

Expectation Maximization – An Example

Amro Al-Baali

December 5, 2020

1 Notations

Let the observations be $X = \{x_1, \dots, x_n\}$ be the set of realizations of the random variable $\underline{x} \in \mathbb{R}$. Additionally, let $Z = \{z_1, \dots, z_n\}$ be the set of realizations of $\underline{z} \in \{0, 1\}$ which is the “missing” data of the problem.¹ The set of complete data is denoted by $Y = \{X, Z\}$. Furthermore, let the two PDF that \underline{x} is distributed from be denoted by $f_1(x; \lambda_1)$ and $f_2(x; \lambda_2)$, where λ_i is the set of parameters for each of the two PDFs.

The set of unknown parameters are $\theta = \{\lambda_1, \lambda_2, \pi_1\}$.

2 Problem statement

Let $\underline{x} \in \mathbb{R}$ be a random variable² that can be sampled from one of two distributions $f_1(x; \lambda_1)$ and $f_2(x; \lambda_2)$.³ Whether \underline{x} will be sampled from $f_1(x; \lambda_1)$ or $f_2(x; \lambda_2)$ will depend on another random variable $\underline{z} \in \{0, 1\}$. Specifically, the conditional PDF of \underline{x} given \underline{z} is

$$f(x | z; \lambda) = \begin{cases} f_1(x; \lambda_1), & z = 0, \\ f_2(x; \lambda_2), & z = 1, \end{cases} \quad (1)$$

where $\lambda = \{\lambda_1, \lambda_2\}$. Let the PMF of \underline{z} be given by

$$p(z) = \begin{cases} \pi_1, & z = 0, \\ 1 - \pi_1, & z = 1, \\ 0, & \text{otherwise.} \end{cases} \quad (2)$$

The problem is to estimate the parameters θ without having known π_1 . The set of unknown parameters will be denoted by $\theta = \{\lambda, \pi_1\}$.

3 Doing “the math”

The PDF $f(y; \theta)$ is needed in the expectation step. The PDF is given by

$$f(y; \theta) = f(x, z; \theta) \quad (3)$$

$$= f(x | z; \theta) p(z; \theta). \quad (4)$$

¹ The data is “missing” in the sense that, if this data was available, then the problem would be significantly simplified.

² In this document a single random variable will be used. It is possible to generalize to multivariate random variables.

³ In this document, we’ll assume that there are only two possible distributions but the idea generalizes to multiple distributions.

Plugging (1) and (2) into (4) gives

$$f(x, z; \boldsymbol{\theta}) = f(x | z; \boldsymbol{\theta}) p(z; \boldsymbol{\theta}) \quad (5)$$

$$= \begin{cases} f_1(x; \boldsymbol{\lambda}_1) \pi_1, & z = 0, \\ f_2(x; \boldsymbol{\lambda}_2) (1 - \pi_1), & z = 1, \end{cases} \quad (6)$$

which can be rewritten⁴ as

$$f(x, z; \boldsymbol{\theta}) = (\pi_1 f_1(x; \boldsymbol{\lambda}_1))^{1-z} ((1 - \pi_1) f_2(x; \boldsymbol{\lambda}_2))^z. \quad (7)$$

The marginal PDF on \underline{x} is obtained by marginalizing out \underline{z} from $f(x, z; \boldsymbol{\theta})$ to give

$$f(x; \boldsymbol{\theta}) = \sum_{i=0}^1 f(x, z = i; \boldsymbol{\theta}) \quad (8)$$

$$= \pi_1 f_1(x; \boldsymbol{\theta}) + (1 - \pi_1) f_2(x; \boldsymbol{\theta}) \quad (9)$$

$$= \pi_1 f_1(x; \boldsymbol{\lambda}_1) + (1 - \pi_1) f_2(x; \boldsymbol{\lambda}_2). \quad (10)$$

