**	**Unsupervised**
Documents	Ham or spam?	What is it about?
Pictures	Does it include a cat?	Which pictures are similar?
Real-valued	What is the response?	Is there a smaller representation?
Categorical	Which response category?	Association rules?
Tweets	About Obama?	Is it trending?
Others?	???	???

Unsupervised Learning

We have studied 2 areas in unsupervised learning:

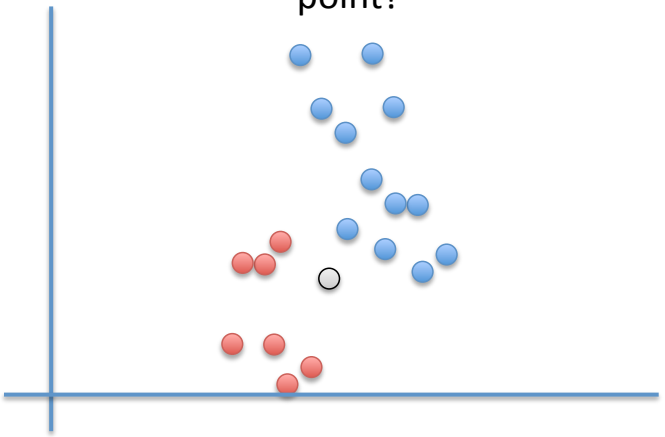
- clustering
- dimension reduction

Today we add a 3rd:

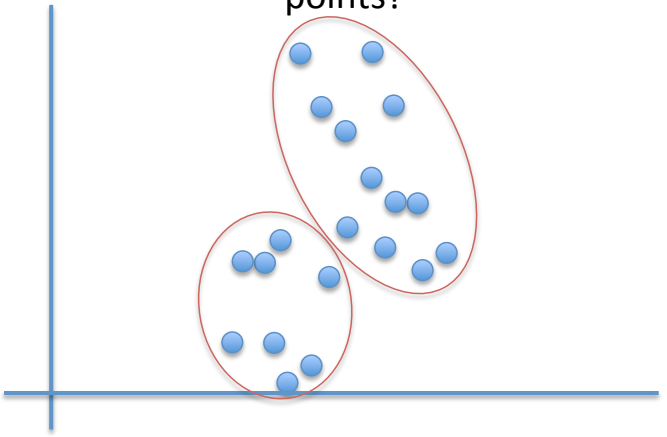
- association rules

Clustering

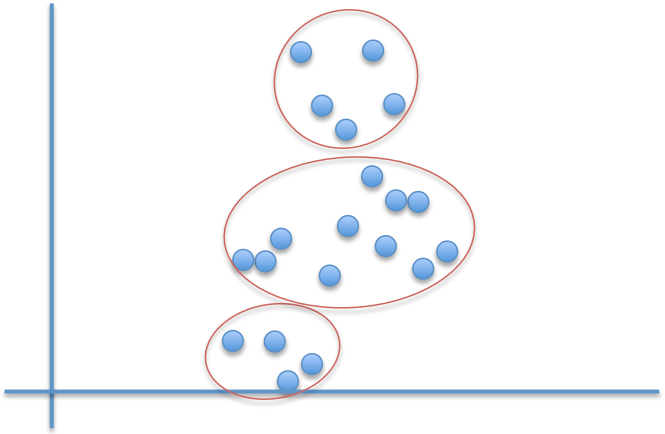
Classification: what is label of new point?



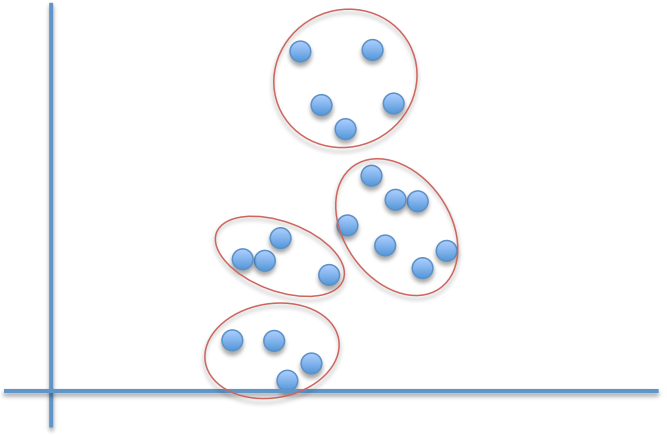
Clustering: how should we group these points?



