Assignment 3

Gengcong Yan - 1009903
CS-E4650 Methods of Data mining

November 14, 2021

1 Task 1

Table 1: Evaluation of rules

num	rule	fr_X	fr_{XC}	lift	leverage	$n \times MI$
1	smoking → AD	300	125	1.389	0.035	19.436
2	stress → AD	500	150	1.000	0.000	0.000
3	higheducation → ¬ AD	500	400	1.143	0.050	34.852
4	tea → ¬ AD	342	240	1.003	0.001	0.005
5	turmeric → ¬ AD	2	2	1.429	0.001	1.030
6	female → ¬ AD	500	352	1.006	0.002	0.055
7	female, stress → AD	260	100	1.282	0.022	8.400
8	berries, apples → AD	120	32	0.889	-0.004	0.531
9	smoking, tea → AD	240	100	1.389	0.028	14.202
10	smoking, higheducation → AD	80	32	1.333	0.008	2.847
11	stress, smoking → AD	200	100	1.667	0.040	32.268
12	female, higheducation → ¬ AD	251	203	1.155	0.273	14.462

1.1 Task1.a

Leverage or lift values for all rules are shown in Table.1. we should be careful the $\neg AD$ and AD represent different fraction in data. $P(AD) = 0.3, P(\neg AD) = 0.7$ will be used in computation of evaluation. The process code of leverage and life are below.

Rule 2 and 8 are pruned out based on leverage and lift,because they doesn't express positive statistical dependence with $leverage \leq 1.0$ and $leverage \leq 0.0$.

```
# calculation of leverage and lift

frx=[300,500,500,342,2,500,260,120,240,80,200,251]
frxc=[125,150,400,240,2,352,100,32,100,32,100,203]

pc=[0.3,0.3,0.7,0.7,0.7,0.7,0.3,0.3,0.3,0.3,0.3,0.7]

px=np.array(frx)/1000
pxc=np.array(frxc)/1000

lift=pxc/(px*pc)
leverage=pxc-(px*pc)
```

1.2 Task1.b

Mutual information MI are shown in Table.1. According to MI, rules 2, 4, 5, 6 and 8 are pruned out because $n \times MI < 1.5$. The process code are below.

```python
def MI(px,pc,pxc,N):
    pnx,pnc=1-px,1-pc

    pnxc=pc-pxc
    pxnc=px-pxc
    pnxnc=1-pxc-pnxc-pxnc

    #print(pnxc,pxnc)

    part1=(pxc**pxc)*(pxnc**pxnc)*(pnxc**pnxc)*(pnxnc**pnxnc)

    part2=(px**px)*(pnx**pnx)*(pc**pc)*(pnc**pnc)

    M=math.log(part1/part2,2)
    return M

N=1000
MIs=[]
for i in range(len(px)):
    re=MI(px[i],pc[i],pxc[i],N)
    MIs.append(N*re)
print(MIs)
```

1.3 Task1.c

For now, the remaining rules after pruned out are $[1, 3, 7, 9, 10, 11, 12]$. According to the hint on the question, if the probability of a subset of rules is greater than the probability of the set of rule itself, we can believe it can be judged as an overfitting rule. The computation of rules are shown below:

$$
\begin{align}
P(1) = 0.417 = P(9) = 0.417 \tag{1}\\
P(1) = 0.417 > P(10) = 0.400 \tag{2}\\
P(1) = 0.417 > P(11) = 0.500 \tag{3}\\
P(3) = 0.8 < P(12) = 0.81 \tag{4}\\
\tag{5}
\end{align}
$$

We think rules 9, 10, 11 are overfitting rules. smoking takes up more influences than other factors in these rules. After all pruned out process, The remaining rules are **1, 3, 7, 12**.

1.4 Task1.d

Based on still existing rules 1, 3, 7, 12. My conclusion is that smoking and stress are likely to lead to Alzheimer's disease, but those with higher education are likely to be much less likely to get Alzheimer's disease. Female who are under stress may be more likely to develop the disease. But current rule does not reflect specific rules for male alone, so it is not very convincing in terms of gender factors.

If you want to avoid getting Alzheimer's disease, we should not smoke, and keep the mood happy to reduce stress, and then keep learning throughout life to keep the mind active.

