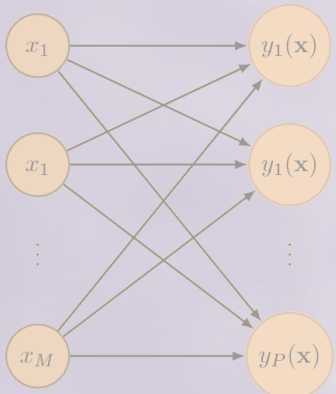


1 — Expectation Maximization

Daniel Khashabi [1]
KHASHAB2@ILLINOIS.EDU

1.1 Introduction

Consider the problem of parameter learning by maximizing the likelihood of the observations for random variables $(X, Y) \sim \{(x_i, y_i)\}_{i=1}^n$. Assume we model the joint distribution between X and Y using $p(X, Y; \theta)$ which is resulted designer's domain knowledge, and is characterized by the parameter θ.

$$\mathcal{L}(\theta) = \log \prod_{i=1}^{n} p(X, Y; \theta) = \sum_{i=1}^{n} \log p(X, Y; \theta)$$

To find the ML estimated parameters, one can maximize $\mathcal{L}(\theta)$ with respect to the model parameters θ. But what if we don't observe anything from Y? We call this scenario the "missing data" case. Assume the following definition of the likelihood,

$$\mathcal{L}(\theta) = \log \prod_{i=1}^{n} p(X; \theta) = \log \prod_{i=1}^{n} \sum_{Y} p(X, Y; \theta) = \sum_{i=1}^{n} \log \sum_{Y} p(X, Y; \theta)$$

$$\mathcal{L}(\theta) = \sum_{i=1}^{n} \log \sum_{Y} p(X, Y; \theta) \tag{1.1}$$

Performing parameter learning in the case of missing data (latent variables) by maximizing 1.1 is not trivial. But EM introduced a formalized way to approximate the maximization, by maximizing the lower-bound on this function.

[1] This is part of my notes; to find the complete list of notes visit `http://web.engr.illinois.edu/~khashab2/learn.html`. This document is updated on January 28, 2014.

1.2 Expectation-Maximization

We first introduce the algorithm, and then analyze its properties. The EM algorithm is the following,

Theorem 1.1 — The EM algorithm. Performing the following iterative steps, will result on the local maximizer of Equation 1.1.

- **Initialization:** Initialize the parameters of the mode θ.
- Repeat until convergence:

1. **Expectation:** Find the expected likelihood, $\mathcal{L}_{EM}$.

$$\mathcal{L}_{EM}(\theta;\theta_n) = \mathbb{E}_{\mathbf{Y}|\mathbf{X},\theta_n}\left[\log p\left(\mathbf{X},\mathbf{y}|\theta\right)\right] \tag{1.2}$$

2. **Maximization:** Maximize the EM objective $\mathcal{L}_{EM}$ with respect to θ.

We prove this iterations will converge to the maximization of Equation 1.1 in several steps.

■ **Lemma 1.1** The EM objective in Equation 1.2 is a lower bound on Equation 1.1.

$$\mathcal{L}_{EM}(\theta;\theta_n) \leq \mathcal{L}(\theta)$$

Proof. TODO 1 ■

[More explanation and analysis here]

■ **Example 1.1** Consider the following simple mixture model:

$$\begin{cases} p_1(x;\theta) = e^{-g_1(x;\theta)} \\ p_2(x;\theta) = e^{-g_2(x;\theta)} \\ p(x;\theta) = \mu_1 p_1 + (1-\mu_1)p_2 \end{cases}$$

Given the set of observations, $x_1,\ldots,x_n$, we want to estimate parameters of this mixture model $\Theta = \{\theta_1,\theta_2,\mu_1\}$.
The conventional maximum likelihood aims at solving the following problem:

$$\Theta := \arg\max_{\Theta}\left\{\sum_i \log\left[\mu_1 e^{-g_1(x_i;\theta)} + (1-\mu_1)e^{-g_2(x_i;\theta)}\right]\right\}$$

this optimization is slightly hard, as we have logarithm of some summation. We will add some additional variables to the model to make the optimization steps more explicit. We assume that each data is coming from one specific component. For that, we define additional variable to specify the component from which the sample is coming from:

$$\delta_i = \begin{cases} 1 & \text{if the sample is coming from the first component} \\ 0 & \text{if the sample is coming from the second component} \end{cases}$$

The complete data likelihood is:

$$\mathcal{L}(\Theta) = \sum_i \log p(x_i, \delta_i|\Theta) = \sum_i \log p(x_i|\delta_i, \Theta) + \log p(\delta_i|\Theta)$$

Sometimes people call the above likelihood $F(\Theta, \delta)$, since we don't know the δ_i values and they need to be estimated. Therefore we can't compute the about likelihood. But we can assume a parametric form for distribution of δ_i, given an estimate to parameters Θ^n (at the n-th step), which we denote with $p(\delta|\Theta^n, X)$. To remove the random variable δ from the full-model likelihood we marginalize over δ:

$$Q(\Theta; \Theta^n) = \mathbb{E}_{p(\delta|\Theta,X)}\left[F(\Theta, \delta)\right]$$

