Expectation maximization

Herman Kamper

2024-01,

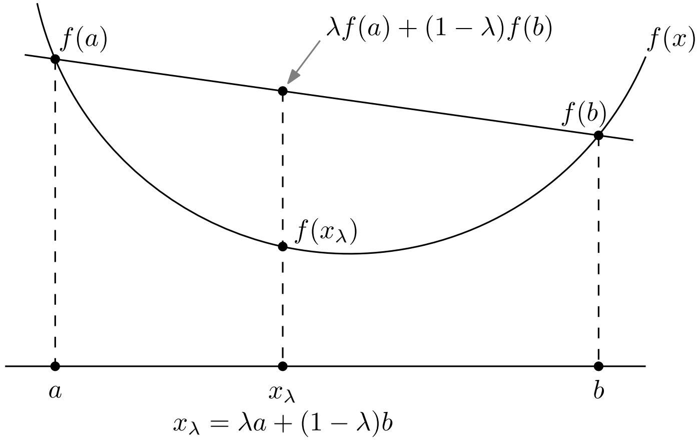
Preliminaries

Jensen's inequality

A function $f(x)$ is convex over (a, b) if every chord lies on or above the function:

$$f\left(\lambda a + (1-\lambda)b\right) \leq \lambda f(a) + (1-\lambda) f(b)$$

with $0 \leq \lambda \leq 1$.


$f(a)$
$\lambda f(a) + (1-\lambda) f(b)$
$f(x)$
$f(b)$
$f(x_\lambda)$
a
x_λ
b
$x_\lambda = \lambda a + (1-\lambda)b$

Jensen's inequality generalizes this definition. For any convex function $f(x)$:

$$f(\sum_{k=1}^{K} \lambda_k x_k) \leq \sum_{k=1}^{K} \lambda_k f(x_k)$$

with $\lambda_k \geq 1$ and $\sum_{k=1}^{K} \lambda_k = 1$.

Easy to see for $K = 2$: Just the definition of a convex function.

For $K > 2$: Proved by induction (Wikipedia).

Kullback-Leibler divergence

The Kullback-Leibler (KL) divergence gives a measure of how different one probability distribution is compared to another reference distribution.

For discrete distributions with probability mass functions P and Q over outcomes $k = 1, 2, \ldots, K$, the KL divergence is defined as

$$D_{\mathrm{KL}}(P \| Q) \triangleq \sum_{k=1}^{K} P(x = k) \log \frac{Q(x = k)}{P(x = k)}$$

For continuous distributions with probability density functions p and q, the KL divergence is

$$D_{\mathrm{KL}}(p \| q) \triangleq \int_x p(x) \log \frac{q(x)}{p(x)} \, \mathrm{d}x$$

Using Jensen's inequality, we can show that

$$D_{\mathrm{KL}}(p \| q) \geq 0$$

with it being zero only if the distributions are identical.

The KL divergence isn't strictly a distance:

$$D_{\mathrm{KL}}(p \| q) \neq D_{\mathrm{KL}}(q \| p)$$

Expectation maximization

We have a model where each observed variable $\mathbf{x}^{(n)}$ depends on a hidden variable $\mathbf{z}^{(n)}$. We want to maximize the log likelihood of the parameters:[1]

$$\begin{aligned} L(\boldsymbol{\theta}) &= \sum_{n=1}^{N} \log p_{\boldsymbol{\theta}}(\mathbf{x}^{(n)}) \\ &= \sum_{n=1}^{N} \log \left[\sum_{\mathbf{z}^{(n)}} p_{\boldsymbol{\theta}}(\mathbf{x}^{(n)}, \mathbf{z}^{(n)}) \right] \end{aligned}$$

The log of a sum cannot be pushed into the sum (as is the case with the log of a product). So we typically can't find a closed-form solution for the parameters by setting $\frac{\partial L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = 0$.