Table 2: Evaluation of remaining rules

num	rule	fr_X	fr_{XC}	lift	leverage	$n \times MI$
1	smoking → AD	300	125	1.389	0.035	19.436
3	higheducation → ¬ AD	500	400	1.143	0.050	34.852
7	female, stress → AD	260	100	1.282	0.022	8.400
12	female, higheducation → ¬ AD	251	203	1.155	0.273	14.462

1.5 Task1.e

1. Rule 5 turmeric→ ¬ AD. There are only 2 examples in a dataset of size 1000. although the precision and lift can be high, it's not trustworthy. The size of rule 5 takes up too little of whole dataset.

2. Rule 3 higher education→ ¬ AD; Rule 10 smoking, higher education→ AD. Rule 10 express strong dependence of smoking and higher education leads to AD, but seen from rule 3, higher education are negatively associated with AD. Statistical dependence is not a monotonic property.

3. Rule 1 smoking→ AD; Rule 10 smoking, higher education→ AD. if you only see rule 10, you may think a smoking person without education won't get AD. But truth is rule 10 is overfitting with rule 1. if rule 1 is strong, no wonder rule 10 will lead to strong, since rule 1 is part of rule 10. Smoking is a more important factor here.

2 Task 2

2.1 Task2.a

We need to extract binary features from the Rat data and save the data in transaction form. From Exercise 3 Task 2, there are some features are already extracted,but remaining features should extract by ourselves. And after extraction of features, we convert the original rat data into transaction form for searching association rules with Kingfisher later.

The rules of extracting data features (some are original from *description.txt* in E3T2) are shown below:

1. Ratid (not needed in rule discovery, but doesn't harm, helps to identify rats if needed; note that each id is now a binary attribute)
2. **Summer** if day 116-300 and **winter** if day 1-115 or 301-365.
3. **Freezer** if day=0
4. **weightlow** if weight≤162 and **weightnormal** otherwise (these probably correspond to puppies/young and adult rats)
5. Gender. **female, male**
6. Femstate. Accoring to 1-4, the features are **pregnant, nursing, pregnant+nursing, femstate_neither**.
7. **Liversmall** if liverind≤0.037 and **liverlarge** if liverind>0.064, in between are **livernormal**.
8. **Batlow** if batind≤0.00067 and **bathigh** if batind>0.00184, in between are **batnormal**.
9. **tailshort** if tailind ≤ 0.74 and **taillong** if tailind > 0.85,in between are **tailnormal**.
10. **wild** if place=1-3, and **lab** otherwise.
11. **adrenalsmall** if ADWBind≤0.21 and **adrenalarge** if ADWBind>0.48, in between are **adrenalnormal**.
12. **BMIsmall** if BMI ≤ 0.47 and **BMIlarge** if BMI > 0.75, in between are **BMInormal**.
13. **heartsmall** if heartind ≤0.0036 **heartlarge** if heartind >0.0046, in between are **heartnormal**.
14. **appsmall** if appind ≤0.0104, **applarge** if appind >0.0167, in between are **appnormal**.
15. **gonsmall** if gonfatind ≤0.0018, **gonfat** if gonfatind >0.0103.in between are **gonnormal**.
16. sulcer. **sulcermild** ≤ 4, **sulcerserious** > 4.

17. kmethod. **According to 1-5, one attribute for each method**.

18. blength. **blengthshort** ≤ 19.5 , **blengthlong** > 22.5, in between are **blengthnormal**.

So after extraction of features, we obtain [**summer, winter, freezer, weightlow, weightnormal, female, male, pregnant, nursing, pregnant+nursing, femstateneither, liversmall, liverlarge, batlow, batnormal, bathigh, tailshort, tailnormal, taillong, wild, lab, adrenalsmall, andadrenalarge, adrenalnormal, BMIsmall, BMIlarge, BMInormal, heartsmall, heartlarge, heartnormal, appsmall, applarge, appnormal, gonsmall, gonfat, gonnormal, sulcermild, sulcerserious, kmethod1, kmethod2, kmethod3, kmethod4, kmethod5, blengthshort, blengthnormal, blengthlong**], totally 46 features.