Clustering: or is this the right grouping?

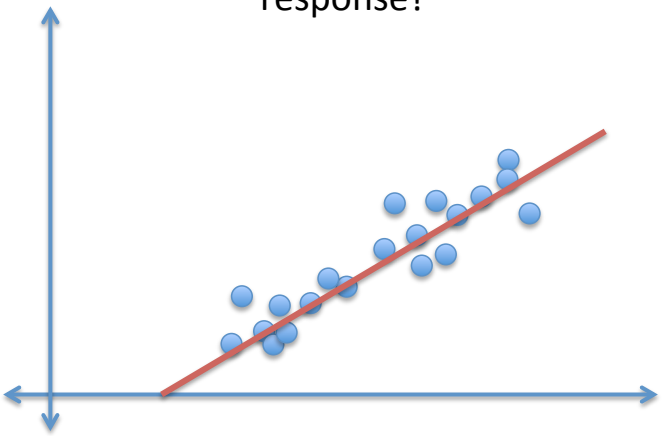


Clustering: what about this?

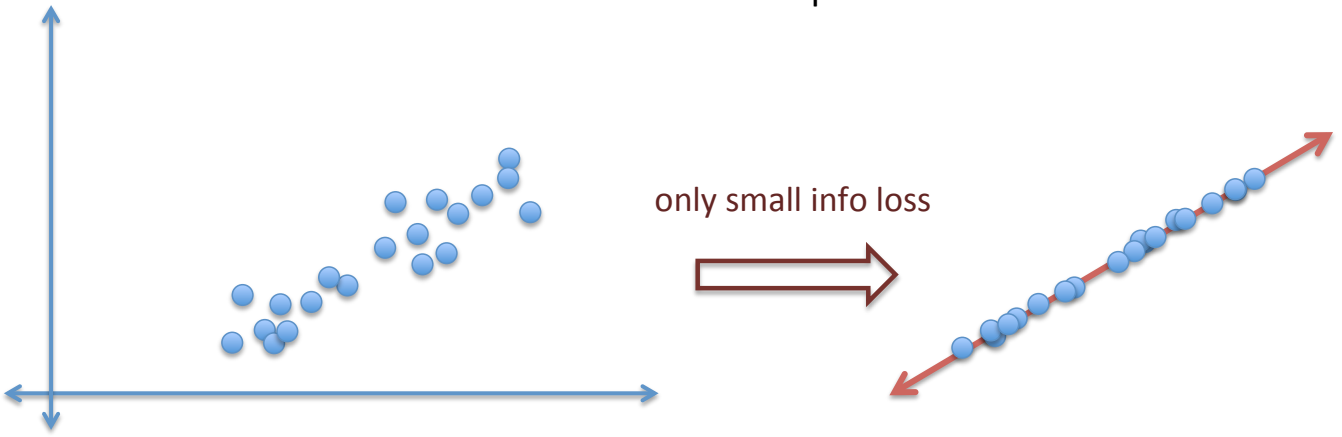


Dimension Reduction

