Graphical Models

Data Mining 2015: assignment 2

Sebastiaan Jong (5546303) & Bas Geerts (5568978)

1 Data Analysis

Questions:

(a) The data contains 10 columns, hence we have 10 nodes. Every possible configuration of edges between these 10 nodes is a graphical model. Looking at an 10 by 10 adjacency matrix, it is clear that $9+8+7+\cdots+2+1=45$ variables are required to define the graph. We can ignore most positions in the matrix, since it is symmetric and because nodes can not have edges to themselves. Over 45 binary variables there are 2^{45} different configurations.

(b) The amount of parameters in a saturated model are is equal to the amount of cells in the table of counts, since we make no independence assumptions. The amount of cells in the table of counts is equal to the product over all possibilities per variable. For this dataset, that is $7 \cdot 2 \cdot 2 \cdot 2 \cdot 3 \cdot 6 \cdot 4 \cdot 3 \cdot 5 \cdot 2 = 155520$ cells.

(c) Performing a forward-backward search with the BIC score function on this data starting from the empty graph gives us a model with a score of 15841.66 and 13 cliques. The cliques are listed in Table 1. The graphical model is shown in Figure 1.

$\{1,8\}$	$\{1,9\}$	$\{2,8\}$	$\{2,9\}$	$\{2,10\}$
$\{3,6\}$	$\{4,6\}$	$\{5,6\}$	$\{5,10\}$	$\{6,7\}$
$\{6,8\}$	$\{6,9\}$	$\{1,3,10\}$		

Table 1: Cliques found in (c).

(d) Based on the independence graph found in (c), we can state that *income* $\perp\!\!\!\perp$ *gender* $|$ *ninsclass*. To predict whether someone survives, we need the variables *ca*, *age* and *meansbp1*. The variable *death* has no edges to other nodes, so it is independent of the rest of the graph when these three variables are given.

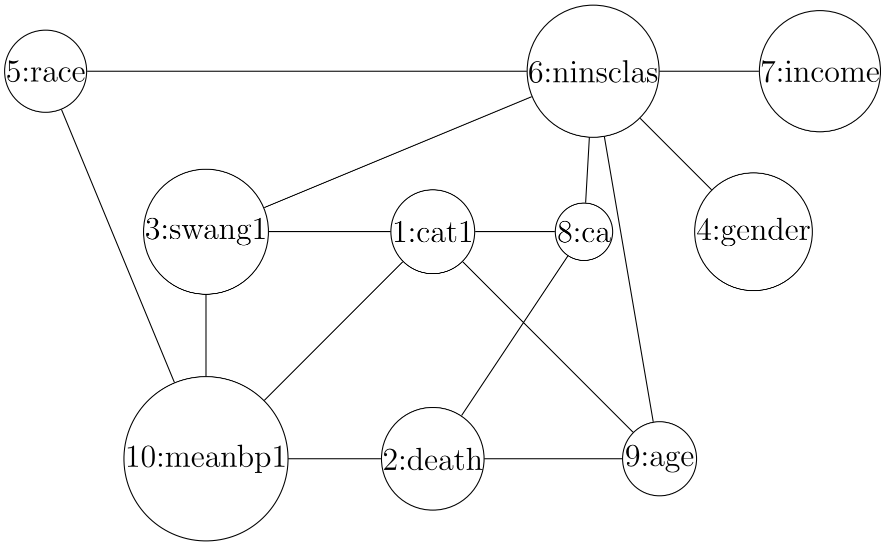

Figure 1: Independence graph found in (c).

(e) Performing a forward-backward search with the BIC score function on this data starting from the complete graph gives us a model with a score of 15850.53 and 15 cliques. The cliques are listed in Table 2. The graphical model is shown in Figure 2. One of the major differences of this graph compared to the one from (c) is that both income and now both have more than one edge. The score of these models is nearly the same.

$\{1,8\}$	$\{1,3,10\}$	$\{2,7\}$	$\{2,8\}$	$\{2,9\}$
$\{2,10\}$	$\{3,4\}$	$\{3,9\}$	$\{4,6\}$	$\{5,7\}$
$\{5,9\}$	$\{5,10\}$	$\{6,7\}$	$\{6,8\}$	$\{6,9\}$

Table 2: Cliques found in (e).

(f) Starting the local search with an empty graph and using the AIC scoring function a local optimum of 14278.21 was found by the algorithm. The 14 cliques for this search can be seen in Table 3. Using the complete graph, exactly the same score and cliques were found.

$\{4,5,6\}$	$\{4,6,8\}$	$\{1,4,8\}$	$\{1,4,10\}$	$\{4,5,10\}$
$\{2,7\}$	$\{5,6,7\}$	$\{1,2,8\}$	$\{1,2,9\}$	$\{1,3,9\}$
$\{3,6,9\}$	$\{5,6,9\}$	$\{1,2,10\}$	$\{1,3,10\}$	

Table 3: Cliques found in (f) while searching from an empty graph.

