Expectation Maximization with Coin Toss example

AntiCodeOn

October 20, 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. After each random selection, the given coin has been tossed ten times and outcomes were recorded. We were given the outcomes of these 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 to 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:

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

This is a word by word definition[1] copied from [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 (Equation (2)) means: Set up the model for the likelihood function pretending that you know what is the value of Θ. In the actual solution to the coin toss problem, we randomly instantiate its values to $\Theta_i \in [0, 1]$ where i is the i-th component of the vector Θ, i.e. bias of the i-th coin. Once we have defined this equation (model) we take the expectation over it to find the probability of each possible value of Z, given Θ. Then we compute a better estimate for the parameters Θ using these probabilities (Equation (3)) . We iterate these procedure until convergence.

2.1 Problem setup

Vector $\mathbf{z}$ 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.

$$z_n = \begin{bmatrix} z_{n1} \\ z_{n2} \end{bmatrix} \in \left\{ \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \end{bmatrix} \right\} \qquad (4)$$

As $\mathbf{z}$ uses 1-of-K representation we can write probability distribution of the event $\mathbf{z_n}$ (which coin we selected in the n-th event) in a following way

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

with requirement that $\sum_k \pi_k = 1$.

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 governed by Bernoulli distribution

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

---

[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.

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

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

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

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

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

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

Finally, combining the (5), (7) and (10) we get the equation

$$P = \prod_{n=1}^{5} \prod_{k=1}^{2} \prod_{j=1}^{10} [\Theta_k^{x_n^j}(1 - \Theta_k^{1-x_n^j})]^{z_{nk}} \pi_k^{z_{nk}} \tag{11}$$

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

$$log(P) = \sum_{n=1}^{5} \sum_{k=1}^{2} z_{nk}[\sum_{j=1}^{10} log(\Theta_k^{x_n^j}(1 - \Theta_k^{1-x_n^j})) + log(\pi_k)] \tag{12}$$

The final step is taking the expectation over hidden random variable z. 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

$$E_{Z|X}[log(P)] =$$

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

Conditional expectation gives us 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 $E(number-on-dice) = (1 + 2 + 3 + 4 + 5 + 6)/6 = 3.5$. But if I told you the parity of the outcome, in that case we would calculate the expectation $E(odd) = (1 + 3 + 5)/3 = 3$ and $E(even) = (2 + 4 + 6)/3 = 4$.

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 Θ 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.

$$E_{Z|X}[z_{nk}] = \sum_z z_{nk} p(z_n | \{x_n^1 \cdots x_n^{10}\}; \Theta) \tag{14}$$

Using the Bayes formula, the posterior of z_n given values of the n-th event is

$$p(z_n | \{x_n^1 \cdots x_n^{10}\}; \Theta) = \frac{p(\{x_n^1 \cdots x_n^{10}\} | z_n; \Theta) p(z_n)}{p(\{x_n^1 \cdots x_n^{10}\}; \Theta)} \tag{15}$$

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

$$p(z_n | x_n; \Theta) = \frac{\prod_{k=1}^{2} [\pi_k p(x_n; \Theta_k)]^{z_{nk}}}{\sum_{z_m} \prod_{m=1}^{2} [\pi_m p(x_n; \Theta_m)]^{z_{nm}}} \tag{16}$$

Going back to our expectation formula

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

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

$$E_{Z|X}[z_{nk}] = \frac{\pi_k p(x_n; \Theta)}{\sum_{m=1}^{2} \pi_m p(x_n; \Theta_m)} \tag{18}$$

As the probability of choosing both coins is equal, we know that the $\pi_k = 0.5$ in both cases. Thus, we can eliminate it from our expression. Also, we fix the Θ_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 and vary Θ_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.

$$maxL(\Theta) = E_{Z|X}[p(X_1, \ldots, X_5, z_1, \ldots, z_5|\Theta] \tag{19}$$

$$= \sum_{n=1}^{5} \sum_{j=1}^{10} \sum_{k=1}^{2} E_{Z|X}[z_{nk}]log(\Theta_k^{x_n^j}(1 - \Theta_k^{1-x_n^j})) + C \tag{20}$$

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

For a single coin k

$$\frac{dL(\Theta)}{d\Theta_k} = \sum_{n=1}^{5} \sum_{j=1}^{10} E_{Z|X}[z_{nk}](\frac{x_n^j}{\Theta_k} - \frac{1 - x_n^j}{1 - \Theta_k}) = 0 \tag{22}$$

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

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

and finally, since Θ's index does not appear in the summation indexes, we can simply divide to obtain its (maximal) value

$$\Theta_k = \frac{\sum_{n=1}^{5} E_{Z|X}[z_{nk}]\sum_{j=1}^{10} x_n^j}{\sum_{n=1}^{5} E_{Z|X}[z_{nk}]\sum_{j=1}^{10} 1} \tag{24}$$

The new parameter value is a weighted sum of the expectations - in the numerator we are taking into account only head outcomes and in the denominator both head and tail outcomes.

EM is sort of like trying to move 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.

3   EM algorithm

4   Results

In order to verify our result we have compared it to [2] where the problem is originally described. We chose the ssame values Θ=[0.6, 0.5] and we iterate

EM steps until convergence. Here we define convergence in a following way. We choose a threshold T. EM procedure is continued until change of either of $\mathbf{\Theta_k}$ in subsequent iterations is greater than value T.

References

[1] Expectation Maximization algorithm wiki page

[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