4 The expectation step

From ⁵ and ⁶, the function to be maximized is $Q(\boldsymbol{\theta} | \boldsymbol{\theta}^{(j)})$,⁷ which is the expectation of the log-likelihood of the complete data. That is,

$$Q(\boldsymbol{\theta} | \boldsymbol{\theta}^{(j)}) = \mathbb{E}_{\underline{Y}} [\log f(\underline{Y}; \boldsymbol{\theta}) | X, \boldsymbol{\theta}^{(j)}]. \quad (11)$$

The log-likelihood function of the complete data is given by

$$\begin{aligned} \log f(\underline{Y}; \boldsymbol{\theta}) &= \sum_{i=1}^n \log \left((\pi_1 f_1(\underline{x}_i; \boldsymbol{\lambda}_1))^{1-\underline{z}_i} \right. \\ &\quad \left. \cdot ((1 - \pi_1) f_2(\underline{x}_i; \boldsymbol{\lambda}_2))^{\underline{z}_i} \right) \end{aligned} \quad (12)$$

$$\begin{aligned} &= \sum_{i=1}^n (1 - \underline{z}_i) (\log \pi_1 + \log f_1(\underline{x}_i; \boldsymbol{\lambda}_1)) \\ &\quad + \sum_{i=1}^n \underline{z}_i (\log (1 - \pi_1) + \log f_2(\underline{x}_i; \boldsymbol{\lambda}_2)) \end{aligned} \quad (13)$$

$$\begin{aligned} &= \sum_{i=1}^n (1 - \underline{z}_i) \log f_1(\underline{x}_i; \boldsymbol{\lambda}_1) + \underline{z}_i \log f_2(\underline{x}_i; \boldsymbol{\lambda}_2) \\ &\quad + \sum_{i=1}^n \underline{z}_i \log (1 - \pi_1) + (1 - \underline{z}_i) \log \pi_1. \end{aligned} \quad (14)$$

⁴ This may be confusing at first, by simply replace z with 0 or 1 and the expression (6) will be exactly recovered.

⁵ Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York, New York, NY, 2009. DOI: 10.1007/978-0-387-84858-7

⁶ Larry Wasserman. *All of Statistics: A Concise Course in Statistical Inference*. Springer Texts in Statistics. Springer New York, New York, NY, 2004. DOI: 10.1007/978-0-387-21736-9

⁷ $\boldsymbol{\theta}^{(j)}$ is the j th estimate of $\boldsymbol{\theta}$.

The conditional expectation (11) can then be expanded to give

$$Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(j)}) = \mathbb{E}_{\underline{Y}} \left[\log f(\underline{Y}; \boldsymbol{\theta}) \mid X, \boldsymbol{\theta}^{(j)} \right] \quad (15)$$

$$= \mathbb{E}_{\underline{Z}} \left[\log f(\underline{X}, \underline{Z}; \boldsymbol{\theta}) \mid \underline{X} = X, \boldsymbol{\theta}^{(j)} \right] \quad (16)$$

$$\begin{aligned} &= \sum_{i=1}^n \left(1 - \mathbb{E} \left[z_i \mid X, \boldsymbol{\theta}^{(j)} \right] \right) \log f_1(x_i; \boldsymbol{\lambda}_1) + \mathbb{E} \left[z_i \mid X, \boldsymbol{\theta}^{(j)} \right] \log f_2(x_i; \boldsymbol{\lambda}_2) \\ &\quad + \sum_{i=1}^n \mathbb{E} \left[z_i \mid X, \boldsymbol{\theta}^{(j)} \right] \log(1 - \pi_1) + \left(1 - \mathbb{E} \left[z_i \mid X, \boldsymbol{\theta}^{(j)} \right] \right) \log \pi_1. \end{aligned} \quad (17)$$

