EAS 507 : Statistical Data Mining II

Homework 3

Shubham Sharma (Person No.: 50290293, Class No. 43)

April 29, 2019

Question 1: Consider the cad1 data set in the package gRbase. There are 236 observations on fourteen variables from the Danish Heart Clinic. A structural learning algorithm has identified the optimal network as given below. For simplicity, not all of them are represented in the network.

The dataset $cad1$ has 236 observations of 14 variables which is a cross classified table with observational data from a Danish heart clinic.

a) Construct this network in R, and infer the Conditional Probability Tables using the cad1 data. (Hint: the function xtabs may be used). Identify any d- separations in the graph.

The network constructed in R is shown in figure 1. The d-separation identified from the graph are as follows:

1. Sex and $Hyperchol$ given $Smoker$ and $SuffHeartF$
2. $Inherit$ and $Hyperchol$ given $Smoker$ where smoker is a common cause
3. $Inherit$ and Sex given $Smoker$ because of indirect evidential effect of smoker
4. CAD and $Smoker$ given $Inherit$, $Hyperchol$
5. Sex and $Inherit$ given $Smoker$ where smoker has an indirect Causal Effect
6. Sex and $SuffHeartF$
7. $Smoker$ and $SuffHeartF$
8. $Inherit$ and $SuffHeartF$

The compiled network has 4 cliques and the maximal clique size is 3.

b) Suppose it is known that a new observation is female with Hypercholesterolemia (high cholesterol). Absorb this evidence into the graph, and revise the probabilities. How does the probability of heart-failure and coronary artery disease (CAD) change after this information is taken into account?

As we see the marginal probabilities before and after absorbing the given evidence in figure 2, we find that the marginal probability for having CAD increases from 0.46 to 0.61 and also for having heart failure ($SuffHeartF$), it increases from 0.29 to 0.38. Thus, given the evidence that a new person is female with Hypercholesterolemia (high cholesterol) we see that it increases the probability for coronary artery disease as well as the chances of heart failure. Here, we use a combination of evidential as well as causal reasoning to infer our results.

Similarly, the joint and the conditional probabilities before and after absorbing the given evidence is shown in figure 3 and 4 respectively.

c) Simulate a new data set with 5 observations conditional upon this new information. Present this new data in a table. Using the new data set and the predict function to estimate the probability of Smoker and CAD given the other variables in your model.

The simulated data corresponding to our propagated network is shown in figure 5. The predictions for the simulated observations are shown in figure 6. The Probabilities for the simulated data for $Smoker$

