Expectation-maximization

The **Expectation-maximization algorithm** (EM algorithm) is used to find a maximum likelihood estimate for the parameters θ of a model when we have only observed some of the variables. That is, we have observed data $X = x$ (i.e. a set of random variables X taking on values x) and not observed the values of some set of latent variables $Z = z$. That is, our full model is characterized by some joint density/mass function $p(x, z\,;\theta)$ where θ parameterizes the model. Since we have not observed z, our likelihood function must marginalize over z:

$$\begin{aligned} l(\theta) &:= \log p(x;\theta) \\ &= \log \int p(x,z;\theta)\, dz \end{aligned}$$

In practice, maximizing this function over θ may be difficult to do analytically due to this integral. For such situations, the EM algorithm may provide a method for computing a local maximum of this function with respect to θ.

Description

The EM algorithm alternates between two steps: an expectation-step (E-step) and a maximization-step (M-step). Throughout execution, the algorithm maintains an estimate of the parameters θ that is updated on each iteration. As the algorithm progresses, this estimate of θ converges to a local maximum of $l(\theta)$. Let t denote a given step of the algorithm. On the tth step, we'll denote θ_t as the tth estimate of θ.

Each step works as follows: On the tth E-step, the algorithm formulates a function of θ called the Q-function, denoted $Q_t(\theta)$. Then, on the next M-step, the algorithm computes θ_{t+1} as the value of θ that maximizes $Q_t(\theta)$. This alternation between formulating a Q-function and maximizing that Q-function are repeated until the estimate of θ converges. As we will prove later, not only is this process guaranteed to converge, but it will converge to a local maximum of $l(\theta)$.

The details of the algorithm are as follows: We begin by initializing our iteration-counter to $t := 0$ and refer to our current estimate of θ as θ_t. We set θ_0 to an arbitrary value. EM then iterates between the following two steps until the estimate of θ converges:

1. **E-Step:**

   Compute the conditional probability $p(z \mid x\,;\theta_t)$. From this calculation, formulate

the Q-function

$$\begin{aligned} Q_t(\theta) &:= E_{Z|x;\theta_t}\left[\log p(x, z\,;\theta)\right] \\ &= \int p(z \mid x\,;\theta_t) \log p(x, z\,;\theta)\, dz \end{aligned}$$

The brunt of the computation in this step is usually in the computation of $p(z \mid x\,;\theta_t)$.

2. **M-Step:**

   Choose θ_{t+1} such that it maximizes the Q-function that was formulated in the E-Step:

$$\theta_{t+1} := \operatorname*{argmax}_{\theta} Q_t(\theta)$$

The EM algorithm as coordinate ascent on the evidence lower bound

We will show that the EM algorithm can be understood as a **coordinate ascent** algorithm for maximizing a lower bound of $p(x\,;\theta)$.

Recall, that for some distribution $q(z)$ over the latent data, the marginal log-likelihood $p(x\,;\theta)$ can be decomposed into the KL-divergence between $q(z)$ and $p(x \mid z\,;\theta)$ and the evidence lower bound (ELBO):

$$\begin{aligned} \log p(x\,;\theta) &= KL(q(z) \,||\, p(z \mid x\,;\theta)) + \text{ELBO} \\ &= KL(q(z) \,||\, p(z \mid x)) + E_{Z\sim q}\left[\log \frac{p(Z, x\,;\theta)}{q(Z)}\right] \end{aligned}$$

We can write the ELBO as a function of two objects: the parameters θ and the distribution q

$$F(q, \theta) = E_{Z\sim q}\left[\log \frac{p(Z, x\,;\theta)}{q(Z)}\right]$$

In order to maximize $\log p(x\,;\theta)$, our strategy will be to maximize $F(q, \theta)$. One approach is to use an optimization technique known as **coordinate ascent**. Given a function $f(a, b)$ that takes two coordinates a and b, the idea of coordinate ascent is that we will iteratively fix a, then choose b that maximizes this function. We then fix b and choose a that maximizes the function. These two steps are repeated until convergence. We will show that performing coordinate ascent on $F(q, \theta)$ yields the EM algorithm.

