Association Rules

Statistical Data Mining II

Spring 2016

Rachael Hageman Blair

Outline

- Description of the problem.
- Association Rules
- Market Basket
- Apriori algorithm
- Extensions and Caution
- Example

Introduction

Previous concern: Developing a model to predict a response.

In particular, suppose we have one or more response variables: $Y = \left(Y_1, Y_2, \ldots, Y_m\right)$ for a given set of input variables $X^T = \left(X_1, X_2, \ldots, X_p\right)$. This data is divided up into training/test data.

Objective – find a model that minimizes the test error in the sense of a loss function $L(y, \hat{y}) = (y - \hat{y})$, where $\hat{y}$ are the model predictions of the response.

*This is supervised learning.

Introduction

- Supervised learning can be framed as a functional approximation problem or a density estimation problem.

What do we mean?

Suppose that (X,Y) are random variables represented by joint probability density $\Pr(X,Y)$, we are essentially interested in the conditional density $\Pr(Y \mid X)$.

We try to approximate: $\mu(x) = \underset{\theta}{\operatorname{argmin}} E_{Y|X} L(Y,\theta)$.

Introduction

- **Why is this "not so bad"?**

$$\Pr(X,Y) = \Pr\left(Y \mid X\right)\Pr(X)$$

- Y is typically low dimensional (often one-dimensional).
- $\Pr(X)$ - is usually of no concern.

*There are many approaches to supervised learning that are selected according to the nature of Y, and the dimension of X.

Introduction

Who is in charge/ the supervisor?

- Y - provides the answers, and allows us to measure loss/accuracy. The response is the "teacher", "trainer" or "supervisor".
- Without the response, we can't measure accuracy to a "gold standard" within our model.
- In many ways, the task of unsupervised learning is harder, because the level of uncertainty is so severe. The questions are also not well formed.
- Performance is a matter of opinion.

Introduction

- **The setup:** We have N observations of a random p vector X having joint density $\Pr(X)$.
- **The goal:** directly infer properties of the probability distribution without any help from a supervisor providing the correct answers or degree of error per observation.

Common issues:

- The dimensionality is often more severe.

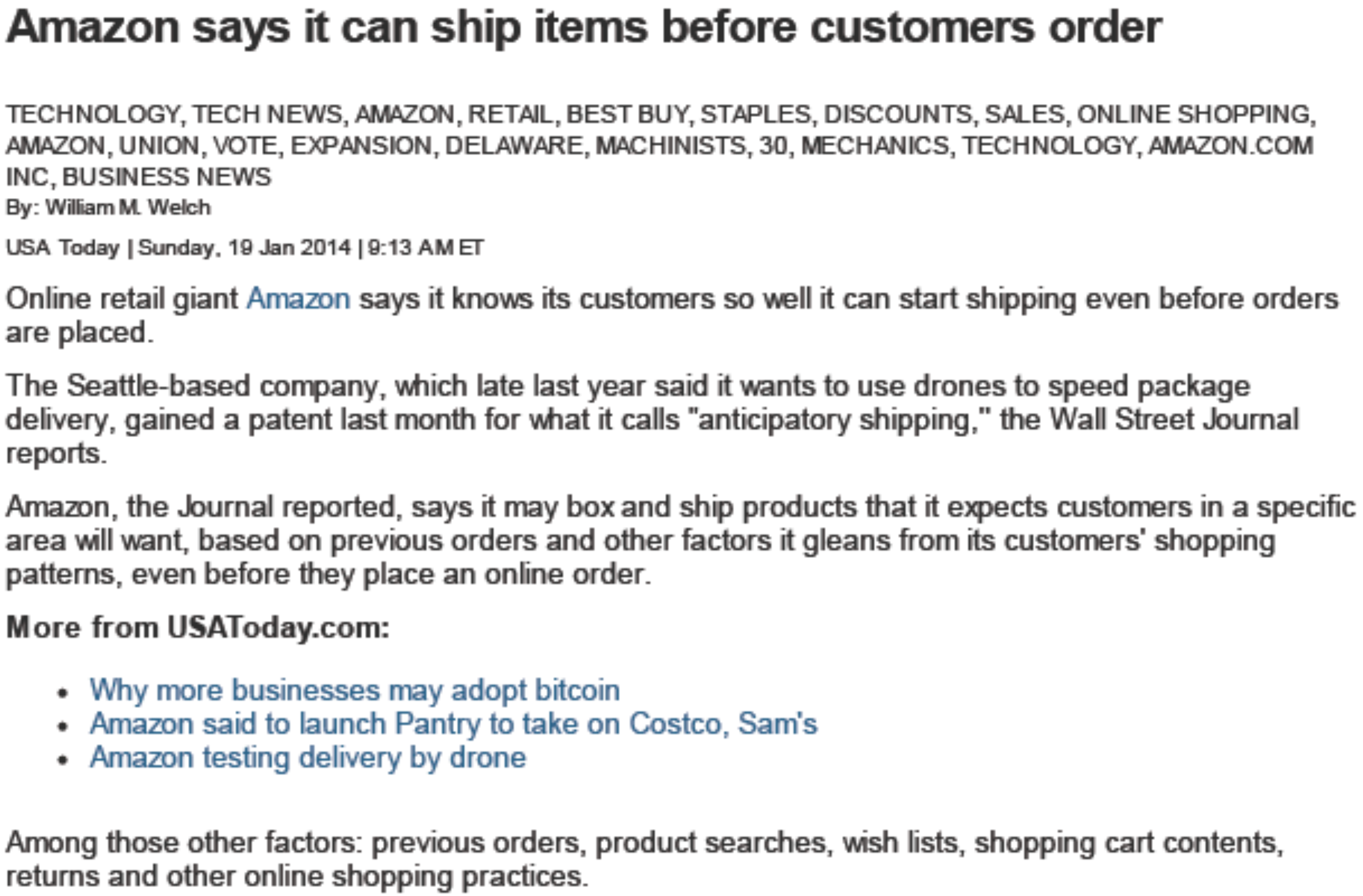
- Desired properties are ill-defined.

- Have to consider what is happening to X_i relative to the entire set of variables X .

Introduction

- We can get at the density properties in low dimensions.
- In higher dimensions we rely on “descriptive statistics” to characterize different aspects of the density.
  - ✓ Identify low-dimensional manifolds within the high dimensional space: PCA, multidimensional scaling, self-organizing maps etc. Manifolds are mined for associations, suggest smaller sets of latent variables.
  - ✓ Mixture type modeling: Clustering aims to identify modules or patterns, these modules suggest $\Pr(X)$ may be modeled as a mixture of simpler densities.
  - ✓ Association rules: Also attempt to characterize $\Pr(X)$ for a special situation of high dimensional binary data.

