

Finding Meaningful Associations in Retail Data

M medium.com/weekly-data-science/finding-meaningful-associations-in-retail-data-6ababd031171

June 21, 2018



What is an association rule?

An *association rule* implies that a particular item is likely to occur given the presence of some itemset.

An *association rule* implies that a particular item is likely to occur given the presence of some itemset.

Let I be the itemset {eggs, flour}. Let j be the item {milk}. Then $I \rightarrow j$ is an association rule that implies that {milk} is likely to occur in a shopping cart if {eggs, flour} occurs.

On its own, this association rule doesn't tell us anything. We want to measure how significant or important our rule is. After all, maybe the appearance of milk is not affected by the presence of eggs and flour. Or maybe shoppers are *less likely* to buy milk when they buy eggs and flour.

We need some way of capturing this information.

Confidence

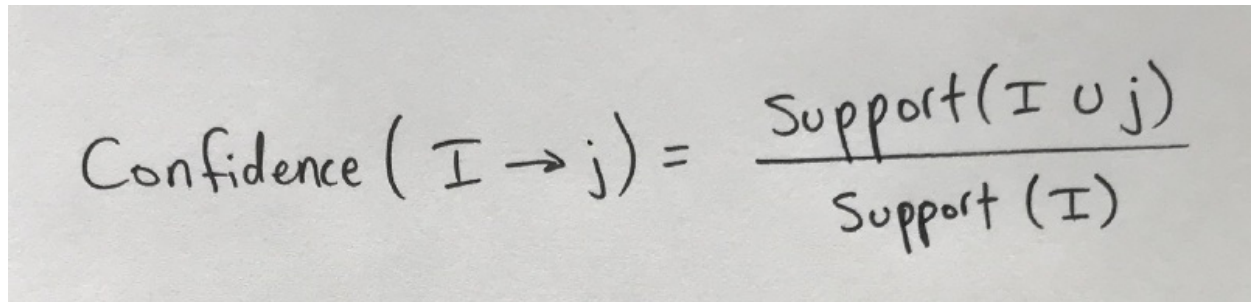
The *confidence* of an association rule is an important building block. It represents the ratio between the number of times that I and J have appeared together and the number of times that I has appeared on its own.

Confidence essentially answers the question: *Out of all the times that people bought eggs and flour, how many times did they also buy milk?* Or, more precisely: *What proportion of the time?*

Confidence essentially answers the question: *Out of all the times that people bought eggs and flour, how many times did they also buy milk?*

Mathematically, we can say:

$$\text{Confidence}(I \rightarrow J) = \text{Support}(I \cup J) / \text{Support}(I)$$

A photograph of a piece of paper with the formula for confidence written in black ink. The formula is: Confidence (I → j) = Support (I ∪ j) / Support (I). The handwriting is clear and legible.

Definition of confidence

In our example, $I \cup J$ is the set {eggs, flour, milk}.

