Statistical Data Mining II
Homework 4

Due: Friday May 6th (11:59 pm)
50 points

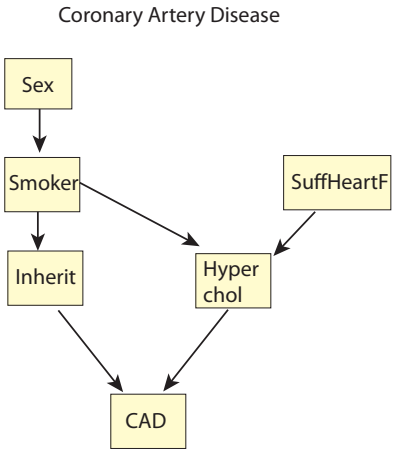
Directions: Complete **all five** exercises. Submit all source codes with write up. Do not embed your code in your write up. Please visit "homework guidelines" on UB learns for detailed information.

1) (10 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 variables in the network are represented.

   a) Construct this network in R, and infer the Conditional Probability Tables using the cad1 data. 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 new observations conditional upon this new evidence (in part B). Save this new data as a *.txt file, and submit it with your assignment. Using the new data set determine the joint distribution of "Smoker" and "CAD" given this evidence.
   (hint: use "simulate" in the gRain package (for help: >?simulate.grain))



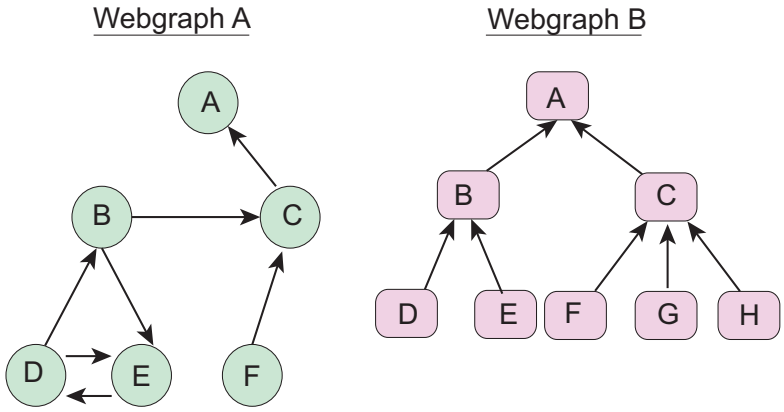
2) (10 points) The data "heart.txt" available on UB Learns, was generated by the "reinis" data in the gRbase library. Data was collected at the beginning of a 15 year follow-up study of probable risk factors (yes implies risk) for coronary

thrombosis. Data are from all men employed in a car factory. The variables are: smoking, strenuous mental work, strenuous physical work, systolic blood pressure, ratio of lipoproteins, Family anamnesis of coronary heart disease. (See ?reinis for details on the variables.). The variables systolic blood pressure, and ratio of lipoproteins are clinical risk measurements (markers) for heart disease.

a) Learn the structure of a network, which is both likely and sensible in terms of the variable meanings.

b) Based on your model, who is more likely to develop high systolic blood pressure (risk = yes), a person with strenuous mental work, or one with strenuous physical work, or both?

(3) (10 points) Consider the following webgraphs.



(a) Compute the PageRank vector of Webgraph A for damping constants p = 0.05, 0.25, 0.50, 0.75, and 0.95. How sensitive is the PageRank vector, and overall ranking of importance, to the damping constant? Does the relative ranking of importance according to PageRank support your intuition?

(b) Compute the PageRank vector of Webgraph B for damping constant p = 0.15. Interpret your results in terms of the relationship between the number of incoming links that each node has. Does the relative ranking of importance according to PageRank support your intuition?

(4) (10 points) The sinking of the Titanic is a famous event in history. The titanic data (UB learns) was collected by the British Board of Trade to investigate the sinking. Many well-known facts—from the proportions of first-class passengers to the 'women and children first' policy, and the fact that that policy was not entirely successful in saving the women and children in the third class—are reflected in the survival rates for various classes of passenger.

You have been petitioned to investigate this data. Analyze this data with tool(s) that we learned in STA546. Summarize your findings for British Board of Trade. Is their evidence that "women and children" were the first evacuated? What characteristics/demographics are more likely in surviving passengers? What characteristics/demographics are more likely in passengers that perished? How do your results support the popular movie "Titanic"? For example, what is the probability that Rose (1st class adult and female) and (3rd class adult and male) would not survive?

(5) (10 points) Data released from the US department of Commerce, Bureau of the Census is available in R

- data(state)
- ?state

a) Focus on the data {Population, Income, Illiteracy, Life Exp, Murder, HS Grad, Frost, Area}. Cluster this data using at least two tools learned in STA 546 (e.g., k-means, hierarchical clustering, SOM, PCA). Keep the class labels (region, or state name) in mind, but do not use them in the modeling. Report your detailed findings.

b) Build a Gaussian Graphical Model using the Graphical Lasso for the 8 predictors mentioned in Part A. What do you find for different penalties, and how does it compliment (and/or contradict) your results in part A?

Extra Credit +5

(Exercise 17.1 in Textbook) For the Markov graph of Figure 17.8, list all of the implied conditional independence relations and find the maximal cliques.

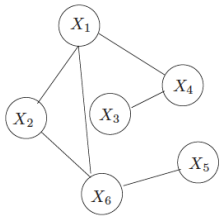


FIGURE 17.8.