Association Rule Mining



Amazon says it can ship items before customers order

TECHNOLOGY, TECH NEWS, AMAZON, RETAIL, BEST BUY, STAPLES, DISCOUNTS, SALES, ONLINE SHOPPING, AMAZON, UNION, VOTE, EXPANSION, DELAWARE, MACHINISTS, 30, MECHANICS, TECHNOLOGY, AMAZON.COM INC, BUSINESS NEWS

By: William M. Welch

USA Today | Sunday, 19 Jan 2014 | 9:13 AM ET

Online retail giant Amazon says it knows its customers so well it can start shipping even before orders are placed.

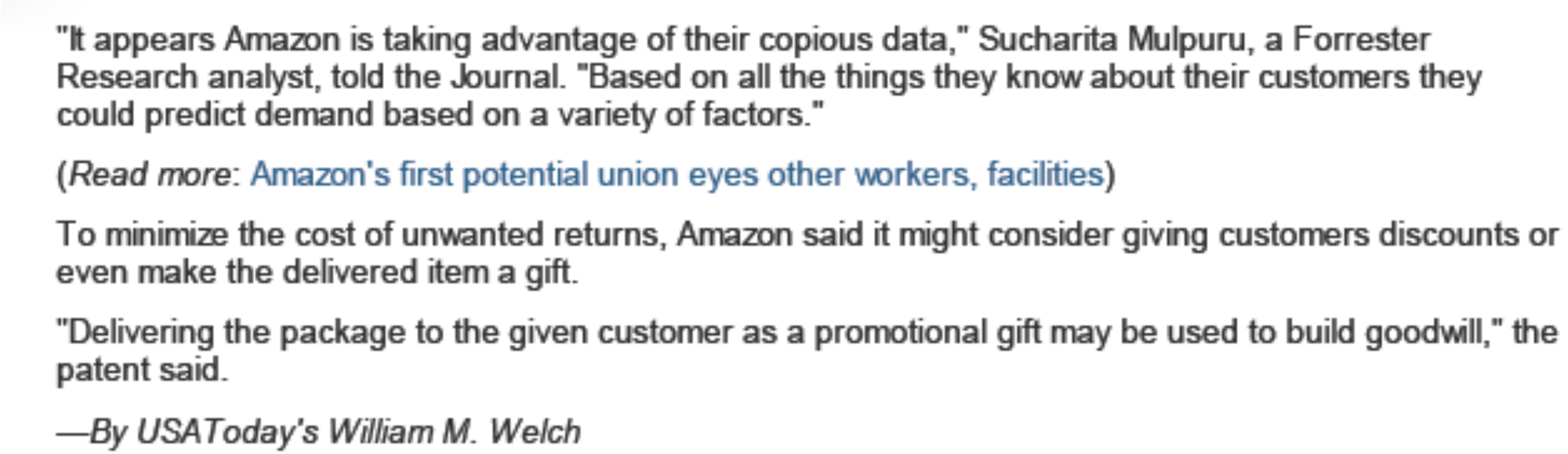
The Seattle-based company, which late last year said it wants to use drones to speed package delivery, gained a patent last month for what it calls "anticipatory shipping," the Wall Street Journal reports.

Amazon, the Journal reported, says it may box and ship products that it expects customers in a specific area will want, based on previous orders and other factors it gleans from its customers' shopping patterns, even before they place an online order.

More from USAToday.com:

- Why more businesses may adopt bitcoin
- Amazon said to launch Pantry to take on Costco, Sam's
- Amazon testing delivery by drone

Among those other factors: previous orders, product searches, wish lists, shopping cart contents, returns and other online shopping practices.



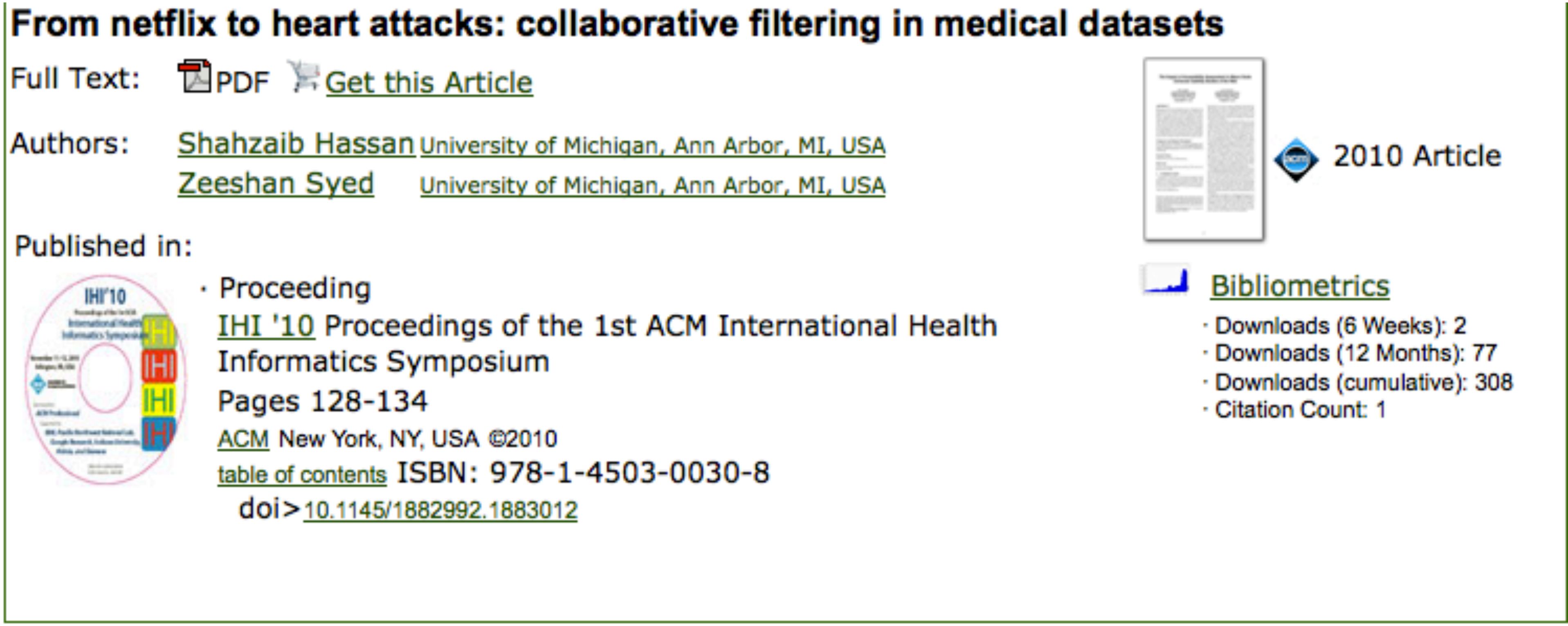
"It appears Amazon is taking advantage of their copious data," Sucharita Mulpuru, a Forrester Research analyst, told the Journal. "Based on all the things they know about their customers they could predict demand based on a variety of factors."

