The “Apriori” Algorithm

Stephen Vardeman

Analytics Iowa LLC

ISU Statistics and IMSE

The “apriori” algorithm

The so-called "apriori algorithm can be used to produce all item sets $\mathcal{S}$ of support at least t. (These can then be examined to find potentially interesting association rules by breaking them into two pieces $\mathcal{S}_1$ and $\mathcal{S}_2$.) It operates as follows.

1. Pass through all p items

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

   identifying those s_j that individually have support/prevalence

   $$\frac{1}{N} \cdot \# \{i \mid x_{ij} = 1\}$$

   at least t and place them in the set

   $$\mathcal{S}_1^t = \{\text{item sets of size 1 with support at least } t\}$$

The "apriori" algorithm step 2

2. For each $s_j \in \mathcal{S}_1^t$ check to see which two-element item sets

$$\left\{s_j, s_{j'}\right\}_{j' \neq j \text{ and } s_{j'} \in \mathcal{S}_1^t}$$

have support/prevalence

$$\frac{1}{N} \cdot \# \left\{i \,|\, x_{ij} x_{ij'} = 1\right\}$$

at least t and place them in the set

$$\mathcal{S}_2^t = \{\text{item sets of size 2 with support at least } t\}$$

$\vdots$

The "apriori" algorithm termination

m. For each $\left\{ \overbrace{s_j, s_{j'}, \ldots}^{m-1 \text{ entries}} \right\} \in \mathcal{S}^t_{m-1}$ check to see which m-element item sets

$$\{s_j, s_{j'}, \ldots\} \cup \{s_{j^*}\} \text{ for } j^* \notin \{j, j', \ldots\} \text{ and } s_{j^*} \in \mathcal{S}^t_1$$

have support/prevalence

$$\frac{1}{N} \cdot \# \{ i \mid x_{ij} x_{ij'} \cdots x_{ij^*} = 1 \}$$

at least t and place them in the set

$$\mathcal{S}^t_m = \{\text{item sets of size } m \text{ with support at least } t\}$$

This algorithm terminates when at some stage m the set $\mathcal{S}^t_m$ is empty.

Use of the results

Then a sensible set of item sets (to consider for making association rules) is $\mathcal{S}^t = \cup \mathcal{S}^t_m$, the set of all item sets with prevalence in the training data of at least t. Apparently for commercial databases of "typical size," unless t is very small it is feasible to use this algorithm to find $\mathcal{S}^t$.

It is apparently also possible to use a variant of the apriori algorithm to find all association rules based on item sets in $\mathcal{S}^t$ with confidence at least c. This then produces a database of association rules that can be queried by a user wishing to identify useful structure in the database/training data set.

Use of the results cont.

In a more statistical vein, one can adopt from $\mathcal{S}^t$ some consequent of interest $\mathcal{S}^{**} = \{s_1^{**}, s_2^{**}, \ldots, s_l^{**}\}$ and consider modeling of the binary variable

$$I\,[\text{all items in } \mathcal{S}^{**} \text{ are in a transaction}] = \prod_{j \text{ s.t. } s_j \in \mathcal{S}^{**}} x_j$$

on the basis of some non-overlapping set of variables related to an antecedent $\mathcal{S}^*$ (disjoint from $\mathcal{S}^{**}$ belonging to $\mathcal{S}^t$). For example, a natural possibility is to use logistic regression based on the set of variables x_j with $s_j \in \mathcal{S}^*$ to look for items (or sets of items if products of these indicators are employed) that are associated with "large" (or "increased") probabilities of the consequent.