Final Homework Set

Directions: Complete two excercises.

1. 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 (3rd class adult and male) would not survive?

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

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

3. Consider the MovieLense data that is available in the *recommenderlab* package. The data was collected through the MovieLens web site during a seven-month, and contains about 100,000 ratings (1-5) from 943 users on 1664 movies. See the help file on the data to understand how to best manipulate the object.

   Design and evaluate your own recommendation system based on the following principles:
   For each user i and each movie j they did not see, find the k most similar users to i who have seen j and then use them to infer the user i's rating on the movie. Handle all exceptions in a reasonable way and report your strategy if you did so; e.g., if you cannot find k users for some movie j, then take all users that have seen it.

4. *****Only available if you completed exercise 3*****
   Continuing from question 3. Test the performance of your system using either a test/training approach, or cross-validation. You may look to the recommenderlab paper that we discussed in class for additional guidance and strategy.

   For example, if you divide your data into $K = 5$ folds, then this means that you will run your algorithm from exactly 5 times. Each time, you will be use the training partition to make predictions for each user on all terms rated in the test partition (by that user). When you complete all K iterations, you will have a large number of user-movie pairs from the test partitions, which you can use to evaluate the performance of your system. (This is only an example, other approaches are possible).

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

   (a) Focus on the data Population, Income, Illiteracy, Life Exp, Murder, HS Grad, Frost, Area. Cluster this data using hierarchical clustering. Keep the class labels (region, or state name) in mind, but do not use them in the modeling. Report your detailed findings. ** You may have done this step in an earlier assignment.

   (b) Focus on the data Population, Income, Illiteracy, Life Exp, Murder, HS Grad, Frost, Area. Cluster this data using SOM. Keep the class labels (region, or state name) in mind, but do not use them in the modeling. Report your detailed findings. ** You may have done this step in an earlier assignment.

   (c) Build a Gaussian Graphical Model using the Graphical Lasso for Population, Income, Illiteracy, Life Exp, Murder, HS Grad, Frost, Area. What do you find for different penalties, and how does it compliment (and/or contradict) your results in part A and B?

   (d) Describe some of the advantages of clustering to GGMs in the context of this problem, and more generally.

6. Consider the Parkinsons Telemonitoring Dataset on the UCI Machine learning repository. This data set was developed with 10 medical centers. Together with a corporation, they developed a telemonitoring device to record speech signals of patients for the prediction of clinical Parkinson?s disease symptom scores on a UPDRS scale. This data is designed for supervised learning, however, we are going to *pretend* that there are no labels/ no response variable.

   (a) Perform basic exploratory data analysis, and present your results. Cluster this *clean data* over using a sensible subset of variables using two methods described in class.

   (b) Fit a Bayesian Network using this data. Include "motorUPDRS" and "totalUPDRS", but not both, and force this variable to be the bottom node of the network.

   (c) A collaborator asks you to characterize "Jitter" related variables for a new patient that has a relatively high UPDRS score (two standard deviations above the mean). Use your Bayesian Network to answer this question.