Methods of data mining

Assignment 3

Task 1

Miro-Markus Nikula

712686

Task1.

For this task I first tried solving the problems with some python code, but I made so many mistakes and got suspicious results that I got tired of trying to debug my solutions that I just decided to calculate most of the probabilities and frequencies that I used to construct the below tables by hand. Attached is Task1.py that only is responsible for calculating the lift, mutual information and conditional mutual information. For the probabilities I wrote for nearly all rules X->C a 2x2 table displaying combinations of X, not X, C and not C such as in the lecture 7 slides.

a)

Num	Lift (γ)
1	1.39
2	1.0
3	1.14
4	1.003
5	1.43
6	1.006
7	1.29
8	0.89
9	1.39
10	1.33
11	1.67
12	1.16

Rule 2 has a lift value of 1.0 which means that it does not express any statistical dependance and rule 8 expresses negative statistical dependance so these rules are pruned out.

b)

Num	n * MI
1	132.48
3	34.85
4	0.005
5	1.03
6	0.05
7	8.40
9	14.20
10	2.85
11	32.27
12	14.46

Remove the rules that have n *MI < 1.5

c) Following are calculated conditional mutual information values * 1000. For rules 1 and 3, calculating this value is not possible.

Num	n * MI(c)
1	-
3	-
7	14.99
9	-1.6 *10^-13
10	0.090
11	45.64
12	19.46

We once again prune out the rules that do not satisfy the criteria given in the task. First, we prune out the ones that have a conditional mutual information * 1000 of under 0.5. For the ones that remain we will calculate conditional mutual information again, this time switching X and Q. Ie. if a rule is stress, smoking -> AD, we will calculate for smoking, stress -> AD.

Num	n * MI(c)
1	-
3	-
7	13.54
11	12.83
12	0.17

We can see that switching X and Q had little effect on rule 7, but it did make CMI for rules 11 and 12 significantly smaller. Filter again on the same criterion and we end up with only 4 rules:

Num	Rule
1	smoking → AD
3	higheducation → ¬ AD
7	female, stress → AD
11	stress, smoking → AD

d) With my very limited insight on Alzheimer's disease, I'll try analyzing whether these rules make sense or not. Smoking increasing the risk for AD definitely does and so does high education with reducing it. The CMI-value of rule 11 decreased a lot when flipping stress and smoking and although it is still relatively high, I reckon that stress's main effect here on the risk of getting AD is that stress increases smoking. In addition, stress itself was statistically independent. This fact, however, makes the rule number 7 seem suspicious for me, but doing some research I found that stress actually does increase the risk of getting AD for middle aged women.

If I were to try avoiding Alzheimer's disease depending on these rules and calculations, I would at least avoid smoking. Also, high education seems to somewhat correlate with not getting AD as likely. A doctor's degree as such most likely doesn't protect from the disease, but rather the increased thinking and challenging your brain and memory more likely does it. It could be even due to higher standard of living for highly educated people in contrast to the ones that are not that might reduce stress, which for women can increase the risk of AD.

e) i. **Rule 5** has a lift of 1.43, but since there only are two samples and that eating turmeric decreases the risk of AD sounds quite suspicious it most likely won't hold in future data.

   ii. **Stress** is statistically independent, but when combined with females, it suddenly does matter.

   iii. **Rule 11** The same example of smoking + stress as before. Stress doesn't increase the risk of AD, but it does increase the need for smoking, which then affects strongly on the risk of getting AD.