MDM Assignment 2

Marco Vassena 4110161
Philipp Hausmann 4003373

October 28, 2014

Contents

1 Problem description **1**

2 Analysis **1**

1 Problem description

The goal of this assignment is to implement and evaluate a tree classification algorithm. In addition to that, we will also discuss two different implementation styles and their performance implications.

2 Analysis

(a) A graphical model is totally represented by its independence graph, because all the constraints (u-terms set to 0) can be read from it. Therefore counting the number of graphical models comes down to counting the number of its independence graph, which are undirected graphs with as many nodes as variables in the model. In a graph with k nodes, there are $\binom{k}{2}$ possible undirected edges:

$$\binom{k}{2} = \frac{k!}{2!(k-2)!} = \frac{k(k-1)}{2}$$

The first node of an edge can be chosen among k elements, the second node among the remaining $k-1$ (we do not allow self loops), finally we divide by 2 because the edges are undirected, thus an edge and the edge obtained swapping the vertices count as one. The number of different

variable	# values
cat1	9
death	2
swang1	2
gender	2
race	3
ninsclas	6
income	4
ca	3
age	5
meanbp1	2

Table 1: Number of levels for each variable of the model

graphs with k nodes corresponds to the total number of subsets of the set of all possible edges. A set with n elements has 2^n subsets, because each element can be included or not in any subset. Therefore the number of graphical models for a model with k variables is $2^{\binom{k}{2}}$. This data set contains 10 nodes, therefore the number of different graphical models is:

$$2^{\binom{10}{2}} = 35184372088832$$

(b) Each cell of the table of counts contains the number of rows in a the data set whose attributes have a certain combination of values. Therefore the number of cells of a table of counts correspond to the total number of possible configurations of the attributes values, which is the product of the number of possible values of each attribute. We report in table 1, the variables of the model, with the number of its possible values.

Therefore the number of cells of the table of counts for this data set is given by $9 \times 2^3 \times 3 \times 6 \times 4 \times 3 \times 5 \times 2$, which is equal to 155520.

The saturated model does not assume any (conditional) independency among the variables, thus all its probabilities are estimated counting how many times a certain combination of values occurs, divided by the total number of observation:

$$\hat{p}(x_1 \dots x_n) = \frac{n(x_1 \dots x_n)}{N}$$

Therefore we need to consider again all the possible configurations of values. However since all the probabilities must sum to one, we can leave out one of those.

N	variable
1	cat1
2	death
3	swang1
4	gender
5	race
6	ninsclas
7	income
8	ca
9	age
10	meanbp1

Table 2: Numbering used to identify the variables of the model

The number of parameters of the saturated models is the number of cells of the table of counts minus 1, which gives 155519 parameters.

(c) The nodes have been numbered according to table 2. The cliques of the resulting model can be seen in table 3. Figure 1 shows the independence graph of the model.

	cliques
1	{1, 3, 10}
2	{3, 6}
3	{4, 6}
4	{5, 6}
5	{5, 10}
6	{6, 7}
7	{1, 8}
8	{2, 8}
9	{6, 8}
10	{1, 9}
11	{2, 9}
12	{6, 9}
13	{2, 10}

Table 3: The cliques of the resulting model for question (c)

(d) The Pairwaise Markov property states that any two non-adjacent variables are conditionally independent given all the others. Since variable