Now the EM updates are the followings:

$$\begin{cases} \text{E: Estimate the distribution of } p(\delta|\Theta^n, X), \text{ and find } Q(\Theta; \Theta^n). \\ \text{M: Find } \Theta^{n+1} := \arg\max_\Theta Q(\Theta; \Theta^n) \end{cases}$$

Let's simplify this. We know:

$$\begin{aligned} \mathcal{L}(\Theta) &= \sum_i \left[\log p(x_i|\delta_i, \Theta) + \log p(\delta_i|\Theta)\right] && (1.3) \\ &= \sum_i \left[-\delta_i g_1(x_i; \theta_1) - (1-\delta_i) g_2(x_i; \theta_2) + \delta_i \ln \mu_1 + (1-\delta_i)\ln(1-\mu_1)\right] && (1.4) \end{aligned}$$

Now we need to estimate $p(\delta|\Theta^n, X)$. Based on its definition we have:

$$\begin{aligned} p(\delta_i = 1|\Theta^n, x_i) &= \frac{p(\delta_i = 1, x_i|\Theta^n)}{p(x_i, \delta_i = 1|\Theta^n) + p(x_i, \delta_i = 0|\Theta^n)} \\ &= \frac{p(x_i|\delta_i = 1, \Theta^n)p(\delta_i = 1|\Theta^n)}{p(x_i|\delta_i = 1, \Theta^n)p(\delta_i = 1|\Theta^n) + p(x_i|\delta_i = 0, \Theta^n)p(\delta_i = 0|\Theta^n)} \\ &= \frac{\mu_1^n e^{-g_1(x_i;\theta_1^n)}}{\mu_1^n e^{-g_1(x_i;\theta_1^n)} + (1-\mu_1^n)e^{-g_2(x_i;\theta_2^n)}} \end{aligned}$$

If you are not convinced that this is a good estimation for $p(\delta_i|\Theta^n, x_i)$, we an derive it in different way. Consider $F(\Theta, \delta)$. Whatever distribution we choose for δ it needs to maximize this function. Thus we take differentiation with respect to δ (Note that this differentiation is with respect to a function, which is called *functional derivative*).

$$\nabla_{\delta_i} F(\Theta^n, \delta) = 0$$

$$\begin{aligned} \Rightarrow \nabla_{\delta_i} F(\Theta^n, \delta) = &- \left[\log \delta_i - \log(1-\delta_i)\right] \\ &+ \left[-g_1(x_i; \theta_1^n) + \log \mu_1^n\right] \\ &+ \left[-g_2(x_i; \theta_2^n) + \log(1-\mu_1^n)\right] = 0 \end{aligned}$$

Which will result in the same distribution for $p(\delta_i|\Theta^n, x_i)$. Given this closed form estimation for $p(\delta_i|\Theta^n, x_i)$ it is easy to find $Q(\Theta; \Theta^n)$, by plugging it into Equation 1.3, and maximizing it with respect Θ. ■

Observation 1. *One interpretation of EM is coordinate-descent optimization, over latent variable coordinate, and the observation coordinates. In the previous example we solved the E-step with another optimization over the latent variables.*

■ **Example 1.2 — A simple Bayesian network(from Roth (2012)).** Assume that a set of 3-dimensional points (x, y, z) is generated according to the following probabilistic generative model over Boolean variables $X, Y, Z \in \{0, 1\}$:

$$Y \leftarrow X \rightarrow Z$$

1. What parameters from the table bellow will you need to estimate in order to completely define the model?

(1) `P(X=1)`	(2) `P(Y=1)`	(3) `P(Z=1)`	
(4) `P(X|Y=b)`	(5) `P(X|Z=b)`	(6) `P(Y|X=b)`	(7) `P(Y|Z=b)`
(8) `P(Z|X=b)`	(9) `P(Z|Y=b)`	(10) `P(X|Y=b,Z=c)`	(11) 3

   Answer: Based on the above generative model we could write the joint distribution as following:

$$p(X, Y, Z) = p(X).p(Y|X).p(Z|X).$$

   So we need to have (1), (6), (8). For this problem in order to find the whole joint distribution we need to know five parameters. For simplicity we denote the parameters using the following:

$$\begin{aligned} p(X = 1) &= \alpha \\ p(Y = 1|X = 1) &= a_1 \\ p(Y = 1|X = 0) &= a_2 \\ p(Z = 1|X = 1) &= b_1 \\ p(Z = 1|X = 0) &= b_2 \end{aligned}$$

   Then we have:

$$\begin{aligned} p(X = x) &= \alpha^x(1 - \alpha)^{1-x} \\ p(Y = y|X = 1) &= a_1^y(1 - a_1)^{1-y} \\ p(Y = y|X = 0) &= a_2^y(1 - a_2)^{1-y} \\ p(Z = z|X = 1) &= b_1^z(1 - b_1)^{1-z} \\ p(Z = z|X = 0) &= b_2^z(1 - b_2)^{1-z} \end{aligned}$$

