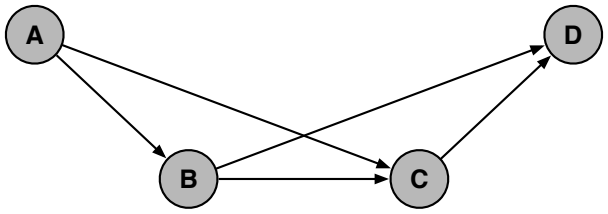
CSE 150. Assignment 5 *Spring 2017*

Out: *Tue May 16*
Due: *Tue May 23*

5.1 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 equality-testing functions, such as:

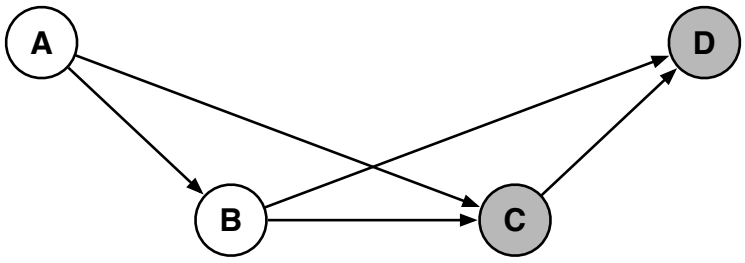
$$I(a, a_t) = \begin{cases} 1 \text{ if } a = a_t, \\ 0 \text{ if } a \neq a_t. \end{cases}$$

For example, in terms of this 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.

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

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

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



(b) **Posterior probability**

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

(c) **Posterior probability**

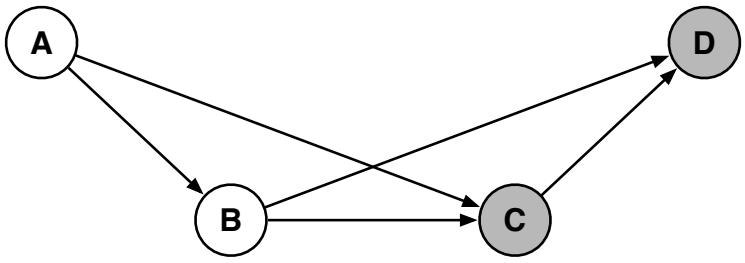
Compute the posterior probabilities $P(a|c, d)$ and $P(b|c, d)$ in terms of your answer from part (b). In other words, in this problem, you may assume that $P(a, b|c, d)$ is given.

(d) **Log-likelihood**

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

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

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 part (b) can be used by an EM algorithm to estimate CPTs that maximize the log-likelihood from part (c). 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 $P(a, b|c_t, d_t)$, $P(a|c_t, d_t)$, and $P(b|c_t, d_t)$, as well as the functions $I(c, c_t)$, and $I(d, d_t)$.

$P(A=a) \leftarrow$ _____

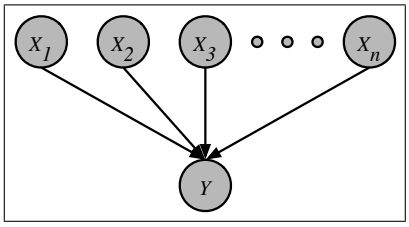
$P(B=b|A=a) \leftarrow$ _____

$P(C=c|A=a, B=b) \leftarrow$ _____

$P(D=d|B=b, C=c) \leftarrow$ _____

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 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 log (conditional) 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 *hw5_noisyOr_x.txt* and *hw5_noisyOr_y.txt* for this homework assignment, and use the EM algorithm to estimate the parameters p_i. The data set has $T=267$ examples over $n=23$ inputs. For those interested, more information about this data set is available here:

http://archive.ics.uci.edu/ml/datasets/SPECT+Heart

However, be sure to use the data files provided on Piazza, as they have been specially assembled for this assignment. The EM update for this model is given by:

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

where T_i is the number of examples in which $X_i = 1$. Initialize all $p_i = \frac{1}{n}$ and perform 512 iterations of the EM algorithm. At each iteration, compute the log conditional likelihood shown above. (If you have implemented the EM algorithm correctly, this log conditional likelihood will always increase from one iteration to the next.) Also compute the number of mistakes $M \leq T$ made by the model at each iteration; a mistake occurs either when $y_t = 0$ and $P(y_t = 1|\vec{x}_t) \geq 0.5$ (indicating a false positive) or when $y_t = 1$ and $P(y_t = 1|\vec{x}_t) \leq 0.5$ (indicating a false negative). The number of mistakes should generally decrease as the model is trained, though it is not guaranteed to do so at each iteration.

Turn in all of the following:

(a) a hard-copy print-out of your source code

(b) a plot or print-out of your final estimated values for p_i

(c) a completed version of the following table

iteration	number of mistakes M	log conditional likelihood $\mathcal{L}$
0	195	-1.045
1	60	
2		-0.4108
4		
8		
16		
32		
64		
128	36	
256		
512		-0.3100

You should use the already completed entries of this table to check your work. As always you may program in the language of your choice.

5.3 EM algorithm for binary matrix completion

In this problem you will use the EM algorithm to build a simple movie recommendation system. Download the files *hw5_movieTitles.txt*, *hw5_studentPIDs.txt*, and *hw5_movieRatings.txt*. The last of these files contains a 258×50 matrix of zeros, ones, and missing elements denoted by question marks. The $\langle i, j \rangle^{\text{th}}$ element in this matrix contains the i^{th} student's rating of the j^{th} movie, according to the following key:

1	recommended,
0	not recommend,
?	not seen.

