

Data Mining (W4240 Section 001)

A Priori

Giovanni Motta

Columbia University, Department of Statistics

December 7, 2015

Outline

Conceptual Review and Setup

Association Rules

Probabilistic Association Rules

The A Priori Algorithm

A Priori in R

Outline

Conceptual Review and Setup

Association Rules

Probabilistic Association Rules

The A Priori Algorithm














































A Priori in R

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	<div><div></div><div>Not Interested</div></div>
Bio-Dome	<div><div></div><div>Not Interested</div></div>
Revenge of the Pink Panther	<div><div></div><div>Not Interested</div></div>
Joe Versus the Volcano	<div><div></div><div>Not Interested</div></div>
Sky Captain and the World of Tomorrow	<div><div></div><div>Not Interested</div></div>
Roman Holiday	<div><div></div><div>Not Interested</div></div>
Porky's	<div><div></div><div>Not Interested</div></div>
Signs	<div><div></div><div>Not Interested</div></div>
Men in Black II	<div><div></div><div>Not Interested</div></div>

Document Collections



Homer, *Odyssey*

Agamemnon

Search

("Agamemnon", "Hom. Od. 9.1", "denarius")

[\[advanced search\]](#) [\[view abbreviations\]](#)

Your current position in the text is marked in blue. Click anywhere in the line to jump to another position.

[Hide browse bar](#)

book:
card:

This text is part of:

[Greek and Roman Materials](#)
[Greek Hexameter](#)
[Greek Poetry](#)
[Greek Texts](#)
[Homer](#)
[Homer, Odyssey](#)

View text chunked by:

book : line

Table of Contents:

▼ [book 1](#)
[lines 1-43](#)
[lines 44-79](#)
[lines 80-124](#)
[lines 125-177](#)
[lines 178-229](#)
[lines 230-279](#)
[lines 280-324](#)
[lines 325-364](#)
[lines 365-420](#)
[lines 421ff.](#)
▶ [book 2](#)

Hom. Od. 1.1

Click on a word to bring up parses, dictionary entries, and frequency statistics

ἄνδρα μοι ἔννεπε, μοῦσα, πολύτροπον, ὃς μάλα πολλὰ
πλάγχθη, ἐπεὶ Τροίης ἱερὸν πτολίεθρον ἔπερσεν:
πολλῶν δ' ἀνθρώπων ἴδεν ἄστεα καὶ νόον ἔγνω,
πολλὰ δ' ὃ γ' ἐν πόντῳ πάθεν ἄλγεα ὃν κατὰ θυμόν,
ἀρνύμενος ἥν τε ψυχὴν καὶ νόστον ἐταίρων.
ἀλλ' οὐδ' ὧς ἐτάρους ἐρρύσατο, ἰέμενός περ:
αὐτῶν γὰρ σφετέρῃσιν ἀτασθαλίῃσιν ὄλοντο,
νήπιοι, οἳ κατὰ βοῦς Ὑπερίονος Ἥελίοιο
ῥισιον: αὐτὰρ ὁ τοῖσιν ἀφείλετο νόστιμον ἦμαρ.
τῶν ἀμόθεν γε, θεά, θύγατερ Διός, εἰπέ καὶ ἡμῖν.

ἐνθ' ἄλλοι μὲν πάντες, ὅσοι φύγον αἰπὺν δλεθρον,
οἴκοι ἔσαν, πόλεμόν τε πεφευγότες ἡδὲ θάλασσαν:
τὸν δ' οἷον νόστου κεκρημένον ἡδὲ γυναικὸς
νύμφη πτόντ' ἔρυκε Καλυψὶ δῖα θεάων
ἐν στήθεσι γλαφυροῖσι, λιλαιομένη πόσιν εἶναι.
ἀλλ' ὅτε δὴ ἔτος ἦλθε περιπλομένω ἐνιαυτῶν,
τῷ οἱ ἐπεκλώσαντο θεοὶ οἰκόνδε νέεσθαι
εἰς Ἰθάκην, οὐδ' ἔνθα πεφυγμένος ἦεν ἀέθλων
καὶ μετὰ οἷσι φίλοισι. θεοὶ δ' ἐλέαινον ἅπαντες
νόσφι Ποσειδάωνος; ὃ δ' ἄσπερχές μενείωνεν
ἀντιθέψ' Ὀδυσσεὶ πάρος ἦν γαῖαν ἰκέσθαι.