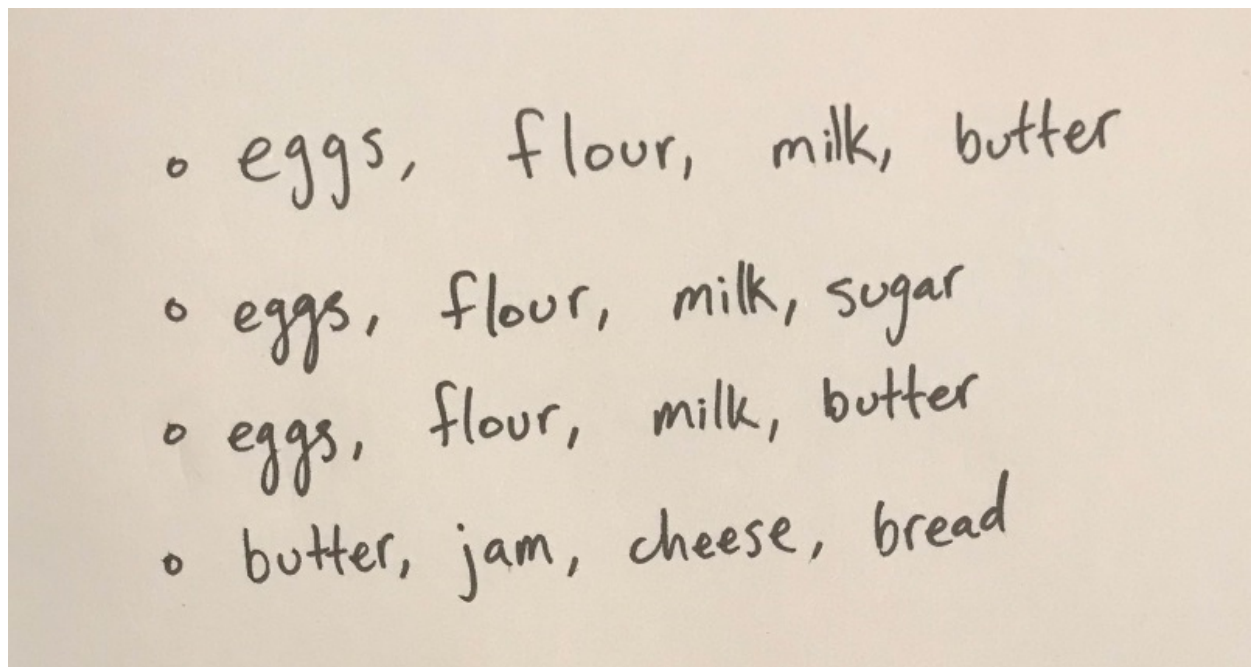
Note that every time the set {eggs, flour, milk} occurs, the set {eggs, flour} also occurs. This means that $\text{Support}(I) \geq \text{Support}(I \cup J) \geq 0$, since supports must be greater than or equal to zero.

Thus, we know that $1 \geq \text{Confidence}(I \rightarrow J) \geq 0$.

Practice with Confidence

Let's work through an example together.

What is the confidence of the association rule {eggs, flour} \rightarrow {milk} in the dataset below?



First, notice that:

$$\text{Support}(\{\text{eggs, flour}\}) = 3$$

and

$$\text{Support}(\{\text{eggs, flour, milk}\}) = 3$$

Next, calculate the confidence:

$$\text{Confidence}(\{\text{eggs, flour}\} \rightarrow \{\text{milk}\}) = \text{Support}(\{\text{eggs, flour, milk}\}) / \text{Support}(\{\text{eggs, flour}\})$$

$$\text{Confidence}(\{\text{eggs, flour}\} \rightarrow \{\text{milk}\}) = 3 / 3 = 1$$

Nice! Our rule has the highest possible confidence value. That must be good, right? ...
Actually, not quite.

What if {milk} occurs frequently independently of whether {eggs, flour} is present? Then we might observe a large confidence value despite the fact that I and J are independent.

What if people always buy milk?

Consider the case where every shopping cart contains milk.

Remember, confidence answers the question: *What proportion of the time that I appears does J also appear?*

If people always buy milk, the answer will be 1. We can clearly see that in this case, buying eggs and flour does not influence people's decision to purchase milk, so this confidence score is not particularly helpful.

