Chapter 9. Market-Basket Analysis

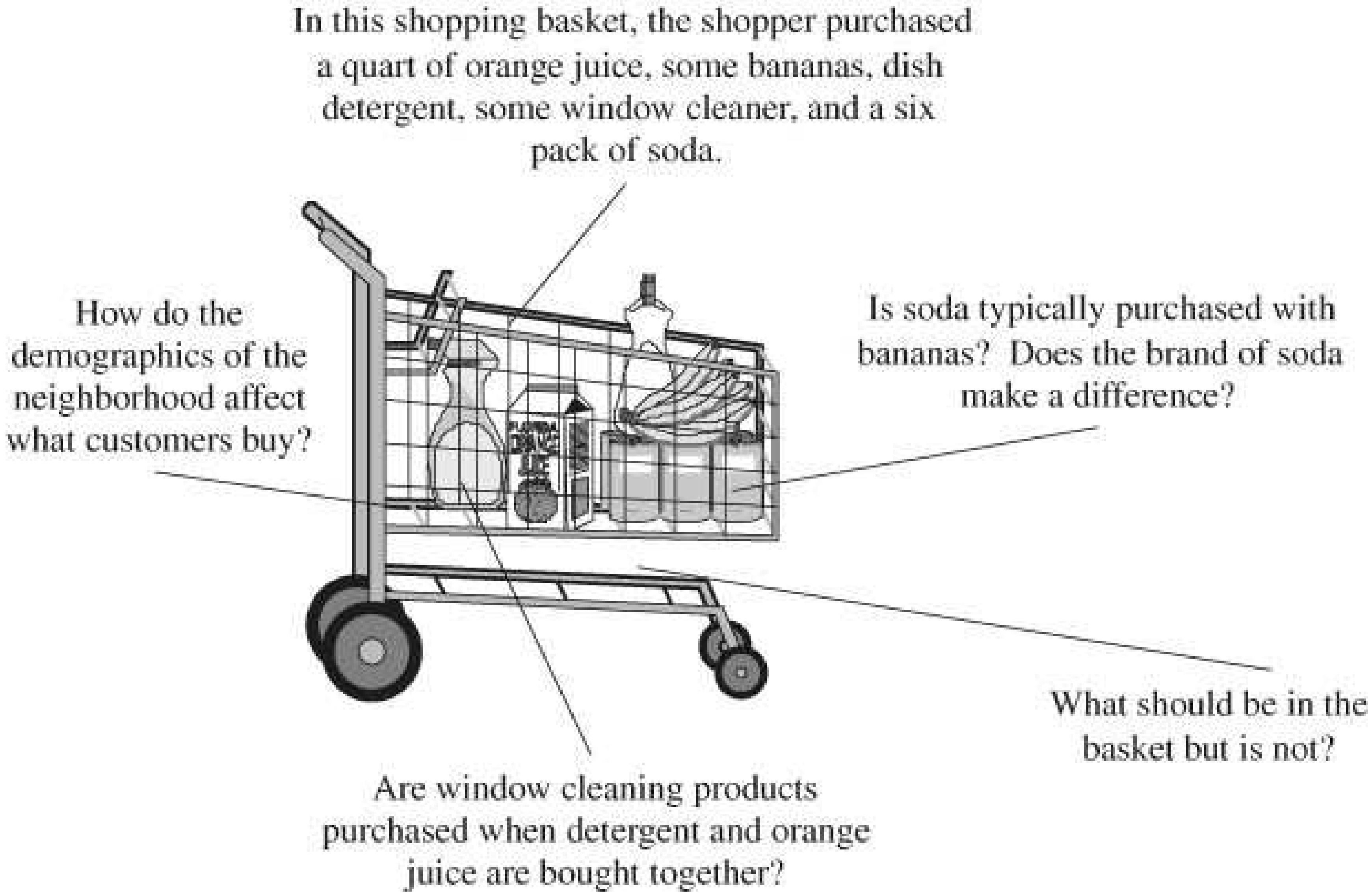
Goal: identify co-occurring items.

Techniques: quantify so-called support, confidence and lift etc important concepts in basket analysis, the Apriori Algorithm

Potential applications: stocking shelves, cross-marketing in sales promotions, catalog design, and consumer segmentation based on buying patterns.

