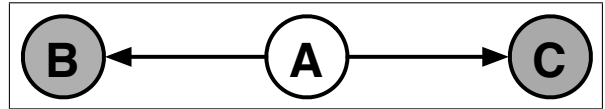

CSE 150. Assignment 4

Out: *Wed May 7*

Due: *Wed May 14*

4.1 EM algorithm for co-occurrence data



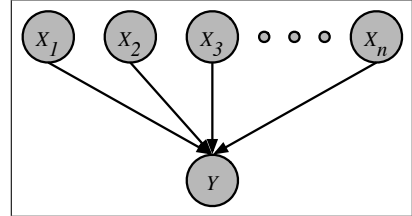
The belief network above is often used to model different types of data that co-occur—for example, images and text on the same web page, or verbs and objects in the same sentence. The co-occurring data types are modeled by the observed nodes B and C . The hidden node A is used to represent or discover different clusters of co-occurrences—for example, groups of verbs that take similar objects.

- Compute the posterior probability $P(A|B, C)$ in terms of the CPTs of this belief network—namely, $P(A)$, $P(B|A)$, and $P(C|A)$.
 - Consider an incomplete data set of T observations $\{b_t, c_t\}_{t=1}^T$ over the nodes B and C of the network, assumed to be sampled i.i.d from the belief network's joint distribution. Compute the log-likelihood of this data set, namely $\mathcal{L} = \sum_t \log P(B=b_t, C=c_t)$, in terms of the CPTs of the belief network.
 - For the data set in part (b), write out the EM updates for re-estimating the CPT attached to each node in this belief network: $P(A)$, $P(B|A)$, and $P(C|A)$. Express the updates in terms of the posterior probabilities $P(a|b_t, c_t)$, used as shorthand notation for $P(A=a|B=b_t, C=c_t)$. Your answer may also use binary indicator functions, such as $I(b, b_t)$, to indicate whether the node B is instantiated to the value b in the t th example. Simplify the updates as much as possible.
-

4.2 EM algorithm for noisy-OR

Consider the belief network on the right, with binary random variables $X \in \{0, 1\}^n$ and $Y \in \{0, 1\}$ and a noisy-OR conditional probability table (CPT). 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 (normalized) conditional log-likelihood is given by:

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

Download the data files on the course web site, and use the EM algorithm to estimate the parameters p_i . The data set has $T = 10000$ examples over $n = 20$ inputs. The EM update is given by:

$$p_i \leftarrow \frac{1}{T_i} \sum_t \frac{y_t x_{it} p_i}{[1 - \prod_{i=1}^n (1 - p_i)^{x_{it}}]},$$

where T_i is the number of examples in which $X_i = 1$. Initialize all $p_i = 0.5$ 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}	-1.6076	-1.1311	?	?	?	?	?	-0.5287

Use the already completed entries of this table to check your work. You may program in the language of your choice (though MATLAB is highly recommended).
