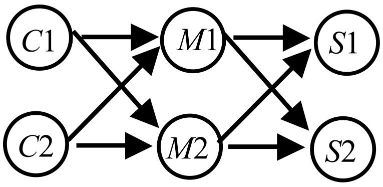
Midterm Algorithmical and Statistical Modeling, Fall 2010

Herbert's part.

Problem 9. This is a little excursion into the domain of *diagnostic reasoning*. Assume you are a doctor and a patient consults you, complaining about pain in the abdomen and fatigue. These are *symptoms*, which we model by random variables $S1$ (abdomen pain) and $S2$ (fatigue). Simplifying the world of medicine a bit, we assume that both are binary variables, that is, a patient either has pain in the abdomen ($S1 = 1$) or not ($S1 = 0$); similar for $S2$. We simplify the world a bit more and assume that there are (only) two possible *causes* that can lead to this sort of pain and/or fatigue, namely $C1$ = gastric ulcers and $C2$ = food poisoning. Again we take these to be binary. Furthermore, these causes do not directly cause the symptoms, but take effect only through *mediating factors*, let's say $M1$ = increased intestinal acidity and $M2$ = intestinal bleeding (I made this up, please don't start treating yourself based on this medical wisdom!). Again, for simplicity we assume that $M1$ and $M2$ are binary. In total we thus have a medical model comprising six binary RVs arranged in a causal cascade structured as in the directed graph below, where an arrow $X \rightarrow Y$ means "X has a direct causal effect on Y".



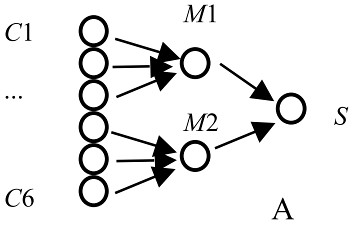
Thus, according to this model, the probability that a patient shows symptom Si ($i = 1, 2$) is determined by the values that $M1$ and $M2$ take; and the probability that a patient has a positive mediating factor Mi is determined by the values that $C1$ and $C2$ take. Furthermore, we assume that $C1$ is independent of $C2$. We finally assume that the prior probabilities $P(Ci = 1)$ in the population, and the conditional probabilities of the kind $P(Mi = 1 \mid C1 = x, C2 = y)$ and $P(Si = 1 \mid M1 = x, M2 = y)$, where $i = 1, 2$; $x, y \in \{0, 1\}$, are known from clinical studies. These quantities are the model parameters.

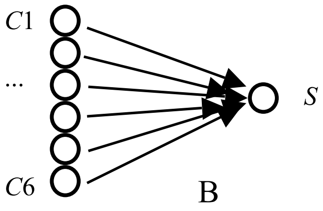
(a, 4 pts) Describe in plain English a reasonable choice for the domain Ω of the underlying probability space. (One sentence should be enough.)

(b, 8 pts) Give a formula for computing $P(S1 =1)$ from the known model parameters.

(c, 10 pts) Give a procedure for computing $P(C1 = 1 \mid S1 = 1, S2 = 1)$ for a patient who complains about both abdominal pain and fatigue.

(d, 6 pts) Now consider the structures of two more complex models for explaining a single symptom S in terms of six possible causes $C1 - C6$. Model A has two mediating factors, model B has none (see figure below). Using the same kind of model parameters as in the simple example above (i.e., population probabilities of causes and assorted conditional probabilities), how many parameters do models A vs. B have?





(e, 12 pts) Now assume that the parameters of these two models A and B have to be estimated from empirical observations. In an expensive clinical survey (sponsored by a pharmaceutical company -- we are getting realistic), from the population a sample of size N of randomly selected people is drawn and for each of them all of $C1 - C6$ and S, $M1$, $M2$ are recorded (for estimating parameters of model B the *Mi* observations are irrelevant). Each data point is thus a 9-dimensional binary vector for estimating A, and a 7-dim vector for B. Explain qualitatively in plain English how you would determine an appropriate sample size *N* for the estimation of A vs. B, and highlight what difference in appropriate *N* between the two cases you would expect. Note: this is a difficult and involved (but important) issue. Don't attempt to be exhaustive or rigorous – the subject is a research field of its own. I only want to see that you discern some of the difficulties inherent in this business of fixing an appropriate sample size, and how it depends on the assumed model structure. Target size of your explanation: about as long as the statement of this point **(e)**.

(f, 12 pts) Now we get even more practical. From thinking about **(d)** and **(e)** you have seen that model A seems preferable over B because it can be estimated from a smaller dataset. In practice, one often assumes models with a hierarchical structure as in A, *without making any commitments about the nature of the mediating variables*. One simply assumes that such $M1$, $M2$ exist and are coupled into a causal cascade as in the figure above, but one is (and remains) ignorant about their clinical / biological nature. In estimating the parameters for a type A model, one thus has only observations of $C1$ – $C6$ and S, i.e. 7-dim data points, and one has committed to the causal structure as in the figure above. Explain in plain English how you would set up an algorithm for estimating the parameters of a type A model from such data. Target size of your explanation: same as for **(e)**.

