Decision rules

Using python and R

Iván Andrés Trujillo
Dayana Castro G.

trujilloiv@javeriana.edu.co
dayana.castro@usco.edu.co

Itemset and binary representation

A itemset is collection of products.

id	itemsets
1	{a,b,c}
2	{x,y}
3	{x,y,z}

id	a	b	c	x	y	z
1	1	1	1	0	0	0
2	0	0	0	1	1	0
3	0	0	0	1	1	1

We can said that a k itemset is a itemset with k elements for instance a 2 itemset could be $\{x, y\}$.

Formal representation

let be I the set of all items, T the set of transactions therefore $T \subseteq I$ and D is database.

α

α is calculated as the number of transaction of a itemset in database. For the itemset X is defined as:

$$\alpha(X) = |t|X \subseteq t, t \in T| \tag{1}$$

The number of transaction in which the itemset it is subset of it. $|.|$ will be used to determine the cardinality (number of elements) in the sets.

Association rule

If a person buy a $item_i$ and $item_j$ then imply that carry out the $item_k$.

$$\{item_i, item_j\} \longrightarrow \{item_k\} \tag{2}$$

in general terms we can said that association rule is a implication:

$$X \longrightarrow Y \tag{3}$$

where X and Y are disjoints itemsets; $X \cap Y = \emptyset$.

Support $s()$

we define that D is dataset, teherefore the support is defined as.

$$s(X \longrightarrow Y) = \frac{\alpha(X \cup Y)}{|D|} \tag{4}$$

$s(\{x, y\})$

id	itemsets
1	{a,b,c}
2	{x,y}
3	{x,y,z}

According to the definition of support, $s(\{x, y\}) = \frac{\alpha(\{x,y\})}{|D|} = \frac{2}{3}$. Given that $\{x, y\}$ appear two times in database, and are three transactions in the dataset.

Confidence

The confidence is therefore:

$$c(x \longrightarrow y) = \frac{\alpha(x \cup y)}{\alpha(x)} \tag{5}$$

Note here that is as a conditional probability, means, the probability of occur x or y given the number of times that occur x. $P(y \mid x)$.

Intuition about support and confidence

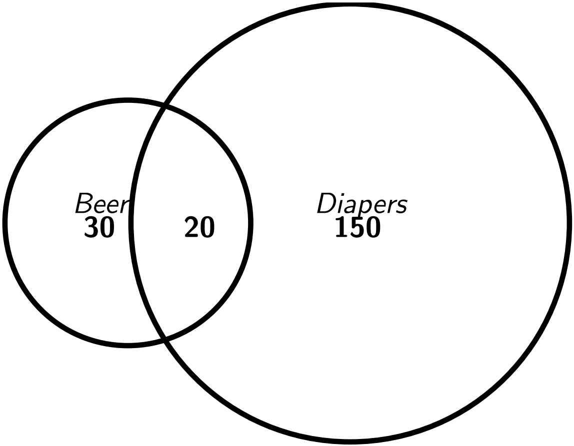
Support is interpreted as the "importance of the rule" in the bussiness environment and confidence it is the reability of the rule.

Confidence

Considerations

In the following rule $X \longrightarrow Y$ could appear a hihger confidence due that the support of the rigth side is higher $s(Y)$ independently of $s(X)$. Prior probability of see it is high.

Biased confidence



The rule $\{Beer\} \longrightarrow \{Diapers\}$ have a confidence of $\frac{20}{50} = 40\%$.

Lift

Lift is defined as

$$l(x \longrightarrow y) = \frac{s(x \longrightarrow y)}{s(x)s(y)} \tag{6}$$

the independent occurrence, note that lift it is measure of actual confidence regard to the expected confidence (the random occurrence of x and y). if lift is equal to one then is not correlated.

The rule is good if improve the decision regarding a random decision. This measure allow us have a insigth of the relevance of the rule.
In other words lift is the increase in the empirical probability of see the consequent given that known the antecedent.

$$l(x \longrightarrow y) = \frac{P(y \mid x)}{p(y)} \tag{7}$$

Suppose that the lift give us $\frac{0.8}{0.5}$ now the probability of carry out y increase of 0.5 to 0.8.

The data mining purpose is find all possible combination of decision rules that satisfy $c(x_i \longrightarrow y_i > limit)$ and also to support, the limits could be differ.
If we have $\{a, b\} \longrightarrow \{c\}$. we search in transactions a, b, c not matter until now the order. now we call **frequent itemset** a the collection of rules that the limits.

Counting

The association rules of items $\{s,t,w,x,y,z\}$ selecting two items $\{s,t\}$ then we can construct:

$$
\begin{aligned}
\{s,t\} &\longrightarrow \{w\} \\
\{s,t\} &\longrightarrow \{x\} \\
\{s,t\} &\longrightarrow \{y\} \\
&\vdots \\
\{s,t\} &\longrightarrow \{w,x\} \\
\{s,t\} &\longrightarrow \{w,y\} \\
&\vdots \\
\{s,t\} &\longrightarrow \{w,x,y,z\}
\end{aligned}
\tag{8}
$$

Note that for the pair there are $\sum_{i=1}^{4}\binom{4}{i}$ possible combinations in the rigth hand.

Counting

In summary, we can select two items in $\binom{6}{2}$ different ways for le left side of $X \longrightarrow Y$, and by each option there are $\sum_{i=1}^{4} \binom{4}{i}$ in the rigth side. Generally we can said that for n items;

$$\underbrace{\{a, b, ..., k\}}_{\text{k items}} \longrightarrow \underbrace{\{h, i, ..., z\}}_{\text{n-k items}}$$

there are $\binom{n}{k} \sum_{i}^{n-k} \binom{n-k}{i}$ decision rules for k items of n.