Interest

The notion of *interest* can help us here. It extends the idea of confidence by subtracting out the proportion of baskets in which J appears.

$$\text{Interest}(I \rightarrow j) = \text{Confidence}(I \rightarrow j) - \text{Pr}(j),$$

$$\text{where } \text{Pr}(j) = \text{Support}(j) / (\text{Num. Baskets})$$

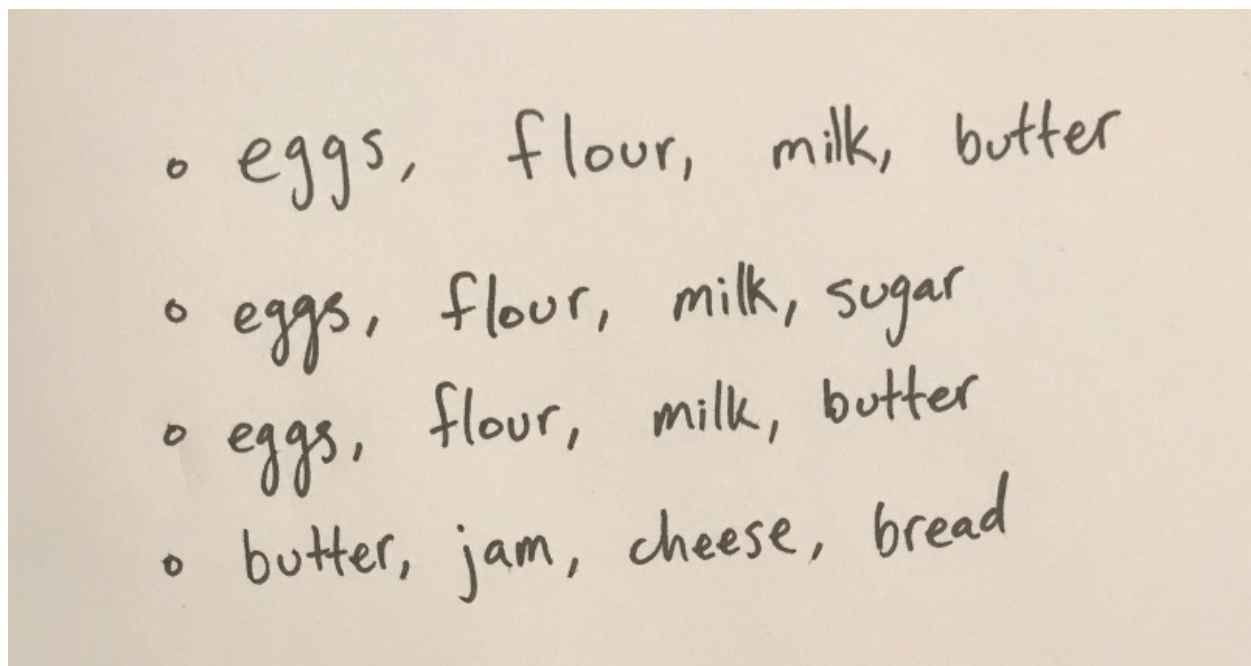
By doing this, we account for the fact that j may occur independently of I .

Since confidence is a number in the range $[0, 1]$ and the proportion of baskets in which j occurs is also a number in the range $[0, 1]$, the interest of an association rule is in the range $[-1, 1]$.

When an association rule has an interest close to 0, it indicates that the presence of I does not imply much about the presence of j . When a rule's interest has an absolute value that is relatively large (typically above 0.5 or so), it indicates that this is a meaningful association in the dataset. Negative interest indicates that the presence of I discourages the presence of j , and positive interest indicates that the presence of I encourages the presence of j .

Practice with Interest

What is the interest of the association rule $\{\text{eggs, flour}\} \rightarrow \{\text{milk}\}$ in the dataset below?
(same dataset as before)



We know from before that the confidence of this rule is 1. We can see that $\{\text{milk}\}$ shows up in 3 out of the 4 baskets, so $\text{Pr}(j) = 3/4$, or 0.75.

Thus, $\text{Interest}(\{\text{eggs, flour}\} \rightarrow \{\text{milk}\}) = 1 - 0.75 = 0.25$.