Further Reading:

Provost and Fawcett (2013): Chapter 9.



In this shopping basket, the shopper purchased a quart of orange juice, some bananas, dish detergent, some window cleaner, and a six pack of soda.

How do the demographics of the neighborhood affect what customers buy?

Is soda typically purchased with bananas? Does the brand of soda make a difference?

What should be in the basket but is not?

Are window cleaning products purchased when detergent and orange juice are bought together?

Each customer purchases a different set of products, in different quantities, at different times.

Market basket analysis uses the information about what customers purchase to provide insight into who they are and why they make certain purchases. Market basket analysis provides insight into the merchandise by telling us which products tend to be purchased together and which are most amenable to promotion.

This information is actionable: it can suggest new store layouts; it can determine which products to put on special sales; it can indicate when to issue coupons, and so on.

When this data can be tied to individual customers through a loyalty card or Web site registration, it becomes even more valuable.

Other applications: Items purchased on a credit card, such as rental cars and hotel rooms, provide insight into the next product that customers are likely to purchase.

Optional services purchased by telecommunications customers (call waiting, call forwarding, DSL, speed call, and so on) help determine how to bundle these services together to maximize revenue.

Banking services used by retail customers (money market accounts, investment services, car loans, and so on) identify customers likely to want other services.

Unusual combinations of insurance claims can be a sign of fraud and can spark further investigation.

Medical patient histories can give indications of likely complications based on certain combinations of treatments.

Suppose a supermarket sells p items in total, has recorded n transactions.

Denote each transaction with a p-vector with components 0 or 1, i.e.

$$\begin{aligned} x_{ij} &= 1 \ \text{ if } i\text{-th transaction contains a purchase of } j\text{-th item,} \\ &\quad \text{0 otherwise} \end{aligned}$$

Thus i-th transaction is represented by $\mathbf{x}_i = (x_{i1}, \cdots, x_{ip})'$, a vector with most components equal to 0.

Let $\mathcal{J} \subset \{1, \cdots, p\}$, indicating a subset of the items sold in the supermarket.

We call $\mathcal{J}$ an ‘item set’, the number of the elements in $\mathcal{J}$ is called the ‘size’ of $\mathcal{J}$.

There are in total $2^p - 1$ item sets.

For $p = 10000$, 2^p can be regarded as infinity!

Support for index set $\mathcal{J}$:

$$T(\mathcal{J}) = \frac{1}{n}\sum_{i=1}^{n}\prod_{j\in\mathcal{J}} x_{ij}$$

i.e. $T(\mathcal{J})$ is the proportion of the transactions which contain all the items in set $\mathcal{J}$.

Note. $T(\mathcal{J})$ is the (estimated) probability for the event that the items in the set $\mathcal{J}$ are purchased together.

Basket Analysis: to identify all the set $\mathcal{J}$ with $T(\mathcal{J}) \ge t$, where $t \in (0, 1)$ is a constant.

Typically t is taken as a small constant such as 0.05 for big supermarket data.

The problem looks simple, at least conceptually. However it is computationally infeasible to search over all possible item sets, as typically $p \approx 10^4$ and $n \approx 10^8$ for big supermarkets.

The Apriori Algorithm. It makes the search feasible if the number of the item sets satisfying the condition $T(\mathcal{J}) > t$ is a small fraction of 2^p (i.e. t cannot be too small).

A simple fact: If item set $\mathcal{K}$ is a subset of item set $\mathcal{J}$, $T(\mathcal{K}) \geq T(\mathcal{J})$.

Key Idea of the Apriori Algorithm:

- The 1st pass over the data computes the support of all single-item sets, and discards those with support smaller than t.

- The 2nd pass computes the support of all item sets of size 2 that are formed from pairs of the single item sets surviving the 1st pass, and discards those with support smaller than t.

- For $m \geq 3$, the m-th pass computes the support of all item sets of size m that are formed from a surviving item set of size $m - 1$ and a surviving single item set, and discards those with support smaller than t.

There are many additional tricks to increase the speed and convergence in the Apriori Algorithm; see Agrawal *et al.*(1995). *Fast discovery of association rules*, Advances in Knowledge Discovery and Data Mining, AAAI/MIT Press, Cambridge, MA.