(*Read more*: Amazon's first potential union eyes other workers, facilities)

To minimize the cost of unwanted returns, Amazon said it might consider giving customers discounts or even make the delivered item a gift.

"Delivering the package to the given customer as a promotional gift may be used to build goodwill," the patent said.

—By USAToday's William M. Welch



From netflix to heart attacks: collaborative filtering in medical datasets

Full Text: PDF Get this Article

Authors: Shahzaib Hassan University of Michigan, Ann Arbor, MI, USA
Zeeshan Syed University of Michigan, Ann Arbor, MI, USA

2010 Article

Published in:

- Proceeding
  IHI '10 Proceedings of the 1st ACM International Health Informatics Symposium
  Pages 128-134
  ACM New York, NY, USA ©2010
  table of contents ISBN: 978-1-4503-0030-8
  doi>10.1145/1882992.1883012

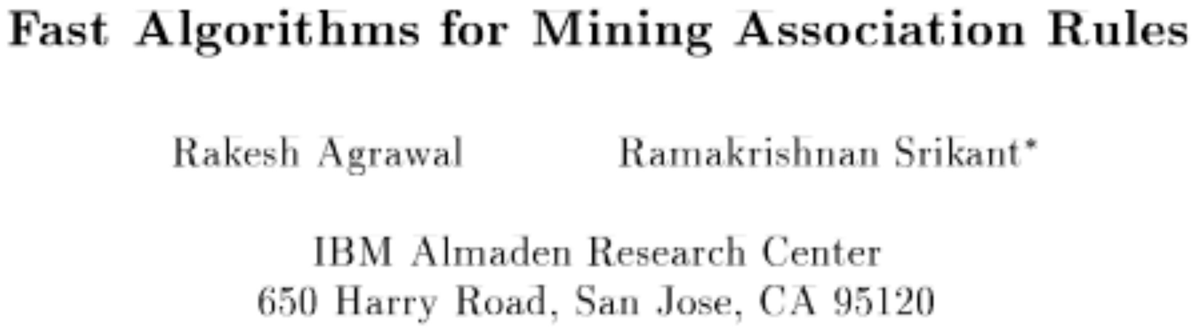
Bibliometrics

- Downloads (6 Weeks): 2
- Downloads (12 Months): 77
- Downloads (cumulative): 308
- Citation Count: 1

Association Rule Mining

- **Objective:** to characterize $\Pr(X)$ for a special situation of high dimensional binary data.
- An important data mining model studied by the database and data mining community.
- Roots in mining commercial databases.
- Original method proposed by Agrawal et al in 1993

  (cited ~15K times)



Fast Algorithms for Mining Association Rules

Rakesh Agrawal Ramakrishnan Srikant*

IBM Almaden Research Center
650 Harry Road, San Jose, CA 95120

Association Rule Mining

Problem: find joint values of the variables $X = \left(X_1, X_2, \ldots, X_p\right)$ that appear most frequently in the database. *Most commonly* applied to binary data, aka, "market basket" analysis.

- Observations -> sales transactions
- For observation *i*, each variable X_j is assigned a 1 if bought.
- Those variables with joint values, are very informative.

  e.g., suggestions, sales, bundles, marketing, promotions.
- Application of this type known as "Market Basket Analysis".

Association Rules

Retail: each customer purchases different sets of products, different quantities, at different times.

Goal of Market Basket Analysis:

Use the information as much as possible!

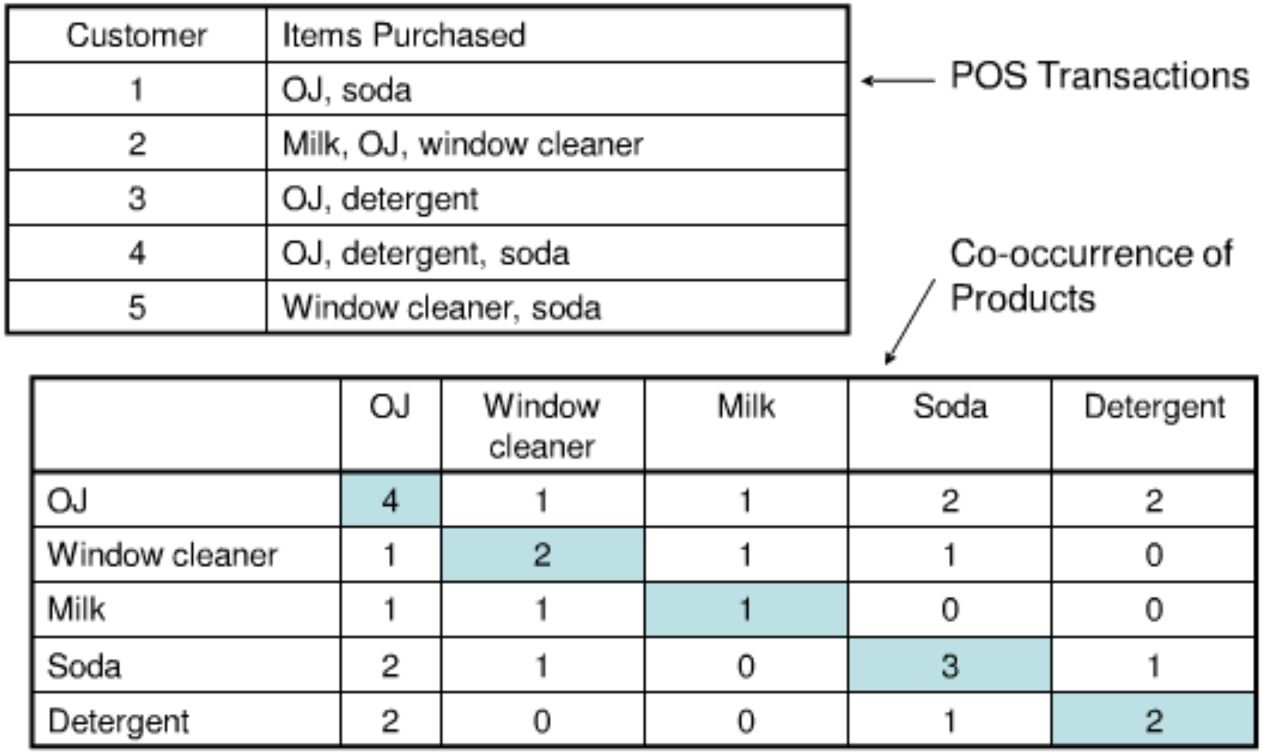
Identify customers

- Understand why they make certain purchases
- Gain insight about merchandise
  - Fast and Slow movers
  - Products which are purchased together
  - Products that may benefit from promotion
- Take action:
  - Store layout
  - Custom coupons
  - Loyalty card

Association Rules

Market Basket Analysis:

General set of techniques, which focus on “point of sale” transaction data consisting of customers, orders (basic purchase data), and items (merchandise/service purchased).



