Statistical Data Mining II
Homework 3
Due 4/26

40 points

Directions: Submit all source codes with write up. You must provide thorough explanations with output. See “homework guidelines” on UB learns for detailed information.

1) (20 points) 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.

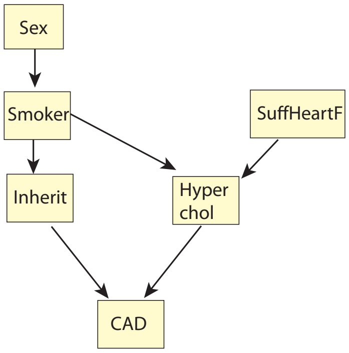
   a) Construct this network in R, and infer the Conditional Probability Tables using the cad1 data. (Hint: the function xtabs may be used). Identify any d-separations in the graph.

   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?

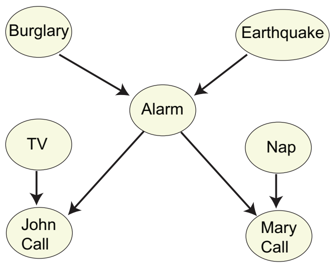
   c) Simulate a new data set with 5 observations conditional upon this new information. Present this new data in a table. Using the new data set and the “predict” function to estimate the probability of “Smoker” and “CAD” given the other variables in your model.

   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.

Coronary Artery Disease



2) (10 points, adopted from exercise 3.11 in Koller et al.) Consider the following famous Bayesian Network by Judea Pearl.

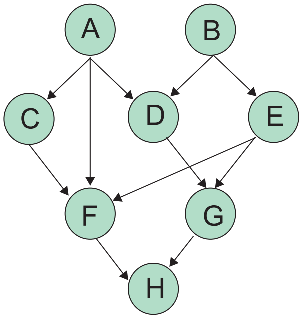

The Alarm Network

The network is set up to answer questions of the following type:
"I'm 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 $P_B(\mathrm{B, E, T, N, J, M})$. Be sure to get all dependencies that remain from the original graph.

3) (10 points) Determine if the following statements are "TRUE OR FALSE" based on the DAG.



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