The Apriori algorithm represents one of the major advances in data mining technology.

Association rules: defined for each high support item set $\mathcal{J}$ with size at least 2.

Partition $\mathcal{J}$ into two non-overlapping subsets $\mathcal{A}$ and $\mathcal{B}$, i.e. $\mathcal{J} = \mathcal{A} \cup \mathcal{B}$ and $\mathcal{A} \cap \mathcal{B}$ is an empty set.

Confidence (or **Predictability**): The confidence of $\mathcal{B}$ from $\mathcal{A}$ is defined as

$$C(\mathcal{A} \Rightarrow \mathcal{B}) = T(\mathcal{A} \cup \mathcal{B}) \Big/ T(\mathcal{A}).$$

Note. $C(\mathcal{A} \Rightarrow \mathcal{B})$ is the (estimated) conditional probability of $\mathcal{B}$ given $\mathcal{A}$. $T(\mathcal{A} \cup \mathcal{B})$ is an estimate for $P(\mathcal{A}\mathcal{B}) \neq P(\mathcal{A} \cup \mathcal{B})$.

Lift: The lift of of $\mathcal{B}$ from $\mathcal{A}$ is defined as

$$L(\mathcal{A} \Rightarrow \mathcal{B}) = C(\mathcal{A} \Rightarrow \mathcal{B}) \Big/ T(\mathcal{B}).$$

Note. (i) $L(\mathcal{A} \Rightarrow \mathcal{B})$ is the ratio of probability of the event that all items in $\mathcal{A}$ and $\mathcal{B}$ are purchased together to the product of the probabilities of two events: (i) all items in $\mathcal{A}$ are purchased together, (ii) all items in $\mathcal{B}$ are purchased together

(ii) $T(\mathcal{B})$ is also called ‘expected confidence’. Thus the lift is the confidence divided by the expected confidence.

(iii) $L(\mathcal{A} \Rightarrow \mathcal{B}) = L(\mathcal{B} \Rightarrow \mathcal{A})$

For example, let $\mathcal{J} = \{\texttt{peanutbutter}, \texttt{jelly}, \texttt{bread}\}$ with support 0.03. Let $\mathcal{A} = \{\texttt{peanutbutter}, \texttt{jelly}\}$ and $\mathcal{B} = \{\texttt{bread}\}$. Suppose $T(\mathcal{A}) = 0.04$. Then the confidence is

$$C(\mathcal{A} \Rightarrow \mathcal{B}) = 0.03/0.04 = 75\%.$$

Hence when peanut butter and jelly were purchased, 75% of the time bread was also purchased.

If $T(\mathcal{B}) = 0.4$, i.e. the 40% purchases include bread, the lift for $\{\texttt{bread}\}$ by $\{\texttt{peanut butter, jelly}\}$ is

$$L(\mathcal{A} \Rightarrow \mathcal{B}) = 0.75/0.4 = 1.875.$$

Example. Consider $n = 9404$ questionnaires filled out by shopping mall customers in the San Francisco Bay Area (Impact Resources, Inc., Columbus OH, 1987). Here we only use answers to the first 14 questions, relating to demographics. These questions are listed below.

The data consist of a mixture of ordinal and (unordered) categorical variables, many of the latter having more than a few values. There are many missing values.

After removing observations with missing values, each ordinal predictor was cut at its median and coded by two dummy variables; each categorical predictor with k categories was coded by k dummy variables. This resulted in a 6876$\times$50 matrix of 6876 observations on 50 dummy variables (i.e. each of 6876 questionnaires is represented by a vector with 50 components being either 0 or 1).

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

With $t = 0.1$, the 6288 item sets with size not greater than 5 were found. Understanding this large set of rules is itself a challenging data analysis task.

Figure 14.2 the relative frequency of each dummy variable in the data (top) and the association rules (bottom). Prevalent categories tend to appear more often in the rules, for example, the first category in language (English). However, others such as occupation are under-represented, with the exception of the first and fifth level.