2.2 Task2.b

The threshold of Kingfisher search in rat data are $In(P_F)$ is set to -50 with option -M. The number of rules is not defined, so it will output by default max 100 rules having $ln(P_F) \leq -50$. It searches for both positive and negative dependencies. An upperbound for the number of attributes is set to 700 with option -k. From all rules generated in program, we have chosen to keep the rules of practical significance in the following Table.3.

Table 3: Selected rules in rat data

Rule	fr_X	Cf	γ	δ	$In(P_F)$
1 weightlow → blengthshort	99	1.0	3.484	0.126	-153.7
2 lab → kmethod4	34	1.0	16.5	0.057	-125.6
3 femstate_neither gonfat heartsmall blengthnormal → lab	33	0.846	13.962	0.055	-104.4
4 femstate_neither weightnormal gonfat heartsmall → lab	33	0.846	13.962	0.055	-104.4
6 BMIsmall → weightlow	63	0.808	4.577	0.088	-96.9
7 female liversmall gonfat blengthnormal → lab	29	0.967	15.95	0.048	-95.6
8 female liversmall gonfat blengthnormal → kmethod4	29	0.967	15.95	0.048	-95.6
9 female liversmall gonfat heartsmall → lab	29	0.935	15.543	0.048	-92.9
10 female sulcermild → femstate_neither	155	0.442	1.598	0.103	-89.6
11 sulcermild weightnormal → blengthnormal	291	0.634	1.222	0.094	-86.9
12 blengthnormal → weightnormal	291	1.0	1.241	0.092	-83.9
13 heartlarge blengthshort → BMIsmall	53	0.639	4.593	0.074	-75.7

2.3 Task2.c

Now from the rules in the table above, we can summarize some interesting findings. The features in these rules 1, 7, 13, 14 all have some linear relationship with each other. It is natural to think that after body length becomes larger, the weight must also become larger, so the BMI will also change with it. These characteristics will change together in general. From rule 2, we can speculate that the rats produced in the experimental environment are killed by specific way. This is because rats can be killed more easily in a laboratory environment, whereas rats in a wild environment may not be treated according to laboratory methods because of bacteria and other problems. From rules 2,3,8,9, we can observe that rats in a laboratory environment, even though their weight and height were normal, were more likely to have smaller hearts and to have more gonadal fat suggesting their excessive food intake. My guess is that they do not have enough space to move around in the lab and are fed regularly every day. So they became lazy and inactive and their hearts did not get exercise. Compared to the rats in the wild environment, they had a better food intake situation every day. From rule 13, we can observe that female rats with mild stomach ulcer are likely not in any specific female state, indicating Rats who are pregnant or breastfeeding are more likely to develop stomach ulcers.

Table 4: Count Example 1

Num	Itemset
1	{A,C}
2	{B,C}
3	{A,B}
4	{C}
5	{A,B,C}

Table 5: Count Example 2

Num	Itemset
1	{A,B}
2	{C}
3	{C}
4	{C}
5	{A,B,C}

3 Task3

3.1 Task3.a

We use the examples in following tables explaining maximal and closed cases.

1. Table 4 Example 1 - Maximal frequent sets. We can see set $\{A, B\}$ are not maximal frequent set, because its superset $\{A, B, C\}$ is frequent. We compute $lift(A \to B) = \frac{2/5}{3/5*3/5} = 10/9 > 1$. So the positive statistical associations $A \to B$ is ignored in this case.

2. Table 5 Example 2 - Closed frequent set. We can see set $\{A, B\}$ are not close frequent set, because its superset $\{A, B, C\}$, $P(ABC) = 1/5 = P(AB)$. But we compute $lift(A \to B) = \frac{2/5}{2/5*2/5} = 5/2 > 1$. So the positive statistical associations $A \to B$ is ignored in this case.

3. Table 6 Example 3 - 0-free frequent set. We can see set $\{A, B\}$ are not 0-free frequent set, because its subset set $P(A) = P(B) = 3/4 = P(AB)$. But we compute $lift(A \to B) = \frac{3/4}{3/4*3/4} = 4/3 > 1$. So the positive statistical associations $A \to B$ is ignored in this case.

3.2 Task3.b

In this task, we need to find all maximal sets in example, whose associations rules are not correct. But all correct associations rules are derived from other sets. The example 4 shows in Table 7.