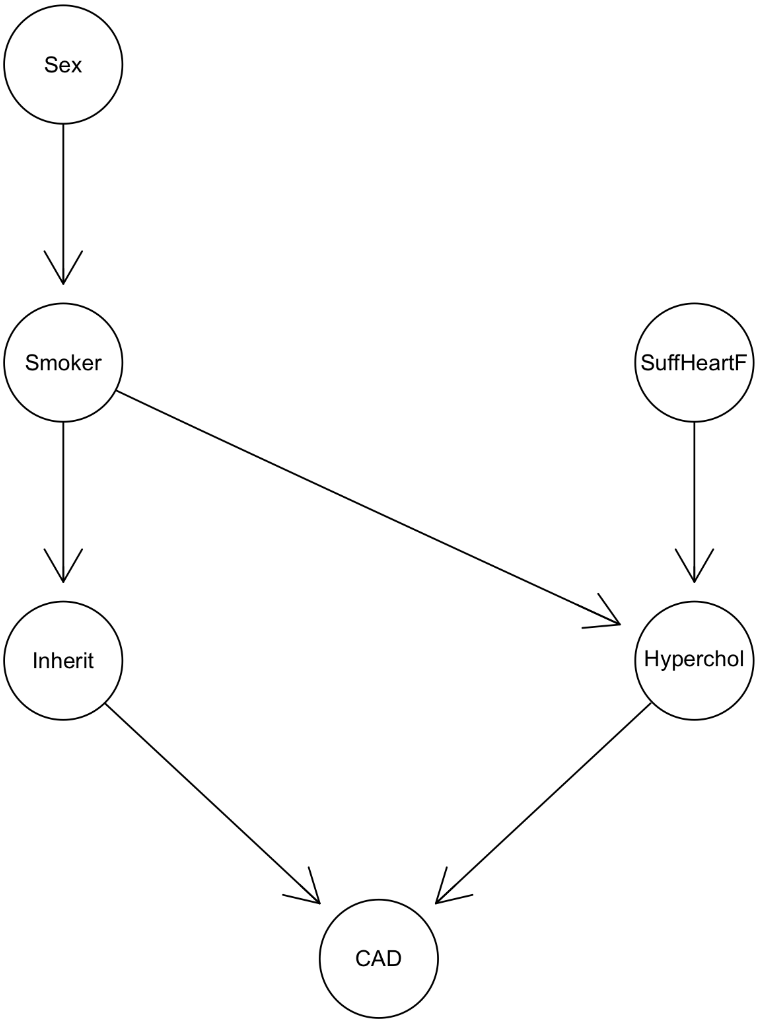
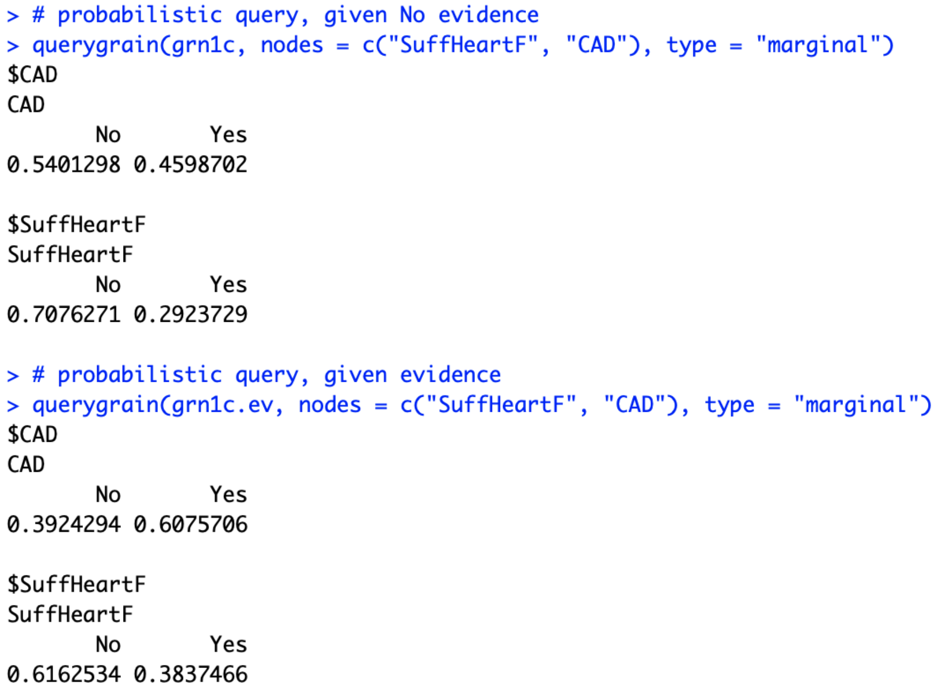
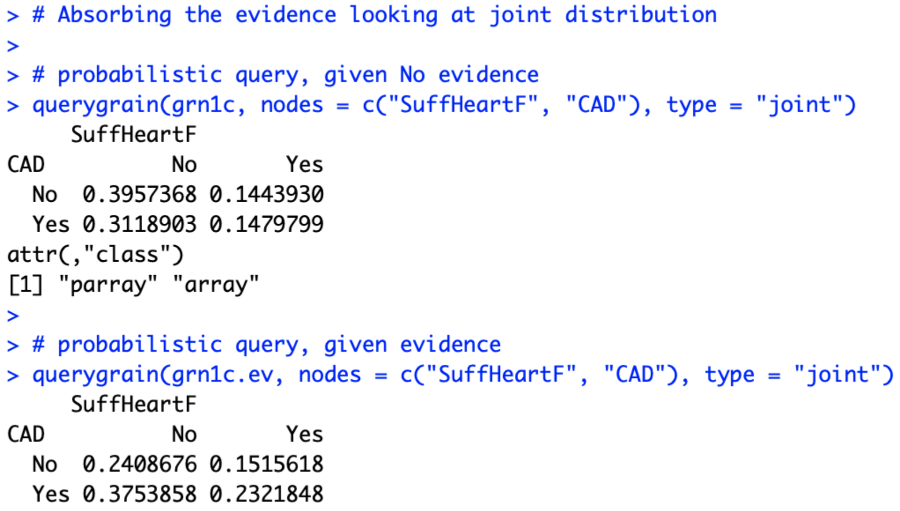