At iteration t let θ_t be the current value for θ. Then, given θ_t, we will find $q(z)$ that maximizes $F(q, \theta)$. That is we solve for

$$q_t := \operatorname*{argmax}_{q} F(q, \theta_t)$$

As Theorem 1 shows, the solution for q_t to this optimization problem is $q_t := p(z \mid x\,; \theta_t)$, which is precisely the distribution that we compute in the E-step. To formulate the next step in coordinate ascent, we hold q_t fixed to form the function $F(q_t, \theta)$. As shown in Theorem 2, finding θ that maximizes $F(q_t, \theta)$ is equivalent to finding θ that maximizes the Q-function $Q_t(\theta)$. This is precisely the M-step. That is,

$$\begin{aligned}\theta_{t+1} &:= \operatorname*{argmax}_{\theta} F(q_t, \theta) \\ &= \operatorname*{argmax}_{\theta} Q_t(\theta)\end{aligned}$$

Theorem 1 *Given the ELBO $F(q, \theta)$ for the marginal likelihood $\log p(x\,; \theta)$ of a model $p(x, z\,; \theta)$ where $X = x$ is the observed data, $Z = z$ is the latent data, θ are the parameters, and q is a distribution of Z, if we fix θ at θ_t, then*

$$\operatorname*{argmax}_{q} F(q, \theta_t) = p(z \mid x\,; \theta_t)$$

Proof:

$$\begin{aligned}\operatorname*{argmax}_{q} F(q, \theta_t) &:= \operatorname*{argmax}_{q} E_{Z\sim q}\left[\log \frac{p(Z, x\,; \theta_t)}{q(Z)}\right] \\ &= \operatorname*{argmax}_{q} E_{Z\sim q}\left[\log \frac{p(Z \mid x\,; \theta_t)p(x\,; \theta_t)}{q(Z)}\right] \\ &= \operatorname*{argmax}_{q} \left(E_{Z\sim q}\left[\log \frac{p(Z \mid x\,; \theta_t)}{q(Z)}\right] + E_{Z\sim q}\left[\log p(x\,; \theta_t)\right]\right) \\ &= \operatorname*{argmax}_{q} \left(-KL(q(z) \mid p(z \mid x\,; \theta_t)) + \log p(x\,; \theta_t)\right) \\ &= \operatorname*{argmax}_{q} -KL(q(z) \mid p(z \mid x\,; \theta_t)) \\ &= p(z \mid x\,; \theta_t)\end{aligned}$$

□

Theorem 2 *Given the ELBO $F(q, \theta)$ for the marginal likelihood $\log p(x\,; \theta)$ of a model $p(x, z\,; \theta)$ where $X = x$ is the observed data, $Z = z$ is the latent data, θ are the parameters, and q is a distribution of Z, if we fix q at $q_t := p(z \mid x\,; \theta_t)$, then*

$$\operatorname*{argmax}_{\theta} F(q_t, \theta) = \operatorname*{argmax}_{\theta} Q_t(\theta)$$

where $Q_t(\theta)$ is the Q-function.

Proof:

$$\begin{aligned}
\underset{\theta}{\text{argmax}}\, F(q_t, \theta) &:= \underset{\theta}{\text{argmax}} E_{Z\sim q_t}\left[\log \frac{p(Z, x\,;\theta)}{q_t(Z)}\right] \\
&= \underset{\theta}{\text{argmax}}\; E_{Z|x;\theta}\left[\log \frac{p(Z, x\,;\theta)}{p(Z \mid x\,;\theta_t)}\right] \\
&= \underset{\theta}{\text{argmax}}\; \left(E_{Z|x;\theta}\left[\log p(Z, x\,;\theta)\right] - E_{Z|x;\theta}\left[p(Z \mid x\,;\theta_t)\right]\right) \\
&= \underset{\theta}{\text{argmax}}\; E_{Z|x;\theta}\left[\log p(Z, x\,;\theta)\right] \\
&= \underset{\theta}{\text{argmax}}\; Q_t(\theta)
\end{aligned}$$

