Association Rules/Market Basket Analysis

Stephen Vardeman

Analytics Iowa LLC

ISU Statistics and IMSE

Setup

Consider N transaction records, each of which may or may not include each one of items

$$s_1, s_2, \ldots, s_p$$

For two disjoint sets of items

$$\mathcal{S}_1 = \{s_{11}, s_{12}, \ldots, s_{1k_1}\} \quad \text{and} \quad \mathcal{S}_2 = \{s_{21}, s_{22}, \ldots, s_{2k_2}\}$$

consider transactions that

1. include all items in $\mathcal{S}_1$,
2. include all items in $\mathcal{S}_2$, or
3. include all items in $\mathcal{S} = \mathcal{S}_1 \cup \mathcal{S}_2$.

Setup cont.

In applications of this formalism to "market-basket analysis" it is common to call $\mathcal{S}, \mathcal{S}_1$, and $\mathcal{S}_2$ **item sets** and the statement

"the transaction includes all of both item set $\mathcal{S}_1$ and item set $\mathcal{S}_2$"

a **conjunctive rule**. It is then common to further talk about **association rules** of the form

$$\mathcal{S}_1 \Longrightarrow \mathcal{S}_2$$

and to consider quantitative measures associated with them. $\mathcal{S}_1$ is called the **antecedent** and $\mathcal{S}_2$ is called the **consequent** in the rule.

Properties of association rules

Define indicator variables

$$I_{ij} = I\,[\text{transaction } i \text{ includes all of item set } \mathcal{S}_j]$$

for $i = 1, \ldots, N$ and $j = 1, 2$. For the association rule $\mathcal{S}_1 \Longrightarrow \mathcal{S}_2$,

1. the **support** of the rule (also the support of the item set $\mathcal{S}$) is

$$\frac{1}{N} \sum_{i=1}^{N} I_{i1} I_{i2}$$

   (the relative frequency with which the full item set is seen),

2. the **confidence** or **predictability** of the rule is

$$\frac{\sum_{i=1}^{N} I_{i1} I_{i2}}{\sum_{i=1}^{N} I_{i1}}$$

   (the relative frequency with which the full item set $\mathcal{S}$ is seen in the cases that exhibit the smaller item set $\mathcal{S}_1$),

Properties of association rules cont.

3. the "**expected confidence**" of the rule is

$$\frac{1}{N}\sum_{i=1}^{N} I_{i2}$$

(the relative frequency with which item set $\mathcal{S}_2$ is seen in the training cases), and

4. the **lift** of the rule is

$$\frac{\textit{confidence}}{\textit{expected confidence}} = \frac{N\sum_{i=1}^{N} I_{i1}I_{i2}}{\left(\sum_{i=1}^{N} I_{i1}\right)\left(\sum_{i=1}^{N} I_{i2}\right)}$$

(a measure of association).

Interpretations

If one thinks of the N transactions in hand as a random sample from some distribution on item sets or equivalently, a distribution for

$$\mathbf{x} = (I\,[\text{transaction includes } s_1], \ldots, I\,[\text{transaction includes } s_p]) \quad ,$$

lets I_1 stand for the event that all items in $\mathcal{S}_1$ are in the set, I_2 stand for the event that all items in $\mathcal{S}_2$ are in the set, and I stand for the event that all items in $\mathcal{S} = \mathcal{S}_1 \cup \mathcal{S}_2$ are in the set, then

1. the support of the rule is an estimate of $P(I)$,
2. the confidence is an estimate of $P(I_2|I_1)$,
3. the expected confidence is an estimate of $P(I_2)$, and
4. the lift is an estimate of the ratio $P(I_1 \text{ and } I_2) \,/\, (P(I_1) \cdot P(I_2))$.

Looking for informative association rules

Usually one is concerned with rules having large support (so that the estimates can be expected to be reliable). Further, large confidence or lift are of interest, as these indicate that the rule will be useful in understanding how the presence or absence of various items are related in the training data. Standard practice is to identify a large number of promising item sets and association rules, and to make a database of association rules that can be queried as (for example):

> *"Find all rules in which YYY is the consequent that have confidence over 70% and support more than 1%."*

Potentially interesting association rules

We haven't yet addressed where one gets appropriate item sets $\mathcal{S}$ and how one uses them to produce ($\mathcal{S}_1$ and $\mathcal{S}_2$ and) corresponding association rules. In answer to the second of these questions, one might say "consider all $2^{|\mathcal{S}|} - 2$ association rules that can be associated with a given item set." But what are "interesting" item sets $\mathcal{S}$, or how does one find a potentially useful set of such?