POS Transactions

Customer	Items Purchased
1	OJ, soda
2	Milk, OJ, window cleaner
3	OJ, detergent
4	OJ, detergent, soda
5	Window cleaner, soda

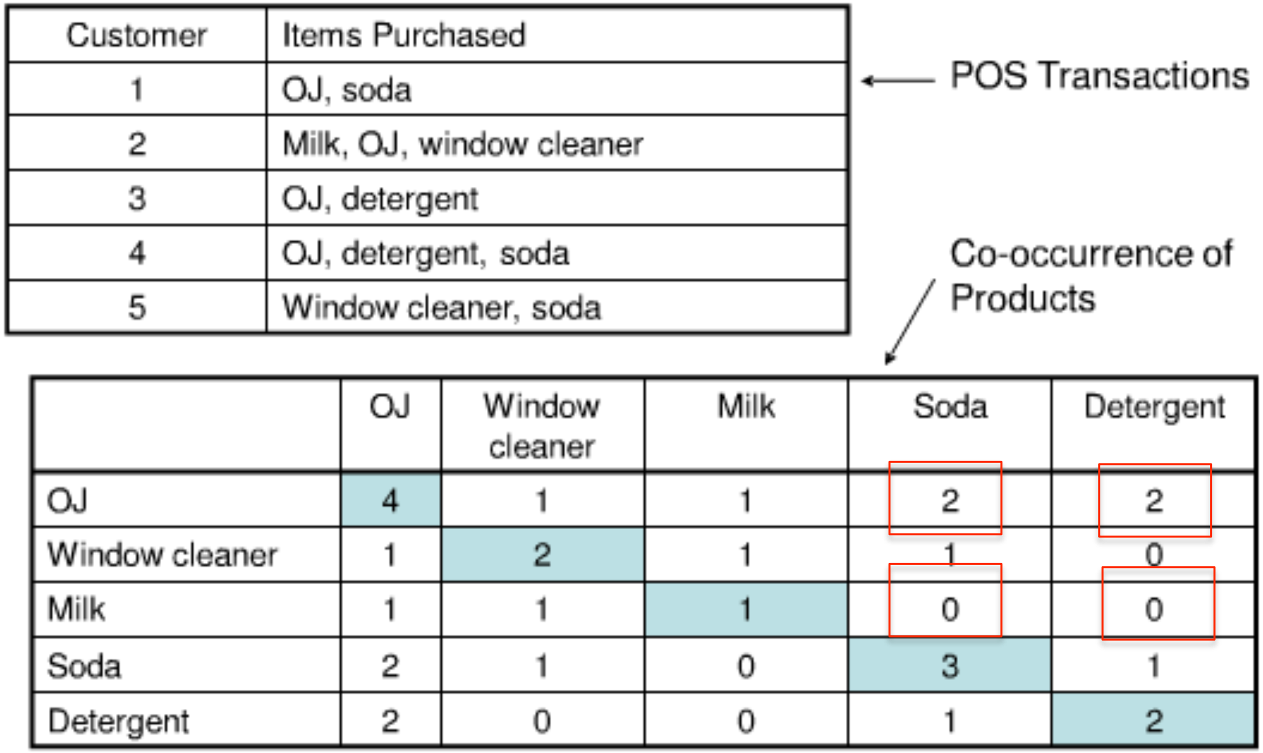
Co-occurrence of Products

	OJ	Window cleaner	Milk	Soda	Detergent
OJ	4	1	1	2	2
Window cleaner	1	2	1	1	0
Milk	1	1	1	0	0
Soda	2	1	0	3	1
Detergent	2	0	0	1	2

Association Rules

Market Basket Analysis:

General set of techniques, which focus on “point of sale” transaction data consisting of customers, orders (basic purchase data), and items (merchandise/service purchased).



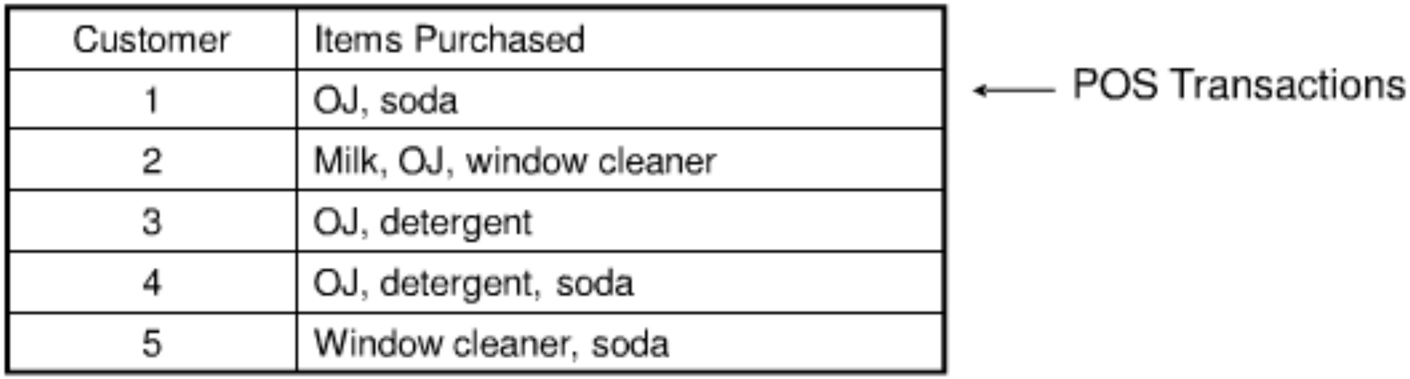
POS Transactions

Customer	Items Purchased
1	OJ, soda
2	Milk, OJ, window cleaner
3	OJ, detergent
4	OJ, detergent, soda
5	Window cleaner, soda

Co-occurrence of Products

	OJ	Window cleaner	Milk	Soda	Detergent
OJ	4	1	1	2	2
Window cleaner	1	2	1	1	0
Milk	1	1	1	0	0
Soda	2	1	0	3	1
Detergent	2	0	0	1	2

Association Rules



Customer	Items Purchased
1	OJ, soda
2	Milk, OJ, window cleaner
3	OJ, detergent
4	OJ, detergent, soda
5	Window cleaner, soda

← POS Transactions

- Will a customer purchasing soda, then purchases OJ?

  (2 out of 3 soda purchasers, also purchase OJ: 2/3~ 67%)

- Will a customer purchasing OJ, then purchase soda?

  (2 out of 4 OJ purchasers, also purchase soda: 1/2~ 50%)

Confidence: Ratio of the number of transactions, similar to a conditional probability.

Association Rules – In more statistical terms

The problem: Find a collection of prototype X-values $v_1, v_2, \ldots, v_L$ for the feature vector X, such that the probability density $\Pr(v_l)$ evaluated at these values is relatively large.

Intuitive Solution: estimates of $\Pr(v_l)$ can be made from the fraction of observations for which $X = v_l$.