□

Convergence to a local maximum of the likelihood function

In the previous section, we showed that the EM algorithm maximizes a lower bound of the evidence; however, we have not shown that by doing so, the EM algorithm will converge to a local maximum of the likelihood function. We can show that this is indeed the case, by realizing a few facts:

1. At iteration t, the function F touches the likelihood function when evaluated at θ_t. That is,

   $$F(q_t, \theta) = l(\theta_t)$$

   which follows from the decomposition of the likelihood into the KL-divergence and ELBO evaluated at θ_t:

   $$\begin{aligned}
   l(\theta) &:= KL(q_t(z) \,||\, p(z \mid x\,;\theta_t)) + F(q_t, \theta_t) \\
   &= KL(p(z \mid x\,;\theta_t) \,||\, p(z \mid x\,;\theta)) + F(q_t, \theta) \\
   \implies l(\theta_t) &= KL(p(z \mid x\,;\theta_t) \,||\, p(z \mid x\,;\theta_t)) + F(q_t, \theta_t) \\
   &= F(q_t, \theta_t) \qquad \text{The KL-divergence term is zero}
   \end{aligned}$$

2. With q fixed at q_t, the F function at θ_{t+1} is greater than at θ_t. That is,

   $$F(q_t, \theta_{t+1}) >= F(q_t, \theta_t)$$

   which follows from the fact that $\theta_{t+1} := \underset{\theta}{\text{argmax}}\, F(q_t, \theta)$.

3. F is bounded above by the likelihood function as shown by the derivation of the ELBO

With these facts we can reason that the EM algorithm converges on a local maximum of $l(\theta)$ by "crawling" up the likelihood surface. That is, on each iteration t of the algorithm, $F(q_t, \theta)$ lies at or below the likelihood surface, but touches it at θ_t (Figure 1). That is, $F(q_t, \theta_t) = l(\theta_t)$. Then, since $F(q_t, \theta_{t+1}) > F(q_t, \theta_t)$, it follows that θ_{t+1} increases the likelihood function. Furthermore, on the next iteration, we can evaluate the likelihood function at θ_{t+1} via $F(q_{t+1}, \theta_{t+1})$. Thus, each next iteration's value of $F(q_{t+1}, \theta_{t+1})$ is a new value on the likelihood surface which is greater than the current iteration's value for $F(q_t, \theta_t)$. This process is visualized in Figure 2.

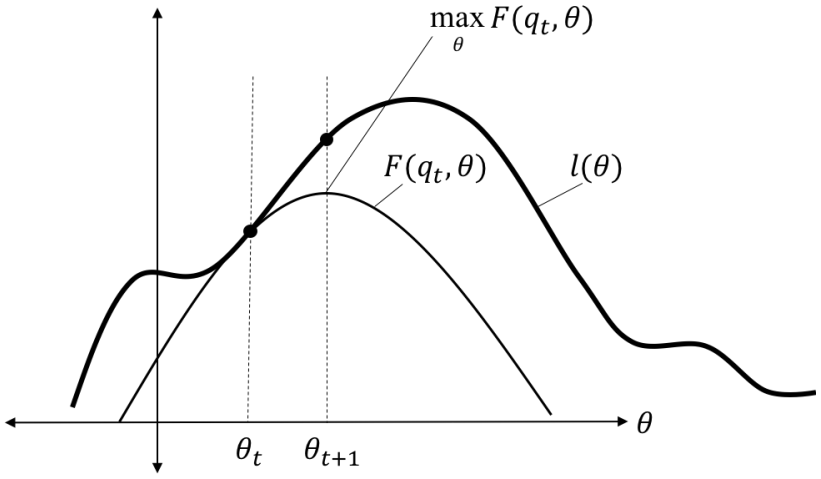

Figure 1: The F function is bounded above by $l(\theta)$ and equals the likelihood function at θ_t.

