Market Basket

Sonal Ghanshani

Terminologies

- **Items** are the objects that we are identifying associations between. For an online retailer, each **item** is a product in the shop. For a publisher, each **item** might be an article, a blog post, a video etc. A group of items is an **item set**.
- An *itemset* is the set of items a customer buys at the same time.
- **Transactions** are instances of groups of items co-occuring together. For an online retailer, a **transaction** is, generally, a, transaction. For a publisher, a transaction might be the group of articles read in a single visit to the website. (It is up to the analyst to define over what period to measure a **transaction**.) For each **transaction**, then, we have an **item set**.
- **Rules** are statements of the form:

$$\{\text{flour,sugar}\} \Rightarrow \{\text{eggs}\}$$

- The output of a market basket analysis is generally a set of rules, that we can then exploit to make business decisions (related to marketing or product placement, for example).

Terminologies

- The **support** of an item or item set is the fraction of transactions in our data set that contain that item or item set. In general, it is nice to identify rules that have a high support, as these will be applicable to a large number of transactions.

$$\frac{\text{Number of times itemsets are together}}{\text{Total number of transactions}}$$

- The *confidence* is the conditional probability that the items will be purchased together.
- The **confidence** of a rule is the likelihood that it is true for a new transaction that contains the items on the LHS of the rule. (I.e. it is the probability that the transaction *also* contains the item(s) on the RHS.)
- Formally: $\text{confidence}(i_m \Rightarrow i_n) = \text{support}(i_m \cup i_n) / \text{support}(i_m)$

Terminologies

- The **lift** of a rule is the ratio of the support of the items on the LHS of the rule co-occuring with items on the RHS divided by probability that the LHS and RHS co-occur if the two are independent.

$$\text{lift}(i_m \Rightarrow i_n) = \text{support}(i_m \cup i_n) / (\text{support}(i_m) \times \text{support}(i_n))$$

- If lift is greater than 1, it suggests that the presence of the items on the LHS has increased the probability that the items on the right hand side will occur on this transaction.
- If the lift is below 1, it suggests that the presence of the items on the LHS make the probability that the items on the RHS will be part of the transaction *lower*.
- If the lift is 1, it suggests that the presence of items on the LHS and RHS really are independent: knowing that the items on the LHS are present makes **no** difference to the probability that items will occur on the RHS.

Rules

- When we perform market basket analysis, then, we are looking for rules with a lift of more than one. Rules with higher confidence are ones where the probability of an item appearing on the RHS is high given the presence of the items on the LHS.
- It is also preferable (higher value) to action rules that have a high support - as these will be applicable to a larger number of transactions.