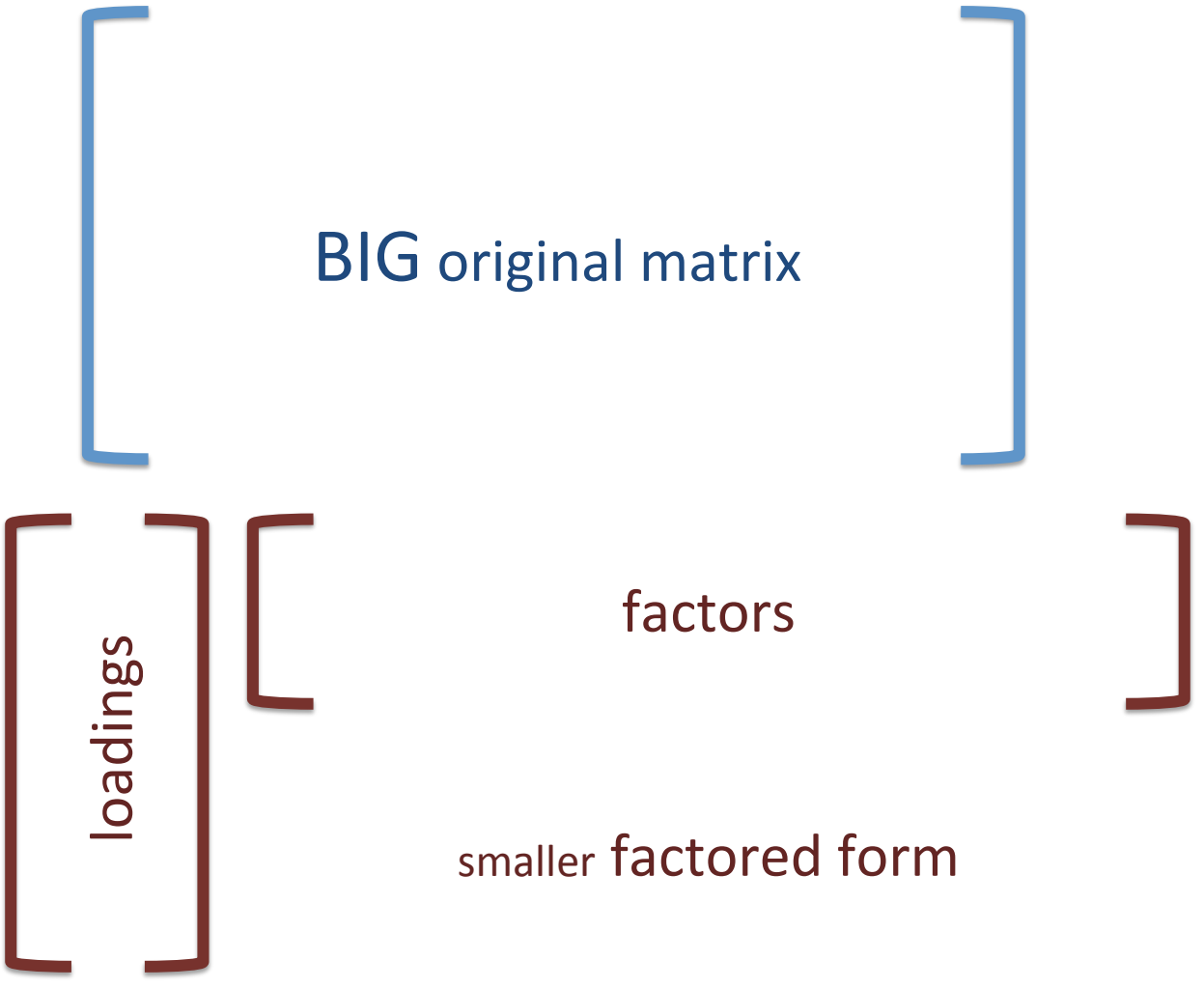
Regression: given covariates, what is response?



Dimension Reduction: can we find a lower dimensional data representation?