English (Samuel Butler)

[focus](#) [load](#)

English (1919)

[focus](#) [load](#)

Notes (W. Walter Merry, James Riddell, D. B. Monro, 1886)

[focus](#) [show](#)

References (24 total)

[hide](#)

- 5 Cross-references to this page (10):
 - Aristotle, *Rhetoric*, [Aristot., Rh. 3.14](#)
 - Sulpicia, *Carmina Omnia*, [1](#)
 - Thomas W. Allen, E. E. Sikes, *Commentary on the Homeric Hymns*, [BIBLIOGRAPHY](#)
 - W. Walter Merry, James Riddell, D. B. Monro, *Commentary on the Odyssey* (1886), [1.100](#)
 - W. Walter Merry, James Riddell, D. B. Monro, *Commentary on the Odyssey* (1886), [1.128](#)
 - Walter Leaf, *Commentary on the Iliad* (1900), [1.1](#)
 - Walter Leaf, *Commentary on the Iliad* (1900), [2.464](#)
 - Basil Lanneau Gildersleeve, *Syntax of Classical Greek*, [3](#)
 - Basil Lanneau Gildersleeve, *Syntax of Classical Greek*, [3](#)
 - Thomas D. Seymour, *Commentary on Homer's Iliad, Books I-III*, [1.3](#)
- 10 Cross-references in notes to this page (1):
 - Polybius, *Historiae*, [An Historian Needs Practical Experience](#)
- 15 Cross-references in general dictionaries to this page (10):
 - LSI, [ἐννεπε](#)
 - LSI, [ἐντροπον](#)
 - LSI, [εἶπν](#)
 - LSI, [ἐγνω](#)
 - LSI, [ἐταῖρον](#)
 - LSI, [ἐπέρσεν](#)
 - LSI, [ἐπεί](#)
 - LSI, [ἐπεί](#)
 - LSI, [ἐπεί](#)
 - LSI, [ἐπεί](#)
- 20 Cross-references in text-specific dictionaries to this page (3):
 - Autenrieth, [Μολον](#)

Images



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

Unsupervised Learning

We have studied 2 areas in unsupervised learning:

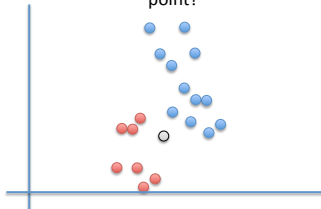
- ▶ clustering
- ▶ dimension reduction

Today we add a 3rd:

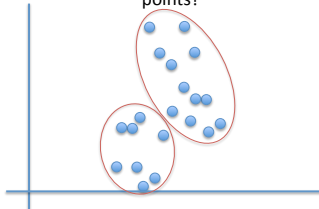
- ▶ association rules

Clustering

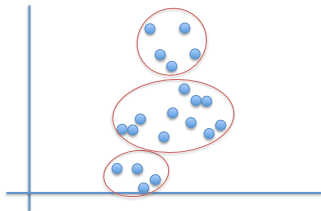
Classification: what is label of new point?



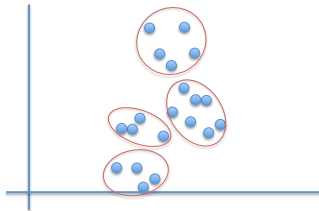
Clustering: how should we group these points?



Clustering: or is this the right grouping?

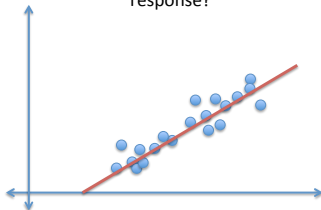


Clustering: what about this?

