Data Analytics
VC 607.916 & 607.926

Dmitri Blueschke

Department of Economics

Summer Term 2023

Chapter 3 – Association rule learning

Association Rule Learning

Association rule learning is a popular, unsupervised learning technique for discovering interesting relations between variables based on transactions involving them in large databases. As it is often used for identifying shopping patterns, it is also known as *market basket analysis*.

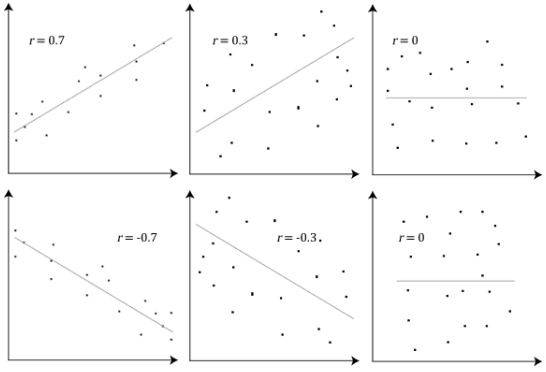
Example: the rule [o n i o n s , p o t a t o e s] $\Rightarrow$ [b u r g e r]

- If a customer buys onions and potatoes together, they are likely to also buy hamburger meat.
- Such information can be used as the basis for decisions about marketing activities such as, e.g., promotional pricing or product placements but also for logistical reasons, web usage mining and so on.

* Note: implication here is *co-occurrence* and not causality.

Association Rule Learning vs Correlation

Statistical relation between two numerical variables: correlation coefficient



Association rule learning:

- nonnumerical variables
- several variables

Association Rule Analysis in Recommendation Systems

- Recommendation systems use association rule analysis to decide on which things you might be interested in (e.g. Netflix). Netflix monitors every interaction with their app and then they use this information to produce customized content for everyone. Based upon your watch history they recommend movies you might like.
- According to Netflix, more than 80% of watched content is based on the service's recommendations.

* These rules does not necessarily extract users preference, but rather find relationships between set of elements present in different entries in data-sets.

Association Rule Learning

A rule is a notation that represents which items are frequently *bought* with what items.

Consider item sets A and B in a set of T transactions of a given database. An association rule $A \rightarrow B$ describes that item set B is purchased if item set A is purchased. A is called antecedent or left-hand-side (LHS) and B consequent or right-hand-side (RHS).

In order to select interesting rules from the set of all possible rules three parameters are used:

- **Support**: It indicates how frequently an item set appears in the data set.
- **Confidence**: It says how likely item set B is purchased when item set A is purchased.
- **Lift**: It says how likely item set B is purchased when item set A is purchased while controlling for how popular item set B is.

Support

Support (*supp*) is the fraction of the total number of transactions in which the item set occurs.

For item sets A and B we calculate:
$frq(A)$ number of transaction including A, $frq(B)$ number of transaction including B, $frq(A, B)$ number of transaction including A and B (T total number of transactions).

$$supp(A) = \frac{frq(A)}{T},$$
$$supp(B) = \frac{frq(B)}{T},$$
$$supp(A, B) = \frac{frq(A, B)}{T}.$$

Support

- If an item set happens to have a very low support, we do not have enough information on the relationship and hence no conclusions can be drawn from such a rule.
- Thus, one might want to consider only the item sets which occur at least in x% of transactions, e.g., minimal support = 0.01.

$\rightarrow$ example in R

Confidence

Confidence (*conf*) is an indication of how often the rule has been found to be true. Given the inclusion of a certain item set, how confident are we that another specific item set was bought with it.

The confidence value of a rule, $A \rightarrow B$, is the proportion of the transactions that contains item set A which also contains item set B.

$$conf(A \rightarrow B) = \frac{supp(A, B)}{supp(A)}$$

Note: Confidence can be interpreted as an estimate of the conditional probability $P(E_B|E_A)$, the probability of finding the RHS of the rule in transactions under the condition that these transactions also contain the LHS.

Confidence

$$conf(A \rightarrow B) = \frac{supp(A, B)}{supp(A)}$$

Example: Consider the following association rule:
computer $\rightarrow$ *antivirus software* with confidence = 50%.
A confidence of 50% means that 50% of the customers who purchased a computer also bought the software.

$\rightarrow$ example in R

Support and Confidence

Consider the following association rule:
computer $\rightarrow$ *antivirus software* [support = 3%; confidence = 50%]

A support of 3% for this association rule means that 3% of all the transactions under analysis show that computer and antivirus software are purchased together. A confidence of 50% means that 50% of the customers who purchased a computer also bought the software.

Typically, association rules are considered interesting if they satisfy both a minimum support threshold and a minimum confidence threshold.

$\rightarrow$ example in R

Confidence

Consider 'retail' example: What about *milk* $\rightarrow$ *dog food*? Confidence for this rule is also relatively high since *milk* is such a frequent item and is present in nearly every other transaction.

