Expectation Maximization – An Example

Amro Al-Baali

December 6, 2020

1 Notations

Let the observations be $X = \{x_1, \ldots, x_n\}$ be the set of realizations of the random variable $\underline{x} \in \mathbb{R}$. Additionally, let $Z = \{z_1, \ldots, z_n\}$ 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 $f_1(x; \boldsymbol{\lambda}_1)$ and $f_2(x; \boldsymbol{\lambda}_2)$, where $\boldsymbol{\lambda}_i$ is the set of parameters for each of the two PDFs.

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

The set of unknown parameters are $\boldsymbol{\theta} = \{\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2, \pi_1\}$.

2 Problem statement

Let $\underline{x} \in \mathbb{R}$ be a random variable[2] that can be sampled from one of two distributions $f_1(x; \boldsymbol{\lambda}_1)$ and $f_2(x; \boldsymbol{\lambda}_2)$.[3] Whether $\underline{x}$ will be sampled from $f_1(x; \boldsymbol{\lambda}_1)$ or $f_2(x; \boldsymbol{\lambda}_2)$ will depend on another random variable $\underline{z} \in \{0, 1\}$. Specifically, the conditional PDF of $\underline{x}$ given $\underline{z}$ is

[2] In this document a single random variable will be used. It is possible to generalization 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.

$$f(x \mid z; \boldsymbol{\lambda}) = \begin{cases} f_1(x; \boldsymbol{\lambda}_1), & z = 0, \\ f_2(x; \boldsymbol{\lambda}_2), & z = 1, \end{cases} \tag{1}$$

where $\boldsymbol{\lambda} = \{\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2\}$. Let the PMF of $\underline{z}$ be given by

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

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

3 Doing "the math"

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

$$f(y; \boldsymbol{\theta}) = f(x, z; \boldsymbol{\theta}) \tag{3}$$

$$= f(x \mid z; \boldsymbol{\theta})\, p(z; \boldsymbol{\theta}). \tag{4}$$

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

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

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

which can be rewritten[4] as

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

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

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

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

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

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

4 *The expectation step*

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

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

[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 jth estimate of $\boldsymbol{\theta}$.

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

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

$$= \sum_{i=1}^{n} (1-\underline{z}_i)\left(\log \pi_1 + \log f_1(\underline{x}_i; \boldsymbol{\lambda}_1)\right) + \sum_{i=1}^{n} \underline{z}_i \left(\log(1-\pi_1) + \log f_2(\underline{x}_i; \boldsymbol{\lambda}_2)\right) \tag{13}$$

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

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

$$
\begin{aligned}
Q\left(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(j)}\right) &= \mathbb{E}_{\underline{Y}}\left[\log f\left(\underline{Y};\boldsymbol{\theta}\right) \mid X, \boldsymbol{\theta}^{(j)}\right] && (15)\\
&= \mathbb{E}_{\underline{Z}}\left[\log f\left(\underline{X},\underline{Z};\boldsymbol{\theta}\right) \mid \underline{X} = X, \boldsymbol{\theta}^{(j)}\right] && (16)\\
&= \sum_{i=1}^{n}\left(1-\mathbb{E}\left[\underline{z}_i \mid X,\boldsymbol{\theta}^{(j)}\right]\right)\log f_1\left(x_i;\boldsymbol{\lambda}_1\right) + \mathbb{E}\left[\underline{z}_i \mid X,\boldsymbol{\theta}^{(j)}\right]\log f_2\left(x_i;\boldsymbol{\lambda}_2\right)\\
&\qquad + \sum_{i=1}^{n}\mathbb{E}\left[\underline{z}_i \mid X,\boldsymbol{\theta}^{(j)}\right]\log\left(1-\pi_1\right) + \left(1-\mathbb{E}\left[\underline{z}_i \mid X,\boldsymbol{\theta}^{(j)}\right]\right)\log\pi_1. && (17)
\end{aligned}
$$

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

$$
\begin{aligned}
p\left(z_i \mid X\right) &= \frac{f\left(X \mid z_i\right)p\left(z_i\right)}{f\left(X\right)} && (18)\\
&= \frac{f\left(x_1\right)\cdots f\left(x_i \mid z_i\right)\cdots f\left(x_n\right)p\left(z_i\right)}{f\left(x_1\right)\cdots f\left(x_i\right)\cdots f\left(x_n\right)} && (19)\\
&= \frac{f\left(x_i \mid z_i\right)p\left(z_i\right)}{f\left(x_i\right)} && (20)\\
&= p\left(z_i \mid x_i\right), && (21)
\end{aligned}
$$

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

$$
\begin{aligned}
\mathbb{E}\left[\underline{z}_i \mid x_i, \boldsymbol{\theta}^{(j)}\right] &= \sum_{k=0}^{1}\frac{f\left(x_i, z;\boldsymbol{\theta}^{(j)}\right)}{f\left(x_i;\boldsymbol{\theta}^{(j)}\right)}z_k && (22)\\
&= \frac{1}{f\left(x_i;\boldsymbol{\theta}^{(j)}\right)}\sum_{k=0}^{1} f\left(x_i, z;\boldsymbol{\theta}^{(j)}\right)z_k && (23)\\
&= \frac{1}{f\left(x_i;\boldsymbol{\theta}^{(j)}\right)}\left(f_1\left(x_i;\boldsymbol{\lambda}_1^{(j)}\right)\pi_1^{(j)}(0)\right.\\
&\qquad \left. + f_2\left(x_i;\boldsymbol{\lambda}_2^{(j)}\right)(1-\pi_1^{(j)})(1)\right) && (24)\\
&= \frac{f_2\left(x_i;\boldsymbol{\lambda}_2^{(j)}\right)\left(1-\pi_1^{(j)}\right)}{\pi_1^{(j)}f_1\left(x_i;\boldsymbol{\lambda}^{(j)}\right) + \left(1-\pi_1^{(j)}\right)f_2\left(x_i;\boldsymbol{\lambda}_2^{(j)}\right)}. && (25)
\end{aligned}
$$