Figure 1: Network constructed in R



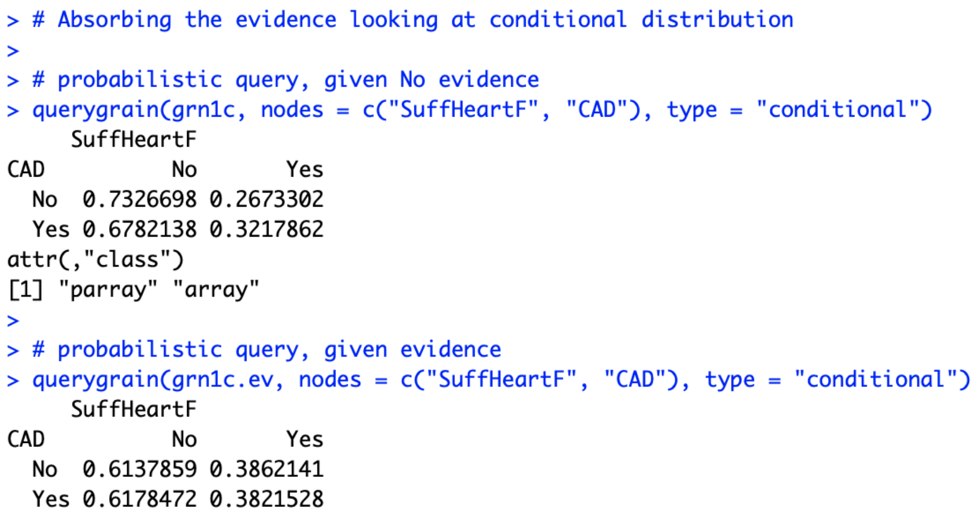
```
> # probabilistic query, given No evidence
> querygrain(grn1c, nodes = c("SuffHeartF", "CAD"), type = "marginal")
$CAD
CAD
       No       Yes 
0.5401298 0.4598702 

$SuffHeartF
SuffHeartF
       No       Yes 
0.7076271 0.2923729 

> # probabilistic query, given evidence
> querygrain(grn1c.ev, nodes = c("SuffHeartF", "CAD"), type = "marginal")
$CAD
CAD
       No       Yes 
0.3924294 0.6075706 

$SuffHeartF
SuffHeartF
       No       Yes 
0.6162534 0.3837466 
```

Figure 2: Marginal Probabilities before and after absorbing the given evidence



```
> # Absorbing the evidence looking at joint distribution
> 
> # probabilistic query, given No evidence
> querygrain(grn1c, nodes = c("SuffHeartF", "CAD"), type = "joint")
     SuffHeartF
CAD          No       Yes
  No  0.3957368 0.1443930
  Yes 0.3118903 0.1479799
attr(,"class")
[1] "parray" "array" 
> 
> # probabilistic query, given evidence
> querygrain(grn1c.ev, nodes = c("SuffHeartF", "CAD"), type = "joint")
     SuffHeartF
CAD          No       Yes
  No  0.2408676 0.1515618
  Yes 0.3753858 0.2321848
```