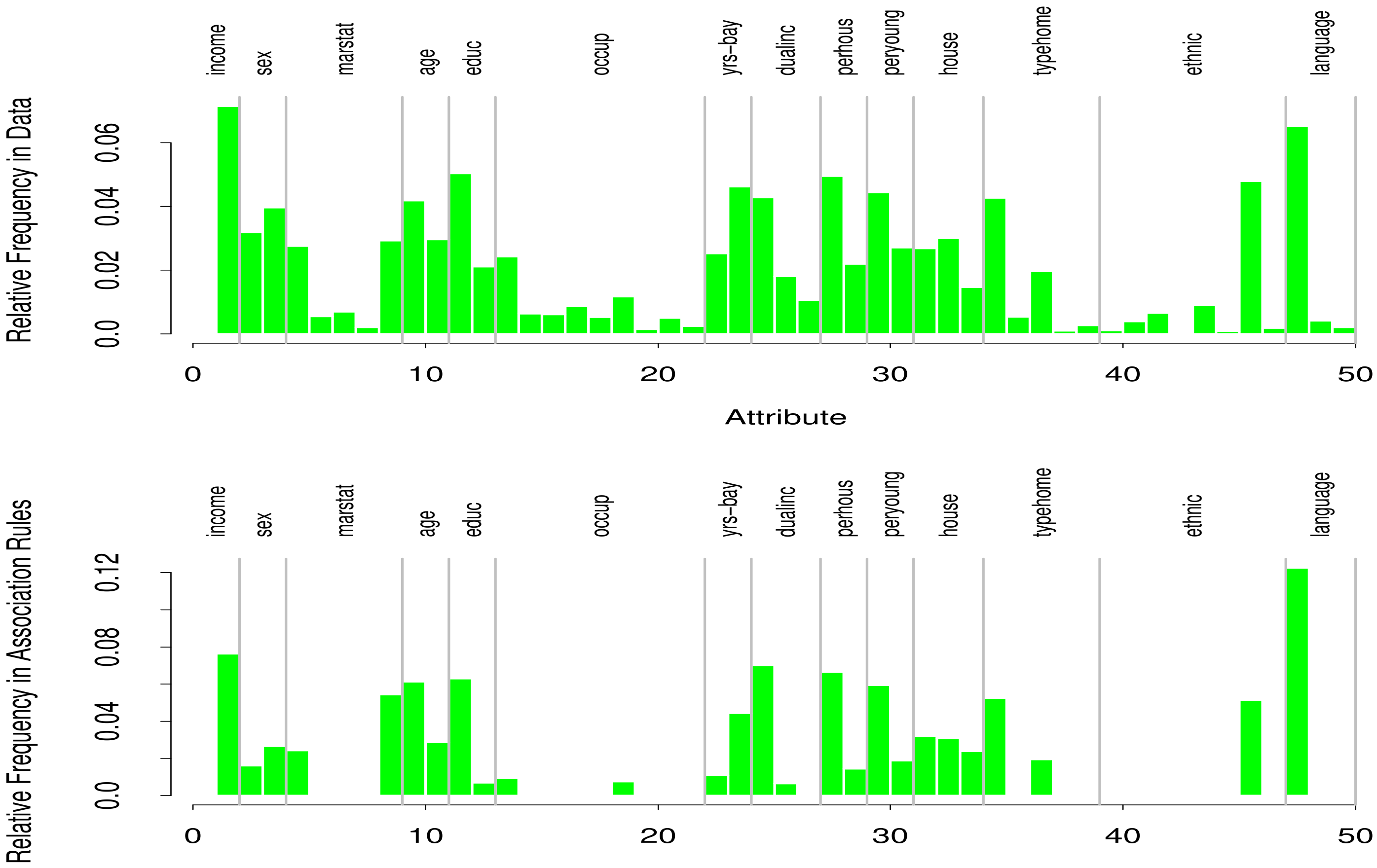


FIGURE 14.2. *Market basket analysis: relative frequency of each dummy variable (coding an input category) in the data (top), and the association rules found by the Apriori algorithm (bottom).*

Three examples of Association Rule found by the Apriori algorithm

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

$\mathcal{A}$	$\mathcal{B}$
number in household =1 number of children = 0	language in home = English

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

$\mathcal{A}$	$\mathcal{B}$
language in home = English householder status = own occupation = professional/managerial	income $\geq$ \$ 40,000

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

$\mathcal{A}$	$\mathcal{B}$
language in home = English	education $\notin$ {college graduate, graduate study}
income $<$ \$ 40,000	
number of children = 0	

Association Rules

There is an intuitive appeal to an association rule because it expresses how tangible products and services group together.