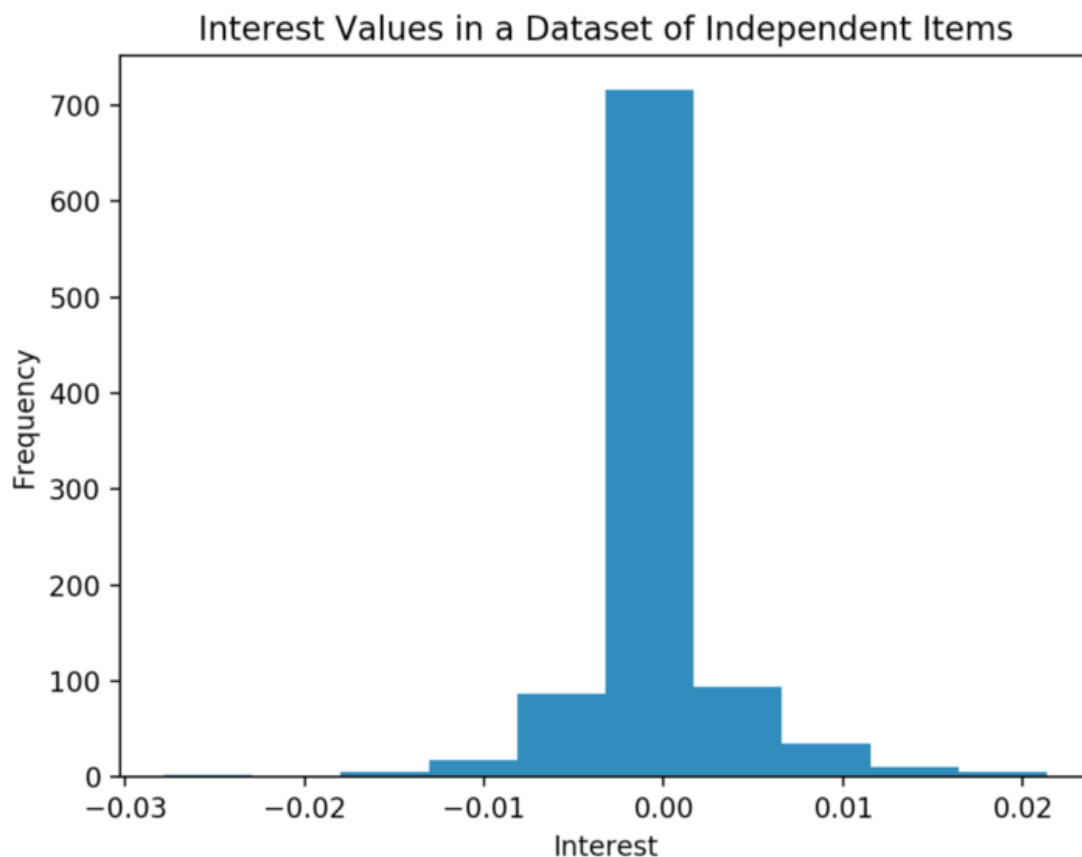
Note that if milk had appeared in every basket, the rule would have an interest of 0.

Thus, we can see that although our rule had high confidence, it's not actually a meaningful association in the dataset, as indicated by its low interest.

A closer look at one of the example datasets

If we look at the example dataset with 100,000 baskets and 5,000 possible items, we'll see that almost all of the association rules have an interest of zero. This is to be expected,

since I generated the dataset using [this script](#), in which items are added to baskets independently of each other.



As expected, these are a bunch of meaningless associations.

Despite the fact that items are independent of each other, we can see that there are some association rules with non-zero interest scores. This happens purely by chance and does not imply that the rules are meaningful. That's why we need to ensure that rules have relatively high interest scores before deeming them meaningful associations.

Want to practice your skills?

Try finding meaningful associations in the example datasets I've set up [on Github](#), or check out [Instacart's Market-Basket Analysis Challenge](#) on Kaggle.

Coming Soon

I'll be writing a quick post detailing why the expected interest of $I \rightarrow j$ is 0 when I and j are independent. Keep an eye out!