Intuition

When to use the EM algorithm

The EM algorithm is a good choice for performing maximum likelihood estimation when the log-likelihood function $l(\theta)$ is difficult to maximize over θ, but the Q-function is easier. The likelihood function is often difficult to maximize due to the presence of the log of an integral:

$$l(\theta) = \log \int p(x, z; \theta)\, dz$$

That is, taking the derivative/gradient of this function with respect to θ, setting it to zero, and then solving for θ is often very difficult to do analytically. If the integral and

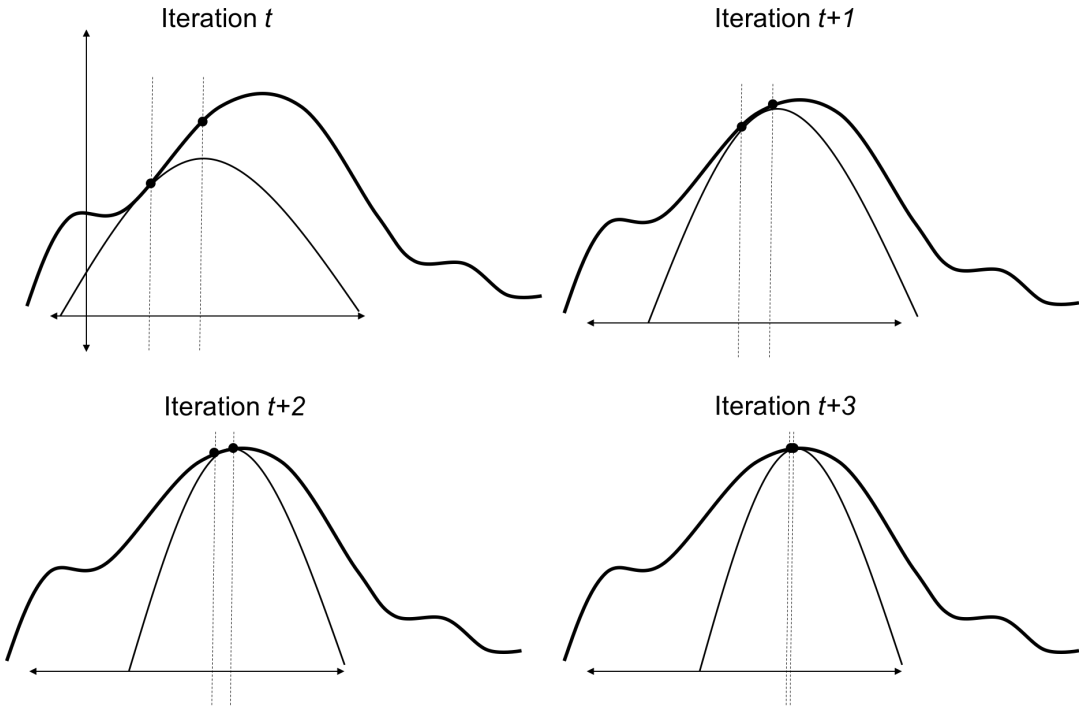

Figure 2: Each formulation of the F function's maximum is closer to a local optimum on the likelihood surface than the previous formulation's. The algorithm runs until convergence.

the log were flipped, this function might be much easier to maximize analytically. The Q-function is just such a function

$$Q_t(\theta) := \int f(z) \log p(x, z\,; \theta)\, dz$$

where $f(z) := p(z \mid x\,; \theta_t)$. Thus, we see that the EM algorithm converts a difficult optimization problem to a series of easier optimization problems where each sub-problem is made easier by flipping the integral and log in the objective function.

