CS-E4650 Methods of Data mining

Exercise 3 / Autumn 2023

3.1 Azheimer disease association rules: basic evaluation

Learning goal: Unconditional evaluation of statistical 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 content is also given in the latex table format in the home page. You can use it as the input for your program and for presenting the results (just add new columns for measures).

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.)

Table 1: Candidate rules $\mathbf{X} \rightarrow C{=}c$, $c \in \{0, 1\}$, related to C = Alzheimer's disease. $fr_X = fr(\mathbf{X})$, $fr_{XC} = fr(\mathbf{X}C{=}c)$.

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

a) Calculate leverage and lift values for all rules. Prune out rules that do not express positive statistical dependence.

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

3.2 Alzheimer disease association rules: further evaluation

Learning goal: Conditional evaluation and interpretation of statistical association rules.

In this task, the same association rules (Table 1) are evaluated further. You can now evaluate only those rules that remained significant after task 1.

a) 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 $P(C{=}c|\mathbf{Y}) \geq P(C{=}c|\mathbf{X})$ (no improvement) or the improvement is not sufficient, $n \cdot MI_C < 0.5$ (i.e., $MI_C < 0.0005$).

b) 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?

c) 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.

3.3 Monotonic upperbounds for leverage

Learning goal: Concept of monotonicity; how to prove and utilize monotonic upperbounds of goodness measures in the search.

In this task, you should prove monotonic upperbounds for the leverage (δ) and then design an association rule discovery algorithm that utilizes the

bounds. In the following, $\mathbf{R}$ denotes the set of all attributes, $C \in \mathbf{R}$ is a single attribute, $\mathbf{X} \subseteq \mathbf{R} \setminus \{C\}$ and $\mathbf{Q} \subseteq \mathbf{R} \setminus \mathbf{X} \setminus \{C\}$ are sets of attributes.

Hint: Remember that $\delta(\mathbf{X}, C) \leq \min\{P(\mathbf{X})P(\neg C), P(\neg X)P(C)\}$.

a) Show that $\delta(\mathbf{X}, C) \leq P(C)P(\neg C)$.

b) Show that $\delta(\mathbf{XQ}, C) \leq P(\mathbf{X})P(\neg C)$.

c) Show that $\delta(\mathbf{XQ}, C) \leq P(\mathbf{X}C)P(\neg C)$.

d) Describe a search algorithm that finds the top-K best association rules of the form $\mathbf{Y} \rightarrow A$, where $A \in \mathbf{R}$ and $\mathbf{Y} \subseteq \mathbf{R} \setminus \{A\}$, using leverage as the goodness function. In addition, leverages of all discovered rules should be at least $\min_\delta > 0$, where $\min_\delta$ is an arbitrary user provided threshold. (If you cannot find K rules, then return as many as you can find). It suffices to describe the algorithm in a general level, but tell how the three upperbounds are utilized in pruning the search space.

3.4 Homework: Bird associations

Learning goals: Mining association rules in practice; making efficient data mining pipelines

This is an explorative task, where you should invent good features to extract from the extended bird data and then search and analyze association rules. The pattern discovery process is iterative, and you will very likely experiment with multiple versions of feature extraction. Therefore, it is recommended to do the preparations well and make a shell script that speeds up the process. You can find instructions and hints in MyCourses (instructionsforkingfisher.pdf). You can find an extended version of the bird species data, **birdspeciesv2.csv**, and its description in MyCourses.

a) Extract good features for association discovery from the bird data. You can find interesting associations only, if the involved properties are captured by features! It is suggested to proceed iteratively, from easier to more difficult features:

   - Group, habitat and diet can be used as such (just list group and all elements of habitat and diet in the transaction).
   - For most binary features (like long-billed), you can use only the Yes-values (list attribute "long-billed" in the transaction, but forget its opposite, "non-long-billed"). The only exception is field sim, where both values are interesting (if genders look similar or different).
   - For multi-valued categorical features, you can create one attribute for each value.
   - Invent some informative features from the spring and autumn migrations times (fields "arrives" and "leaves"), e.g., describing that migration starts early or ends late.
   - Invent how to handle numerical features. Usually, only the extremes are interesting, like laying relatively few eggs or many eggs.

b) Search association rules with Kingfisher. You may need to search quite many rules (e.g., 300) to find more versatile rules, since there will be many variants of similar associations. Try to find rules that describe different aspects of the data, like different groups, appearance, diet, habits, environment, etc, but remember that all attributes do not necessarily participate any significant associations.

c) Report the most significant and interesting rules. The idea is not to list all rules, but group rules and describe the information they reveal (e.g., what things are associated to scolopacidae or plunge-divers). Tell also which features seem to be irrelevant (did not occur in any rules).

Parts of the report:

1. Cover page: title (course name and assignment number), names and student ids of all participants of the team.

2. Section 1 "Methods": Describe very briefly the methods: what programming language you used for feature extraction, what were the parameter settings for Kingfisher, if you used any constraints etc. However, do not describe the feature extraction here.

3. Section 2 "Feature extraction": Describe compactly but carefully what features you extracted. You can, e.g., use a list or a table that tells the original feature, new attributes, and how they were extracted. Describe carefully non-trivial extraction (like handling numerical values or migration month ranges).

4. Section 3: "Results": Describe the most significant and interesting associations you discovered (see above).

5. Section "Appendix": Include here the code of your feature extraction program.

Produce a pdf report including all parts and submit it in MyCourses before the deadline. Recall that this task is homework that is done in groups of 2–3 students – you cannot do the task alone or in a larger group. You can search collaborators in zulip, exercise sessions, or ask help from the TAs.

Appendix A: Required equations of mutual information

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

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

The base of the logarithm is not fixed (usually 2 or e), but in this task you are asked to use the 2-based logarithm for better comparison of results. 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.