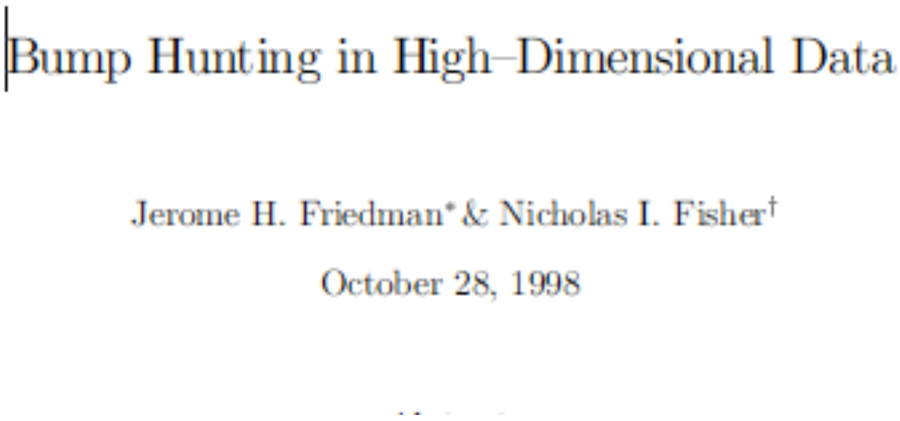
- The problem – in almost all cases, there is a large number of observations, and the number of observations for which $X = v_l$ is almost always too small to be reliable.

*Must recast the problem, and its solution.

Association Rules

More generally known as “bump hunting” or “mode finding”.

- One seeks a set of sub-regions of the input space within which the value of the output is considerably larger (or smaller) than its average value over the entire input domain.



Bump Hunting in High–Dimensional Data

Jerome H. Friedman[*] & Nicholas I. Fisher[†]

October 28, 1998

Association Rules

Subtle but important difference:

- Instead of seeking values x where $\Pr(x)$ is large, we are looking for regions of the X-space with high probability relative to **size** and $\underline{\textbf{support}}$.

S_j - denotes all possible values of the *j*th variable

s_j - be a subset of the all possible values, $s_j \subseteq S_j$

for quantitative – interval, qualitative – value.

Modified Goal: Find a subset of variable values $s_1,\ldots,s_p$ such that the probability of each of the variables simultaneously assuming a value within its respective subset:

$$\Pr\left[\underline{\bigcap_{j=1}^{p}\left(X_j \in s_j\right)}\right],$$

is relatively large.

conjunctive rule

Market Blanket Analysis

- We want to find regions to maximize

$$\Pr\left[\bigcap_{j=1}^{p}\left(X_j \in s_j\right)\right]$$

- **Several approaches** ~ but can’t handle the magnitude of real world problems (p=10^4, N=10^8).

- In order to tackle these “real world problems”, more simplifications have to be made.

Market Blanket Analysis

Simplification (further):

Only consider two types of subsets, “all” or “one”, either s_j consists of a single value of X_j, $s_j = v_{0j}$, or it consists of the entire set of values that X_j can assume.

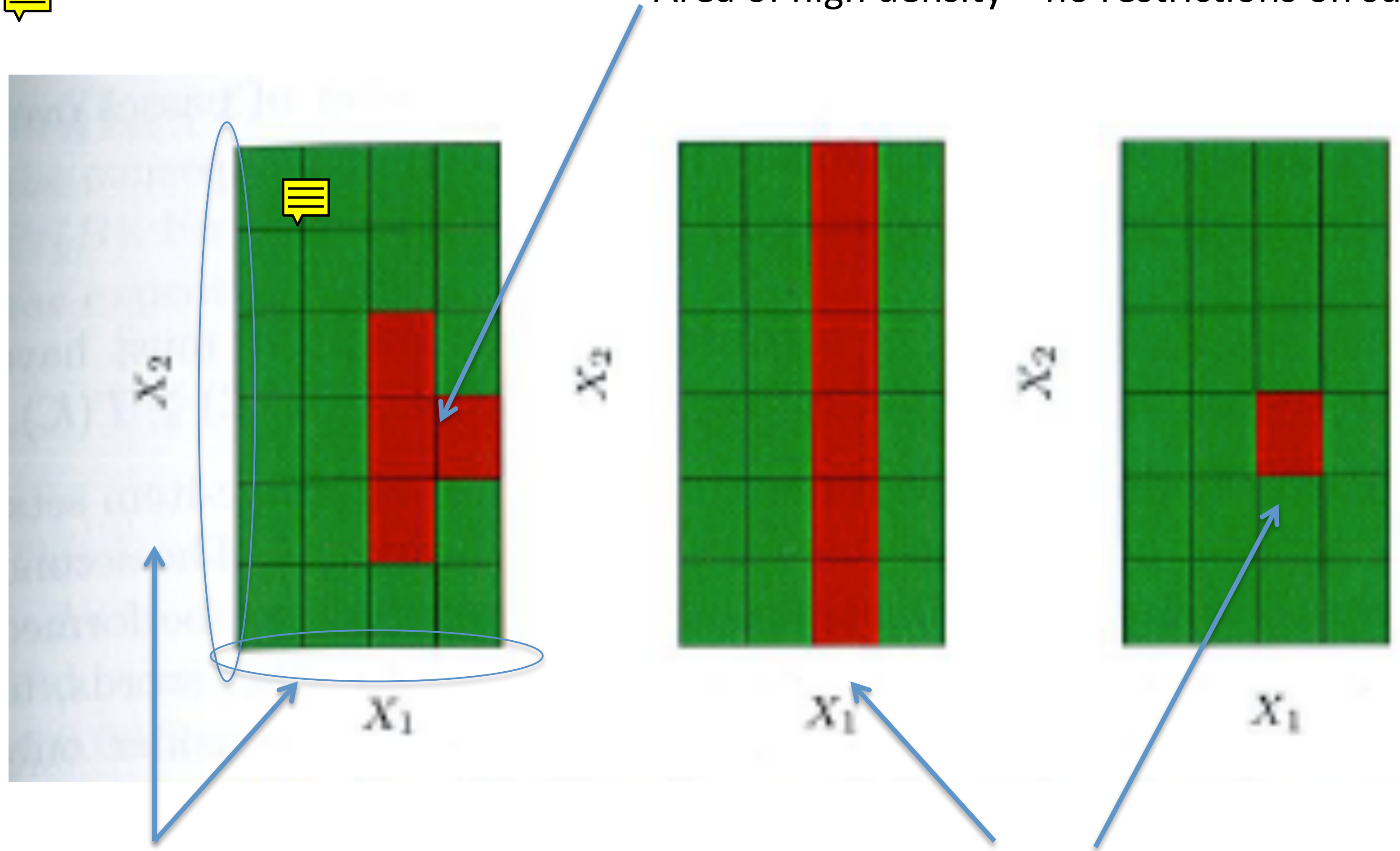
The problem is “simplified” to finding subsets of integers $j \in J$, such that:

$$\Pr\left[\bigcap_{j \in J}\left(X_j = v_{0j}\right)\right]$$

Is large.

