CS-E4650 Methods of Data mining
Home assignments 3

Deadline Sun 14.11.2021 23:30. Maximum 90 points + 10 extra points.
Submit your solutions early, since MyCourses can get stuck, if many people submit simultaneously. If your solutions are in before 24:00, there is no penalty, but after that **-10% penalty**. The ultimate deadline (with -10% penalty) is Tue 16.11. 23:30.

For each task, write a **pdf report** where you present everything **clearly and compactly**. The idea is that your solution is understandable based on the report alone, including all middle steps. **Note that notebooks are not accepted as reports.** By default, the source code is not opened or executed to see if you had actually done something correctly, if it is missing or incorrect in the report. Put possible source code into a **zip** package that is attached (you can include the pdf also in zip, it doesn't harm, but **remember to submit it also separately**). The main reason for submitting the source code is checking that you have done the work independently and using allowed functions (if asked to implement something on your own, without library functions). Include your name and student number in the report and code header.

There is an optional latex template for the report, including examples, in the MyCourses page, section Assignments.

Note This time we have only three tasks but they are larger and give more points. In addition, there is one extra task for those who have extra time and interest.

---

1. **(max 30p)** In this task, you should evaluate association rules.

   Let us consider a database consisting of $n = 1000$ patients (50% female, 50% male), 30% of them with Alzheimer's disease (AD). The database contains information on patients and their life style like smoking status, diet, use of natural products, stress and education levels. Table 1 lists some candidate rules related to AD. The required equations for mutual information are given in Appendix 1. (Note that we will use $n \cdot MI$ because it is easier to interpret.)

   You can present the calculated measures in one table in your report. The latex source of Table 1 (ADruletable.tex) is given on the homepage, if you want to utilize it in your report or scripts. Remember to tell how you made calculations and attach possible source code. (Scripts or a programmable calculator are recommended.)

   a) (6p) Calculate leverage or lift values for all rules. Prune out rules that do not express positive statistical dependence.

   b) (6p) Evaluate mutual information MI of remaining rules (report $n \cdot MI$ values) and prune out rules where $n \cdot MI < 1.5$ (i.e., $MI < 0.0015$).

   c) (8p)Evaluate overfitting among remaining rules using value-based interpretation and conditional mutual information MI_C: Rule $\mathbf{X} \rightarrow C{=}c$ is pruned out if there exists some $\mathbf{Y} \subsetneq \mathbf{X}$, such that for $\mathbf{X} \rightarrow C{=}c$ either $P(C{=}c|\mathbf{Y}) \geq P(C{=}c|\mathbf{X})$ or the improvement is not sufficient, $n \cdot MI_C < 0.5$ (i.e., $MI_C < 0.0005$).

   d) (4p) What are your conclusions based on the remaining association rules? What would you recommend to do if one would like to avoid Alzheimer's disease?

Table 1: Candidate rules $\mathbf{X} \rightarrow C{=}c$ related to C = Alzheimer's disease. $fr_X = fr(\mathbf{X})$, $fr_{XC} = fr(\mathbf{X}C)$.

num	rule	fr_X	fr_{XC}
1	smoking $\rightarrow$ AD	300	125
2	stress $\rightarrow$ AD	500	150
3	higheducation $\rightarrow \neg$ AD	500	400
4	tea $\rightarrow \neg$ AD	342	240
5	turmeric $\rightarrow \neg$ AD	2	2
6	female $\rightarrow \neg$ AD	500	352
7	female, stress $\rightarrow$ AD	260	100
8	berries, apples $\rightarrow$ AD	120	32
9	smoking, tea $\rightarrow$ AD	240	100
10	smoking, higheducation $\rightarrow$ AD	80	32
11	stress, smoking $\rightarrow$ AD	200	100
12	female, higheducation $\rightarrow \neg$ AD	251	203

e) (6p)Give example rules (among all 12 rules) that demonstrate the following things. Explain your choices briefly (why they demonstrate something). One example suffices for each part.

i) An association rule may have high precision and lift but still lack validity (unlikely hold in future data).

ii) Statistical dependence is not a monotonic property. I.e., a rule can express strong dependence, even if more general rules express independence or opposite dependence (positive instead of negative or negative instead of positive).

iii) Overfitted rules can lead to wrong conclusions.