(g) The BIC scoring function penalizes large models more severely than AIC. This also can be seen in the results obtained in (f) compared to (c) and (e). In (e), we found

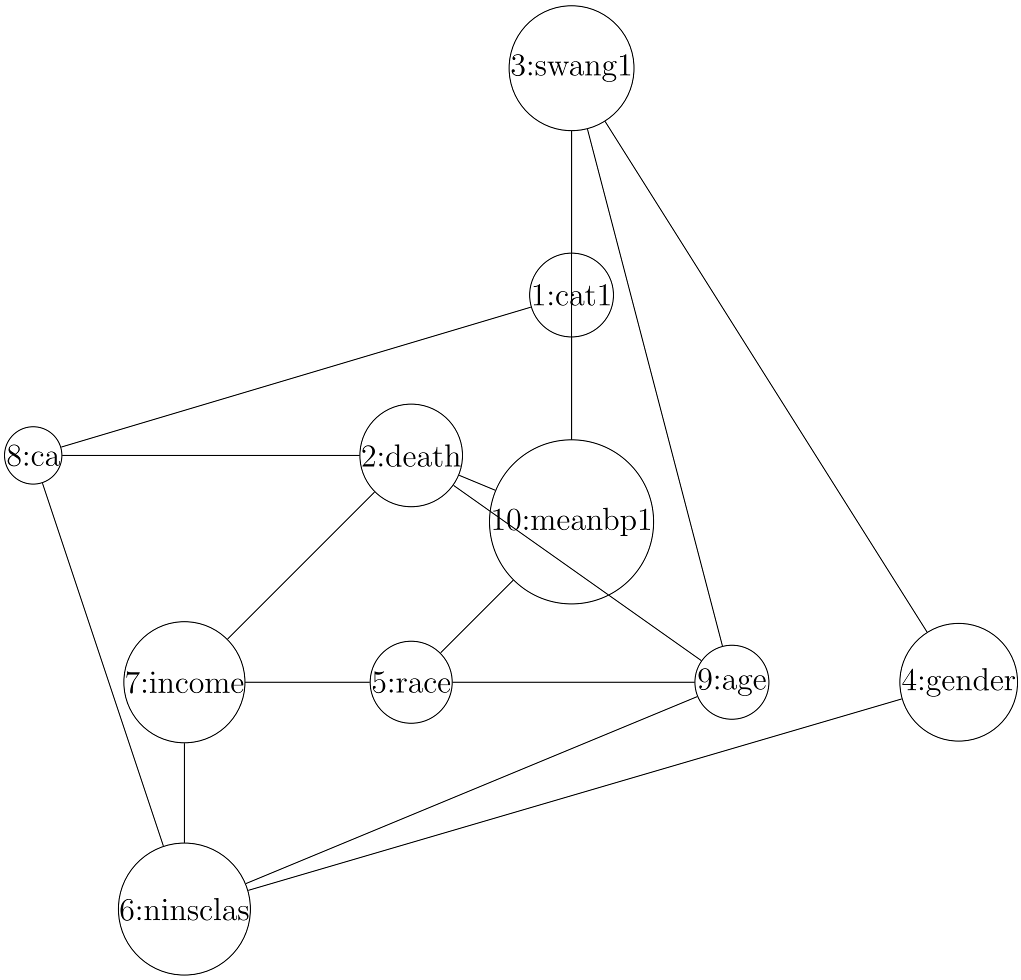

Figure 2: Independence graph found in (e).

several cliques of size 3, since the AIC function is more willing to add additional edges. Cliques found in (f) and (c) mainly are of size 2.

(h) For this experiment we used 487254080, which is a randomly generated number from an online service. As parameters we will try all $prob = \{0.2, 0.4, 0.6, 0.8\}$ with 10 restarts for both scoring functions. The idea behind these parameters is as follows. It is fairly pointless to try multiple restart values, the *gm.restart* function is seeded, so a call with $nstart = 10$ and $nstart = 5$ will cause the second call to generate the same 5 graphs as the first 5 in the first call. Since we use $nstart = 10$, it is also unlikely that trying probability values closer to eachother will yield any interesting results. There will not be much difference between 10 graphs generated with $prob = 0.5$ and $prob = 0.4$.

For the BIC scoring function, all runs ended with a score of 15783.74. During the earlier experiments without restarts, the best score found with BIC was 15841.66 starting from an empty graph. From the complete graph, a score of 15850.53 was achieved. These results tell us that starting from an empty of complete graph both result in finding a local optimum. It is quite likely that 15783.74 is the global optimum, considering the large amount of restarts. However, to be sure of this we would have to use an exact algorithm.
The amount of steps required to find the best score differ wildy for each probability. The runs with $prob = 0.2$ and $prob = 0.4$ only required 19 and 16 steps respectively. Meanwhile, the runs with $prob = 0.6$ and $prob = 0.8$ required 27 and 33 steps.

