EM algorithm is an **construct** and this note is designed for deriving an example o using EM-algorithm to better understand how it works.

EM Algorithm For Binomial Mixture Model

Given two coins with unknown probabilities of heads θ_1 and θ_2 respectively, the first coin is chosen with probability π_1 and the second one with probability $1 - \pi_1$. The chosen coin is flipped once, and the outcome is 0 or 1. Performing this random experiment for N trials independently, the outcomes are recorded as dataset $X = \{x_i\}_{i=1}^N$.

(a) Let's understand from a probabilistic perspective if we want to know how a single random variable of work under this setting by writing down the expression for the log-likelihood $\log p(X|\theta_1, \theta_2, \pi_1)$.

Solution

For the probability of observing a single observation x_i from the random variable of X, the likelihood can be expressed as the probability of seeing π_1 with head θ_1 plus the probability of seeing π_2, or just $(1 - \pi_1)$, with head θ_2:

$$p(x_i \mid \theta_1, \theta_2, \pi_1) = \pi_1 p(x_i \mid \theta_1) + (1 - \pi_1) p(x_i \mid \theta_2)$$

And the probability of a single i random variable follows a binomial distribution, which is:

$$p(x_i \mid \theta_k) = \theta_k^{x_i}(1 - \theta_k)^{1-x_i}$$

Combining together, for N independent trials, the likelihood of the dataset $X = \{x_i\}_{i=1}^N$ is:

$$p(X \mid \theta_1, \theta_2, \pi_1) = \prod_{i=1}^{N} \left[\pi_1 \theta_1^{x_i}(1 - \theta_1)^{1-x_i} + (1 - \pi_1)\theta_2^{x_i}(1 - \theta_2)^{1-x_i}\right]$$

Taking the logarithm:

$$\log p(X \mid \theta_1, \theta_2, \pi_1) = \sum_{i=1}^{N} \log \left[\pi_1 \theta_1^{x_i}(1 - \theta_1)^{1-x_i} + (1 - \pi_1)\theta_2^{x_i}(1 - \theta_2)^{1-x_i}\right]$$

This is nice and easy to solve, but we will make it complicated.

Next, we introduce the latent variable for the EM algorithm. Let $z_i = (z_{1i}, z_{2i})$ be an indicator vector for each observation x_i, such that $z_{ki} = 1$ if the k-th coin is chosen, and 0 otherwise, $k = \{1, 2\}$. For the dataset, we have $Z = \{z_i\}_{i=1}^N$.

(b) Write down the expression for the log-likelihood $\log p(X, Z|\theta_1, \theta_2, \pi_1)$.

Solution

Notice that for this question, there is a few "dimension", there is the probability of head, the probability of seeing coin 1 or coin 2, and there is the variable of seeing what the $k^{\mathbf{th}}$ coin is. So we are making this problem of talking about just one random variable of observing x_i from X into a chain of random variable of observing x_i from X given that we are looking at $z_{ki} = 1$ trial.

To incorporate the latent variable $Z = \{z_i\}_{i=1}^N$, where $z_i = (z_{1i}, z_{2i})$ is an indicator vector such that $z_{ki} = 1$ if the k-th coin is chosen and 0 otherwise ($k \in \{1, 2\}$), we need to enumerate over all the possible combination between Z and X. Furthermore, z_{ki} need to serve as an indicator of whether the function takes value at all for the i^{th} Z latent variable. We can utilize properties of exponential where if $z_{ki} = 1$, the function contributes and if $z_{ki} = 0$, then the function takes 1 and does not contribute. This can be written as.

$$p(X, Z \mid \theta_1, \theta_2, \pi_1) = \prod_{i=1}^{N} \prod_{k=1}^{2} \left[\pi_k \theta_k^{x_i} (1 - \theta_k)^{1-x_i}\right]^{z_{ki}}$$

We can take the log-likelihood by the following:

$$\log p(X, Z \mid \theta_1, \theta_2, \pi_1) = \sum_{i=1}^{N} \sum_{k=1}^{2} z_{ki} \big[\log \pi_k + x_i \log \theta_k + (1 - x_i) \log(1 - \theta_k) \big]$$

Here, π_k represents the **prior probability** of selecting the k-th coin (or just in general how likely it is to select the k^{th} coin (not in terms of the chain of trial but which number of coin is selected)). This whole expression can be deemed as taking the **expectation** with regarding to the latent distribution of z. However, this problem becomes intractable, which is why we need to use EM to solve it.

Remember that in the **most generalized version of EM**, we have an hidden Z distribution that we don't know, we assume that our data distribution X depends on this hidden distribution of Z. Since we don't know about this hidden, we can't just maximize this partial log-likelihood directly (problem becomes intractable), which is why we want to **infer** what such Z distribution is (E-step), then maximize (M-step) it.

(a) Expect an q (expected posterior) distribution from what we know in our data.

(b) Maximize under the assumption that our q distribution is correct.

