Expectation Maximization with Coin Toss example

Marijo Simunovic

October 18, 2017

1 Coin toss problem description

This documents contains detailed solution for the problem described in [2]. In summary, someone choose randomly one of two biased coins five times. Each time, selected coin has been tossed and outcomes are recorded. We are given the outcomes of this five events but not the identities of coins whom each event belongs. Also, coin biases are unknown. Our goal is to somehow infer these parameters from the given data. In order to solve the problem we resort the Expectation Maximization (EM) algorithm.

2 Theoretical background of EM

Given the statistical model which generates a set $\mathbf{X}$ of observed data, a set of unobserved latent data or missing values $\mathbf{Z}$, and a vector of unknown parameters Θ, along with a likelihood function $L(\boldsymbol{\Theta}; \boldsymbol{X}, \boldsymbol{Z}) = p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\Theta})$, the maximum likelihood estimate (MLE) of the unknown parameters is determined by the marginal likelihood of the observed data

$$L(\boldsymbol{\Theta}; \boldsymbol{X}) = p(\boldsymbol{X}|\boldsymbol{\Theta}) = \int p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\Theta}) d\boldsymbol{Z} \tag{1}$$

The EM algorithm seeks to find the MLE of the marginal likelihood by iteratively applying these two steps:

Expectation step (E step): Calculate the expected value of the log likelihood function, with respect to the conditional distribution of $\mathbf{Z}$ given $\mathbf{X}$ under the current estimate of the parameters $\boldsymbol{\Theta}^{(t)}$:

$$\boldsymbol{Q}(\boldsymbol{\Theta}|\boldsymbol{\Theta}^{(t)}) = \boldsymbol{E}_{\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\Theta}^{(t)}}[\boldsymbol{log} L(\boldsymbol{\Theta}; \boldsymbol{X}, \boldsymbol{Z})] \tag{2}$$

Maximization step (M step): Find the parameter that maximizes this quantity:

$$\boldsymbol{\Theta}^{(t+1)} = \boldsymbol{\arg\min_{\Theta}} \boldsymbol{Q}(\boldsymbol{\Theta}|\boldsymbol{\Theta}^{(t)}) \tag{3}$$

This is a word by word definition copied from [1][1]. If you don't understand what the above means at first, don't worry. It is merely due to the compactness of the mathematical language. The first term means: Set up the model for the likelihood function pretending that you know what is the value of $\boldsymbol{\Theta}$. In the actual solution, we randomly instantiate its values to $\boldsymbol{\Theta_i} \in \boldsymbol{[0, 1]}$ where i

[1] I choose to take this particular definition (out of many out there) because it is the first one I actually understood. I don't think that this is a better or worse than the others, it's just that my brain was finally ready.

is the i-th component of the vector $\boldsymbol{\Theta}$. Once we have defined this equation (model) we take the expectation over it to find the probability of each possible value of $\boldsymbol{Z}$, given $\boldsymbol{\Theta}$. Then we compute a better estimate for the parameters $\boldsymbol{\Theta}$ using these probabilities. We iterate these procedure until convergence.

2.1 Math Mode

Vector **z** corresponds to the event of choosing one of the two given coins. It is bivariate, the values of each of the two possible outcomes are mutually exclusive.

$$\boldsymbol{z_n} = \begin{bmatrix} \boldsymbol{z_{n1}} \\ \boldsymbol{z_{n2}} \end{bmatrix} \in \left\{ \begin{bmatrix} \mathbf{1} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \mathbf{0} \\ \mathbf{1} \end{bmatrix} \right\} \tag{4}$$

Probability of the event $\boldsymbol{z_n}$ (which coin we selected in the n-th event) is given by

$$\boldsymbol{p(z_n)} = \prod_{\boldsymbol{k=1}}^{\mathbf{2}} \boldsymbol{\pi_k^{z_{nk}}} \tag{5}$$

with requirement that $\sum_{\boldsymbol{k}} \boldsymbol{\pi_k} = 1$.

