# Machine Learning 1 Week 9 Assignment:

# Association rules

Mihaly Hanics, MS in Social Data Science

Task 1: Answer questions

1) What is the primary goal of ARM (Association Rule Mining)?

To find logical (sort of) relationships, a.k.a. association rules between data.

2) Define the terms 'antecedent' and 'consequent' in the context of association rules. Provide a simple example to illustrate these terms.

If we see that some set of items, objects, data etc. appearing together in the data commonly apply some other set of items appearing too, then we can write this as L → R, where L is the first set of items, and R is the second set, meaning if all items in L are "present", then probably items in R are also present. Here, the left side set L is the antecedent, whilst the right side set R is the consequent.

3) Explain the concept of support in Association Rule Mining. What is it's range of values, what does a high or low support value indicate?

Support of an itemset is the amount of times the itemset appears in the database as "one set" (so the amount of rows where all items in the set are present, in the same row), divided by the total amount of data. The range of values is the closed interval of [0,1], it is 0 when the the itemset has no instances in the data, and 1 when all instances include the itemset. For a rule L → R, the support value is the fraction of the instances where both L and R are present. High support indicates that an itemset appears in a large fraction of the data, and low support means that it only appears for a small percentage.

4) What is confidence, it's range of values, and what does a high or low confidence value indicate?

Confidence for an association rule L → R is the number of instances where L and R appear in the dataset divided by the number of instances where L appears. Again, can be anything between 0 and 1 (well, any rational number), a high value indicates that when L is present, commonly R also present, whereas a low value indicates that when L is present then R is usually not present.

5) What are the main steps of the Apriori algorithm?

Firstly, find all supported items (sets of size one, so only one item), the set containing these sets (itemsets of size 1) will be $L_1$. Then in each step, we first create a candidate list by assigning all (unincluded) items separately, once, to all groups in $L_{k-1}$ and then from this set, we cut the ones that have low support. We do this till for some $k$, $L_k$ is empty.

6) What is a “frequent itemset”?

It is the same as supported itemset: a set of items which appear more frequently in the database than a given value *minsup.*

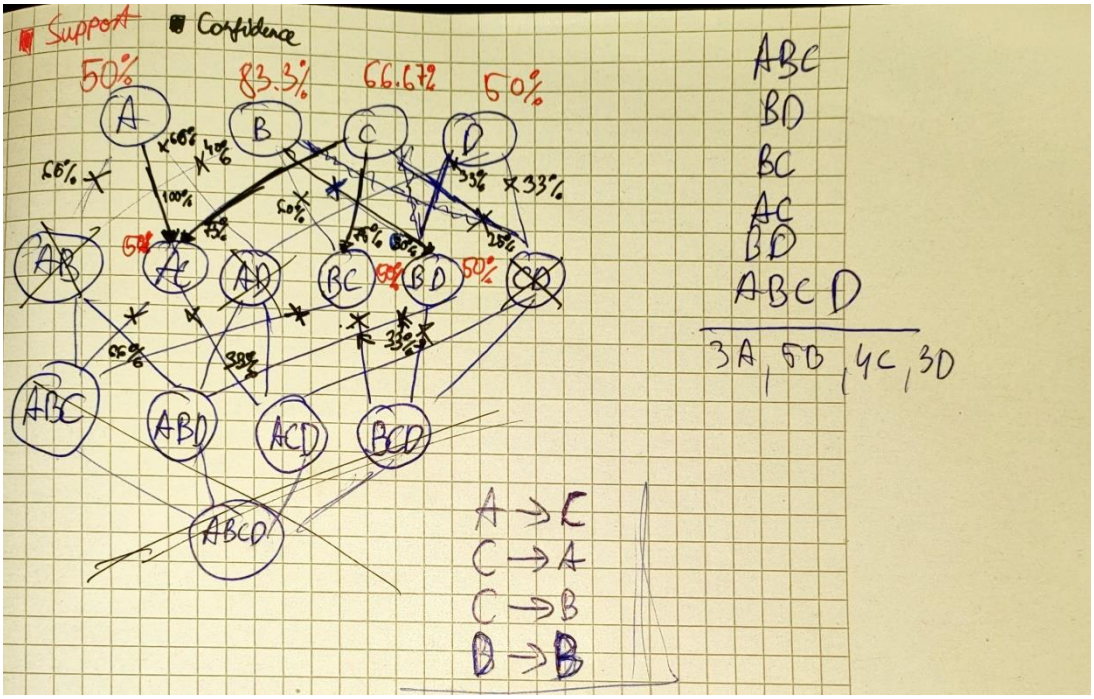
7) What is the role of “support-based pruning” when generating itemsets? How does it contribute to the efficiency of the mining process?

To create itemsets that are frequent enough in the database. This way we can quickly, step by step disregard itemsets that are not common, making computation much faster. (Otherwise we’d have to look at all itemsets of $n$ elements, that is $2^n$.

8) What is the advantage of rule interestingness measures Lift and Leverage as compared to Confidence?

They are more statistical measures, describing how “outstanding” the occurrence of a union of two sets is compared to if they were independent sets. They are better at suggesting statistical connections (rules).

Task 2:




Task 3: I have found that in the Food dataset, only at minimum support 0.01% does the algorithm start to give associations. Some of these even have 100% confidence, for example: Cheese, Muffins, TV Dinner →Dried Fruit (surprising), or Personal Hygiene, Cooking Oil, Fresh Fruit → Fresh Vegetables. In the Congressional Voting Records dataset, the best find was physician-fee-freeze=n, aid-to-nicaraguan-contras=y → adoption-of-the-budget-resolution=y, with support of around 4.5%. I found the 4% minimal support to be appropriate.

I’ve found that for dense data, we rather use “y/n”, and not amount of transactions, as is the case for transactional data.