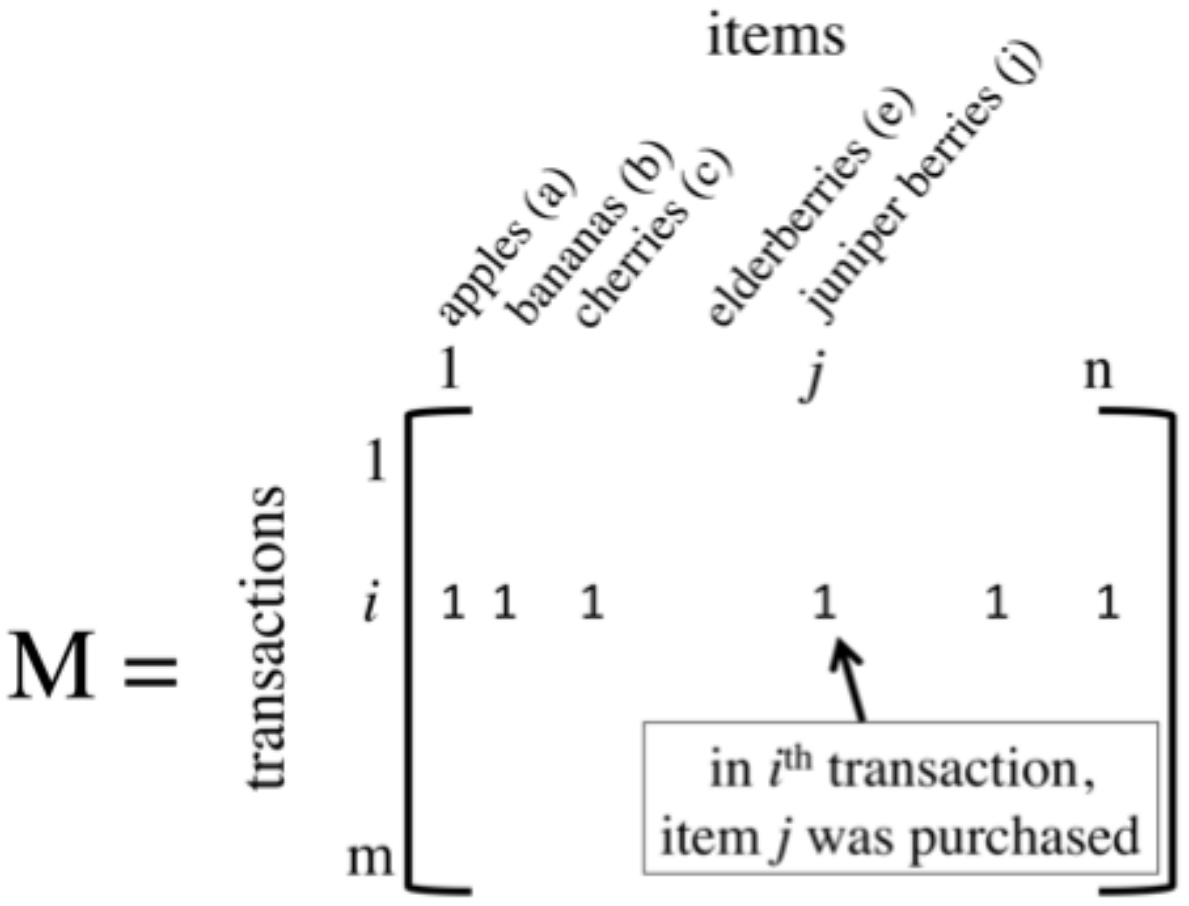
Dimension Reduction



Outline

Association Rules

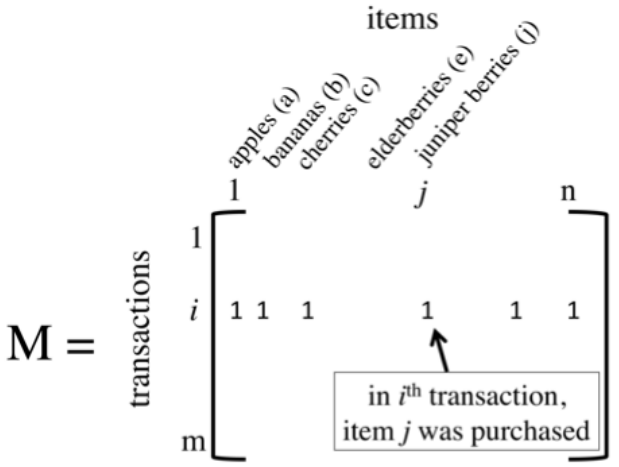
Have a basket of items:



Important terminology for today: *itemset* = a set of items

Association Rules

Have a basket of items:



Key data mining question: If I have itemset a, will I also have itemset b, denoted $a \rightarrow b$?