(a) **Sanity check**

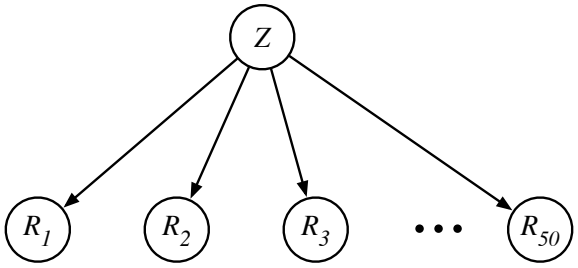
Compute the mean popularity rating of each movie, given by the simple ratio

$$\frac{\text{number of students who recommended the movie}}{\text{number of students who saw the movie}},$$

and sort the movies by this ratio. Print out the movie titles from least popular (*The Last Airbender*) to most popular (*Inception*). Note how well these rankings do or do not corresponding to your individual preferences.

(b) **Likelihood**

Now you will learn a naive Bayes model of these movie ratings, represented by the belief network shown below, with hidden variable $Z \in \{1, 2, \ldots, k\}$ and partially observed binary variables $R_1, R_2, \ldots, R_{50}$ (corresponding to movie ratings).



This model assumes that there are k different types of movie-goers, and that the i^{th} type of movie-goer—who represents a fraction $P(Z=i)$ of the overall population—likes the j^{th} movie with conditional probability $P(R_j=1|Z=i)$. Let Ω_t denote the set of movies seen (and hence rated) by the t^{th} student. Show that the likelihood of the t^{th} student's ratings is given by

$$P\left(\left\{R_j = r_j^{(t)}\right\}_{j\in\Omega_t}\right) \;=\; \sum_{i=1}^{k} P(Z=i) \prod_{j\in\Omega_t} P\left(R_j = r_j^{(t)} \,\middle|\, Z=i\right).$$

(c) **E-step**

The E-step of this model is to compute, for each student, the posterior probability that he or she corresponds to a particular type of movie-goer. Show that

$$P\left(Z=i \,\middle|\, \left\{R_j=r_j^{(t)}\right\}_{j\in\Omega_t}\right) \;=\; \frac{P(Z=i)\,\prod_{j\in\Omega_t} P\left(R_j=r_j^{(t)}\,\middle|\, Z=i\right)}{\sum_{i'=1}^{k} P(Z=i')\,\prod_{j\in\Omega_t} P\left(R_j=r_j^{(t)}\,\middle|\, Z=i'\right)}.$$

(d) **M-step**

The M-step of the model is to re-estimate the probabilities $P(Z=i)$ and $P(R_j=1|Z=i)$ that define the CPTs of the belief network. As shorthand, let

$$\rho_{it} \;=\; P\left(Z=i \,\middle|\, \left\{R_j=r_j^{(t)}\right\}_{j\in\Omega_t}\right)$$

denote the probabilities computed in the E-step of the algorithm. Also, let $T = 258$ denote the number of students. Show that the EM updates are given by

$$\begin{aligned} P(Z=i) \;&\leftarrow\; \frac{1}{T}\sum_{t=1}^{T}\rho_{it}, \\ P(R_j=1|Z=i) \;&\leftarrow\; \frac{\sum_{\{t|j\in\Omega_t\}}\rho_{it}\,I\left(r_j^{(t)},1\right)+\sum_{\{t|j\notin\Omega_t\}}\rho_{it}\,P(R_j=1|Z=i)}{\sum_{t=1}^{T}\rho_{it}}. \end{aligned}$$

(e) **Implementation**

Download the files *hw5_probZ_init.txt* and *hw5_probRgivenZ_init.txt*, and use them to initialize the probabilities $P(Z=i)$ and $P(R_j=1|Z=i)$ for a model with $k=4$ types[1] of movie-goers. Run 128 iterations of the EM algorithm, computing the (normalized) log-likelihood

$$\mathcal{L} \;=\; \frac{1}{T}\sum_{t=1}^{T}\log P\left(\left\{R_j=r_j^{(t)}\right\}_{j\in\Omega_t}\right)$$

at each iteration. Does your log-likelihood increase (i.e., become less negative) at each iteration? Fill in a completed version of the following table, using the already provided entries to check your work:

iteration	log-likelihood $\mathcal{L}$
0	-23.4375
1	-13.8082
2	
4	
8	
16	-10.8541
32	
64	
128	

[1]There is nothing special about these initial values or the choice of $k=4$, and you should feel free to experiment with other choices.

(f) **Personal movie recommendations**

Find your student PID in *hw5_studentPIDs.txt* to determine the row of the ratings matrix that stores your personal data. Compute the posterior probability in part (c) for this row from your trained model, and then compute your *expected* ratings on the movies *you haven't yet seen*:

$$P\left(R_\ell = 1 \,\middle|\, \left\{R_j = r_j^{(t)}\right\}_{j \in \Omega_t}\right) \;=\; \sum_{i=1}^{k} P\left(Z = i \,\middle|\, \left\{R_j = r_j^{(t)}\right\}_{j \in \Omega_t}\right) P(R_\ell = 1 | Z = i) \quad \text{for } \ell \notin \Omega_t.$$

Print out the list of these (unseen) movie sorted by their expected ratings. Does this list seem to reflect your personal tastes better than the list in part (a)? Hopefully it does (although our data set is obviously *far* smaller and more incomplete than the data sets at companies like Netflix or Amazon).

Note: if you didn't previously complete the survey, then you will need to hard-code your ratings in order to answer this question.

(g) **Source code**

Turn in a hard-copy printout of your source code for all parts of this problem. As usual, you may program in the language of your choice.
