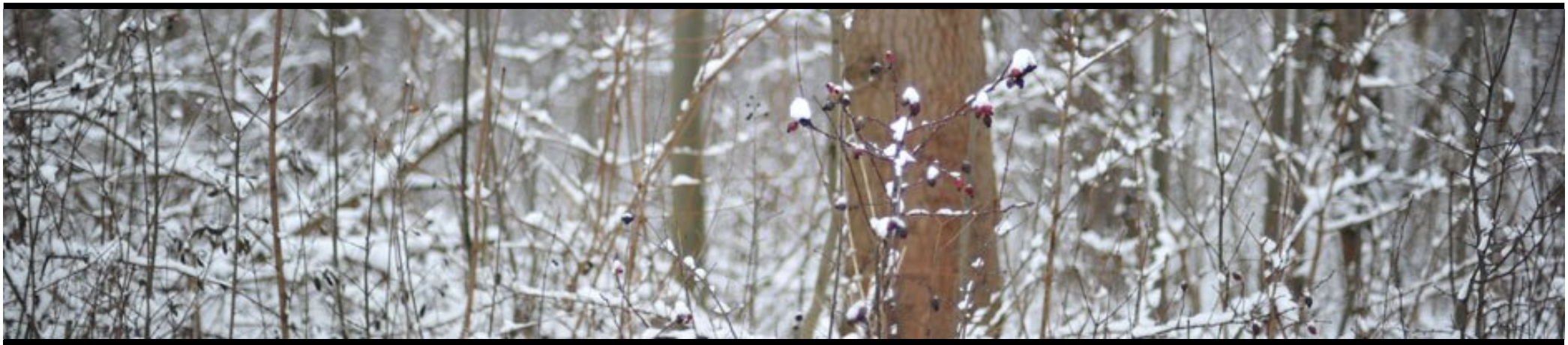
Justin Domke's Weblog



Expectation Maximization

Expectation-Maximization (EM) is a general technique for maximum likelihood learning in the presence of counfounding (aka hidden, nuisance) variables. Consider fiting a distribution $p(a,b)$ of two variables. However, we want to fit p to make $p(a)$ accurate, not the joint distribution $p(a,b)$. (There are several reasons we might want to do this– the most common is that we only have training data for a.) So, write down the log-likelihood for some data element $\hat{a}$.

Search

Archives
- September 2015
- December 2014
- February 2014
- January 2014
- September 2013
- September 2012
- January 2012
- November 2011

$$L(\hat{a}) = \log p(\hat{a})$$

$$= \log(\textstyle\sum_b p(\hat{a}, b))$$

$$= \log(\textstyle\sum_b r(b|\hat{a}) p(\hat{a}, b) / r(b|\hat{a}))$$

In the last step, $r(b|\hat{a})$ is any valid distribution over b. (The reason for introducing this will be clear shortly.)

Now, recall Jensen's inequality. For any distribution $q(x)$, and concave function $f(x)$,

$$f(\textstyle\sum_x q(x)x) \geq \sum_x q(x) f(x).$$

Applying this to the last expression we had for $L(\hat{x})$, we get a lower bound.

$$L(\hat{a}) \geq Q(\hat{a})$$

$$Q(\hat{a}) = \textstyle\sum_b r(b|\hat{a})(\log p(\hat{a}, b) - \log r(b|\hat{a}))$$

The basic idea of EM is very simple. Instead of directly trying to maximize L, instead maximize the lower bound, Q. This is accomplished in two steps.

- Maximize Q with respect to r. This is called the "Expectation" (E) step.
- Maximize Q with respect to p. This is called the "Maximization" (M) step.

One performs the M-step basically like normal (joint) maximum likelihood learning. That is, one fits p to maximize

- October 2011
- July 2011
- May 2011
- March 2011
- November 2009
- October 2009
- August 2009
- June 2009
- May 2009
- March 2009
- February 2009
- January 2009
- December 2008
- November 2008
- October 2008
- August 2008
- July 2008
- June 2008

Meta

- Register
- Log in

$\sum_{\hat{a}} \sum_b r(b|\hat{a}) \log p(\hat{a}, b),$

which is basically just weighted maximum likelihood learning.

The E-step needs a bit more explanation. (Why is it called the "expectation" step?) It can be shown, by setting up a Lagrange multiplier problem, that the optimal r will in fact be $r(b|\hat{a}) = p(b|\hat{a})$. Moreover, it is easy to show that, after substituting this value of r, $Q(\hat{a}) = L(\hat{a})$! Of course, once we start messing around with p in the M step while holding r constant, this will no longer be true, but we can see that at convergence, the bound will be tight. This, EM truly maximizes L, not an approximation to it.

Leave a Reply

Enter your comment here...