Each single coin toss is independent on the other tosses. It only depends on the coin we choose previously. Probability of the single coin toss outcome given coin k is

$$\boldsymbol{p(x_n^j | z_n, \Theta)} = \prod_{\boldsymbol{k=1}}^{\mathbf{2}} [\boldsymbol{\Theta_k^{x_n^j}(1 - \Theta_k^{1-x_n^j})}]^{\boldsymbol{z_{nk}}} \tag{6}$$

where $\boldsymbol{x_n^j = 1}$ if j-th outcome was head and $\boldsymbol{x_n^j = 0}$ if j-th outcome was tail. We can now determine the probability of each of our 5 events

$$\boldsymbol{p(\{x_n^1 \cdots x_n^{10}\} | z_n, \Theta)} = \prod_{\boldsymbol{j=1}}^{\mathbf{10}} \boldsymbol{p(x_n^j | z_n, \Theta)} \tag{7}$$

Joint probability of the coin toss outcomes and coin selections is given by:

$$\boldsymbol{p(X_1, \ldots, X_5, z_1, \ldots, z_5 | \Theta)} \tag{8}$$

$$\boldsymbol{= p(\{x_1^1 \cdots x_1^{10}\}, \ldots, \{x_5^1 \cdots x_5^{10}\}, z_1, \ldots, z_5 | \Theta)} \tag{9}$$

$$\boldsymbol{= p(\{x_1^1 \cdots x_1^{10}\}, \ldots, \{x_5^1 \cdots x_5^{10}\} | z_1, \ldots, z_5, \Theta) p(z_1, \ldots, z_5)} \tag{10}$$

$$= \prod_{\boldsymbol{n=1}}^{\mathbf{5}} \boldsymbol{p(\{x_n^1 \cdots x_n^{10}\} | z_n, \Theta)} \prod_{\boldsymbol{n=1}}^{\mathbf{5}} \boldsymbol{p(z_n)} \tag{11}$$

Finally, combining the above we get the equation

$$\boldsymbol{p(X_1, \ldots, X_5, z_1, \ldots, z_5 | \Theta)} = \prod_{\boldsymbol{n=1}}^{\mathbf{5}} \prod_{\boldsymbol{j=1}}^{\mathbf{10}} \prod_{\boldsymbol{k=1}}^{\mathbf{2}} [\boldsymbol{\Theta_k^{x_n^j}(1 - \Theta_k^{1-x_n^j})}]^{\boldsymbol{z_{nk}}} \prod_{\boldsymbol{n=1}}^{\mathbf{5}} \prod_{\boldsymbol{k=1}}^{\mathbf{2}} \boldsymbol{\pi_k^{z_{nk}}} \tag{12}$$

Taking the log of the expression above we get rid of the products in exchange of the summations.

$$\boldsymbol{log(P)} = \sum_{\boldsymbol{n=1}}^{\mathbf{5}} \sum_{\boldsymbol{k=1}}^{\mathbf{2}} \boldsymbol{z_{nk}} [\sum_{\boldsymbol{j=1}}^{\mathbf{10}} \boldsymbol{log(\Theta_k^{x_n^j}(1 - \Theta_k^{1-x_n^j})) + log(\pi_k)}] \tag{13}$$

The final step is taking the expectation over the $\boldsymbol{z}$ random variable. In this case, we are using the conditional expectation, as we do not know what was the outcome on each of the step. Using the linearity of the expectation as

$$\boldsymbol{E_{Z|X}[log(P)] = \sum_{n=1}^{5}\sum_{k=1}^{2} E_{Z|X}[z_{nk}][\sum_{j=1}^{10} log(\Theta_k^{x_n^j}(1-\Theta_k^{1-x_n^j})) + log(\pi_k)]} \tag{14}$$

Conditional expectation gives as the expected value of the random experiment in light of the new information on which we are conditioning. Consider throwing an ordinary dice. The expected value of the outcome is $\boldsymbol{E(numberondice) = (1+2+3+4+5+6)/6 = 3.5}$. But if I told you the parity of the outcome, in that case we would calculate the expectation $\boldsymbol{E(odd) = (1+3+5)/3 = 3}$ and $\boldsymbol{E(even) = (2+4+6)/3 = 4}$. We see that it is not enough to express the expectation with the one value only. We can use similar analogy in our experiment. As we have see the outcomes of each of the five events, we will assign bigger value of $\boldsymbol{\Theta}$ to the ones that have had more heads. If there isn't any significant difference in the outcomes we will have a hard time trying to figure out which event belongs to which coin.

$$\boldsymbol{E_{Z|X}[z_{nk}] = \sum_{z} z_{nk} p(z_n|\{x_n^1 \cdots x_n^{10}\}; \Theta)} \tag{15}$$

Using the Bayes formula, the posterior of $\boldsymbol{z_n}$ given values of the n-th event is

$$\boldsymbol{p(z_n|\{x_n^1 \cdots x_n^{10}\}; \Theta) = \frac{p(\{x_n^1 \cdots x_n^{10}\}|z_n; \Theta)p(z_n)}{p(\{x_n^1 \cdots x_n^{10}\}; \Theta)}} \tag{16}$$

We already know the expressions for the terms in numerator, and the expression in denominator is given by marginalizing over all possible outcomes of $\boldsymbol{z}$ (in our, case there are only two, as already mentioned).

