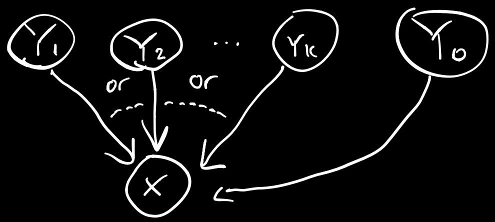
Exercise 19.4

Consider the problem of applying EM to parameter estimation for a variable X whose local probabilistic model is a noisy-or. Assume that X has parents $Y_1, \dots, Y_k$, so that our task for X is to estimate the noise parameters $\lambda_0, \dots, \lambda_k$. Explain how we can use the EM algorithm to accomplish this task. (Hint: Utilize the structural decomposition of the noisy-or node.)

Assume $Y_1, \dots, Y_k$ and X are binary variables.

A noisy-or model means that all of the parents of a single node are independent of eachother.

Our model will look like this: (our <u>structural decomposition</u>)



– leak probability (probability of something random happening)

However, we will need to compute $P(Y_1=0)\cdot P(Y_2=0)$ instead of $P(X|Y_1=0, Y_2=0)$

Thus, our parameters are $\lambda_1 = P(Y_1=1), \lambda_2 = P(Y_2=1), \dots, \lambda_k = P(Y_k=1)$ and $\lambda_0 = \theta$

where θ is an arbitrary probability

The likelihood on a network with 3 parents on X with input $(Y_1=0, Y_2=0, Y_3=1)$

is $(1-\lambda_1)\cdot(1-\lambda_2)(1-\lambda_3)\cdot\lambda_0$ (or $(1-\lambda_0)$)

We can also see we dont need X for any parameters, so there are only k+1 parameters to estimate.

Assume we are in the E-step of EM and we have computed all the likelihoods from a dataset $\mathcal{D}$. We can now compute parameters like so:

$$\lambda_1 = \frac{\text{all weights where } Y_1=1}{\text{all weights computed}}$$

$\vdots$

$$\lambda_k = \frac{\text{all weights where } Y_k=1}{\text{all weights computed}}$$

$$\lambda_0 = \frac{\text{all weights where } Y_0=1}{\text{all weights computed.}}$$

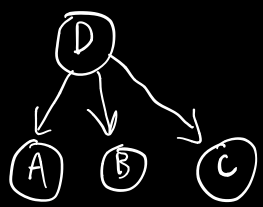
Thus, we can easily perform M-step and estimate our parameters given knowledge X is a noisy-or model.

$P(X|Y_0, Y_1, \dots, Y_k)$ can be computed by setting $P(X=0|Y_0=y_0, \dots, Y_k=y_k)$ to 1 – probability of failure, and $P(X_1=1|Y_0=y_0, \dots, Y_k=y_k) = 1 - P(X=0|Y_0=y_0, \dots, Y_k=y_k)$

- (5 points) Consider the dataset given below. All variables are Binary and take values from the domain $\{0,1\}$. "?" denotes a missing value.

A	B	C	D
0	1	1	?
1	0	0	?
0	1	1	?
1	1	0	?
0	0	1	?
1	1	1	?

Assume that you are learning the parameters of a Bayesian network having the following structure: D is the root node (having no parents). The parent of A is D, B is D and C is D. (Thus the only edges in the Bayesian network are $D \to A$, $D \to B$ and $D \to C$). Starting with probabilities that are initialized uniformly (i.e., all probabilities are initialized to 0.5), calculate the parameters of this Naive Bayes model using the EM algorithm. Stop at convergence or after 3 iterations, whichever is earlier. Does the EM algorithm converge and after how many iterations?



P(D) =

D_0	P_1
0.5	0.5

P(A|D) =

D	A_0	A_1
0	0.5	0.5
1	0.5	0.5

P(B|D) =

D	B_0	B_1
0	0.5	0.5
1	0.5	0.5

P(C|D) =

D	C_0	C_1
0	0.5	0.5
1	0.5	0.5

A	B	C	D	w
0	1	1	0	$0.5^4/0.5^3=0.5$
0	1	1	1	$0.5^4/0.5^3=0.5$
1	0	0	0	$0.5^4/0.5^3=0.5$
1	0	0	1	$0.5^4/0.5^3=0.5$
0	1	1	0	$0.5^4/0.5^3=0.5$
0	1	1	1	$0.5^4/0.5^3=0.5$
1	1	0	0	$0.5^4/0.5^3=0.5$
1	1	0	1	$0.5^4/0.5^3=0.5$
0	0	1	0	$0.5^4/0.5^3=0.5$
0	0	1	1	$0.5^4/0.5^3=0.5$
1	1	1	0	$0.5^4/0.5^3=0.5$
1	1	1	1	$0.5^4/0.5^3=0.5$

