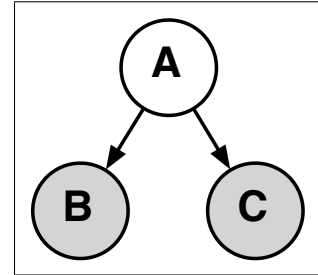

CSE 250a. Assignment 5

Out: *Wed Nov 7*

Due: *Mon Nov 19*

5.1 EM algorithm for co-occurrence data

The belief network on the right 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.



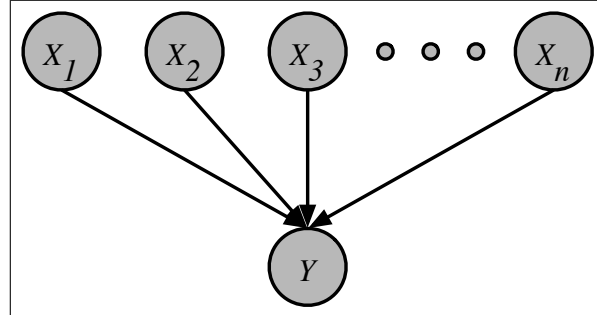
- Compute the posterior probability $P(A|B, C)$ in terms of the CPTs of this belief network—namely, $P(A)$, $P(B|A)$, and $P(C|A)$.
 - Consider an incomplete data set of T observations $\{b_t, c_t\}_{t=1}^T$ over the nodes B and C of the network, assumed to be sampled i.i.d from the belief network's 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.
-

5.2 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 $p_i \in [0, 1]$.



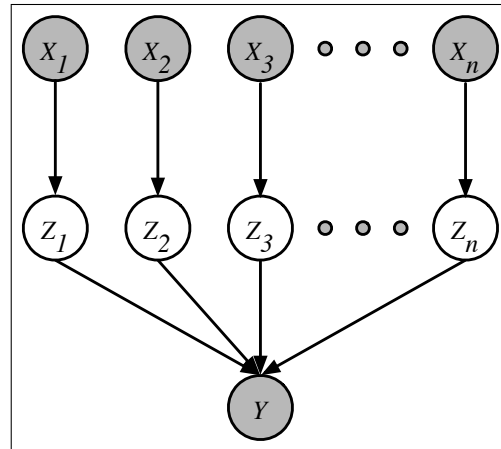
In this problem, you will derive and implement an EM algorithm for estimating the noisy-OR parameters p_i . It may seem that the EM algorithm is not suited to this problem, in which all the nodes are observed, and the CPT has a parameterized form. In fact, the EM algorithm can be applied, but first we must express the model in a different but equivalent form.

Consider the belief network shown to the right. In this network, a binary random variable $Z_i \in \{0, 1\}$ intercedes between each pair of nodes X_i and Y . Suppose that:

$$\begin{aligned} P(Z_i = 1|X_i = 0) &= 0, \\ P(Z_i = 1|X_i = 1) &= p_i. \end{aligned}$$

Also, let the node Y be *determined* by the logical-OR of Z_i . In other words:

$$P(Y = 1|Z) = \begin{cases} 1 & \text{if } Z_i = 1 \text{ for any } i, \\ 0 & \text{if } Z_i = 0 \text{ for all } i. \end{cases}$$



- (a) Show that this “extended” belief network defines the same conditional distribution $P(Y|X)$ as the original one. In particular, starting from

$$P(Y = 1|X) = \sum_{Z \in \{0,1\}^n} P(Y = 1, Z|X),$$

show that the right hand side of this equation reduces to the noisy-OR CPT with parameters p_i . To perform this marginalization, you will need to exploit various conditional independence relations.

- (b) Consider estimating the noisy-OR parameters p_i to maximize the conditional log-likelihood of joint observations of the variables X and Y . This (normalized) conditional log-likelihood is given by:

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

where $(\vec{x}^{(t)}, y^{(t)})$ is the t th joint observation of X and Y . From your result in part (a), it follows that we can estimate the parameters p_i in either the original network or the extended one (since in both networks they would be maximizing the same equation for the conditional log-likelihood).

Notice that in the extended network, we can view X and Y as observed nodes and Z as hidden nodes. Thus in this network, we can use the EM algorithm to estimate each parameter p_i , which simply defines one row of the “look-up” CPT for the node Z_i .

Compute the posterior probability that appears in the E-step of this EM algorithm. In particular, for joint observations $x \in \{0, 1\}^n$ and $y \in \{0, 1\}$, use Bayes rule to show that:

$$P(Z_i = 1, X_i = 1 | X = x, Y = y) = \frac{yx_i p_i}{1 - \prod_j (1 - p_j)^{x_j}}$$

- (c) For the data set $\{\vec{x}^{(t)}, y^{(t)}\}_{t=1}^T$, show that the EM update for the parameters p_i is given by:

$$p_i \leftarrow \frac{1}{T_i} \sum_t P(Z_i = 1, X_i = 1 | X = x^{(t)}, Y = y^{(t)}),$$

where T_i is the number of examples in which $X_i = 1$. (You should derive this update as a special case of the general form presented in lecture.)

- (d) Finally you will estimate the parameters p_i from a data set of $T = 10000$ examples (with $n = 10$). The data is contained in the ASCII files `X.dat` and `Y.dat`, available from the course web site.

Initialize all $p_i = 0.6$ and perform 64 iterations of the EM algorithm. At each iteration, compute the conditional log-likelihood shown in part (b). 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}	-3.3405	-0.7772	?	?	?	?	?	-0.5429

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).