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

The expectation step is then

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

$$= \frac{f_2\left(x_i; \boldsymbol{\lambda}_2^{(j)}\right)\left(1-\pi_1^{(j)}\right)}{\pi_1^{(j)} f_1\left(x_i; \boldsymbol{\lambda}^{(j)}\right) + \left(1-\pi_1^{(j)}\right) f_2\left(x_i; \boldsymbol{\lambda}_2^{(j)}\right)}. \tag{27}$$

where $\hat{z}_i^{(j)}$ will be used from now on for brevity.

5 *Maximization step*

Now that the missing data $\underline{z}_i$ 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 $\hat{Y} = \left\{X, \hat{Z}^{(j)}\right\}$. [8]

[8] Note that $\hat{Z}^{(j)} = \left\{\hat{z}_1^{(j)}, \ldots, \hat{z}_n^{(j)}\right\}$.

Let the PDFs $f_i\left(x; \boldsymbol{\lambda}_i\right)$ be exponentially distributed with parameter λ_i. Then, the PDFs can be written as

$$f_i\left(x; \lambda_i\right) = \begin{cases} \lambda_i e^{-\lambda_i x}, & x \geq 0, \\ 0, & x < 0, \end{cases} \tag{28}$$

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 $Q\left(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(j)}\right)$ from (17). Then,

$$\begin{aligned} Q\left(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(j)}\right) &= \sum_{i=1}^{n}\left(1-\hat{z}_i^{(j)}\right)\log f_1\left(x_i; \boldsymbol{\lambda}_1\right) + \hat{z}_i^{(j)} \log f_2\left(x_i; \boldsymbol{\lambda}_2\right) \\ &\quad + \sum_{i=1}^{n} \hat{z}_i^{(j)} \log\left(1-\pi_1\right) + \left(1-\hat{z}_i^{(j)}\right)\log \pi_1 \qquad (29) \\ &= \sum_{i=1}^{n}\left(1-\hat{z}_i^{(j)}\right)\log\left(\lambda_1 e^{-\lambda_1 x_i}\right) + \hat{z}_i^{(j)} \log\left(\lambda_2 e^{-\lambda_2 x_i}\right) \\ &\quad + \sum_{i=1}^{n} \hat{z}_i^{(j)} \log\left(1-\pi_1\right) + \left(1-\hat{z}_i^{(j)}\right)\log \pi_1 \qquad (30) \\ &= \sum_{i=1}^{n}\left(1-\hat{z}_i^{(j)}\right)\left(\log\lambda_1 - \lambda_1 x_i\right) + \hat{z}_i^{(j)}\left(\log\lambda_2 - \lambda_2 x_i\right) \\ &\quad + \sum_{i=1}^{n} \hat{z}_i^{(j)} \log\left(1-\pi_1\right) + \left(1-\hat{z}_i^{(j)}\right)\log \pi_1. \qquad (31) \end{aligned}$$

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

$$\frac{\partial Q\left(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(j)}\right)}{\partial \lambda_1} = \sum_{i=1}^{n} \left(1 - \hat{z}_i^{(j)}\right) \left(\frac{1}{\lambda_1} - x_i\right) \tag{32}$$

$$= \frac{1}{\lambda_1} \sum_{i=1}^{n} \left(1 - \hat{z}_i^{(j)}\right) - \sum_{i=1}^{n} \left(1 - \hat{z}_i^{(j)}\right) x_i. \tag{33}$$

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

$$\hat{\lambda}_1^{(j+1)} = \frac{\sum_{i=1}^{n} \left(1 - \hat{z}_i^{(j)}\right)}{\sum_{i=1}^{n} \left(1 - \hat{z}_i^{(j)}\right) x_i}. \tag{34}$$

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

$$\hat{\lambda}_2^{(j+1)} = \frac{\sum_{i=1}^{n} \hat{z}_i^{(j)}}{\sum_{i=1}^{n} \hat{z}_i^{(j)} x_i}. \tag{35}$$

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

$$\frac{\partial Q\left(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(j)}\right)}{\partial \pi_1} = \sum_{i=1}^{n} -\hat{z}_i^{(j)} \frac{1}{1 - \pi_1} + \left(1 - \hat{z}_i^{(j)}\right) \frac{1}{\pi} \tag{36}$$

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

$$= \frac{1}{\pi_1 (1 - \pi_1)} \sum_{i=1}^{n} \left(1 - \hat{z}_i^{(j)} - \pi_1\right). \tag{38}$$

Equating to 0 and solving for π_1 gives the next estimate

$$\hat{\pi}_1^{(j+1)} = 1 - \frac{1}{n} \sum_{i=1}^{n} \hat{z}_i^{(j)}. \tag{39}$$

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.