(g, 20 pts) Finally, we get more realistic. Such *causal dependency graphs* as we have considered so far are in fact used as a basis for diagnostic and decision making systems in probabilistic domains (among other, in medical diagnosis, but there are many more application areas). However, in practice these graphs are typically much bigger than the toy examples above, easily comprising hundreds or even thousands of nodes (= random variables). When such causal dependency graphs have no cycles (as in our examples), one speaks of *Bayesian networks*. The root nodes in such graphs correspond to the (ultimate) causes, the leaves to symptoms, and the other nodes to mediating factors, just as in our simple cases above. One usage is to infer back from observed symptoms to probablities of causes (as in **(c)**). Analytical solutions like the one you gave in **(c)** quickly grow too large to be computable due to combinatorial explosion. In that case, one can take resort to *sampling* to arrive at approximations to the correct probabilities. Describe in plain English how you would set up such a sampling scheme to find out the probability $P(C = 1 \mid S_1 = x_1, ..., S_n = x_n)$ of a particular cause C (one of the root nodes in the graph) given values $x_1, ..., x_n$ of some of the symptoms (i.e. some of the leaf nodes). Note that not all symptom nodes need to have known values in a particular diagnostic situation. Assume for simplicity that all RVs are binary. Your solution should contain a specification of the state space over which one samples, and of a reasonable proposal distribution. Target size of your explanation: about as long as this problem statement.

Solution.

(b)

$$P(S1=1,S2=1) = \sum_{x,y,u,v\in\{0,1\}} P(C1=x)P(C2=y)P(M1=u \mid C1=x,C2=y)\cdot$$
$$\cdot P(M2=v \mid C1=x,C2=y)P(S1=1 \mid M1=u,M2=v)P(S2=1 \mid M1=u,M2=v)$$

(c) By Bayes rule, we get

$$P(C1=1 \mid S1=1,S2=1) = \frac{P(S1=1,S2=1 \mid C1=1)P(C1=1)}{P(S1=1,S2=1)}.$$

The denominator can be computed as in **(a)**. $P(C1 = 1)$ is a known model parameter. It remains to compute $P(S1 = 1, S2 = 1 \mid C1 = 1)$, which is a variant of the solution to **(a)**:

$$P(S1=1,S2=1 \mid C1=1) = \sum_{y,u,v\in\{0,1\}} P(C2=y)P(M1=u \mid C1=1,C2=y)\cdot$$
$$\cdot P(M2=v \mid C1=1,C2=y)P(S1=1 \mid M1=u,M2=v)P(S2=1 \mid M1=u,M2=v)$$

(d) Model A has 6 + 2 * 2^3 + 4 = 26 nominal parameters, model B has 6 + 2^6 = 70. Because conditional probabilities sum to one, each conditional probability effectively has one free parameter less, so a more correct account would be for model A: 6 + 2*7 + 3 = 23 and for B: 6 + 63 = 69 parameters.

(e) A very dirty estimate of *N* would rely on the rule of thumb that one needs at least 10 times as many data points as one has model parameters, that is, *N* = 260 for A and *N* = 700 for B. However, the rule of thumb is inappropriate if used in such a global way. What one wants to determine are (for model B) conditional probabilities of the kind $P(S = 1 \mid C1 = x1, ..., C6 = x6)$. We have to use the rule of thumb locally, that is, for each such conditional parameter we need at least 10 data points. The terrible insight is that some of the combinations of $C1 = x1, ..., C6 = x6$ will be very rare in the population. The global sample size *N* would have to rise to astronomical numbers to ensure that for each *C*-combination there are at least 10 data points available. The practical solution taken is to choose a *balanced* sample, that is, one would actively select 10 cases for each value combination of the *Ci*, which sums to 640 carefully selected cases. Such a sample would not give information about the base rates of the *Ci* in the population; these would have to be estimated from a separate, unbalanced sample. For model B, one would similarly take samples that are balanced w.r.t. the combination triples of *C*1-*C*3 and *C*4-*C*6, respectively, which would sum to 160 cases; plus balanced samples for the estimation of the *S*-given-*M* conditionals, which would be another 40.

(f) The natural choice is to invoke an EM algorithm. It would be initialized with made-up values θ_0 for the *S*-given-*M* and the *M*-given-*C* conditional probs. In the E-step, using the data points *D* and θ_0, one would derive the expected values of probabilities of the *M* for each individual person in *D*. From these "complete" data, one would then re-estimate the maximally likely parameters, getting θ_1, etc. Doing this in exact detail is too much for a midterm, but would be a nice homework exercise.