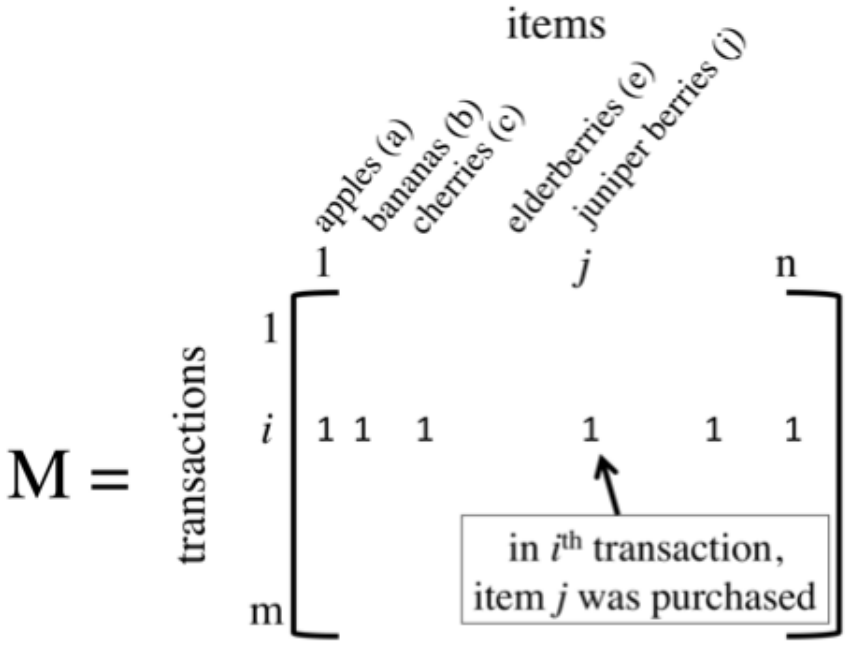
- itemsets are collections of items, indexed by $\{2, 3, 5\}$
- can create itemsets through unions of other itemsets: $\{1, 3\} = \{1\} \cup \{3\}$

Important question for:

- recommender systems
- medical diagnostics
- store layouts
- marketing

Association Rules

Have a basket of items:



So what makes a good rule?

Outline

Association Rules

We are going to use probability based rules:

$$P(b \mid a) = p$$

Use rule *support* and *confidence*:

- *Support* of an itemset: the number of transactions containing an itemset, $\mathrm{Supp}(a)$
- *Confidence* of a rule:

$$\mathrm{Conf}(a \to b) = \frac{\mathrm{Supp}(a \text{ and } b)}{\mathrm{Supp}(a)} = \frac{\#\text{ times } a \text{ and } b \text{ are purchased}}{\#\text{ times } a \text{ is purchased}}$$

$$= \hat{P}(b \mid a)$$

Association Rules

So what makes a good rule?

- *It should be common*: set a minimum level of support
  观测到的a∩b的数量要大于这个临界值

$$\mathrm{Supp}(a \cup b) \geq \theta$$

- *It should be right*: set a minimum level for probability

$$\mathrm{Conf}(a \to b) \geq \mathrm{minconf}$$

These conditions lead to **strong rules**.

Outline

A Priori

We will use the A Priori algorithm to find sets and make strong rules

- A Priori finds all frequent itemsets and their support

$\{1\}$	5
$\{2\}$	6
$\{1,2\}$	4
...	...

- Use the list of frequent itemsets from A Priori to get strong rules

$\{1\} \rightarrow \{2\}$	$4/5$
$\{2\} \rightarrow \{1\}$	$4/6$
...	...

P(“2”|“1”)=#{1,2}/#{1}=4/5

A Priori

A Priori finds frequent itemsets through a method similar to tree construction

- search over all single items; retain in list if support is at least θ
- do until no further itemsets can be made:
  - *Create candidate itemsets:* for each pair of itemsets in list for itemsets with k items, combine if they share $k-1$ items
  - *Prune*: retain if candidate has support at least θ to make list of itemsets with $k+1$ items
  - stop if list for $k+1$ items is empty

Note: this is how we build itemsets. Then:

- Derive candidate rules according to $\hat{P}(b \mid a)$
- Prune all candidate rules with $\hat{P}(b \mid a) < \texttt{minconf}$

A Priori

Example: Start with transaction records

Records
$1, 3, 4$
$2, 3, 4$
$1, 2, 3, 5$
$1, 4, 5$
$1, 2, 4$
$2, 3, 4, 5$
$2, 4, 5$
$2, 3, 4$

Parameters:

- Minimum support: $\theta = 3$
- Minimum confidence: $\text{minconf} = .75$

A Priori

Example: make a list of 1-item sets along with counts (candidate list)

Table: Transactions

Records
$1, 3, 4$
$2, 3, 4$
$1, 2, 3, 5$
$1, 4, 5$
$1, 2, 4$
$2, 3, 4, 5$
$2, 4, 5$
$2, 3, 4$