The expectation $\mathbb{E} \left[z_i \mid X, \boldsymbol{\theta}^{(j)} \right]$ is simplified to $\mathbb{E}_{z_i} \left[z_i \mid x_i, \boldsymbol{\theta}^{(j)} \right]$ since

$$p(z_i \mid X) = \frac{f(X \mid z_i) p(z_i)}{f(X)} \quad (18)$$

$$= \frac{f(x_1) \cdots f(x_i \mid z_i) \cdots f(x_n) p(z_i)}{f(x_1) \cdots f(x_i) \cdots f(x_n)} \quad (19)$$

$$= \frac{f(x_i \mid z_i) p(z_i)}{f(x_i)} \quad (20)$$

$$= p(z_i \mid x_i), \quad (21)$$

where the independence of \underline{x}_j from z_i for $i \neq j$ was used. The expectation is therefore

$$\mathbb{E} \left[z_i \mid x_i, \boldsymbol{\theta}^{(j)} \right] = \sum_{k=0}^1 \frac{f(x_i, z; \boldsymbol{\theta}^{(j)})}{f(x_i; \boldsymbol{\theta}^{(j)})} z_k \quad (22)$$

$$= \frac{1}{f(x_i; \boldsymbol{\theta}^{(j)})} \sum_{k=0}^1 f(x_i, z; \boldsymbol{\theta}^{(j)}) z_k \quad (23)$$

$$\begin{aligned} &= \frac{1}{f(x_i; \boldsymbol{\theta}^{(j)})} \left(f_1(x_i; \boldsymbol{\lambda}_1^{(j)}) \pi_1^{(j)}(0) \right. \\ &\quad \left. + f_2(x_i; \boldsymbol{\lambda}_2^{(j)}) (1 - \pi_1^{(j)})(1) \right) \end{aligned} \quad (24)$$

$$= \frac{f_2(x_i; \boldsymbol{\lambda}_2^{(j)}) (1 - \pi_1^{(j)})}{\pi_1^{(j)} f_1(x_i; \boldsymbol{\lambda}_1^{(j)}) + (1 - \pi_1^{(j)}) f_2(x_i; \boldsymbol{\lambda}_2^{(j)})}. \quad (25)$$

Note that the expectation step does not require us to exploit the two PDFs $f_i(x; \boldsymbol{\lambda}_i)$; just plug in the data and get an estimate of z_i for $i = 1, \dots, n$.

From here onwards, the notation

$$\hat{z}_i^{(j)} := \mathbb{E} \left[z_i \mid x_i, \boldsymbol{\theta}^{(j)} \right] \quad (26)$$

will be used.

5 Maximization step

Now that the missing data z_i is estimated in the expectation step, the next step is to estimate a new set of parameters $\boldsymbol{\theta}^{(j+1)}$ using maximum likelihood (ML) estimator on the log-likelihood function of the complete *estimated* data set $\hat{Y} = \{X, \hat{Z}^{(j)}\}$.⁸

⁸ Note that $\hat{Z}^{(j)} = \{\hat{z}_1^{(j)}, \dots, \hat{z}_n^{(j)}\}$.

Let the PDFs $f_i(x; \lambda_i)$ be exponentially distributed with parameter λ_i . Then, the PDFs can be written as

$$f_i(x; \lambda_i) = \begin{cases} \lambda_i e^{-\lambda_i x}, & x \geq 0, \\ 0, & x < 0, \end{cases} \quad (27)$$

for $i = 1, 2$. The parameters in this case are $\boldsymbol{\theta} = \{\lambda_1, \lambda_2, \pi_1\}$.