While association rules are easy to understand, they are not always useful.

- Actionable Rules: contains high-quality, actionable information, such as

  Wal-Mart customers who purchase Barbie dolls have a 60 percent likelihood of also purchasing one of three types of candy bars.

- Trivial Rules: already known by anyone at all familiar with the business, may simply the reflection of previous marketing campaigns.

Customers who purchase paint buy paint brushes; oil and oil filters are purchased together, as are hamburgers and hamburger buns, and charcoal and lighter fluid.

Customers who purchase maintenance agreements are very likely to purchase large appliances.

- Inexplicable Rules: have no explanation and do not suggest a course of action.

  When a new hardware store opens, one of the most commonly sold items is toilet bowl cleaners – Discovered for new store openings by a large hardware company, intriguing, giving little insight into consumer behavior or the merchandise or suggest further actions

FAMOUS RULES: BEER AND DIAPERS – A famous story in late 1980s when computers were just getting powerful enough to analyze large volumes of transaction data.

Somewhere in the midwest of America, the fact that beer and diapers are selling together was discovered in the transaction data.

This immediately sets marketing minds in motion to figure out what is happening. A flash of insight provides the explanation: beer drinkers do not want to interrupt their enjoyment of televised sports, so they buy diapers to reduce trips to the bathroom – No, that is not it!

The more likely story: families with young children are preparing for the weekend, diapers for the kids and beer for Dad. Dad probably knows that after he has a couple of beers, Mom will change the diapers.

This is a powerful story. Setting aside the analytics, what can a retailer do with this information? There are two competing views. One says to put the beer and diapers close together, so when one is purchased, customers remember to buy the other one. The other says to put them as far apart as possible, so the customer must walk by as many stocked shelves as possible, having the opportunity to buy yet more items. The store could also put higher-margin diapers a bit closer to the beer, although mixing baby products and alcohol would probably be unseemly.

The story was debunked on 16 April 1998 in an article in *Forbes* Magazine called Beer-Diaper Syndrome

Note. Basket analysis may apply to the data collected at different levels, instead of supermarket transaction data only.

Case Study: Spanish or English

A chain of supermarket in Texas use the summarized data to investigate the differences in shopping patterns between Hispanics and non-Hispanics communities

Business problem: should this chain of supermarkets advertise the same products in Spanish as in English?

Data: The accumulated weekly sales of all products in different supermarket stores. In addition, the percentages of different ethnic groups in the catchment area for each store are also available.

Initial analysis found the association rule: the higher the percentage of African-American population, the lower the Hispanic population, and vice versa — Not interesting!

Rephrased business question: What are the differences in products sold in stores with high Hispanic catchment area versus in a low Hispanic catchment area?

This leads to the division of the stores into three groups: *High Hispanic*, *Mixed* and *Not very Hispanic*. Only use the data from High Hispanic and Not very Hispanic stores

Let $\mathcal{A} = \{\texttt{High Hispanic store}\}$ and $\mathcal{B} = \{\texttt{Not very Hispanic store}\}$.

Define *Hispanicity Preference* for each product $\mathcal{I}$ as

$$C(\mathcal{A} \Rightarrow \mathcal{I}) - C(\mathcal{B} \Rightarrow \mathcal{I}),$$

then sort different products according this Hispanicity Preference score.

Different purchase patterns were discovered. For example, non-Hispanics tend to prefer beef which Hispanics prefer pork; non-Hispanics prefer potato chips and French fries as snacks whereas Hispanics prefer corn chips as snacks.

Ads for the 4th July picnics: hamburgers and potato chips in English, and perhaps sausages and Doritos corn chips in Spanish.

The Apriori algorithm is implemented in R package `arules`. Data set `AdultCUI` within the package contains 48,842 responses for a questionnaire with 15 questions. Since each question will be represented by several items (i.e. binary variables), it is necessary to have a structure which can deal large amounts of sparse binary data in an efficient manner. See the example at the bottom of the help menu under `?AdultCUI`.

More detailed information on `arules`, with a data example illustration, can be found at
`https://cran.r-project.org/web/packages/arules/vignettes/arules.pdf`