$$\boldsymbol{p(z_n|x_n; \Theta) = \frac{\prod_{k=1}^{2}[\pi_k p(x_n; \Theta_k)]^{z_{nk}}}{\sum_{z_m}\prod_{m=1}^{2}[\pi_m p(x_n; \Theta_m)]^{z_{nm}}}} \tag{17}$$

Going back to our expectation formula

$$\boldsymbol{E_{Z|X}[z_{nk}] = \frac{\sum_{z_n} z_{nk}\prod_{k=1}^{2}[\pi_k p(x_n; \Theta_k)]^{z_{nk}}}{\sum_{z_m}\prod_{m=1}^{2}[\pi_m p(x_n; \Theta_m)]^{z_{nm}}}} \tag{18}$$

We see that in the numerator those events where $\boldsymbol{z_{nk} = 0}$ will be eliminated, while in the denominator we are still summing over all possibilities.

$$\boldsymbol{E_{Z|X}[z_{nk}] = \frac{\pi_k p(x_n; \Theta)}{\sum_{m=1}^{2}\pi_m p(x_n; \Theta_m)}} \tag{19}$$

As the probability of the choosing both coins is equal, we know that the $\boldsymbol{\pi_k = 0.5}$ in both cases. Thus, we can eliminate it from our expression. Also, we fix the $\boldsymbol{\Theta_k}$ values (first iteration uses randomly initiated values, while the subsequent steps use values calculated in the previous step) in the E step of the EM algorithm.

In the M step, we fix the expectation value calculated in the E step vary $\boldsymbol{\Theta_k}$ values. To find the values which maximize the expectation of log equation, we use the derivation. Second term, after the summation sign becomes constant, so it does not have the effect on the maximization procedure.

$$\boldsymbol{maxL(\Theta) = E_{Z|X}[p(X_1, \ldots, X_5, z_1, \ldots, z_5|\Theta]} \tag{20}$$

$$= \sum_{n=1}^{5} \sum_{j=1}^{10} \sum_{k=1}^{2} \boldsymbol{E_{Z|X}[z_{nk}] log(\Theta_k^{x_n^j}(1-\Theta_k^{1-x_n^j}))} + \boldsymbol{CONST.} \tag{21}$$

$$= \sum_{n=1}^{5} \sum_{j=1}^{10} \sum_{k=1}^{2} \boldsymbol{E_{Z|X}[z_{nk}](x_n^j log(\Theta_k) + (1-x_n^j) log(1-\Theta_k))} + \boldsymbol{CONST.} \tag{22}$$

For a single coin $\boldsymbol{k}$

$$\frac{\boldsymbol{dL(\Theta)}}{\boldsymbol{d\Theta_k}} = \sum_{n=1}^{5} \sum_{j=1}^{10} \boldsymbol{E_{Z|X}[z_{nk}]}(\frac{\boldsymbol{x_n^j}}{\boldsymbol{\Theta_k}} - \frac{\boldsymbol{1-x_n^j}}{\boldsymbol{1-\Theta_k}}) = \boldsymbol{0} \tag{23}$$

Multiplying with both denominators $\boldsymbol{\Theta_k(1-\Theta_k)}$ we get

$$\frac{\boldsymbol{dL(\Theta)}}{\boldsymbol{d\Theta_k}} = \sum_{n=1}^{5} \sum_{j=1}^{10} \boldsymbol{E_{Z|X}[z_{nk}](x_n^j - \Theta_k)} = \boldsymbol{0} \tag{24}$$

and finally, since $\boldsymbol{\Theta}$'s index does not appear in the summation indexes, we can simply divide to obtain its (maximal) value

$$\boldsymbol{\Theta_k} = \frac{\sum_{n=1}^{5} \sum_{j=1}^{10} \boldsymbol{E_{Z|X}[z_{nk}] x_n^j}}{\boldsymbol{10} \sum_{n=1}^{5} \boldsymbol{E_{Z|X}[z_{nk}]}} \tag{25}$$

We see that the result for each coin is weighted expectation of a single coin over a sum of the expectations of both coins.

EM is sort of like moving a heavy table without anyone's help. You push a bit from one side and then from the other. Bit by bit you are reaching the goal.

References

[1] Expectation Maximization algorithm wiki page,https://en.wikipedia.org/wiki/Expectation

[2] Chuong B Do, Serafim Batzoglou, *What is the expectation maximization algorithm*, Nature Biotechnology 26, 897-899, dpi:10.1038/nbt14052008, http://www.nature.com/nbt/journal/v26/n8/full/nbt1406.html