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.

```
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:

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

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 $\rightarrow$ AD	300	125	1.389	0.035	19.436
3	higheducation $\rightarrow \neg$ AD	500	400	1.143	0.050	34.852
7	female, stress $\rightarrow$ AD	260	100	1.282	0.022	8.400
12	female, higheducation $\rightarrow \neg$ AD	251	203	1.155	0.273	14.462

1.5 Task1.e

1. Rule 5 turmeric$\rightarrow \neg$ 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$\rightarrow \neg$ AD; Rule 10 smoking, higher education$\rightarrow$ 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$\rightarrow$ AD; Rule 10 smoking, higher education$\rightarrow$ 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.