Market Blanket Analysis

Area of high density – no restrictions on subsets



X1 has 6 distinct values
X2 has 4 distinct values

Area of high density – restrictions on subsets

Market Blanket Analysis

Standard Formulation:

Assume the support S_j is finite for each variable X_j. We can use a *dummy* coding system. Let $Z_1,\ldots,Z_K$ denote the dummy scheme encoding a variable for each of the values v_{lj} attainable by each of the original variables $X_1,\ldots,X_p$.

Note that there are K dummy variables:

$$K = \sum_{j=1}^{p} \left| S_j \right|.$$

Number of distinct values attainable by Xj

Each dummy variable is assigned a $Z_K = 1$ if the variable with which it is associated with takes on the value which Z_K is assigned.

Market Blanket Analysis

Standard Formulation:

We have transformed the problem. Now, we are after a subset of integers $\mathrm{K}^* \subset \{1,\ldots,K\}$ such that:

$$\Pr\left[\bigcap_{k\in\mathrm{K}}(Z_k=1)\right]=\Pr\left[\prod_{k\in\mathrm{K}}Z_k=1\right]$$

is large.

*The standard formulation of the market basket problem.

“Item set”

Number of variables Zk in an “item set” is its size.

Market Blanket Analysis

Standard Formulation:

The estimated value of this probability is simple the fraction of observations in the data base for which $\Pr\left[\bigcap_{k\in\mathrm{K}}(Z_k = 1)\right] = \Pr\left[\prod_{k\in\mathrm{K}} Z_k = 1\right]$ is true, that is,

$$\hat{\Pr}\left[\prod_{k\in\mathrm{K}} Z_k = 1\right] = \frac{1}{N}\sum_{i=1}^{N}\prod_{k\in K} z_{ik}.$$

This is the "support" or "prevalence" $T\left(\mathrm{K}^*\right)$ of the item set K^*.

An observation i for which $\prod_{k\in\mathrm{K}} z_{ik} = 1$ is said to "contain" item set K^*.

Market Blanket Analysis

In Practice:

A lower bound in the search is set, and we seek all item sets K_l , such that, a certain support t in the database is obtained:

$$\left\{\mathrm{K}_l^* \mid T\left(\mathrm{K}_l^*\right) > t\right\}.$$

Adjustment to make the problem tractable.

******Our first brush with the curse of dimensionality*

Apriori Algorithm

In Practice:

A lower bound in the search is set, and we seek all item sets K_l , such that, a certain support t in the database is obtained:

★ $$\left\{ \mathrm{K}_l^* \mid T\left(\mathrm{K}_l^*\right) > t \right\}.$$

Adjustment to make the problem tractable.

Apriori algorithm – a method for learning associations in the dataset according to ★ , exploits database properties for efficiency.

Apriori Algorithm

For a given support threshold t :

- The cardinality $\left|\left\{\mathrm{K}_l^* \mid T\left(\mathrm{K}_l^*\right) > t\right\}\right|$ is relatively small.
- Any item set L consisting of a subset of the items in K^* must have support greater than or equal to that of K^*, that is:

$$L \subseteq \mathrm{K}^* \Rightarrow T\left(L\right) \geq T(\mathrm{K}^*).$$

Apriori algorithm:

Consecutive scans over the database, trimming the database each time based on support, and as increasing the item set. Continue until all candidate rules have support less that the specified value t .

***Critical Point:** Highly efficient (fast and cheap), can deal with data sets that cannot be handled in computer memory.

Apriori Algorithm

Apriori algorithm :

1. Compute the support of all single-item sets. Those whose support is less than a threshold are discarded.
2. Compute the support of all item sets of size two that can be formed from pairs of the single items that survive step one.
3. Compute the support of all item sets of size three that can be formed from pairs of the single items that survive step two.

Continue until all items are at or above the threshold value.

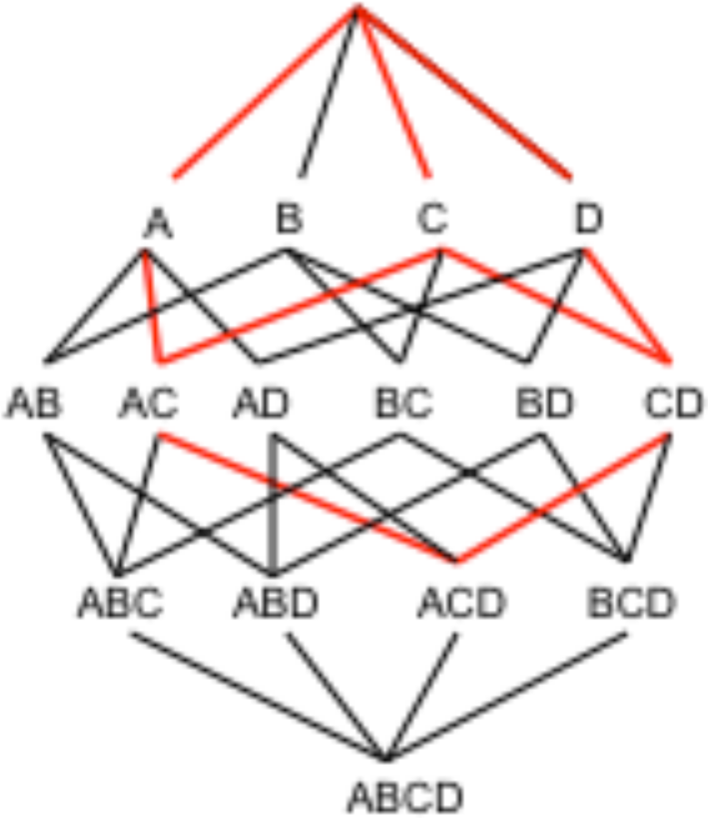
*Note that the apriori algorithm requires only one pass over the data for each value $\left|K^*\right|$, this is what makes it fast.

Apriori Algorithm

Apriori algorithm :

- Progressively identifies large item sets of different sizes.

- Exploiting the nested properties of item sets.

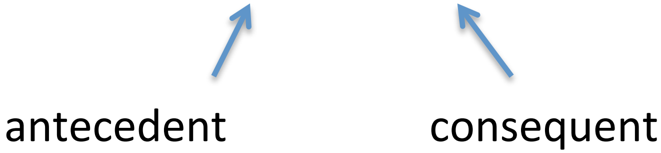


Apriori requires only one pass over the data for each K (itemset). Critical, because transaction data may not fit into computer memory.