2. You are given a sample of m data points sampled independently at random. However, when the observations are given to you, the value of X is always omitted. Hence, you get to see $\{(y^1, z^1), \ldots, (y^m, z^m)\}$. In order to estimate the parameters you identified in part (a), in the course of this

question you will derive update rules for them via the EM algorithm for the given model.
Express $\Pr(y^j, z^j)$ for an observed sample (y^j, z^j) in terms of the unknown parameters.
Answer: We can use the joint distribution and integrate out the unseen variables:

$$\begin{aligned} Pr(y^j, z^j) &= \sum_i Pr(X = x^i, y^j, z^j) \\ &= \sum_i p(X = x^i).p(Y = y^j|X = x^i).p(Z = z^j|X = x^i) \end{aligned}$$

We can replace each term with its own parameter denoted in the previous part. -Let $p_i^j = Pr(X{=}i|y^j, z^j)$ be the probability that hidden variable X has the value $i \in \{0, 1\}$ for an observation $(y^j, z^j), j \in \{1, \ldots, m\}$. Express p_i^j in terms of the unknown parameters.
Answer: Using the Bayes law we could express the probability like this:

$$\begin{aligned} p_i^j = Pr(X{=}i|y^j, z^j) &= Pr(y^j, z^j|X{=}i).Pr(y^j, z^j)/Pr(X{=}i) \\ &= Pr(y^j|X{=}i).Pr(z^j|X{=}i).Pr(y^j, z^j)/Pr(X{=}i) \end{aligned}$$

Each of the terms are previously calculated.

3. Let (x^j, y^j, z^j) represent the completed j^{th} example, $j \in \{1, \ldots, m\}$. Derive an expression for the expected log likelihood (LL) of the completed data set $\{(x^j, y^j, z^j)\}_{j=1}^m$, given the parameters in (a).
Answer:
$LL = \sum_j \log p(x^j, y^j, z^j) = \sum_j \log p(X = x^j) + \sum_j \log p(Y = y^j|X = x^j) + \sum_j \log p(Z = z^j|X = x^j)$ - Maximize LL, and determine update rules for *any two* unknown parameters of your choice (from those you identified in part (a)).
Answer: Because we don't haven't seen the latent variable values x we cannot explicitly plug in their values into the log-likelihood. But instead we need to takes its expectation with respect to the variable. This corresponds to the E-step in EM algorithm:

$$Q(y, z) = \sum_x p(x|y, z) \log p(x, y, z)$$

In the next step we shall maximize the expected likelihood with respect to the parameters:

$$\theta = \arg\max_\theta Q(y, z)$$

We can calculate the expectation of the whole-data likelihood as follows:

$$Q = \mathbb{E}\left[\sum_j \log p(x^j, y^j, z^j)\right] = \sum_j \mathbb{E}\left[\log p(x^j, y^j, z^j)\right] = \sum_j [p_1^j A + p_0^j B]$$

Where,

$$A = y^j \log b_1 + (1 - y^j) \log(1 - b_1) + z^j \log a_1 + (1 - z^j) \log(1 - a_1) + \log \alpha$$

$$B = y^j \log b_2 + (1 - y^j) \log(1 - b_2) + z^j \log a_2 + (1 - z^j) \log(1 - a_2) + \log(1 - \alpha)$$

To maximize the above expression, we must take derivitive with respect to the parameters:

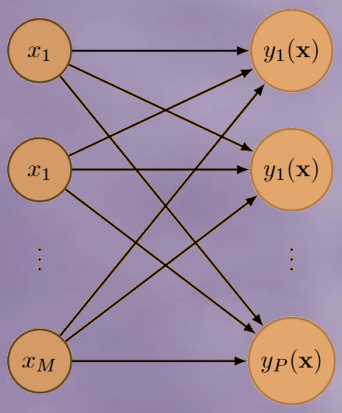
$$\frac{\partial Q}{\partial \alpha} = \sum_j p_1^j \frac{1}{\alpha} - \sum_j p_0^j \frac{1}{1 - \alpha} = 0 \Rightarrow \alpha = \frac{\sum_j p_1^j}{\sum_j p_1^j + \sum_j p_0^j}$$

$$\frac{\partial Q}{\partial b_1} = \sum_j p_1^j y^j \frac{1}{b_1} - \sum_j p_1^j (1 - y^j) \frac{1}{1 - b_1} = 0 \Rightarrow b_1 = \frac{\sum_j p_1^j y^j}{\sum_j p_1^j}$$

■

1.3 Bibliographical notes

Some examples are from David Forsyth's optimization class at UIUC.



Bibliography

Dan Roth. Cs546 at uiuc: Machine learning for nlp, homework. 2012. URL http://l2r.cs.uiuc.edu/~danr/Teaching/CS546-13/.