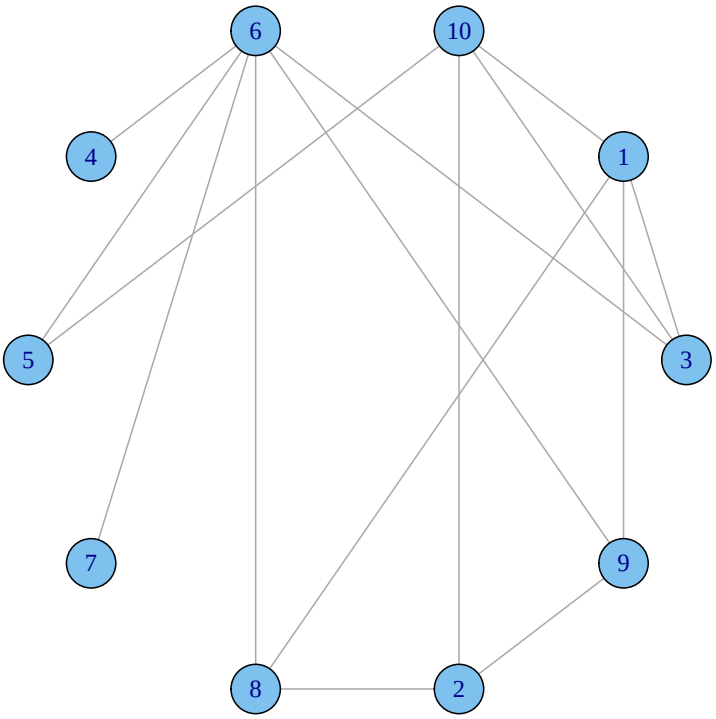

Figure 1: The independence graph for question (c)

node 4 (variable *gender*) and node 7 (variable *income*) are not adjacent (the edge $(4, 7)$ is not present in the graph), we can conclude that:

$$X_4 \perp\!\!\!\perp X_7 | (X_1, X_2, X_3, X_5, X_6, X_7, X_8, X_9, X_{10})$$

Using the Global Markov property we can derive a stronger statement. Node 4 is separated from node 7 by node 6, because any path from 4 to 7 passes through node 6, therefore *gender* is independent from *income* given *ninsclas* (6).

$$X_4 \perp\!\!\!\perp X_7 | X_6$$

By the Local Markov property a variable, given its closed neighbourhood, its independent from all remaining vertices. Therefore, in order to estimate the variable *death* (2) (which determines whether someone survives), it is sufficient to consider its adjacents variables: *ca* (8), *age* (9), *meanbp1* (10).

(e) The model found starting from the saturated model (complete graph) is slightly worse than the model found starting from the independence model (empty graph). The BIC score of the model is 15851, whereas the BIC score of the model found in (c) is 15842. The model found in (e) contains more cliques than the model found (c), namely: $\{2, 7\}, \{3, 4\}, \{3, 9\}, \{5, 7\}, \{5, 9\}$. However model (c) contains the following cliques, which are not included in (e): $\{1, 9\}, \{3, 6\}, \{5, 6\}$

Because of the additional nodes departing from 4 and 7 we can derive weaker independence statements about *income* and *gender*, namely:

$$X_4 \perp\!\!\!\perp X_7 | (X_6, X_3)$$

$$X_4 \perp\!\!\!\perp X_7 | (X_2, X_5, X_6)$$

Also for predicting variable X_2 *death* we need to consider a larger set of variables, because its neighborhood includes also *income* (7), in addition to *ca* (8), *age* (9), *meanbp1* (10) as before.

We would like to point out that during the analysis, the IPF algorithm used in the hill-climbing search, often failed to converge, producing therefore models of lower quality.

(f) The scores of the models starting from the complete and empty graphs are equal and are 14278. In addition, both models have the same cliques as can be seen in table 5 and 6, and therefore also the same graph. This indicates that the same local optima is found from both starting points.

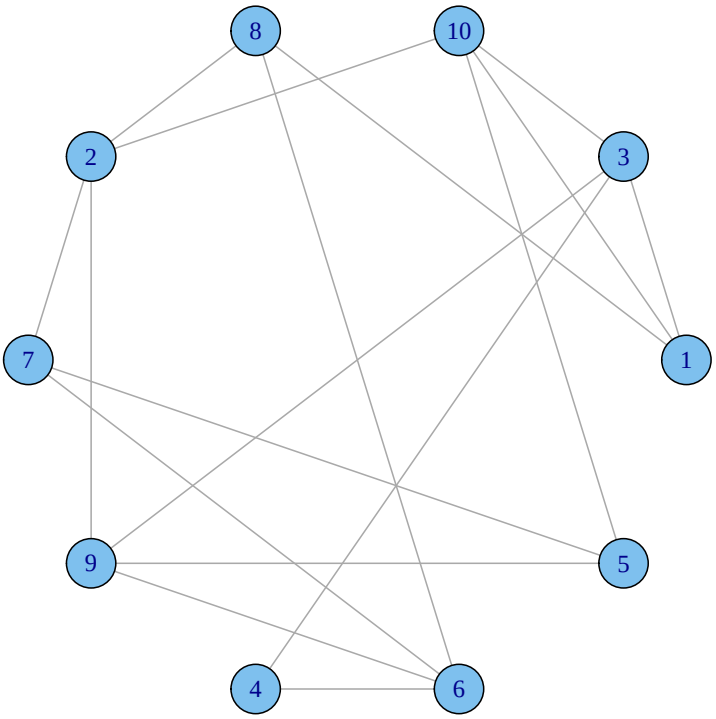

Figure 2: The independence graph for question (e)