(c) **E-step:** Let $\theta_1^{t-1}, \theta_2^{t-1}, \pi_1^{t-1}$ be the parameter estimation given by the $t-1$ iteration of the EM algorithm. Derive $p(z_{ki} = 1 | x_i, \theta_1^{t-1}, \theta_2^{t-1}, \pi_1^{t-1})$, $k = \{1, 2\}$.

Solution

In the E-step (usually the hard part), we want to derive the **posterior probability** (given all observation, how likely it is for coin k to be selected at trial i). We want to know the probability of the latent being 1 (number k^{th} coin getting chosen) given the random variable (observation), and previous probability of coin-1-head, coin-2-head, and coin-1-showing. We can decompose the previous notion by using **Bayes' Rule**:

$$p(z_{ki} = 1 \mid x_i, \theta_1^{(t-1)}, \theta_2^{(t-1)}, \pi_1^{(t-1)}) = \frac{p(z_{ki} = 1, x_i \mid \theta_1^{(t-1)}, \theta_2^{(t-1)}, \pi_1^{(t-1)})}{p(x_i \mid \theta_1^{(t-1)}, \theta_2^{(t-1)}, \pi_1^{(t-1)})}$$

We can derive the **numerator** by looking at the **joint distribution** (seeing k^{th} coin with the observation) through using prior probabilistic distribution of seeing the k^{th} coin in the **previous** trial ($\pi_k^{(t-1)}$) and the likelihood ($p(x_i \mid \theta_k^{(t-1)})$) derived from the observation of the previous trial:

$$p(z_{ki} = 1, x_i \mid \theta_1^{(t-1)}, \theta_2^{(t-1)}, \pi_1^{(t-1)}) = \pi_k^{(t-1)} \cdot p(x_i \mid \theta_k^{(t-1)})$$

where the likelihood is simply expressed as a Bernoulli distribution (since we are talking about the probability of seeing certain variable in a sequence of binary decisions):

$$p(x_i \mid \theta_k^{(t-1)}) = (\theta_k^{(t-1)})^{x_i} (1 - \theta_k^{(t-1)})^{1-x_i}$$

Notice that this expression is highly alike the probability distribution that we derived earlier, just that this is a particular instance in the chain now instead of the general expression we described earlier.

$$p(X, Z \mid \theta_1, \theta_2, \pi_1) = \prod_{i=1}^{N} \prod_{k=1}^{2} \left[\pi_k \theta_k^{x_i} (1 - \theta_k)^{1-x_i}\right]^{z_{ki}}$$

Continues on next page...

Solution

Continues from previous page...
Now we have the joint distribution, we need to focus on the **denominator**, the **marginal probability** $p(x_i \mid \theta_1^{(t-1)}, \theta_2^{(t-1)}, \pi_1^{(t-1)})$, which is the **total probability** of observing x_i (which we have derived the general expression earlier in the joint distribution already):

$$p(x_i \mid \theta_1^{(t-1)}, \theta_2^{(t-1)}, \pi_1^{(t-1)}) = \sum_{k=1}^{2} p(z_{ki} = 1, x_i \mid \theta_1^{(t-1)}, \theta_2^{(t-1)}, \pi_1^{(t-1)})$$

Notice that this inner component is something that we have derived before, which is:

$$p(x_i \mid \theta_1^{(t-1)}, \theta_2^{(t-1)}, \pi_1^{(t-1)}) = \sum_{k=1}^{2} \pi_k^{(t-1)} \cdot p(x_i \mid \theta_k^{(t-1)})$$

This is sort of summing all the prior probabilistic distribution of seeing coin k with a likelihood weighting term.

$$p(x_i \mid \theta_1^{(t-1)}, \theta_2^{(t-1)}, \pi_1^{(t-1)}) = \sum_{k=1}^{2} \pi_k^{(t-1)} \cdot p(x_i \mid \theta_k^{(t-1)}).$$

Specifically for $k = 2$ condition :

$$p(x_i \mid \theta_1^{(t-1)}, \theta_2^{(t-1)}, \pi_1^{(t-1)}) = \Big[(\pi_1^{(t-1)}) \cdot p(x_i \mid \theta_1^{(t-1)})\Big] + \Big[(1 - \pi_1^{(t-1)}) \cdot p(x_i \mid \theta_2^{(t-1)})\Big]$$

Substituting the above expressions we get with marginal distribution and the joint distribution, we would get the following. For $k \in \{1, 2\}$, the posterior probability is:

$$p(z_{ki} = 1 \mid x_i, \theta_1^{(t-1)}, \theta_2^{(t-1)}, \pi_1^{(t-1)}) = \frac{\pi_k^{(t-1)} \cdot (\theta_k^{(t-1)})^{x_i} (1 - \theta_k^{(t-1)})^{1-x_i}}{\sum_{j=1}^{2} \pi_j^{(t-1)} \cdot (\theta_j^{(t-1)})^{x_i} (1 - \theta_j^{(t-1)})^{1-x_i}}$$

and we should construct our E-step based on this expression above.

