CS-E4650 - Assignment 3

Arttu Häkkinen - 596077

November 9, 2021

Task 1

a)

num	rule	fr_X	fr_{XC}	leverage δ
1	smoking $\rightarrow$ AD	300	125	0.0350
3	higheducation $\rightarrow \neg$ AD	500	400	0.0500
4	tea $\rightarrow \neg$ AD	342	240	0.0006
5	turmeric $\rightarrow \neg$ AD	2	2	0.0006
6	female $\rightarrow \neg$ AD	500	352	0.0020
7	female, stress $\rightarrow$ AD	260	100	0.0220
9	smoking, tea $\rightarrow$ AD	240	100	0.0280
10	smoking, higheducation $\rightarrow$ AD	80	32	0.0080
11	stress, smoking $\rightarrow$ AD	200	100	0.0400
12	female, higheducation $\rightarrow \neg$ AD	251	203	0.0273

These were calculated by going through the data row by row and calculating $\delta = P(XC) - P(X)P(C)$. If the value for δ was negative, the row was pruned out.

b)

num	rule	fr_X	fr_{XC}	leverage δ	$n \cdot MI$
1	smoking $\rightarrow$ AD	300	125	0.0350	19.44
3	higheducation $\rightarrow \neg$ AD	500	400	0.0500	34.85
7	female, stress $\rightarrow$ AD	260	100	0.0220	8,40
9	smoking, tea $\rightarrow$ AD	240	100	0.0280	14.20
10	smoking, higheducation $\rightarrow$ AD	80	32	0.0080	2.85
11	stress, smoking $\rightarrow$ AD	200	100	0.0400	32.27
12	female, higheducation $\rightarrow \neg$ AD	251	203	0.0273	14.46

The value for each remaining rows' $n \cdot MI$ was calculated using a script that was programmed to do the same calculations as appendix 1 equation suggests.

All the necessary terms for the equation were computed based on the known values for $P(XC), P(X), P(C)$ for each row. E.g. $P(X\neg C) = P(X) - P(XC)$, $P(\neg XC) = P(C) - P(XC)$ and $P(\neg X\neg C) = 1 - (P(XC) + P(\neg XC) + P(X\neg C))$. Finally, the rows with $n \cdot MI < 1.5$ were dropped/pruned out.

c)

The value for each remaining rows' $n \cdot MI_C$ was calculated using a script that was programmed to do the same calculations as appendix 1 equation suggests. Only difference to the part b was that now we only computed the value for the rows which had a conditioning set of size two. All the necessary terms for the MI_C equation were computed based on the known values for $P(X), P(XQ), P(XC), P(XQC)$ for each row. For more on detailed information about implemetations, see the function for conditional MI from the attached source code. Basically I did the same as in part b but now for three attribute joint probability distribution. Also, it was asked to compare the conditional probabilities of the proper subsets Y_1, Y_2 of the each two variable conditioning set X, and see if any of the proper subsets attributes had greater or equal conditional probability $P(C = c|Y_j), j = 1, 2$ in comparison to $P(C = c|X)$. Finally, if any the following conditions held, the rule was pruned out: $MI_C(X = Y_j, XQ = X) < 0.5, j = 1, 2$ or $P(C = c|Y_j) \geq P(C = c|X), j = 1, 2$. The two remaining rules after this were: "smoking $\rightarrow$ AD" and "higheducation $\rightarrow$ $\neg$ AD".

d)

My main conclusion based on the remaining association rules is that in order to avoid Alzheimer's disease, one should avoid smoking and try to reach for higher education by studying more.

e)

i) Example rule: "stress, smoking $\rightarrow$ AD". Reason: This rule has the second highest leverage value of all rules $\delta = 0.0400$, but still it lacks validity since it has such a low MI_C values with both of the attributes of its conditioning subset.
ii) Example rule: "stress, smoking $\rightarrow$ AD". Reason: This rule has high positive association expressing that the conditioning set X and the consequent set C are strongly statistically dependent, even though the more general rule "stress $\rightarrow$ AD" expresses the opposite dependence having $\delta < 0$.
iii) Example rule: "stress, smoking $\rightarrow$ AD". Reason: This rule has really high mutual information score as well as high positive statistical dependence expressed by the high leverage. Without evaluating overfitting as done in the part c, we would have ended up with the conclusion that stress and smoking is a bad combo when it comes to Alzheimer's disease, which is only partly true,

since stress does actually have negative statistical dependence with Alzheimer's disease.

References