For the ease of reading, the notation (26) will be used to rewrite $Q(\boldsymbol{\theta} | \boldsymbol{\theta}^{(j)})$ from (17). Then,

$$\begin{aligned} Q(\boldsymbol{\theta} | \boldsymbol{\theta}^{(j)}) &= \sum_{i=1}^n \left((1 - \hat{z}_i^{(j)}) \log f_1(x_i; \lambda_1) + \hat{z}_i^{(j)} \log f_2(x_i; \lambda_2) \right. \\ &\quad \left. + \sum_{i=1}^n \hat{z}_i^{(j)} \log(1 - \pi_1) + (1 - \hat{z}_i^{(j)}) \log \pi_1 \right) \end{aligned} \quad (28)$$

$$\begin{aligned} &= \sum_{i=1}^n \left((1 - \hat{z}_i^{(j)}) \log(\lambda_1 e^{-\lambda_1 x_i}) + \hat{z}_i^{(j)} \log(\lambda_2 e^{-\lambda_2 x_i}) \right. \\ &\quad \left. + \sum_{i=1}^n \hat{z}_i^{(j)} \log(1 - \pi_1) + (1 - \hat{z}_i^{(j)}) \log \pi_1 \right) \end{aligned} \quad (29)$$

$$\begin{aligned} &= \sum_{i=1}^n \left((1 - \hat{z}_i^{(j)}) (\log \lambda_1 - \lambda_1 x_i) + \hat{z}_i^{(j)} (\log \lambda_2 - \lambda_2 x_i) \right. \\ &\quad \left. + \sum_{i=1}^n \hat{z}_i^{(j)} \log(1 - \pi_1) + (1 - \hat{z}_i^{(j)}) \log \pi_1. \right) \end{aligned} \quad (30)$$

The function $Q(\boldsymbol{\theta} | \boldsymbol{\theta}^{(j)})$ is to be differentiated with respect to the parameters $\boldsymbol{\theta}$ and be equated to zero in order to solve for the critical points.

$$\frac{\partial Q(\boldsymbol{\theta} | \boldsymbol{\theta}^{(j)})}{\partial \lambda_1} = \sum_{i=1}^n \left((1 - \hat{z}_i^{(j)}) \left(\frac{1}{\lambda_1} - x_i \right) \right) \quad (31)$$

$$= \frac{1}{\lambda_1} \sum_{i=1}^n (1 - \hat{z}_i^{(j)}) - \sum_{i=1}^n (1 - \hat{z}_i^{(j)}) x_i. \quad (32)$$

Equating to partial derivative to zero and solving for λ_1 gives the next estimate

$$\hat{\lambda}_1^{(j+1)} = \frac{\sum_{i=1}^n (1 - \hat{z}_i^{(j)})}{\sum_{i=1}^n (1 - \hat{z}_i^{(j)}) x_i}. \quad (33)$$

The similar procedure can be done for λ_2 which gives the expression

$$\hat{\lambda}_2^{(j+1)} = \frac{\sum_{i=1}^n \hat{z}_i^{(j)}}{\sum_{i=1}^n \hat{z}_i^{(j)} x_i}. \quad (34)$$

Now, differentiate $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(j)})$ with respect to π_1 .

$$\frac{\partial Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(j)})}{\partial \pi_1} = \sum_{i=1}^n -\hat{z}_i^{(j)} \frac{1}{1 - \pi_1} + (1 - \hat{z}_i^{(j)}) \frac{1}{\pi_1} \quad (35)$$

$$= \frac{1}{\pi_1 (1 - \pi_1)} \sum_{i=1}^n (1 - \pi_1) (1 - \hat{z}_i^{(j)}) - \hat{z}_i^{(j)} \pi_1 \quad (36)$$

$$= \frac{1}{\pi_1 (1 - \pi_1)} \sum_{i=1}^n (1 - \hat{z}_i^{(j)} - \pi_1). \quad (37)$$

Equating to 0 and solving for π_1 gives the next estimate

$$\hat{\pi}_1^{(j+1)} = 1 - \frac{1}{n} \sum_{i=1}^n \hat{z}_i^{(j)}. \quad (38)$$

References

- Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York, New York, NY, 2009. DOI: 10.1007/978-0-387-84858-7.
- Larry Wasserman. *All of Statistics: A Concise Course in Statistical Inference*. Springer Texts in Statistics. Springer New York, New York, NY, 2004. DOI: 10.1007/978-0-387-21736-9.