Expectation maximization is an iterative procedure that gets around this problem.

Let's first consider a dataset with a single item $\mathbf{x}^{(n)}$. We use a helper distribution $Q(\mathbf{z})$, which can be any distribution (for now):

$$\begin{aligned} \log p_{\boldsymbol{\theta}}(\mathbf{x}^{(n)}) &= \log \left[\sum_{\mathbf{z}^{(n)}} p_{\boldsymbol{\theta}}(\mathbf{x}^{(n)}, \mathbf{z}^{(n)}) \right] \\ &= \log \left[\sum_{\mathbf{z}^{(n)}} Q(\mathbf{z}^{(n)}) \frac{p_{\boldsymbol{\theta}}(\mathbf{x}^{(n)}, \mathbf{z}^{(n)})}{Q(\mathbf{z}^{(n)})} \right] \end{aligned}$$

Since log is concave, we have from Jensen's inequality:

$$\begin{aligned} \log p_{\boldsymbol{\theta}}(\mathbf{x}^{(n)}) &= \log \left[\sum_{\mathbf{z}^{(n)}} Q(\mathbf{z}^{(n)}) \frac{p_{\boldsymbol{\theta}}(\mathbf{x}^{(n)}, \mathbf{z}^{(n)})}{Q(\mathbf{z}^{(n)})} \right] \\ &\geq \sum_{\mathbf{z}^{(n)}} Q(\mathbf{z}^{(n)}) \log \frac{p_{\boldsymbol{\theta}}(\mathbf{x}^{(n)}, \mathbf{z}^{(n)})}{Q(\mathbf{z}^{(n)})} = J(Q, \boldsymbol{\theta}) \end{aligned}$$

[1] For notational convenience, we assume that $\mathbf{z}$ is discrete and $\mathbf{x}$ continuous.

J is called the evidence lower bound (ELBO):[2] [3]

$$\boxed{\begin{aligned} J(Q, \boldsymbol{\theta}) &= \sum_{\mathbf{z}^{(n)}} Q(\mathbf{z}^{(n)}) \log p_{\boldsymbol{\theta}}(\mathbf{x}^{(n)}, \mathbf{z}^{(n)}) - \sum_{\mathbf{z}^{(n)}} Q(\mathbf{z}^{(n)}) \log Q(\mathbf{z}^{(n)}) \\ &= \mathbb{E}_Q \left[\log p_{\boldsymbol{\theta}}(\mathbf{x}^{(n)} | \mathbf{z}^{(n)})\right] + \mathbb{E}_Q \left[\log p_{\boldsymbol{\theta}}(\mathbf{z}^{(n)})\right] - \mathbb{E}_Q \left[\log Q(\mathbf{z}^{(n)})\right] \end{aligned}}$$

Importantly, the ELBO gives a lower bound for the log likelihood:

$$\log p_{\boldsymbol{\theta}}(\mathbf{x}^{(n)}) \geq J(Q, \boldsymbol{\theta})$$

for *any* choice of Q. (It might be a very crappy lower bound, but it would be a lower bound nevertheless.)

We can choose a Q that makes it easy to maximize $J(Q, \boldsymbol{\theta})$ in terms of $\boldsymbol{\theta}$, and then hope that we thereby push up the log likelihood. But there's no guarantee that we will be improving the log likelihood. Except if we are clever in our choice of Q!