* It does not really matter what you have as RHS. The confidence for an association rule having a very frequent LHS will usually be high.

But we know intuitively that these two products have a weak association and there is something misleading about this high confidence value.

$\Rightarrow$ Considering just the value of confidence limits our capability to make any business inference.

$\Rightarrow$ **Lift** is introduced to overcome this challenge.

Lift

Lift (*lift*) controls for the support of RHS while calculating the conditional probability of occurrence of RHS given LHS.

Lift is the rise in probability of having B on the cart with the knowledge of A being present over the probability of having B on the cart without any knowledge about presence of A. It says how likely item set B is purchased when item set A is purchased while controlling for how popular item set A is.

Lift

Mathematically: Lift is the ratio of the observed support to that expected if A and B were independent or equivalently the ratio of the confidence of the rule to the expected confidence of the RHS item set by independence.

$$lift(A \rightarrow B) = \frac{conf(A \rightarrow B)}{expected\_confidence\_of\_B}$$

with $$expected\_confidence\_of\_B = \frac{frq(B)}{T} = supp(B).$$

Thus:

$$lift(A \rightarrow B) = \frac{conf(A \rightarrow B)}{supp(B)} = \frac{supp(A, B)}{supp(A) \times supp(B)}$$

Lift

Consider $lift(milk \rightarrow cookies)$ in the 'retail' example:

- Probability of having cookies on the cart with the knowledge that milk is present (i.e. $conf(milk \rightarrow cookies)$): $3/9 \approx 0.33$.
- Probability of having cookies on the cart without any knowledge about milk: $4/10 = 0.4$.
- These numbers show that having milk on the cart actually reduces the probability of having cookies on the cart to 0.33 from 0.4!
- This will result in $lift(milk \rightarrow cookies) = \dfrac{0.33}{0.4} \approx 0.83$.

Lift

- If lift $= 1 \Rightarrow$ the possibilities of occurrence of LHS (A) and RHS (B) are independent of each other.
- If lift $< 1 \Rightarrow$ the occurrence of LHS has negative effect on occurrence on RHS and vice versa.
- If lift $> 1 \Rightarrow$ the two occurrences are dependent on one another, and these rules are very useful in determining the RHS in latter cases. It also lets us know to what extent the occurrences are dependent on one another.

Association Rule Learning

- The association rule learning includes a three-step process:
  - to generate an item set like (Bread, Egg, Milk)
  - to generate a rule from each item set like {Bread $\rightarrow$ (Egg, Milk)}, {(Bread, Egg) $\rightarrow$ Milk} etc.
  - to mine the data set using support, confidence and lift parameters.
- Problems by calculating support, confidence and lift (standard methods):
  - risk of finding many spurious associations, i.e. they can occur by-chance
  - mining of rules from large data set can be computationally expensive/inefficient (number of the rules grows exponentially in the number of items).
- Another approach: Algorithmic methods

Association Rule Algorithms

Many algorithms for generating association rules have been proposed.
Some well-known algorithms are:

- Apriori
- Eclat
- FP-Growth

Apriori Algorithm

- Apriori algorithm is the most popular algorithm used for association rule mining. The objective is to find subsets that are common to at least a minimum number of the item sets. A frequent item set is an item set whose support is greater than or equal to minimum support threshold.
- The Apriori property is a downward closure property, which means that all nonempty subsets of a frequent item set must also be frequent.
- Apriori algorithm uses a bottom-up approach; and the size of frequent subsets is gradually increased, from one-item subsets to two-item subsets, then three-item subsets, and so on. Groups of candidates at each level are tested against the data for minimum support.
- Agarwal, R., & Srikant, R. (1994). Fast algorithms for mining association rules. In Proc. of the 20th VLDB Conference (Vol. 487).

Apriori Algorithm

Apriori algorithm in R

Application Areas of Association Rule Learning

Application areas of association rule mining:

- **Market basket analysis:** This is the most typical example of association rule learning as data is easily collected in most supermarkets and online shops. Association rule mining can help to:
  - improve cross-selling (online but also physical trough store layout/product placement)
  - optimize logistics and warehousing
  - improve targeted advertising
  - set up recommendation systems
  - detect fraud
  - ...

Application Areas of Association Rule Learning

- **Medical diagnosis:** Association rules in medical diagnosis can be useful for assisting physicians for curing patients. Using relational association rule mining, we can identify the probability of the occurrence of illness concerning various factors and symptoms. Further, using learning techniques, this interface can be extended by adding new symptoms and defining relationships between the new signs and the corresponding diseases.
- **Census data:** Every government has tonnes of census data. This data can be used to plan efficient public services (education, health, transport) as well as help public businesses (for setting up new factories, shopping malls, and even marketing particular products). It is helpful in supporting sound public policy and bringing forth an efficient functioning of a democratic society.