Itemset	**Support**
$\{1\}$	4
$\{2\}$	6
$\{3\}$	5
$\{4\}$	7
$\{5\}$	4

A Priori

Example: remove from the candidate list those with support less than θ

Table: Itemsets for $k = 1$

Itemset	**Support**
$\{\mathbf{1}\}$	**4**
$\{\mathbf{2}\}$	**6**
$\{\mathbf{3}\}$	**5**
$\{\mathbf{4}\}$	**7**
$\{\mathbf{5}\}$	**4**

A Priori

Example: now generate candidates for 2 item sets

Table: Itemsets for $k = 1$

Itemset	**Support**
$\{\mathbf{1}\}$	**4**
$\{\mathbf{2}\}$	**6**
$\{\mathbf{3}\}$	**5**
$\{\mathbf{4}\}$	**7**
$\{\mathbf{5}\}$	**4**

Itemset	**Support**
$\{1, 2\}$	2
$\{1, 3\}$	2
$\{1, 4\}$	3
$\{1, 5\}$	2
$\{2, 3\}$	4
$\{2, 4\}$	5
$\{2, 5\}$	3
$\{3, 4\}$	4
$\{3, 5\}$	2
$\{4, 5\}$	3

A Priori

Example: remove from the candidate list those with support less than θ

Table: Itemsets for $k = 2$

Itemset	Support
$\{1, 4\}$	3
$\{2, 3\}$	4
$\{2, 4\}$	5
$\{2, 5\}$	3
$\{3, 4\}$	4
$\{4, 5\}$	3

A Priori

Example: now generate candidates for 3 item sets

Table: Itemsets for $k = 2$

Itemset	**Support**
$\{\mathbf{1,4}\}$	**3**
$\{\mathbf{2,3}\}$	**4**
$\{\mathbf{2,4}\}$	**5**
$\{\mathbf{2,5}\}$	**3**
$\{\mathbf{3,4}\}$	**4**
$\{\mathbf{4,5}\}$	**3**

Itemset	**Support**
$\{2,3,4\}$	3
$\{2,4,5\}$	2

- can't combine $\{1,4\}$ and $\{2,4\}$ because $\mathrm{Supp}(\{1,2\}) < \theta$
- same with $\{1,4\}$ and $\{3,4\}$, $\{1,4\}$ and $\{4,5\}$, $\{3,4\}$ and $\{4,5\}$

A Priori

Example: remove from the candidate list those with support less than θ

Table: Itemsets for $k = 3$

Itemset	**Support**
$\{\mathbf{2}, \mathbf{3}, \mathbf{4}\}$	**3**

A Priori

Example: full itemset list generated by A Priori

Table: Itemsets

Itemset	Support
$\{1\}$	4
$\{2\}$	6
$\{3\}$	5
$\{4\}$	7
$\{5\}$	4
$\{1,4\}$	3
$\{2,3\}$	4
$\{2,4\}$	5
$\{2,5\}$	3
$\{3,4\}$	4
$\{4,5\}$	3
$\{2,3,4\}$	3

A Priori: Make Rules

Example: A Priori only makes a list of itemsets along with their support. We use this to candidate rules

Table: Itemsets

Itemset	Support
$\{1\}$	4
$\{2\}$	6
$\{3\}$	5
$\{4\}$	7
$\{5\}$	4
$\{1,4\}$	3
$\{2,3\}$	4
$\{2,4\}$	5
$\{2,5\}$	3
$\{3,4\}$	4
$\{4,5\}$	3
$\{2,3,4\}$	3

$\#\{1,4\}/\#\{1\}=3/4$

Table: Candidate Rules

Rule	Confidence	Rule	Confidence
$1 \to 4$	$3/4$	$2 \to 3$	$4/6$
$2 \to 4$	$5/6$	$2 \to 5$	$3/6$
$3 \to 2$	$4/5$	$3 \to 4$	$4/5$
$4 \to 1$	$3/7$	$4 \to 2$	$5/7$
$4 \to 3$	$4/7$	$4 \to 5$	$3/7$
$5 \to 2$	$3/4$	$5 \to 4$	$3/4$
$2,3 \to 4$	$3/4$	$2,4 \to 3$	$3/5$
$3,4 \to 2$	$3/4$	$2 \to 3,4$	$3/6$
$3 \to 2,4$	$3/5$	$4 \to 2,3$	$3/7$