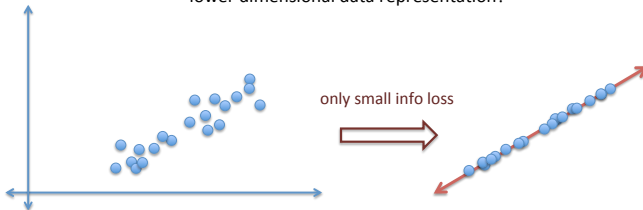


Dimension Reduction

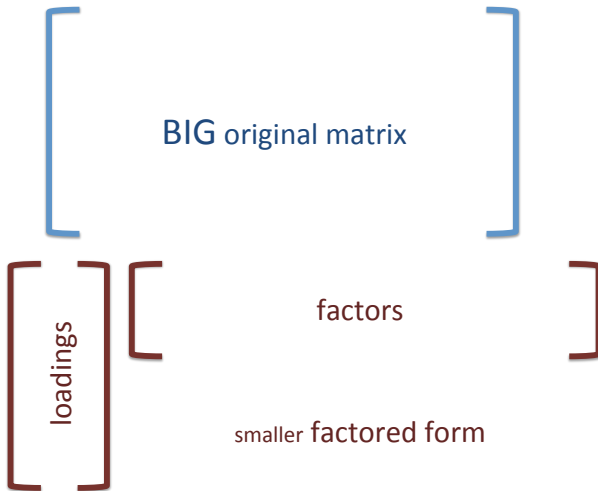
Regression: given covariates, what is response?



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



Dimension Reduction



Outline

Conceptual Review and Setup

Association Rules

Probabilistic Association Rules

The A Priori Algorithm

A Priori in R

Association Rules

Have a basket of items:

$$M = \begin{matrix} & \begin{matrix} \text{apples (a)} \\ \text{bananas (b)} \\ \text{cherries (c)} \\ \text{elderberries (e)} \\ \text{juniper berries (j)} \end{matrix} & \\ \begin{matrix} \text{transactions} \\ 1 \\ i \\ m \end{matrix} & \begin{bmatrix} 1 & & & & \\ & 1 & 1 & 1 & & \\ & & & 1 & & 1 & 1 \end{bmatrix} & \begin{matrix} 1 \\ j \\ n \end{matrix} \end{matrix}$$

in i^{th} transaction,
item j was purchased

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

Association Rules

Have a basket of items:

$$M = \begin{matrix} & \begin{matrix} \text{items} \\ \text{apples (a)} \\ \text{bananas (b)} \\ \text{cherries (c)} \\ \text{elderberries (e)} \\ \text{jumper berries (j)} \end{matrix} & \\ \begin{matrix} \text{transactions} \\ 1 \\ i \\ m \end{matrix} & \begin{bmatrix} 1 & & & & \\ & 1 & 1 & & \\ & & & 1 & & 1 & 1 \\ & & & & & & \end{bmatrix} & \end{matrix}$$

in i^{th} transaction,
item j was purchased

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:

Diagram illustrating a transaction matrix M (rows are transactions, columns are items).

Items: 1: apples (a), 2: bananas (b), 3: cherries (c), 4: elderberries (e), 5: juniper berries (j).

Transactions: 1, i , m .

Matrix entries (1 indicates purchase):

- Transaction 1: 1, 1, 1, 0, 0
- Transaction i : 1, 0, 0, 1, 1
- Transaction m : 0, 0, 0, 0, 1

Callout: in i^{th} transaction, item j was purchased.

So what makes a good rule?

Outline

Conceptual Review and Setup

Association Rules

Probabilistic Association Rules

The A Priori Algorithm

A Priori in R

Association Rules

We are going to use probability based rules:

$$P(b | a) = p$$

Use rule *support* and *confidence*:

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

$$\begin{aligned}\text{Conf}(a \rightarrow b) &= \frac{\text{Supp}(a \star b)}{\text{Supp}(a)} = \frac{\# \text{ times } a \text{ and } b \text{ are purchased}}{\# \text{ times } a \text{ is purchased}} \\ &= \hat{P}(b | a)\end{aligned}$$

Association Rules

So what makes a good rule?

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

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

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

$$\text{Conf}(a \rightarrow b) \geq \text{minconf}$$

These conditions lead to **strong rules**.