Apriori Algorithm

- Each high support item set $\mathcal{K}^*$ returned by the algorithm is cast into a set of "association rules". The items $Z_k, k \in \mathcal{K}^*$, are partitioned into two disjoint subsets, $A \bigcup B = \mathcal{K}^*$, and written:

$$A \Rightarrow B.$$



antecedent consequent

- Association rules have to do with different aspects of this relationship.
- Support of the rule: $T(A \Rightarrow B)$ is the fraction of observations in the union of the antecedent and consequent, equivalently, the support of the item set $\mathcal{K}^*$ from which they are derived.

Apriori Algorithm

Confidence or **Predictability:** support divided by support of the antecedent:

$$C(A \Rightarrow B) = \frac{T(A \Rightarrow B)}{T(A)}$$

which can be viewed as the conditional probability: $\Pr(B \mid A)$, and $\Pr(A)$ is the probability of an item set A occurring in a basket.

. (recall: $\Pr(A \cap B) = \Pr(B \mid A) P(A)$)

Apriori Algorithm

Expected Confidence: the support of the consequent $T\left(B\right)$, which estimates the probability, $\Pr\left(B\right)$.

Lift: An estimate of the association $\Pr\left(A \text{ and } B\right) / \Pr(A)\Pr(B)$.

We denote the lift as the ratio: $L(A \Rightarrow B) = \dfrac{C(A \Rightarrow B)}{T(B)}$.

The “lift” can be thought of as a measure of “improvement”, telling us how much better a rule is at predicting the result than just assuming the result in the first place.

- Lift >1 - the rule is better than predicting the result than guessing.
- Lift <1 – the rule is doing worse than the informed guessing, using a negative rule produces a better result than guessing.

Apriori Algorithm

Example: $\mathrm{K}^* = \{\text{peanut butter, jelly, bread}\}$

- Consider the relationship $\{\text{peanut butter, jelly}\} \Rightarrow \{\text{bread}\}$.

- A support value of 0.03 implies that peanut butter, jelly and bread appear together in 3% of the market baskets.

- Confidence of 0.82 for this rule implies that when peanut butter and jelly were purchased, 82% of the time, bread was also purchased.

- If bread appeared in 43% of all market baskets, then the rule $\{\text{peanut butter, jelly}\} \Rightarrow \{\text{bread}\}$ lift would be 1.95.

Apriori Algorithm

Goal: Produce association rules with high values of support and confidence:

$$\mathrm{A} \Rightarrow \mathrm{B}.$$

Output: All item sets with high support as defined by the support threshold t.

To get at the rules, is another story,

ANOTHER threshold is set wrt to confidence,

$$\{\mathrm{A} \Rightarrow \mathrm{B} \mid \mathrm{C}(\mathrm{A} \Rightarrow \mathrm{B}) > c\}.$$

*Note that for each item set of size $|\mathrm{K}|$, there are $2^{|K|-1}$ rules of the form $\{\mathrm{A} \Rightarrow (\mathrm{K} \setminus A)\}$. (item set 10 -> 512 rules, item set 15 -> 16384).

Apriori Algorithm

- Results are stored/ranked database and queried accordingly.
- All results satisfy:

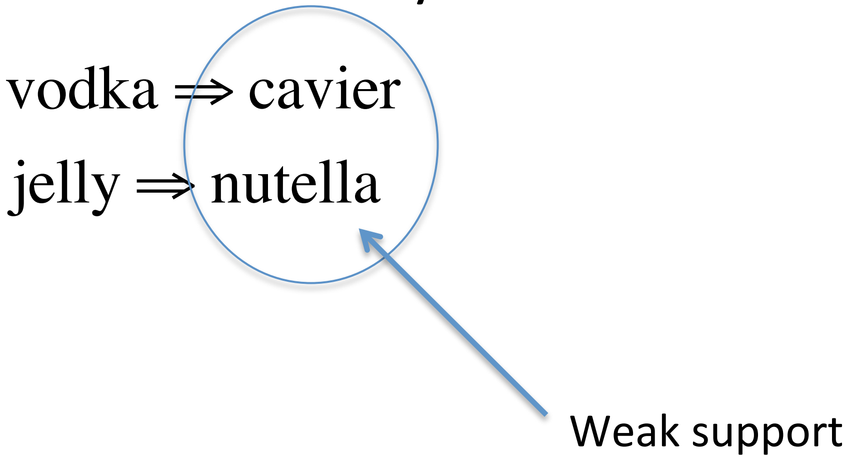
  $$T(\mathrm{A} \Rightarrow \mathrm{B}) \geq \mathrm{t} \quad \text{and} \quad C\left(\mathrm{A} \Rightarrow \mathrm{B}\right) > c\ .$$

- May just look at rankings, lift etc., for more general marketing.
- May look at specific points of intervention. The query may be "consequent driven", which casts the unsupervised problem as supervised one.

  For example: Display all transactions in which ice skates are the consequent and have confidence over 80% and support of more than 2%.

Apriori Algorithm

- Setting the thresholds too low, will make the problem intractable, even for smaller datasets.
- Also note, that if the consequent has small support, relationships which involve it cannot be unveiled by construction.



$$\text{vodka} \Rightarrow \text{cavier}$$

$$\text{jelly} \Rightarrow \text{nutella}$$

Weak support

Model Interpretation: Caution

Other considerations: some of the arguments derived from association rules are not necessarily useful, or may be obvious already.

- “Customers who purchase maintenance agreements also purchase large appliances”. – obvious, but useful
- “Customers who purchase a home mortgage may also buy furniture”. - obvious, but useful
- “Men who buy diapers on Thursdays also buy beer” – nonsense
- “Customers who buy data mining book A, also buy data mining book B” - useful

*No way to drill down without the human expert.

Caution

Rare item problem: if frequencies of items vary a great deal, we will encounter too problems.

- If the support is too high, we will not be able to find rare items.
- Finding rules that involve frequent and rare items, would require major computational demand.

Extensions

- Multiple minimum class supports

  - User can specify different minimum supports to different classes, which effectively assigns a different minimum support to rules.

  e.g., female – minimum support 5%

  male – minimum support 10%.

  - Also we can speed things up by assigning minimum support of 100% to certain classes. Useful in applications.