A Priori: Make Rules

Example: use minimum confidence to prune rules, keeping rules with $\text{minconf} \geq 0.75$

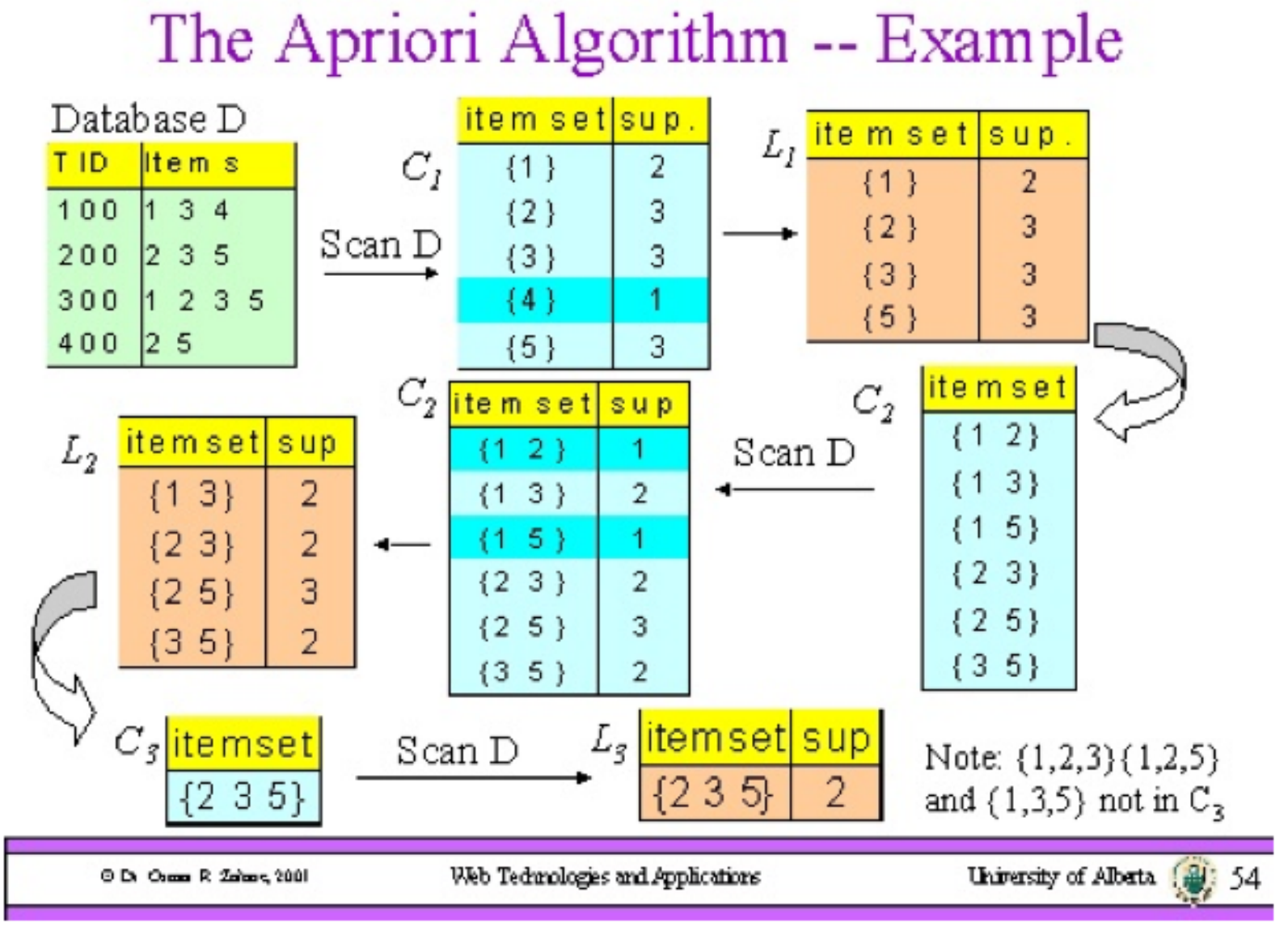
Table: Transactions

Records
$1, 3, 4$
$2, 3, 4$
$1, 2, 3, 5$
$1, 4, 5$
$1, 2, 4$
$2, 3, 4, 5$
$2, 4, 5$
$2, 3, 4$