2. **(max 30p)** This task continues Exercise session 3 task 2, analyzing statistical association rules in the Rat data. The goal is to find some interesting associations, so you need to use some creativity in designing how to extract good binary features. The data set is normalized and cleaned version of the data that we have used before (load ratdataNormChecked.csv from the course homepage).

a) Extract binary features from the Rat data and save the data in transaction form. Report all attributes included in the transaction data and how the were extracted. You can look at an example from Exercise 3 Task 2. You can use the same binarization as in exercises for the features that were included there (see under Exercise 3, description.txt related to rattrans.txt), if you want, but you need to invent binarization for the other features[1]. Note that in the transaction form only positive-valued features are listed, so you will need to create attributes also for interesting negative values (like female and male=¬female). Here you are encouraged to use creativity, since results depend heavily on the features! You should get totally about 40 attributes.

[1]Note: gonfatind and gonind are so similar that you can choose only one of them. Using year is optional, although features derived from it may reveal something interesting.

If you try many binarization approaches, it suffices to report only your final attributes.

b) Transform the attribute (item) names to numerical codes and search association rules with Kingfisher (see Exercise 3). It suffices to search for only positive associations. The goodness measure is $\ln(p_F)$ (program default, natural logarithm of Fisher's p). Search for the top 100–200 rules so that you'll find something interesting (the number of rules needed depends on how good features you have and if you used any options to eliminate trivial associations). Report the measure ($\ln(p_F)$) threshold and possible other non-default parameters you used in the search (the final choice, whose results you report). (Note: you can adjust e.g., the rule complexity or minimum frequency, if you want, but justify your choices.)

Transform the rules into human readable form and report the most interesting (10–20 rules)! You can present them as a table presenting the rule, its frequency, precision (confidence), leverage, lift and significance ($\ln(p_F)$). Tell how you chose interesting rules (no need to interpret, yet). Here are some hints for filtering:

- If the rule is a side product of discretization, it is not interesting.
- If you find a rule pair $\mathbf{X} \to C$ and $\mathbf{Y} \to C$, where $\mathbf{Y} \subsetneq \mathbf{X}$, report just the more significant one.
- If you find many rules from the same set, e.g., $AB \to C$, $AC \to B$ and $BC \to A$, pick up the best one, but tell in the report that all combinations occurred.
- From the biological point of view, most interesting associations would characterize reasons of stomach ulcer, stress (especially ADWBind), general welfare (especially gonind, possibly also batind and BMI), differences between wild and lab rats, differences between places and seasons, effects of gender and motherhood, differences between puppies/youngsters and adult rats, etc.

c) Interpret your results! What do the most interesting associations mean? Here you can group together rules expressing variants of the same association or related to the same attribute of interest (e.g., stomach ulcer) and discuss them together. In the interpretation, you can utilize the known facts told in the data description, but you don't need to study biological explanations from any external sources (unless you want). Just tell your observations and your guesses what they mean. E.g., if you find extremely strong association telling that rats with high BMI have high batind, you can simply speculate that bat is also a type of fat and thus fat rats are likely to have a lot of it.

3. **max 30p** Searching only maximal, closed or 0-free sets (minimal generators) are popular techniques for reducing the number of frequent patterns. (See definitions in the Appendix 2.) In this task you should analyze why it is a bad short-cut to construct statistical association rules only from these condensed representations. Recall that from set $\mathbf{X}$ you can construct rules $\mathbf{X} \setminus \{C\} \to C$ (here it suffices that the the consequent is positive valued).

You can assume that $\min_{fr}$ is so small that it does not exclude any statistical associations (i.e., you could find all statistical associations, if all frequent sets were used).

a) (12p) Show with a counter-example that you can miss **some** positive statistical associations, if you construct rules only from i) maximal, ii) closed, iii) 0-free sets. Explain your example in a sufficient detail.

b) (6p) Show with a counter-example that you can miss **all** statistical associations if you construct rules only from maximal sets. Explain your example in a sufficient detail.

c) (12p) Can you detect overfitted (over-specialized or redundant) rules, if you are given only i) maximal sets, ii) closed sets, iii) 0-free sets? Justify your answer!

4. **Extra task, 10 extra points** Search some interesting application of association mining that has been published as a scientific paper (including arxiv reports). **Important: Give reference to the paper and www address where it can be loaded**.