Extensions

- Consider the following items:

  bread, shoes, clothes

  The user-specified MIS values are as follows:

  Support(*bread*) = 2%   Support(*shoes*) = 0.1%

  Support(*clothes*) = 0.2%

  The following rule doesn't satisfy its minsup:

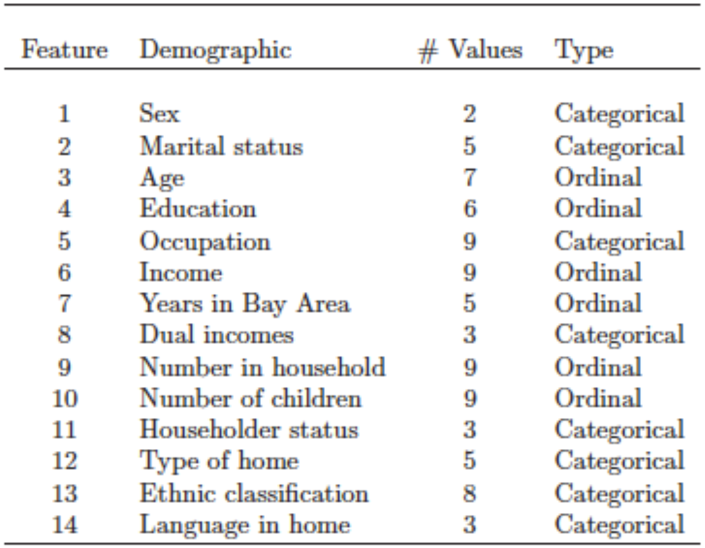
  clothes → *bread* [sup=0.15%,conf =70%]

  The following rule satisfies its minsup:

  clothes → *shoes* [sup=0.25%,conf =75%]

An Example

Dataset: **9,409** questionnaires filled out by shopping mall customers in San Francisco. Utilized answers to the first 14 questions, related to demographics.



Feature	Demographic	# Values	Type
1	Sex	2	Categorical
2	Marital status	5	Categorical
3	Age	7	Ordinal
4	Education	6	Ordinal
5	Occupation	9	Categorical
6	Income	9	Ordinal
7	Years in Bay Area	5	Ordinal
8	Dual incomes	3	Categorical
9	Number in household	9	Ordinal
10	Number of children	9	Ordinal
11	Householder status	3	Categorical
12	Type of home	5	Categorical
13	Ethnic classification	8	Categorical
14	Language in home	3	Categorical

In R package:

```
> library(ElemStatLearn)
> ?marketing
```

An Example

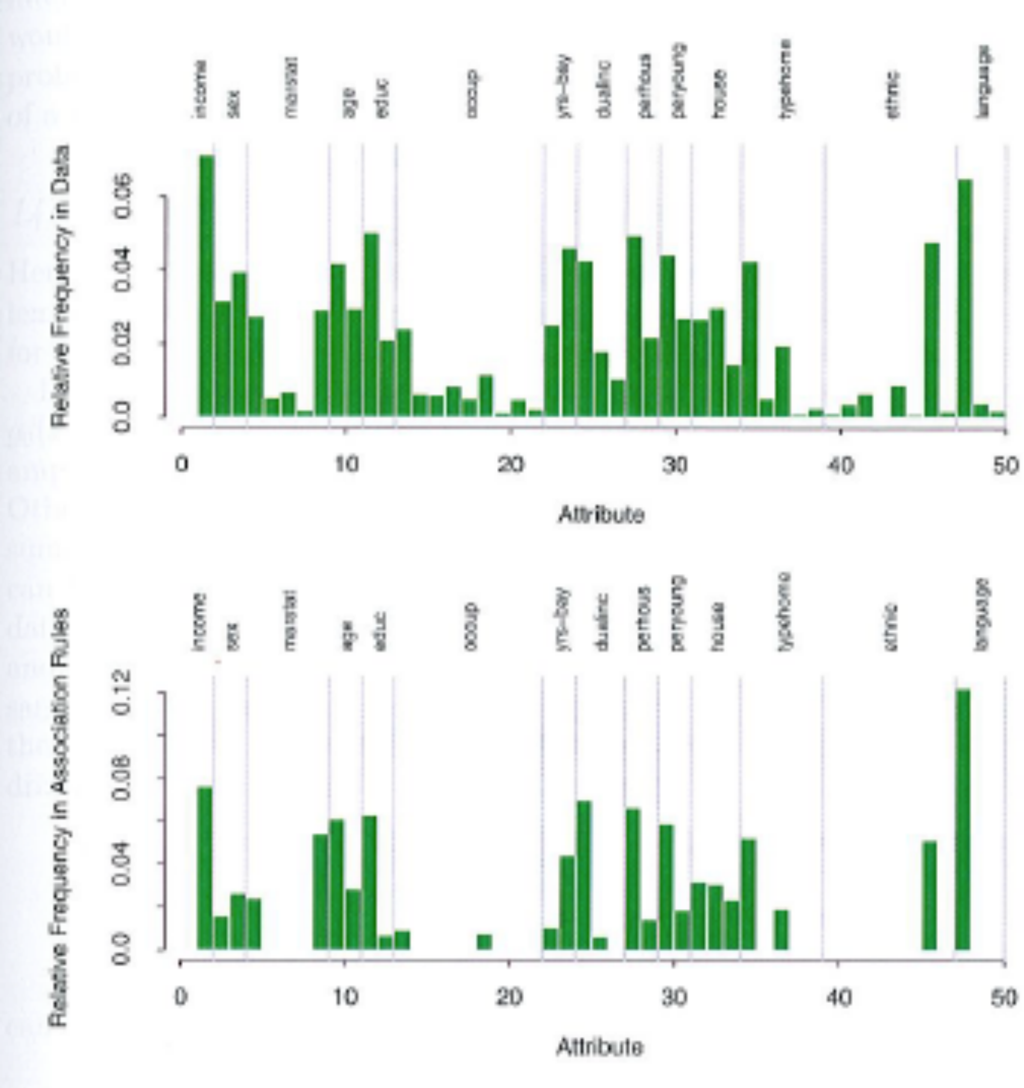
Freeware (http://www.borgelt.net/apriori.html) standalone, C, developed by software engineer.

Alternative package: “arules” -> calls the C implementation by Christian Borgelt.

Preprocessing: Removing observations with missing values, each ordinal predictor was cut at its median and coded by two dummy variables, the resulting dataset is 6876 (obs) x 50 (dummy variables)

An Example

Results: The algorithm found a total of 6288 association rules, involving less than 5 predictors with support of at least 10%.



An Example

Some Association Rules with “high support, confidence and lift”:

- Support 25%, confidence 99.7% and lift 1.03:

$$\left[\begin{array}{rcl} \text{number in household} & = & \mathit{1} \\ \text{number of children} & = & \mathit{0} \end{array}\right] \\ \Downarrow \\ \text{language in home} = \mathit{English}$$

- Support 13.4%, confidence 80.8%, and lift 2.13:

$$\left[\begin{array}{rcl} \text{language in home} & = & \mathit{English} \\ \text{householder status} & = & \mathit{own} \\ \text{occupation} & = & \{\mathit{professional/managerial}\} \end{array}\right] \\ \Downarrow \\ \text{income} \geq \mathit{\$40,000}$$

- Support 26.5%, confidence 82.8% and lift 2.15:

$$\left[\begin{array}{rcl} \text{language in home} & = & \mathit{English} \\ \text{income} & < & \mathit{\$40,000} \\ \text{marital status} & = & \mathit{not\ married} \\ \text{number of children} & = & \mathit{0} \end{array}\right] \\ \Downarrow \\ \text{education} \notin \{\mathit{college\ graduate,\ graduate\ study}\}$$