

8.1 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 *hw8_movies.txt*, *hw8_ids.txt*, and *hw8_ratings.txt*. The last of these files contains a 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

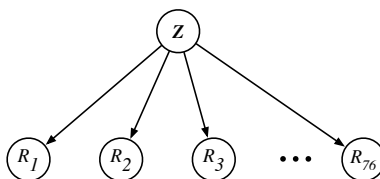
Compute the mean popularity rating of each movie, given by the simple ratio

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

and sort the movies by this ratio. Print out the movie titles from least popular (*I Feel Pretty*) to most popular (*Inception*). Note how well these rankings do or do not correspond 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 $Z \in \{1, 2, \dots, k\}$ and partially observed binary variables R_1, R_2, \dots, R_{76} (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

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

(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

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

(d) **M-step**

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

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

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

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

(e) **Implementation**

Download the files *hw8_probZ_init.txt* and *hw8_probR_init.txt*, and use them to initialize the probabilities $P(Z=i)$ and $P(R_j=1 \mid Z=i)$ for a model with $k=4$ types¹ of movie-goers. Run 256 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:

¹There is nothing special about these initial values or the choice of $k=4$; feel free to experiment with other choices.

iteration	log-likelihood \mathcal{L}
0	-27.0358
1	-17.5604
2	
4	
8	
16	-14.2638
32	
64	
128	
256	

(f) **Personal movie recommendations**

Find your student PID in *hw8_ids.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*:

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

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 complete the survey in time, 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.