Perspectives on the Q-function

In the case in which Z is a discrete random variable, the Q-function can be interpreted as a "generalized" version of the complete data likelihood function. First, let's say that both X and Z are observed where $X = x$ and $Z = z$. Then, we could write the complete data likelihood as

$$\log p(x, z; \theta) := \sum_{z'} \mathbb{1}_{(z=z')} \log p(x, z'\,; \theta)$$

Now, recall the Q-function is

$$Q(\theta) := \sum_{z'} p(z' \mid x\,; \theta) \log p(x, z'\,; \theta)$$

We note that the difference between these two functions is that the indicator variable in the complete data likelihood $\mathbb{1}_{(z=z')}$ is replaced by the probability $p(z \mid x\,; \theta)$. Thus, we can view the Q-function as a sort of generalization of the complete data likelihood. That is, $p(z \mid x\,; \theta)$ acts as a weight for its corresponding term in the summation and measures our current certainty that the hidden data is z'. When we know $Z = z$, all of the weight is assigned to the term in which $z' = z$ and no weight is assigned to the terms in which $z' \neq z$.

Another way to view the Q-function is as a complete data likelihood function over a hypothetical dataset where this hypothetical dataset is generated from a distribution that depends on our current best guess of θ (i.e. θ_t at the tth iteration of the EM algorithm). That is, let us assume that we generate a very large set of values for Z from the distribution specified by $p(z \mid x\,; \theta_t)$. That is, assume we generate

$$z'_1, z'_2, \ldots, z'_N \overset{i.i.d.}{\sim} p(z \mid x\,; \theta_t)$$

For each z'_i, we create a new "sub"-dataset (x, z'_i). Note, that the observed data x is duplicated in each of these sub-datasets. We then merge all of the datasets $(x, z'_1), \ldots, (x, z'_n)$ to form our new hypothetical dataset. Then, the likelihood over these data is

$$l'(\theta) := \prod_{i=1}^{N} p(x, z'_i\,; \theta)$$

Now we will show that maximizing the Q-function is equivalent to maximizing this likelihood function over the hypothetical data. Let $Z' := \{z'_1, z'_2, \ldots, z'_N\}$. For each $z'_i \in Z'$, we will form a complete data set (z'_i, x). Now, let $c_{Z'}(z) := |\{z'_i \in Z' \mid z'_i = z\}$. That is $c_{Z'}(z)$ is the number of items in Z' that equal z. If N is very large, we would expect

$$\frac{c_{Z'}(z)}{\sum_{z^* \in \mathcal{Z}} c_{Z'}(z^*)} = p(z \mid x\,; \theta_t)$$

where $\mathcal{Z}$ is the support of Z's probability mass function. Then we can formulate our optimization problem as

$$\begin{aligned}
\operatorname*{argmax}_\theta l'(\theta) &:= \prod_{i=1}^{N} p(x, z'_i\,; \theta) \\
&= \operatorname*{argmax}_\theta \prod_{z \in \mathcal{Z}} p(x, z\,; \theta)^{c_{Z'}(z)} \\
&= \operatorname*{argmax}_\theta \sum_{z \in \mathcal{Z}} c_{Z'}(z) \log p(x, z\,; \theta) \\
&= \operatorname*{argmax}_\theta \frac{1}{\sum_{z^* \in \mathcal{Z}} c_{Z'}(z^*)} \sum_{z \in \mathcal{Z}} c_{Z'}(z) \log p(x, z\,; \theta) \\
&= \operatorname*{argmax}_\theta \sum_{z \in \mathcal{Z}} p(z \mid x\,; \theta_t) \log p(x, z\,; \theta) \\
&= \operatorname*{argmax}_\theta Q_t(\theta)
\end{aligned}$$

In this light, we can view the EM algorithm as an iterative process, in which we posit a distribution over the hidden data $p(z \mid x\,; \theta_t)$, then generate a hypothetical data set on which we perform maximum likelihood estimation. This new estimate of the parameters, as estimated from the hypothetical dataset, is then used in the next iteration to generate another data set. This process is repeated until the estimates of the parameters converge.