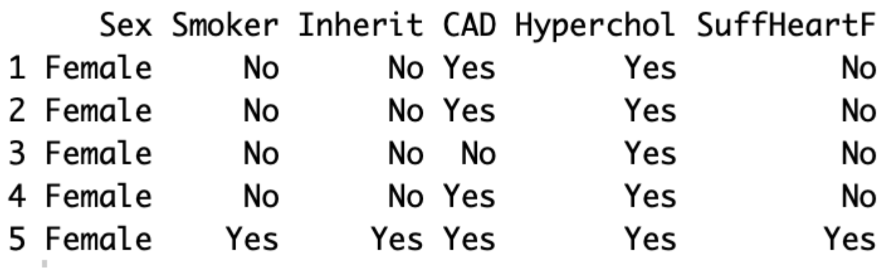
Figure 3: Joint Probabilities before and after absorbing the given evidence



```
> # Absorbing the evidence looking at conditional distribution
>
> # probabilistic query, given No evidence
> querygrain(grn1c, nodes = c("SuffHeartF", "CAD"), type = "conditional")
     SuffHeartF
CAD          No       Yes
  No  0.7326698 0.2673302
  Yes 0.6782138 0.3217862
attr(,"class")
[1] "parray" "array"
>
> # probabilistic query, given evidence
> querygrain(grn1c.ev, nodes = c("SuffHeartF", "CAD"), type = "conditional")
     SuffHeartF
CAD          No       Yes
  No  0.6137859 0.3862141
  Yes 0.6178472 0.3821528
```

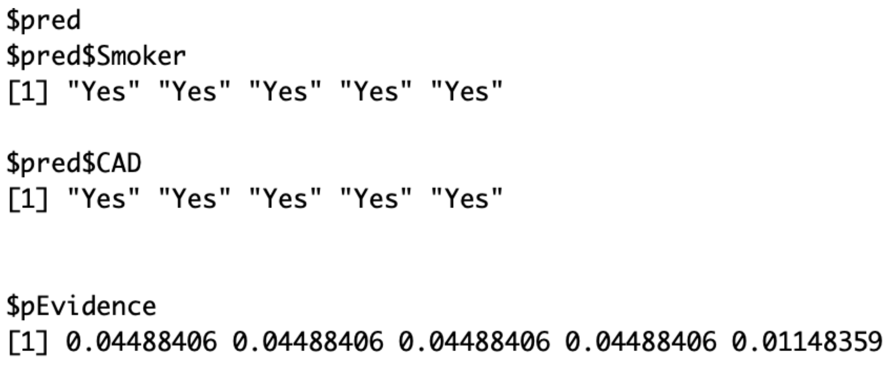
Figure 4: Conditional Probabilities before and after absorbing the given evidence

and CAD are shown in figure 7.



```
     Sex Smoker Inherit CAD Hyperchol SuffHeartF
1 Female     No      No Yes       Yes         No
2 Female     No      No Yes       Yes         No
3 Female     No      No  No       Yes         No
4 Female     No      No Yes       Yes         No
5 Female    Yes     Yes Yes       Yes        Yes
```

Figure 5: Simulated data corresponding to our propagated network

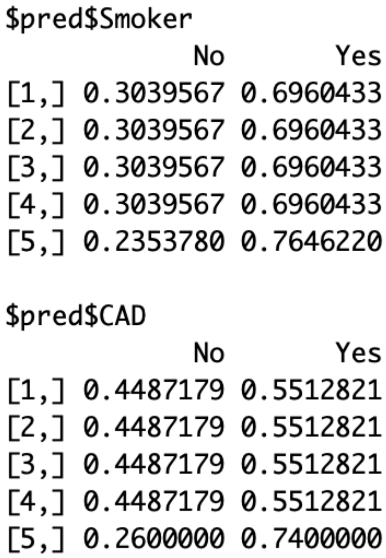


```
$pred
$pred$Smoker
[1] "Yes" "Yes" "Yes" "Yes" "Yes"

$pred$CAD
[1] "Yes" "Yes" "Yes" "Yes" "Yes"


$pEvidence
[1] 0.04488406 0.04488406 0.04488406 0.04488406 0.01148359
```

Figure 6: Predictions for the simulated data

d) Create a new data set, as done in part C, this time with 500 data points. Save this data and submit it with your assignment (form: *.RData or *.txt file). Use this data and the predict function to estimate the probability of Smoker and CAD given the other variables in your model. Calculate the misclassification rate. Comment on the performance of the network for predictive purposes, and what might be done to improve it.

