Final Homework Set

Directions: Complete FOUR exercises.

1. Consider the "cad1" data set in the package gRbase. These observations are from individuals in the Danish Heart Clinic.

   (a) Learn a Bayesian Network using a structural learning knowledge, and prior knowledge obtained through the definitions of the variables in the help files. You do not have to use all of the variables. Make sure to detail your network construction process.

   (b) Construct the above network in R, and infer the Conditional Probability Tables using the cad1 data. (Hint: extractCPT or cptable may be used from the gRain package). Identify any d-separations in the graph.

   (c) 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?

   (d) Simulate a new data set with 100 observations either conditional upon this new information in part (C) using the original paramaterization. Present this new data in a table. Estimate the probability of "Smoker" and "CAD" given the other variables in your model. (Hint: you may try simulate.grain from the gRain package, you may use predict.grain as well).

2. The sinking of the Titanic is a famous event in history. The titanic data 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 class. Summarize your findings for British Board of Trade.

   In your report, please touch on the following questions. Is their evidence that *women and children* were the evacuated first? 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" (1997)? For example, what is the probability that Rose (1st class adult and female) would survive and Jack (3rd class adult and male) would not survive?

3. Specify the structure of a Bayesian Network that contains four nodes $\{W, X, Y, Z\}$ and has satisfies the following set of independencies.

$$
\begin{aligned}
W &\perp X \\
W &\not\perp Z \mid X \\
Z &\perp W \mid Y \\
W &\not\perp Y \\
X &\not\perp Y \\
W &\not\perp X \mid Z \\
X &\perp Z \mid W, Y
\end{aligned}
$$

4. Data released from the US department of Commerce, Bureau of the Census is available in R.
   >data(state)
   >?state

   Build a Gaussian Graphical Model using the Graphical Lasso for the 8 predictors (Population, Income, Illiteracy, Life Exp, Murder, HS Grad, Frost, Area) using a range of penalties. What do you find for different penalties, and how does it compliment (and/or contradict) a model fit with SOM?

5. Write a function to implement single linkage, average linkage and complete linkage agglomerative clustering. Write your function as general as possible, and comment your code. The functions need to be compatible with any dissimilarity. Demo your code with the 'iris' data for each linkage and plot the results.

6. From Probabilistic Graphical Models textbook (Koller).

Exercise 3.5

Consider the Bayesian network of figure 3.14.

Assume that all variables are binary-valued. We do not know the CPDs, but do know how each random variable *qualitatively* affects its children. The influences, shown in the figure, have the following interpretation:

- $X \xrightarrow{+} Y$ means $P(y^1 \mid x^1, \boldsymbol{u}) > P(y^1 \mid x^0, \boldsymbol{u})$, for all values $\boldsymbol{u}$ of Y's other parents.
- $X \xrightarrow{-} Y$ means $P(y^1 \mid x^1, \boldsymbol{u}) < P(y^1 \mid x^0, \boldsymbol{u})$, for all values $\boldsymbol{u}$ of Y's other parents.

We also assume explaining away as the interaction for all cases of intercausal reasoning.

For each of the following pairs of conditional probability queries, use the information in the network to determine if one is larger than the other, if they are equal, or if they are incomparable. For each pair of queries, indicate all relevant active trails, and their direction of influence.

(a)	$P(t^1 \mid d^1)$	$P(t^1)$
(b)	$P(d^1 \mid t^0)$	$P(d^1)$
(c)	$P(h^1 \mid e^1, f^1)$	$P(h^1 \mid e^1)$
(d)	$P(c^1 \mid f^0)$	$P(c^1)$
(e)	$P(c^1 \mid h^0)$	$P(c^1)$
(f)	$P(c^1 \mid h^0, f^0)$	$P(c^1 \mid h^0)$
(g)	$P(d^1 \mid h^1, e^0)$	$P(d^1 \mid h^1)$
(h)	$P(d^1 \mid e^1, f^0, w^1)$	$P(d^1 \mid e^1, f^0)$
(i)	$P(t^1 \mid w^1, f^0)$	$P(t^1 \mid w^1)$

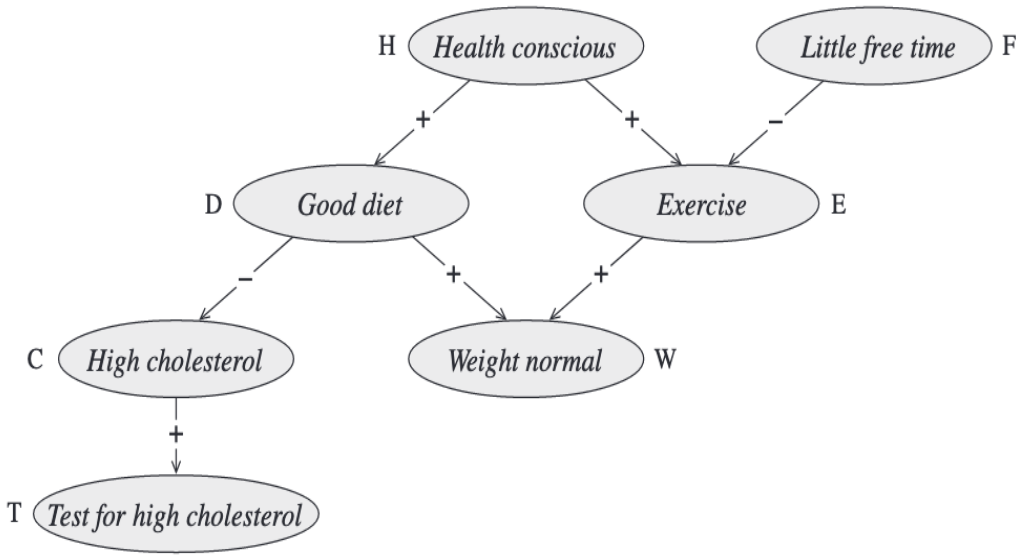

Figure 3.14 A Bayesian network with qualitative influences