We can see $\{A, C, D\}$ and $\{B, C, D\}$ are maximal sets and rules derived from them are $A, C \to D$ and $B, D \to C$. Then we compute $lift(A, C \to D) = \frac{1/6}{1/2*2/3} = 1/2 < 1$ and

Table 6: Count Example 3

Num	Itemset
1	{A,B}
2	{A,B}
3	{C}
4	{A,B,C}

Table 7: Count Example 4

Num	Itemset
1	{A,C}
2	{A,C}
3	{B,D}
4	{B,D}
5	{A,C,D}
6	{B,C,D}

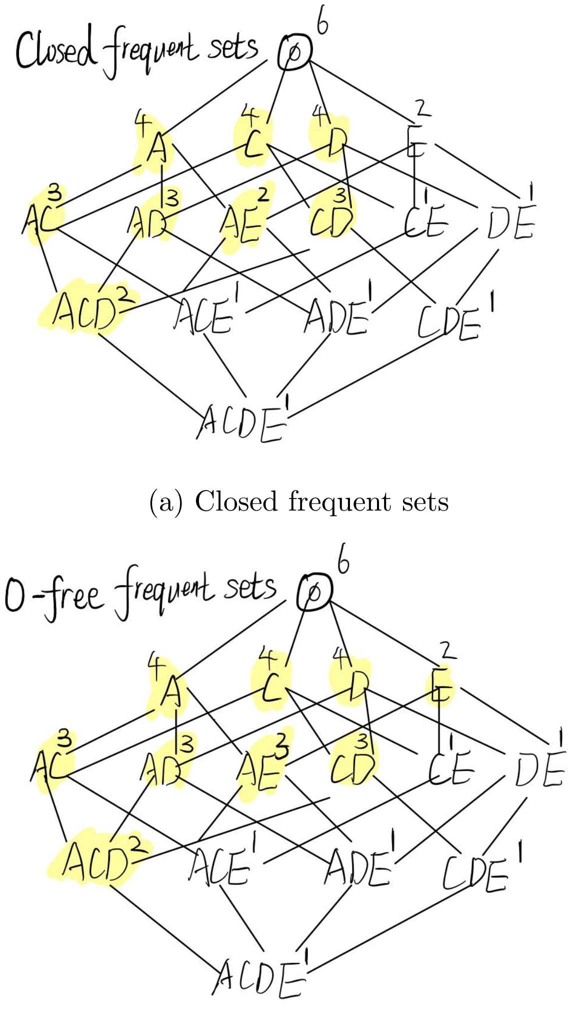
$lift(B, D \rightarrow C) = \frac{1/6}{1/2*2/3} = 1/2 < 1$. So rules in maximal sets are not effective. whereas all effective rules actually are $A \rightarrow C$ and $B \rightarrow D$, $lift(A \rightarrow C) = \frac{1/2}{1/2*2/3} = 3/2 > 1$, $lift(B \rightarrow D) = \frac{1/2}{1/2*2/3} = 3/2 > 1$, which produced from $\{A, C\}$ and $\{B, C\}$ not belonging to maximal sets.

3.3 Task3.c

1. Maximal frequent sets. If only given maximal sets, it's hard to detect overfitted rules in these sets, because The subsets of maximal sets can not be maximal sets, there is no redundant in current rules. For example, if $\{A, B, C\}$ is a maximal set, subsets $\{A, B\}, \{A, C\}, \{B, C\}$ will not be considered association rules. There are no redundant rules derived from $\{A, B, C\}$.

2. Closed frequent sets. The example shows in Fig.1a. It an adaptation from P58 lecture 7. The subsets of closed sets $\{A, C, D\}$ are still closed sets. So there are a lot redundant in the rules derived from all closed sets. $\{A, C, D\}$ will most likely produce overfitted rules.

3. 0-free frequent sets. The example shows in Fig.1b. The subsets of 0-free sets $\{A, C, D\}$ are still 0-free sets. So there are a lot redundant in the rules derived from all 0-free sets. $\{A, C, D\}$ will most likely produce overfitted rules.



(a) Closed frequent sets

(b) 0-free frequent sets

Figure 1: Task3.C Examples
(Adaptation from P58 Lecture 7, $min_{fr} = 1/3$)