The ELBO can also be written as (confirm this for yourself):[4]

$$J(Q, \boldsymbol{\theta}) = -D_{\text{KL}} \left(Q(\mathbf{z}^{(n)}) \parallel P_{\boldsymbol{\theta}}(\mathbf{z}^{(n)} | \mathbf{x}^{(n)})\right) + \log p_{\boldsymbol{\theta}}(\mathbf{x}^{(n)})$$

Expectation step

Let's choose the Q that gives the tightest possible lower bound. Since $D_{\text{KL}} \geq 0$, to get the largest $J(Q, \boldsymbol{\theta})$, we choose

$$Q(\mathbf{z}^{(n)}) = P_{\boldsymbol{\theta}}(\mathbf{z}^{(n)} | \mathbf{x}^{(n)})$$

This could work, but we actually still don't know the optimal value of $\boldsymbol{\theta}$. So instead we use $\boldsymbol{\theta}^{(m)}$, our guess for the optimal $\boldsymbol{\theta}$ at iteration m. We denote this choice of Q as

$$Q_{\boldsymbol{\theta}^{(m)}}(\mathbf{z}^{(n)}) = P_{\boldsymbol{\theta}^{(m)}}(\mathbf{z}^{(n)} | \mathbf{x}^{(n)})$$

[2] Note that in other literature the ELBO is sometimes denoted as Q, but we use Q here for the helper function. These two things are very different.

[3] Also note that J is a functional: It takes another function as an argument.

[4] The first term is a compression term that encourages a compact latent representation while the second is a reconstruction term (showing a trade-off between compression and reconstruction).

Note that if we do this, then for the current $\boldsymbol{\theta}^{(m)}$, the ELBO is equal to the log likelihood:

$$J(Q_{\boldsymbol{\theta}^{(m)}}, \boldsymbol{\theta}^{(m)}) = \log p_{\boldsymbol{\theta}^{(m)}}(\mathbf{x}^{(n)})$$

Maximization step

We now keep Q fixed and maximize the ELBO in terms of $\boldsymbol{\theta}$. At iteration $m+1$:

$$\boldsymbol{\theta}^{(m+1)} = \arg\max_{\boldsymbol{\theta}} J(Q_{\boldsymbol{\theta}^{(m)}}, \boldsymbol{\theta})$$

Recall the first form of the ELBO:

$$J(Q, \boldsymbol{\theta}) = \sum_{\mathbf{z}^{(n)}} Q(\mathbf{z}^{(n)}) \log p_{\boldsymbol{\theta}}(\mathbf{x}^{(n)}, \mathbf{z}^{(n)}) - \sum_{\mathbf{z}^{(n)}} Q(\mathbf{z}^{(n)}) \log Q(\mathbf{z}^{(n)})$$

The second term doesn't depend on $\boldsymbol{\theta}$, so

$$\begin{aligned}\boldsymbol{\theta}^{(m+1)} &= \arg\max_{\boldsymbol{\theta}} \sum_{\mathbf{z}^{(n)}} Q_{\boldsymbol{\theta}^{(m)}}(\mathbf{z}^{(n)}) \log p_{\boldsymbol{\theta}}(\mathbf{x}^{(n)}, \mathbf{z}^{(n)}) \\ &= \arg\max_{\boldsymbol{\theta}} \sum_{\mathbf{z}^{(n)}} P_{\boldsymbol{\theta}^{(m)}}(\mathbf{z}^{(n)} | \mathbf{x}^{(n)}) \log p_{\boldsymbol{\theta}}(\mathbf{x}^{(n)}, \mathbf{z}^{(n)})\end{aligned}$$

Up to now we have been looking at a dataset with a single item, but the steps can be repeated when we have N items. We actually then have to choose a $Q^{(n)}(\mathbf{z}^{(n)})$ separately for each item to get a per-item term $J^{(n)}(Q^{(n)}, \boldsymbol{\theta})$. The sum of these terms give a lower bound on the log likelihood of the parameters:

$$\sum_{n=1}^{N} \log p_{\boldsymbol{\theta}}(\mathbf{x}^{(n)}) \geq \sum_{n=1}^{N} J^{(n)}(Q^{(n)}, \boldsymbol{\theta})$$

or more concisely:

$$L(\boldsymbol{\theta}) \geq J(Q, \boldsymbol{\theta})$$

with Q now being a set of distributions.

