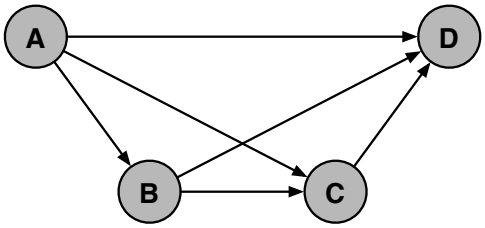
CSE 150. Assignment 4

Out: *Tue Feb 12*
Due: *Tue Feb 19*

4.1 Maximum likelihood estimation



(a) **Complete data**

Consider a complete data set of *i.i.d.* examples $\{a_t, b_t, c_t, d_t\}_{t=1}^{T}$ drawn from the joint distribution of the above belief network. Compute the maximum likelihood estimates of the conditional probability tables (CPTs) shown below for this data set. Express your answers in terms of indicator functions, such as:

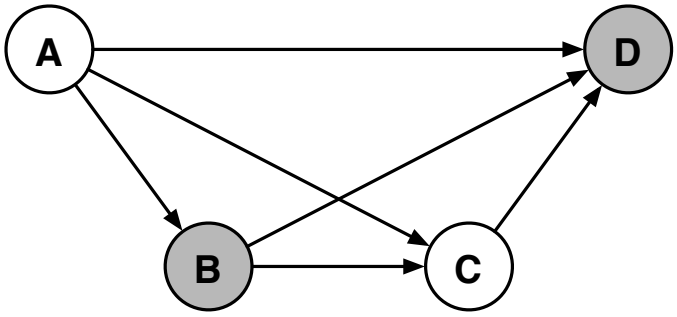
$$I(a, a_t) = \begin{cases} 1 \text{ if } a = a_t, \\ 0 \text{ if } a \neq a_t. \end{cases}$$

For example, in terms of this indicator function, the maximum likelihood estimate for the CPT at node A is given by $P(A=a) = \frac{1}{T}\sum_{t=1}^{T} I(a, a_t)$. Complete the numerators and denominators in the below expressions.

$$P(B=b|A=a) \quad = \quad \text{———————————}$$

$$P(C=c|A=a, B=b) \quad = \quad \text{———————————}$$

$$P(D=d|A=a, B=b, C=c) \quad = \quad \text{———————————}$$



(b) **Posterior probability**

Consider the belief network shown above, with observed nodes B and D and hidden nodes A and C. Compute the posterior probability $P(a, c|b, d)$ in terms of the CPTs of the belief network—that is, in terms of $P(a)$, $P(b|a)$, $P(c|a, b)$ and $P(d|a, b, c)$.

(c) **Posterior probability**

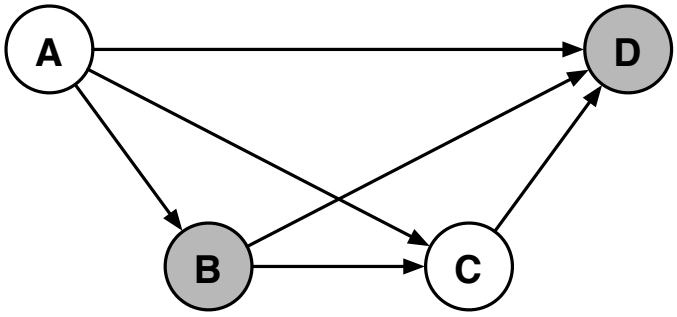
Compute the posterior probabilities $P(a|b, d)$ and $P(c|b, d)$ in terms of your answer from part (b). In other words, in this problem, you may assume that $P(a, c|b, d)$ is given.

(d) **Log-likelihood**

Consider a partially complete data set of *i.i.d.* examples $\{b_t, d_t\}_{t=1}^{T}$ drawn from the joint distribution of the above belief network. The log-likelihood of the data set is given by:

$$\mathcal{L} = \sum_t \log P(B = b_t, D = d_t).$$

Compute this log-likelihood in terms of the CPTs of the belief network. You may re-use work from earlier parts of the problem.



(e) **EM algorithm**

The posterior probabilities from parts (b) and (c) can be used by an EM algorithm to estimate CPTs that maximize the log-likelihood from part (d). Complete the numerator and denominator in the below expressions for the EM update rules. Simplify your answers as much as possible, expressing them in terms of the posterior probabilities $P(a, c|b_t, d_t)$, $P(a|b_t, d_t)$, and $P(c|b_t, d_t)$, as well as the indicator functions $I(b, b_t)$, and $I(d, d_t)$.

$$P(A=a) \leftarrow \frac{\qquad\qquad}{\qquad\qquad}$$

$$P(B=b|A=a) \leftarrow \frac{\qquad\qquad}{\qquad\qquad}$$

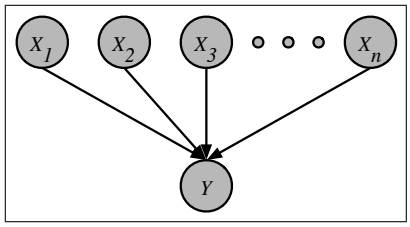
$$P(C=c|A=a, B=b) \leftarrow \frac{\qquad\qquad}{\qquad\qquad}$$

$$P(D=d|A=a, B=b, C=c) \leftarrow \frac{\qquad\qquad}{\qquad\qquad}$$

4.2 EM algorithm for noisy-OR

Suppose that n diseases are known to be partially responsible for an observed symptom of illness. From empirical data (e.g., patient profiles), how can we estimate a model of the relationship between disease and symptom?

Consider the belief network on the right over the diseases $X_i \in \{0, 1\}$ and symptom $Y \in \{0, 1\}$. The noisy-OR CPT is given by:



$$P(Y = 1|X) \;=\; 1 - \prod_{i=1}^{n} (1 - p_i)^{X_i},$$

which is expressed in terms of the noisy-OR parameters $p_i \in [0, 1]$.

In this problem, you will use the EM algorithm derived in class for estimating the noisy-OR parameters p_i. For a data set $\{(\vec{x}_t, y_t)\}_{t=1}^{T}$, the (conditional) log-likelihood is given by:

$$\mathcal{L} \;=\; \sum_{t=1}^{T} \log P(Y = y_t | X = \vec{x}_t).$$

Download the data files on the course web site, and use the EM algorithm to estimate the parameters p_i. The data set has $T = 12000$ patient profiles over $n = 18$ diseases. The EM update is given by:

$$p_i \;\leftarrow\; \frac{1}{T_i} \sum_{t} \frac{y_t x_{it} p_i}{\left[1 - \prod_{j=1}^{n} (1 - p_j)^{x_{jt}}\right]},$$

where T_i is the number of examples in which $X_i = 1$. Initialize all $p_i = 0.2$ and perform 64 iterations of the EM algorithm. At each iteration, compute the conditional log-likelihood shown above. If you have implemented the EM algorithm correctly, this conditional log-likelihood will always increase from one iteration to the next. **Turn in your source code and a completed version of this table:**

#	0	1	2	4	8	16	32	64
$\mathcal{L}$	-7088.1	-6746.8	?	?	?	?	?	-6335.1

Use the already completed entries of this table to check your work. **Also print out the final values that you estimate for the parameters p_i.** You may program in the language of your choice.