$(D=0) = \frac{6 \cdot 0.5}{12 \cdot 0.5} = 0.5$

$P(A=0|D=0) = \frac{3 \cdot 0.5}{6 \cdot 0.5} = 0.5$

$P(A=0|D=1) = \frac{3 \cdot 0.5}{6 \cdot 0.5} = 0.5$

$P(B=0|D=0) = \frac{2 \cdot 0.5}{6 \cdot 0.5} = 0.33$

$P(B=0|D=1) = \frac{2 \cdot 0.5}{6 \cdot 0.5} = 0.33$

$P(C=0|D=0) = \frac{2 \cdot 0.5}{6 \cdot 0.5} = 0.33$

$P(C=0|D=1) = \frac{2 \cdot 0.5}{6 \cdot 0.5} = 0.33$

$P(D) =$

D_0	D_1
0.5	0.5

$P(B|D) =$

D	B_0	B_1
0	0.33	0.66
1	0.33	0.66

$P(A|D) =$

D	A_0	A_1
0	0.5	0.5
1	0.5	0.5

$P(C|D) =$

D	C_0	C_1
0	0.33	0.66
1	0.33	0.66

A	B	C	D	w
0	1	1	0	$0.5^2 \cdot 0.66^2 / (2 \cdot 0.5^2 \cdot 0.66^2) = 0.5$
0	1	1	1	$0.5^2 \cdot 0.66^2 / (2 \cdot 0.5^2 \cdot 0.66^2) = 0.5$
1	0	0	0	$0.5^2 \cdot 0.33^2 / (2 \cdot 0.5^2 \cdot 0.33^2) = 0.5$
1	0	0	1	$0.5^2 \cdot 0.33^2 / (2 \cdot 0.5^2 \cdot 0.33^2) = 0.5$
0	1	1	0	$0.5^2 \cdot 0.66^2 / (2 \cdot 0.5^2 \cdot 0.66^2) = 0.5$
0	1	1	1	$0.5^2 \cdot 0.66^2 / (2 \cdot 0.5^2 \cdot 0.66^2) = 0.5$
1	1	0	0	$0.5^2 \cdot 0.33 \cdot 0.66 / (2 \cdot 0.5^2 \cdot 0.33 \cdot 0.66) = 0.5$
1	1	0	1	$0.5^2 \cdot 0.33 \cdot 0.66 / (2 \cdot 0.5^2 \cdot 0.33 \cdot 0.66) = 0.5$
0	0	1	0	$0.5^2 \cdot 0.33 \cdot 0.66 / (2 \cdot 0.5^2 \cdot 0.33 \cdot 0.66) = 0.5$
0	0	1	1	$0.5^2 \cdot 0.33 \cdot 0.66 / (2 \cdot 0.5^2 \cdot 0.33 \cdot 0.66) = 0.5$
1	1	1	0	$0.5^2 \cdot 0.66^2 / (2 \cdot 0.5^2 \cdot 0.33 \cdot 0.66) = 0.5$
1	1	1	1	$0.5^2 \cdot 0.66^2 / (2 \cdot 0.5^2 \cdot 0.33 \cdot 0.66) = 0.5$

$$(D=0) = \frac{6 \cdot 0.5}{12 \cdot 0.5} = 0.5$$

$$P(A=0|D=0) = \frac{3 \cdot 0.5}{6 \cdot 0.5} = 0.5$$

$$P(A=0|D=1) = \frac{3 \cdot 0.5}{6 \cdot 0.5} = 0.5$$

$$P(B=0|D=0) = \frac{2 \cdot 0.5}{6 \cdot 0.5} = 0.33$$

$$P(B=0|D=1) = \frac{2 \cdot 0.5}{6 \cdot 0.5} = 0.33$$

$$P(C=0|D=0) = \frac{2 \cdot 0.5}{6 \cdot 0.5} = 0.33$$

$$P(C=0|D=1) = \frac{2 \cdot 0.5}{6 \cdot 0.5} = 0.33$$

Thus. EM converges in <u>2 steps</u>

- (5 points) Let us generalize our experience with such datasets, the above Bayesian network and EM with uniform initialization. Assume that you are given a dataset such that D is (always) missing but A, B and C are observed in all examples in the dataset. Assume that you will learn the parameters of the Bayesian network given above using the EM algorithm with uniform initialization. Answer the following questions based on these assumptions.
  - At convergence, what will be the parameters of the Naive Bayes model?
  - After how many iterations will EM converge?

- Under these assumptions, for every variable where n is the size of the dataset

$$P(A=0 \mid D) = \frac{\#\text{ of times } A=0}{n} \qquad P(C=0 \mid D) = \frac{\#\text{ of times } C=0}{n}$$

$$P(B=0 \mid D) = \frac{\#\text{ of times } B=0}{n} \qquad P(D) = \frac{1}{\text{domain of } D} = 0.5 \text{ if } D \text{ is binary}$$

- EM will always converge in this situation in <u>2 iterations</u>