(d) **M-step:** Show that

$$\pi_1^t = \frac{N_1}{N},$$

where N_1 is the number of trials the first coin is chosen in the t-th iteration of the EM algorithm. Notice that π_1^t is essentially the probability of observing coin 1 at trial t. **We essentially want to conduct an MLE on the likelihood function of** $Q(\pi_1, \theta_1, \theta_2)$ (adjusting variables such that we get teh maximum probability of observing π_1).

Solution

To update π_1 in the M-step, we maximize the expected complete data log-likelihood. The complete data log-likelihood is given by (notice that this is sort of taking the **expectation** with regard to the latent distribution):

$$\log p(X, Z \mid \pi_1, \theta_1, \theta_2) = \sum_{i=1}^{N} \sum_{k=1}^{2} z_{ki} \left[\log \pi_k + x_i \log \theta_k + (1 - x_i) \log(1 - \theta_k)\right].$$

Or just that:

$$Q(\pi_1, \theta_1, \theta_2) = \mathbb{E}_{z_{ki}} \left[\log p(X, Z \mid \pi_1, \theta_1, \theta_2)\right]$$

Since the latent variables Z are not observed, we compute the expected complete data log-likelihood over the posterior distribution of Z that we retrieved from the E-step. The posterior probabilities are:

$$q_{ki} = p(z_{ki} = 1 \mid x_i, \pi_1^{t-1}, \theta_1^{t-1}, \theta_2^{t-1})$$

where q_{ki} is our build-up expected value of z_{ki}. Taking the expectation under our q_{ki} distribution, we replace z_{ki} with q_{ki}:

$$Q(\pi_1, \theta_1, \theta_2) = \mathbb{E}_{q_{ki}} \left[\log p(X, Z \mid \pi_1, \theta_1, \theta_2)\right]$$

Substituting the expectation of distribution q_{ki} into the complete data log-likelihood:

$$Q(\pi_1, \theta_1, \theta_2) = \log p(X, Z \mid \pi_1, \theta_1, \theta_2) = \sum_{i=1}^{N} \sum_{k=1}^{2} q_{ki} \left[\log \pi_k + x_i \log \theta_k + (1 - x_i) \log(1 - \theta_k)\right]$$

This Q function represents the expected complete data log-likelihood, which is what we usually maximized during the M-step to update the parameters $\pi_1, \theta_1, \theta_2$.
Continue on next page...

Solution

Continue from previous page...

For the sake of this question, we need simplification. Again, **we essentially want to conduct an MLE on the likelihood function of** $Q(\pi_1, \theta_1, \theta_2)$ (adjusting variables such that we get teh maximum probability of observing π_1). Simplifying $Q(\pi_1, \theta_1, \theta_2)$, we can separate the terms involving π_k, θ_1, and θ_2 since we don't care about the rest k coins.

$$Q(\pi_1, \theta_1, \theta_2) = \sum_{i=1}^{N} \left[q_{1i} \log \pi_1 + q_{2i} \log(1 - \pi_1)\right] + \sum_{i=1}^{N} \sum_{k=1}^{2} q_{ki} \left[x_i \log \theta_k + (1 - x_i) \log(1 - \theta_k)\right].$$

Notice that we have separated out just the terms for only π_q involved in it. So we can write just $Q(\pi_1)$ since we only want to know about π_1^t (the probability of seeing coin 1 at the t^{th} iteration), we can throw away the rest of the terms since we are not talking about any coins that is not 1 and nor are we talking about head or tail probability:

$$Q(\pi_1) = \sum_{i=1}^{N} q_{1i} \log \pi_1 + \sum_{i=1}^{N} q_{2i} \log(1 - \pi_1),$$

Taking the derivative of $Q(\pi_1)$ with respect to π_1:

$$\frac{\partial Q}{\partial \pi_1} = \frac{\sum_{i=1}^{N} q_{1i}}{\pi_1} - \frac{\sum_{i=1}^{N} q_{2i}}{1 - \pi_1}$$

Set $\frac{\partial Q}{\partial \pi_1} = 0$:

$$\pi_1 \sum_{i=1}^{N} q_{2i} = (1 - \pi_1) \sum_{i=1}^{N} q_{1i}$$

Continues on next page...

Solution

Continued from last page...
Since we only have two coins, we can use the fact that the expected number of times that coin 2 would be chosen is the total number of chooses minus the expected number of times that coin 1x is chosen: $\sum_{i=1}^{N} q_{2i} = N - \sum_{i=1}^{N} q_{1i}$. When substituting, we get that:

$$\pi_1(N - \sum_{i=1}^{N} q_{1i}) = (1 - \pi_1) \sum_{i=1}^{N} q_{1i}$$

Expand and collect terms:

$$\pi_1 N = \sum_{i=1}^{N} q_{1i}$$

Solve for π_1:

$$\pi_1 = \frac{\sum_{i=1}^{N} q_{1i}}{N}$$

Define $N_1 = \sum_{i=1}^{N} q_{1i}$, which is the expected number of trials where the first coin is chosen (following problem definition), we would get that:

$$\pi_1^t = \frac{N_1}{N}$$

The proof is thus complete.