The mis-classification rate for Smoker is 35.4 % while for CAD is 40.2 %. Clearly, the performance is not good with an accuracy of 60 - 65 % for the predicted values. More permutations of networks



```
$pred$Smoker
            No       Yes
[1,] 0.3039567 0.6960433
[2,] 0.3039567 0.6960433
[3,] 0.3039567 0.6960433
[4,] 0.3039567 0.6960433
[5,] 0.2353780 0.7646220

$pred$CAD
            No       Yes
[1,] 0.4487179 0.5512821
[2,] 0.4487179 0.5512821
[3,] 0.4487179 0.5512821
[4,] 0.4487179 0.5512821
[5,] 0.2600000 0.7400000
```

Figure 7: Probabilities for the simulated data

can result in better performance since Bayesian model averaging generates an ensemble of possible structures and averages the prediction of all possible structures; due to the immense number of structures, approximations are needed. Also, domain knowledge to incorporate prior information will surely improve the performance.

Question 2: Consider the following famous Bayesian Network by Judea Pearl.

The network is set up to answer questions of the following type: I am at work, neighbor John calls to say my alarm is ringing, but neighbor Mary does not call. Sometimes minor earthquakes set it off. Is there a burglar? One operation on Bayesian Networks that arises in many settings is the marginalization of some node in the network. Let the original Bayesian Network be denoted as B. Construct a Bayesian Network B' over all of the nodes EXCEPT for Alarm that is the minimal I-map for the marginal distribution P_B(B, E, T, N, J, M). Be sure to get all dependencies that remain from the original graph.

The Bayesian Network B' over all of the nodes EXCEPT for Alarm that is the minimal I-map for the marginal distribution P_B(B, E, T, N, J, M) is shown in figure 8.

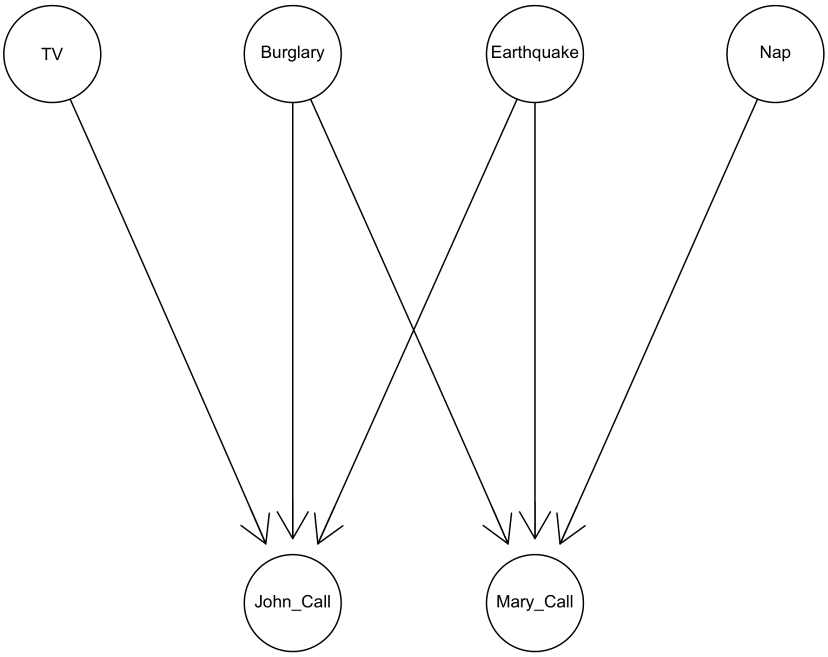


Figure 8: Bayesian Network B'

Question 3: Determine if the following statements are TRUE OR FALSE based on the DAG.

A) C and G are d-separated: False
B) C and E are d-separated: True
C) C and E are d-connected given evidence about G: True
D) A and G are d-connected given evidence about D and E: False
E) A and G are d-connected given evidence on D: True