Coursework Part 2: The Maximally Weighted Spanning Tree

This coursework is concerned with the spanning tree algorithm for finding a singly connected Bayesian network from a data set. The material you need will be completely covered by Lecture 6 of the course. You can use the same skeleton program that was set up for Coursework 1, continuing to fill in code where indicated. You will need to use the function that you wrote to compute the joint probability distribution of a pair of variables. If your original solution was wrong and you are stuck then email me and I will give you some working code to do this (after the hand in date for coursework 1).

Task 2.1: 6 marks

Complete the function "MutualInformation" which calculates the mutual information (or Kullback Leibler divergence) of two variables from their joint probability table. The process involves marginalising the joint probability table and then applying the formula as described in Lecture 6.

Task 2.2: 5 marks

Complete the function "DependencyMatrix" which uses mutual information as a measure and creates a symmetric matrix showing the pairwise dependencies between the variables in a data set.

Task 2.3: 6 marks

Complete the function "DependencyList" which turns the dependency matrix into a list or arcs ordered by their dependency. Your list items should be triplets: [dependency, node1, node2]. Using your dependency list draw by hand the Bayesian network that you would get for the HepatitisC data set. Make an image file of your network (either by scanning your hand drawing or using a drawing package (eg powerpoint or open office)).

Task 2.4: 6 marks (difficult)

Starting with a dependency list find the maximally weighted spanning tree automatically, and append it as a list to your results file. Don't wory about implementing marginal independence to find the causal directions.

Results File

You should finish by writing a main program part at the end of the skeleton file which will add the following items to your results file:

1. A title giving your group members

2. The dependency matrix for the HepatitisC data set

3. The dependency list for the HepatitisC data set.

4. Ths spanning tree found for the HepatitisC data set (if you attempted task 2.4).

Assemble the results listed above with the network picture into a .pdf file (You can do this with word, open office or latex). Rename your python file "DAPICoursework02.py" and submit it and your results file named "DAPIResults02.pdf" using CATE.