   Write a brief explanation (**about 300 words, maximum 400 words**, consisting full sentences), where you describe the application idea by your own words. Include the following elements:

   - What problem is solved? What kind of data is used and what kind of information can be discovered? (If association patterns are used as a subroutine for something else, then tell the ultimate goal.)
   - What type of association patterns are used (sets or rules or something else) and how? What are the criteria for association patterns? (E.g., frequent and confident rules or strong statistical dependencies).
   - Evaluate the idea critically, not only strengths but also potential problems and limitations (e.g., usefulness, quality of discoveries, assumptions).
   - Tell your own ideas, like other areas where a similar approach could applied or how the method could be improved.

   Any decent mini essay with the these elements is given full points. If there is a clear trial, but some elements are missing or the essay is very superficial or contains factual errors, it will receive 5 points. If the source is not a scientific publication or there is no real trial (to understand the idea and write an explanation), it is 0 points, so don't return an essay unless you really want to write it.

   When writing, remember that you need to **write by your own words**, all copying is prohibited (vs. Aalto University Code of Academic Integrity).

   Hint: Make a google search with interesting keywords (e.g., "association mining text data" or "association rule medicine"). If you cannot find anything interesting, you can ask hints from others in the Zulip chat (channel Assignments).

Appendix A: Required equations of mutual information

Mutual information of rule $\mathbf{X} \to C{=}c$ is

$$MI = \log \frac{P(\mathbf{X}C)^{P(\mathbf{X}C)} P(\mathbf{X}\neg C)^{P(\mathbf{X}\neg C)} P(\neg\mathbf{X}C)^{P(\neg\mathbf{X}C)} P(\neg\mathbf{X}\neg C)^{P(\neg\mathbf{X}\neg C)}}{P(\mathbf{X})^{P(\mathbf{X})} P(\neg\mathbf{X})^{P(\neg\mathbf{X})} P(C)^{P(C)} P(\neg C)^{P(\neg C)}}$$

Conditional mutual information for evaluating rule $\mathbf{XQ} \rightarrow C{=}c$ given $\mathbf{X}$ in the value-based interpretation is

$$MI_C = \log \frac{P(\mathbf{X})^{P(\mathbf{X})} P(\mathbf{XQ}C)^{P(\mathbf{XQ}C)} P(\mathbf{XQ}\neg C)^{P(\mathbf{XQ}\neg C)} P(\mathbf{X}\neg\mathbf{Q}C)^{P(\mathbf{X}\neg\mathbf{Q}C)} P(\mathbf{X}\neg\mathbf{Q}\neg C)^{P(\mathbf{X}\neg\mathbf{Q}\neg C)}}{P(\mathbf{XQ})^{P(\mathbf{XQ})} P(\mathbf{X}\neg\mathbf{Q})^{P(\mathbf{X}\neg\mathbf{Q})} P(\mathbf{X}C)^{P(\mathbf{X}C)} P(\mathbf{X}\neg C)^{P(\mathbf{X}\neg C)}}$$

Here log is the 2-based logarithm. Note that in the task you should report $n \cdot MI$ and the thresholds are also given for $n \cdot MI$, where n=data size.

Note also that mutual information doesn't differentiate between positive and negative dependencies. Therefore you need other means to find out if (conditional) dependence is positive or negative.

Appendix B: Definitions of maximal, closed and 0-free frequent sets

Let $\mathbf{X}$ be a frequent itemset, i.e., $P(\mathbf{X}) \geq min_{fr}$. $\mathbf{X}$ is

- maximal, if for all $\mathbf{Y} \supsetneq \mathbf{X}$ $P(\mathbf{Y}) < min_{fr}$;
- closed, if for all $\mathbf{Y} \supsetneq \mathbf{X}$ $P(\mathbf{Y}) < P(\mathbf{X})$;
- 0-free (=maximal generator), if for all $\mathbf{Y} \subsetneq \mathbf{X}$ $P(\mathbf{Y}) > P(\mathbf{X})$.

Note: $\mathbf{X} \subsetneq \mathbf{Y}$ means that $\mathbf{X}$ is a proper subset of $\mathbf{Y}$ (excludes $\mathbf{X} = \mathbf{Y}$).