	cliques
1	{1, 3, 10}
2	{1, 8}
3	{2, 7}
4	{2, 8}
5	{2, 9}
6	{2, 10}
7	{3, 4}
8	{3, 9}
9	{4, 6}
10	{5, 7}
11	{5, 9}
12	{5, 10}
13	{6, 7}
14	{6, 8}
15	{6, 9}

Table 4: The cliques of the resulting model for question (e)

(g) The models computed using BIC score are simpler than those found using AIC. The former have mostly small cliques (2 verteces) whereas the latter tend to have bigger cliques (3 verteces). The reason lies in the difference between BIC and AIC score, namely the weight used for the $dim(M)$ component. This factor represents the number of non-zero u-terms and determines the complexity of the model. For a model M, a small $dim(M)$ means that many u-terms are set to zero, resulting in a independence graph will smaller cliques. For AIC the wheight used is the constant 2, whereas for BIC it is $ln(N)$ where N is the number of observations in the data set. The data set used in this assignment contains 5735 rows, thus $log(5735) \approx 8.654343 > 2$. Using the BIC score the hill-climbing algorithm compensates the worse scores, due to this greater weight producing simpler models, i.e. models with fewer non-zero u-terms (smaller $dim(M)$, which result in smaller cliques.

(h) We have searched for better models using the `restart` approach with different parameters. For each parameter combination we performed 20 restarts, and both backward and forward search was always enabled. The complete results can be seen in table 7.

For the AIC scoring function, the restarting approach yields a minimally better result with a AIC score of 14263 for values of `prob` of $\{25\%, 50\%, 75\%\}$. The difference in AIC score between these three values

	cliques
1	{1, 2, 8}
2	{1, 2, 9}
3	{1, 2, 10}
4	{1, 3, 9}
5	{1, 3, 10}
6	{1, 4, 8}
7	{1, 4, 10}
8	{4, 5, 6}
9	{5, 6, 7}
10	{5, 6, 9}
11	{4, 5, 10}
12	{3, 6, 9}
13	{4, 6, 8}
14	{2, 7}

Table 5: The cliques of the resulting model for question (f), starting from the complete graph.

of `prob` is minimal and may well be just random noise.

For the BIC scoring function, the found models for `prob` values of $\{25\%, 50\%\}$ have a score of 15783. This is a minor improvement over the earlier models, which had a best score of 15841.

References

[1] Mark Hopkins, Erik Reeber, George Forman, Jaap Suermondt. *SPAM E-mail Database*. Hewlett-Packard Labs, 1501 Page Mill Rd., Palo Alto, CA 94304, June-July 1999.

	cliques
1	{1, 2, 8}
2	{1, 2, 9}
3	{1, 2, 10}
4	{2, 7}
5	{1, 3, 9}
6	{1, 3, 10}
7	{3, 6, 9}
8	{4, 5, 6}
9	{4, 6, 8}
10	{1, 4, 8}
11	{1, 4, 10}
12	{4, 5, 10}
13	{5, 6, 7}
14	{5, 6, 9}

Table 6: The cliques of the resulting model for question (f), starting from the empty graph.

	scorefunction	prob	score
1	aic	0	14278.2116691951
2	aic	0.25	14263.9723892169
3	aic	0.5	14263.9723897704
4	aic	0.75	14263.9723892216
5	aic	1	14278.2116692202
6	bic	0	15841.6564202562
7	bic	0.25	15783.7445130578
8	bic	0.5	15783.7445130578
9	bic	0.75	15843.3222805307
10	bic	1	15850.5261994279

Table 7: Result of training models with different parameters for question (h).