The EM algorithm

- E-step:

$$Q^{(n)}_{\boldsymbol{\theta}^{(m)}}(\mathbf{z}^{(n)}) = P_{\boldsymbol{\theta}^{(m)}}(\mathbf{z}^{(n)}|\mathbf{x}^{(n)})$$

$$\text{for } n = 1, 2, \ldots, N$$

- M-step:

$$\begin{aligned}
\boldsymbol{\theta}^{(m+1)} &= \arg\max_{\boldsymbol{\theta}} J(Q_{\boldsymbol{\theta}^{(m)}}, \boldsymbol{\theta}) \\
&= \arg\max_{\boldsymbol{\theta}} \sum_{n=1}^{N} J^{(n)}(Q^{(n)}_{\boldsymbol{\theta}^{(m)}}, \boldsymbol{\theta}) \\
&= \arg\max_{\boldsymbol{\theta}} \sum_{n=1}^{N} \sum_{\mathbf{z}^{(n)}} P_{\boldsymbol{\theta}^{(m)}}(\mathbf{z}^{(n)}|\mathbf{x}^{(n)}) \log p_{\boldsymbol{\theta}}(\mathbf{x}^{(n)}, \mathbf{z}^{(n)})
\end{aligned}$$

Why does EM converge?

I said earlier that just pushing up the lower bound is not guaranteed to push up the log likelihood. But in EM it is actually guaranteed!

This is because after the E-step, the bound is tight:

$$L(\boldsymbol{\theta}^{(m)}) = \sum_{n=1}^{N} \log p_{\boldsymbol{\theta}^{(m)}}(\mathbf{x}^{(n)}) = J(Q_{\boldsymbol{\theta}^{(m)}}, \boldsymbol{\theta}^{(m)})$$

So the ELBO J is always below the log likelihood L, but at $\boldsymbol{\theta}^{(m)}$ they are equal.

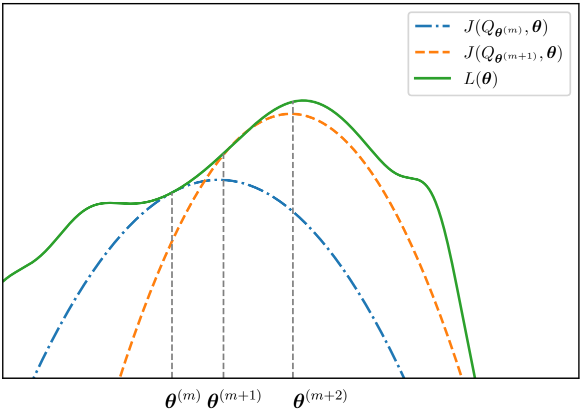
In the M-step, we then maximize J:[5]

- Guaranteed to give a $\boldsymbol{\theta}^{(m+1)}$ with a J greater than or equal to what it was before:

$$J(Q_{\boldsymbol{\theta}^{(m)}}, \boldsymbol{\theta}^{(m+1)}) \geq J(Q_{\boldsymbol{\theta}^{(m)}}, \boldsymbol{\theta}^{(m)})$$

- But because the bound was tight at $\boldsymbol{\theta}^{(m)}$, this means that

$$L(\boldsymbol{\theta}^{(m+1)}) \geq L(\boldsymbol{\theta}^{(m)})$$



[5] Figure reproduced from (Murphy, 2012, Fig. 11.15).

Example: EM for HMMs

The soft EM equations in the the HMM notes can be derived using the procedure above. The math gets quite hairy! But Tang (2021) gives a very complete example. You can check that the equations for the transition probabilities obtained by Tang (2021) matches those in the HMM note.

References

T. Ma and A. Ng, “The EM algorithm”, *Stanford University*, 2019.

K. P. Murphy, *Machine Learning: A Probabilistic Perspective*, 1st ed., 2012.

H. Tang, “Hidden markov models (part 2),” *University of Edinburgh*, 2021.