Outline

Conceptual Review and Setup

Association Rules

Probabilistic Association Rules

The A Priori Algorithm

A Priori in R

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
...	...

$$P(\text{"2"}|\text{"1"}) = \# \{1, 2\} / \# \{1\} = 4/5$$

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 $\hat{P}(b|a)$
- ▶ Prune all candidate rules with $\hat{P}(b|a) < \text{minconf}$

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
{1}	4
{2}	6
{3}	5
{4}	7
{5}	4

A Priori

Example: now generate candidates for 2 item sets

Table: Itemsets for $k = 1$

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

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

- ▶ can't combine $\{1, 4\}$ and $\{2, 4\}$ because $\text{Supp}(\{1, 2\}) < \theta$
- ▶ 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
{2, 3, 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
<u>$1 \rightarrow 4$</u>	<u>$3/4$</u>	$2 \rightarrow 3$	$4/6$
$2 \rightarrow 4$	$5/6$	$2 \rightarrow 5$	$3/6$
$3 \rightarrow 2$	$4/5$	$3 \rightarrow 4$	$4/5$
$4 \rightarrow 1$	$3/7$	$4 \rightarrow 2$	$5/7$
$4 \rightarrow 3$	$4/7$	$4 \rightarrow 5$	$3/7$
$5 \rightarrow 2$	$3/4$	$5 \rightarrow 4$	$3/4$
$2, 3 \rightarrow 4$	$3/4$	$2, 4 \rightarrow 3$	$3/5$
$3, 4 \rightarrow 2$	$3/4$	$2 \rightarrow 3, 4$	$3/6$
$3 \rightarrow 2, 4$	$3/5$	$4 \rightarrow 2, 3$	$3/7$

A Priori: Make Rules

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

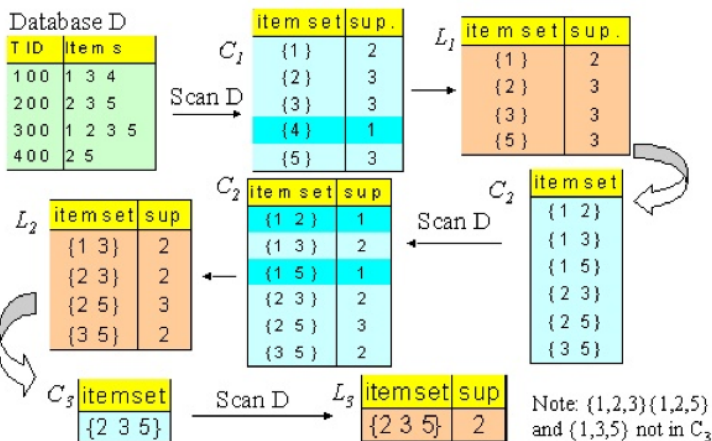
Table: Transactions

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
$1 \rightarrow 4$	$3/4$
$2 \rightarrow 4$	$5/6$
$3 \rightarrow 2$	$4/5$
$3 \rightarrow 4$	$4/5$
$5 \rightarrow 2$	$3/4$
$5 \rightarrow 4$	$3/4$
$2, 3 \rightarrow 4$	$3/4$
$3, 4 \rightarrow 2$	$3/4$

The Apriori Algorithm -- Example



A Priori

How do we choose rules?

- ▶ order by *confidence*

$$\text{Conf}(a \rightarrow b) = \hat{P}(b | a) = \frac{\text{Supp}(a \cup b)}{\text{Supp}(a)}$$

- ▶ order by *lift*

Lift 越大越好

$$\text{Lift}(a \rightarrow b) = \frac{\hat{P}(b | a)}{\hat{P}(b)} = \frac{\text{Supp}(b)}{1 - \frac{\text{Supp}(a \cup b)}{\text{Supp}(a)}}$$

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. $(\#\{j|i1,i2,i3\}/\#\{i1,i2,i3\})/(\#\{j\}/\#\{total\})$

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

- ▶ 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? (**Supervised** Classification? Regression? Groupings?)
- ▶ 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

Conceptual Review and Setup

Association Rules

Probabilistic Association Rules

The A Priori Algorithm

A Priori in R

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 .