Machine Learning: Homework 7

Laurent Hayez

November 14, 2016

Exercise 1. *Build the co-occurrence matrix for the observations.*

Calculate support, confidence, completeness, lift, and leverage for the following rules.

- *Apple* → *Donut*
- *Apple* → *Onion*
- *Sugar* → *Yoghurt*
- *Donut* → *Onion*
- *Donut* → *Raspberry*
- *Onion* → *Raspberry*

Explain these measures (why they are useful, what ranges of numbers they can return, what the values mean).

Use the Apriori algorithm to find frequent item sets. We are only interested in item sets having a support value of at least 50%.

Solution. The co-occurrence matrix is computed in `ML_hayezl_homework7.xlsx`, and is given in Table 1.

The support, confidence, completeness, lift and leverage were computed with the following formulas

$$\text{support} = \frac{N_{\text{both}}}{N_{\text{total}}},$$

$$\text{confidence} = \frac{N_{\text{both}}}{N_{\text{left}}},$$

$$\text{completeness} = \frac{N_{\text{both}}}{N_{\text{right}}},$$
$$\text{lift}(L \to R) = \frac{\text{support}(L \cup R)}{\text{support}(L) \cdot \text{support}(R)},$$
$$\text{leverage}(L \to R) = \text{support}(L \cup R) - (\text{support}(L) \cdot \text{support}(R)).$$

The results for the different rules are displayed in Table 2.

- The **support** of a set I measures the proportion of baskets in which I appears. Hence $\text{support}(I) \in [0,1]$ (or $]0,1]$ to be more precise, because considering items that never appear is not interesting). This measure is useful to know if I appears often or not.
- The **confidence** of a rule measures how reliable a rule is, or in other words, if the rule is $L \to R$, it measures the proportion of the appearance of R when L appears. This measure takes values in $]0,1]$. It is useful to know if some items are correlated, or often bought together.
- The **completeness** of a rule measures the proportion of times L and R happen with respect to R. This measure takes values in $]0,1]$. If the measure is close to 1, it means that R is very correlated with L, as it means that R appears almost always when L appears. It is useful to determine if an item is bought when another item is bought.
- The **lift** of a rule $L \to R$ measure how the appearance of two items at the same time differ from how they would appear if L and R were statistically independent. If L and R are independent, the expectation of L and R appearing together is $|L| \cdot \text{support}(R)$, and we need to compare this to the actual number of time they appear together, i.e., $|L \cup R|$. This measure takes values in $\mathbb{R}_{>0}$, but the interesting values are when $\text{lift}(L \to R) > 1$ because this tells us that L and R are correlated, in the sense that when L is bought, R is also bought.
- The **leverage** of a rule $L \to R$ compares the support of $L \cup R$ and L, R. It gives a measure that tells us whether the elements are associated "by chance". This measure takes values in $R_{>0}$. It measures the proportion of times items are bought together more than if we had chosen them randomly.

We start by creating $L_1 = \{\{\text{Apple}\}, \{\text{Donut}\}, \{\text{Ice-cream}\}, \{\text{Onion}\}, \{\text{Raspberry}\}\}$ which consists of the items that have a support at least 50%. From this set we create

C_2 which consist of the 10 possible unordered pairs of items. We keep the pairs that have a support greater than 50% and we create

$$L_2 = \{\{\text{Apple, Donut}\}, \{\text{Apple, Ice-cream}\}, \{\text{Apple, Onion}\}, \{\text{Apple, Raspberry}\}, \{\text{Donut, Onion}\}, \{\text{Onion, Raspberry}\}\}.$$

From L_2 we can form $C_3 = \{\{\text{A, D, I}\}, \{\text{A, D, O}\}, \{\text{A, D, R}\}, \{\text{D, O, R}\}\}$ where A = Apple, D = Donut, I = Ice-cream, O = Onion, R = Raspberry. The only set with support greater than 50% is {Apple, Donut, Onions} =: L_3, and we can't form any other set. □

Table 1: Co-occurrence matrix for the observation

Co-occurrences	Apple	Donut	Ice-cream	Mango	Onion	Raspberry	Sugar	Tomato	Yoghurt
Apple	6	4	3	2	5	4	2	1	2
Donut	4	4	2	0	3	2	2	1	1
Ice-cream	3	2	3	1	2	2	1	0	1
Mango	2	0	1	2	2	2	0	0	1
Onion	5	3	2	2	5	4	1	1	1
Raspberry	4	2	2	2	4	4	1	0	1
Sugar	2	2	1	0	1	1	2	0	1
Tomato	1	1	0	0	1	0	0	1	0
Yoghurt	2	1	1	1	1	1	1	0	2

Table 2: Support, confidence, completeness, lift and leverage of the different rules

Rules	Apple → Donut	Apple → Onion	Sugar → Yoghurt
Support	0.666666667	0.833333333	0.166666667
Confidence	0.666666667	0.833333333	0.5
Completeness	1	1	0.5
Lift	1	1	1.5
Leverage	0	0	0.055555556
Rules	**Donut → Onion**	**Donut → Raspberry**	**Onion → Raspberry**
Support	0.5	0.333333333	0.666666667
Confidence	0.75	0.5	0.8
Completeness	0.6	0.5	1
Lift	0.9	0.75	1.2
Leverage	-0.055555556	-0.111111111	0.111111111