Counting

The total number of $k-itemset$ that we can construc of n are $\sum_{k=1}^{n}\binom{n}{k}$ therefore the number of total decision rules are considering that $1 \leq k < d$ (Avoiding have empty set in the rigth side).

$$\sum_{k=1}^{n-1}\binom{n}{k}\sum_{i=1}^{n-k}\binom{n-k}{i} \tag{9}$$

To solve this expression take in mind the properties presented in the followig slide:

Brute force approach

Complexity

Proof by binomial theorem;

$$(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^{n-k} y^k \tag{10}$$

Consider the following, $x, y = 1$

$$\sum_{k=1}^{n} \binom{n}{i} = 2^n - 1 \tag{11}$$

$$\binom{n}{0} + \sum_{i=1}^{n-1} \binom{n}{1} + \binom{n}{n} = 2^n \tag{12}$$

Counting

$$\sum_{k=1}^{n-1} \binom{n}{k} \sum_{i=1}^{n-k} \binom{n-k}{i} \tag{13}$$

$$= \sum_{k=1}^{n-1} \binom{n}{k} (2^{n-k} - 1) \tag{14}$$

$$\sum_{k=1}^{n-1} \binom{n}{k} 2^{n-k} - \sum_{k=1}^{n-1} \binom{n}{k} \tag{15}$$

Now we ca stated previously that $\sum_{k=1}^{n-1}\binom{n}{k} = 2^n - 2$.
Note that

$$\sum_{i=1}^{n-1}\binom{n}{k}2^{n-k} = \sum_{i=1}^{n-1}\binom{n}{k}2^{n-k}1^k \tag{16}$$

$$3^n = \binom{n}{0}2^{n-0}1^0 + \sum_{k=1}^{n-1}\binom{n}{k}2^{n-k}1^k + \binom{n}{n}2^{n-n}1^n \tag{17}$$

$$3^n = 2^n + \sum_{k=1}^{n-1}\binom{n}{k}2^{n-k}1^k + 1 \tag{18}$$

Reemplacing in

$$\sum_{k=1}^{n-1}\binom{n}{k}2^{n-k} - \sum_{k=1}^{n-1}\binom{n}{k} \tag{19}$$

$$3^n - 2^n - 1 - (2^n - 2) = 3^n - 2^{n+1} + 1 \tag{20}$$

The total number of association rules with n items.

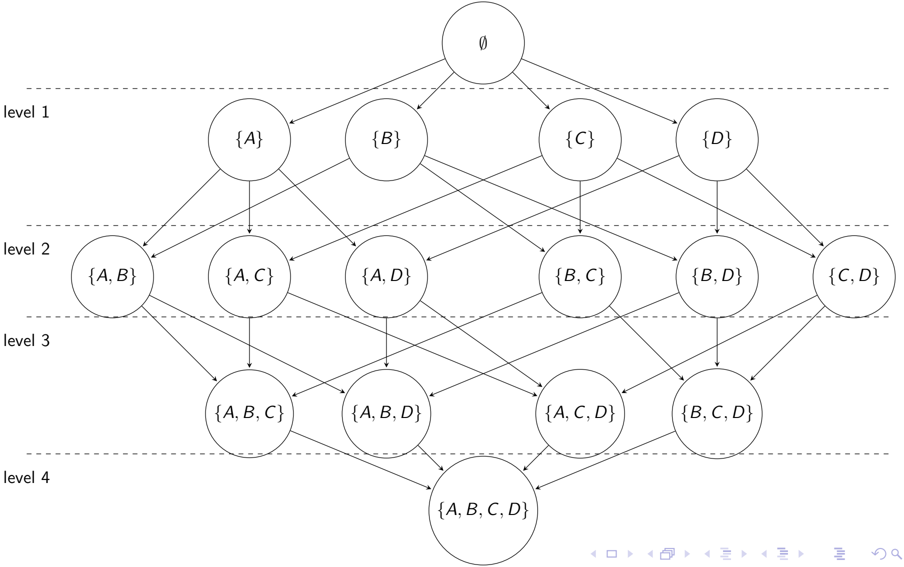
strategies

Principle of monotocity

$$\forall X, Y : (X \subseteq Y) \Rightarrow s(X) \geq s(Y) \tag{21}$$

In the following picture we can see, that if unitary set(level 1); for instance $\{A\}$ not is frequent then, itemset of the another levels. In other words; if a itemset not is frequent, then its superset either.

Lattice structure



A priori algorithm

keep the itemset that satisfied $s(itemset) > minsup$ (You can consider 1-kitemset). After generate the possible association rules, and assesst that $c(itemset) > minconf$. Repeat the process until there are not itemsets.

```
%load_ext rpy2.ipython
```

Lexicographical order

Two important arrays

C_i and L_i iterating over the algorithm C_i keep the possible $i - itemsets$ and L_i keep the frequent $i - itemset$, namely whose support is greater or equal to *minsup*.
why??

Rules

Now the rules are constructed in relation with the combinatios that fill the *minsuport* in the generation of candidates. for instance $\{x, y\}$ is a candidate in c_2 then we could get:

$$\begin{aligned} \{x\} &\longrightarrow \{y\} \\ \{y\} &\longrightarrow \{x\} \end{aligned} \tag{22}$$

for $\{w, x, y\}$ we could get:
$\{w\} \longrightarrow \{x\}$,$\{w, x\} \longrightarrow \{y\}, \ldots, \{x, y\} \longrightarrow \{x\}$ note that finally the rules also have a *minconf* treshold.

References

- Zhang, C., Zhang, S. (Eds.). (2002). Association rule mining: models and algorithms. Berlin, Heidelberg: Springer Berlin Heidelberg.
- Tan, P. N., Steinbach, M., Kumar, V. (2016). Introduction to data mining. Pearson Education India.