The run with the AIC scoring function all ended with the same score aswell. The best score found was 14263.97 for all probabilities tried. Once more, this is lower than the score found in earlier experiments (14278.21). For all runs, the amount of steps required was between 21 and 29. This indicates that this dataset in particular does not contain a lot of local optimums in which the multi-start local search algorithm can get stuck. The best AIC model can be seen in figure 3 and the BIC model in figure 4.

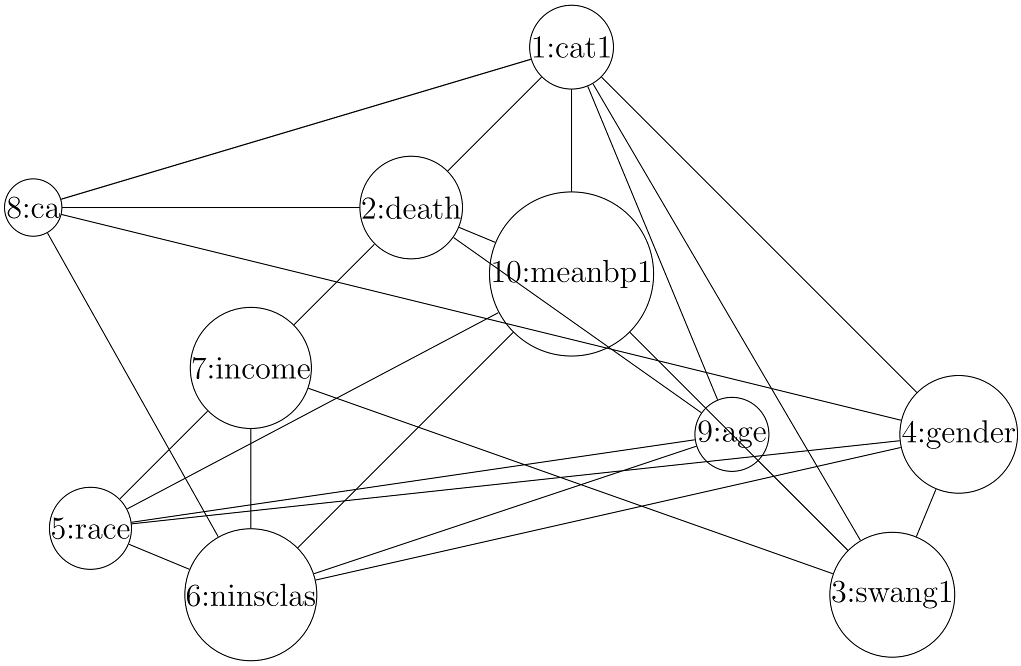

Figure 3: Independence graph found with the AIC function in (h).

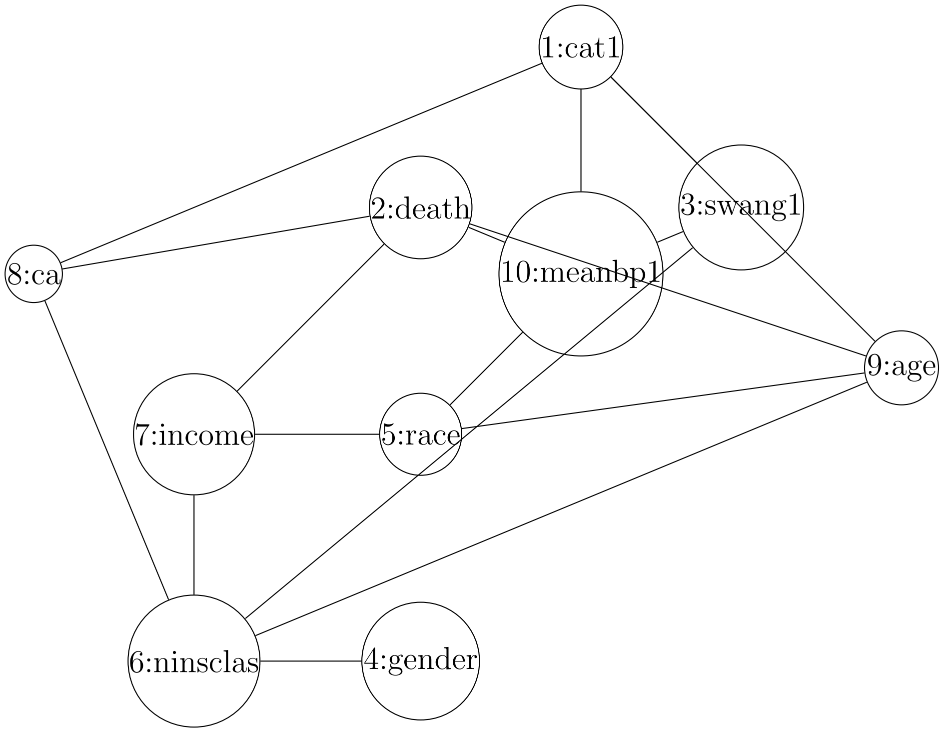

Figure 4: Independence graph found with the BIC function in (h).