Table: Mined Rules

Rule	**Confidence**
$1 \rightarrow 4$	$3/4$
$2 \rightarrow 4$	$5/6$
$3 \rightarrow 2$	$4/5$
$3 \rightarrow 4$	$4/5$
$5 \rightarrow 2$	$3/4$
$5 \rightarrow 4$	$3/4$
$2, 3 \rightarrow 4$	$3/4$
$3, 4 \rightarrow 2$	$3/4$

A Priori


The Apriori Algorithm -- Example
Database D
TID | Items
100 | 1 3 4
200 | 2 3 5
300 | 1 2 3 5
400 | 2 5
Scan D
C_1
itemset | sup.
{1} | 2
{2} | 3
{3} | 3
{4} | 1
{5} | 3
L_1
itemset | sup.
{1} | 2
{2} | 3
{3} | 3
{5} | 3
C_2
itemset
{1 2}
{1 3}
{1 5}
{2 3}
{2 5}
{3 5}
Scan D
C_2
itemset | sup
{1 2} | 1
{1 3} | 2
{1 5} | 1
{2 3} | 2
{2 5} | 3
{3 5} | 2
L_2
itemset | sup
{1 3} | 2
{2 3} | 2
{2 5} | 3
{3 5} | 2
C_3
itemset
{2 3 5}
Scan D
L_3
itemset | sup
{2 3 5} | 2
Note: {1,2,3}{1,2,5} and {1,3,5} not in C_3
© Dr. Osmar R. Zaïane, 2001
Web Technologies and Applications
University of Alberta
54

A Priori

How do we choose rules?

- order by *confidence*

$$\mathrm{Conf}(a \to b) = \hat{P}(b \mid a) = \frac{\mathrm{Supp}(a \cup b)}{\mathrm{Supp}(b)}$$

- order by *lift*

Lift 越大越好

$$\mathrm{Lift}(a \to b) = \frac{\hat{P}(b \mid a)}{\hat{P}(b)} = \frac{\mathrm{Supp}(b)}{1 - \frac{\mathrm{Supp}(a \cup b)}{\mathrm{Supp}(a)}}$$

Lift (from Wikipedia)

- Lift is a measure of the performance of a targeting model (association rule) at predicting or classifying cases as having an enhanced response (with respect to the population as a whole), measured against a random choice targeting model.
- A targeting model is doing a good job if the response within the target is better than the average for the population as a whole. `(#{j|i1,i2,i3}/#{i1,i2,i3})/(#{j}/#{totoal})`

$$\text{Lift} = \frac{\text{target response}}{\text{average response}}$$

- Example: a population has an average response rate of 5%, but a certain model (or rule) has identified a segment with a response rate of 20%. Then that segment would have a lift of $4.0 = \frac{0.2}{0.05}$.

A Priori

Comments:

- Association rules are a conceptual class of algorithms (like clustering)
- A Priori is one of the most popular Association rule algorithms
- (note: A Priori can be seen as a breadth-first search algorithm)

A Priori

Comments:

- Association rules are a conceptual class of algorithms (like clustering)
- A Priori is one of the most popular Association rule algorithms
- (note: A Priori can be seen as a breadth-first search algorithm)

Questions:

- what types of questions does this answer? (Classification? Regression? Groupings?)

Supervised

- will it tell me the list of items that imply a?
- what if we have continuous fields?
- how many computations does it need to do?
- how many rules can it generate?

Outline

A Priori

To implement this in R, use the package arules:

```
> library(arules)
> data("Adult")
> dim(Adult)
[1] 48842   115
> inspect(Adult[1:10,1:10])
> rules <- apriori(Adult,
+             parameter = list(supp = 0.5, conf = 0.9,
+                                 target = "rules"))
> inspect(rules)
```

A Priori

Summary:

- mine association rules from item basket data
- A Priori uses a tree construction like algorithm to create itemsets with large support
- then use the itemsets and support to create rules
- A Priori can produce a huge number of itemsets—bad for making rules

Final-esque question

Table: Transactions

Records
$1, 3$
2
$1, 2, 3$
3
$1, 2, 3$
2
2
3

Use a priori to make a set of rules with support $\geq 30\%$ and confidence ≥ 0.6.