Assumed Density Filtering

Ananth Ranganathan

23rd November, 2004

In Assumed Density filtering (ADF), we choose a distribution that is easy for us to work with and project the true posterior after each measurement update onto this distribution family. The moments of the approximate posterior (that lies in the distribution family of our choice) are found by minimizing the KL-divergence between the true posterior and the distribution of our choice. In the end this comes down to moment matching. This process is derived below.

Since we desire an approximation to the posterior that is simple to handle, the approximating distribution will, most likely, be a distribution with a finite moment vector. This, in turn, means that a good choice for the approximating distribution is from the exponential family of distributions. Hence, we can write the approximating distribution as

$$q(x) \quad = \quad h(x)\exp\left(\phi^T(\theta)u(x)+f(x)+g(y)\right)$$

where $\phi(\theta)$ is the vector of natural parameters and $u(x)$ is the natural statistics vector. If $p(x)$ is the true posterior after a measurement update, to find the parameters of $q(x)$, we minimize the KL-divergence

$$KL(p||q) \quad = \quad \int_x p(x)\log\frac{p(x)}{q(x)}$$

Since we need to find the natural parameters of $q(x)$, we differentiate the KL-divergence wrt $\phi(\theta)$ and set the derivative to zero

$$\begin{aligned}\frac{d}{d\phi}KL(p||q) &= -\frac{d}{d\phi}\int_x \left(p(x)\phi^T(\theta)u(x)+p(x)g(\phi)\right)\\ &= -\int_x\left(p(x)u(x)+p(x)\frac{dg(\phi)}{d\phi}\right)=0 \qquad (1)\end{aligned}$$

Now, since $q(x)$ is a probability distribution

$$\int_x q(x) \quad = \quad 1$$

and differentiating this wrt $\phi(\theta)$, we get

$$\frac{d}{d\phi}\int_x q(x) \quad = \quad 0$$

$$\begin{aligned}\int_x q(x)\left(u(x)+\frac{dg(\phi)}{d\phi}\right) &= 0 \\ \langle u(x)\rangle_{q(x)} &= -\frac{dg(\phi)}{d\phi} \qquad (2)\end{aligned}$$

Plugging this result into (1), we get

$$\langle u(x)\rangle_{q(x)} = \int_x u(x)p(x)$$

and hence, the moments of the distribution $q(x)$ that minimize the KL-divergence to the true posterior are equal to the moments of the true posterior. Hence, assumed density filtering transforms to moment matching.

It can be seen from the above discussion that ADF requires that the moments of the true posterior be calculated relatively efficiently. If this is not the case, then ADF cannot be applied in a straight-forward manner. Note that one of the main differences between ADF and Variational methods is that ADF minimizes the KL-divergence according to the true posterior $KL(p||q)$, while Variational techniques (for eg., Variational Bayes EM) minimize the reverse KL-divergence $KL(q||p)$. The reverse KL-divergence is employed since it is assumed that the true posterior is intractable and hence, its moments cannot be calculated. Thus Variational techniques can be used in such scenarios where ADF may be inapplicable.

A specific example of ADF where a Gaussian mixture posterior is projected onto a single Gaussian is given in [2]. ADF is also the underlying principle in the Boyen-Koller algorithm, given in [1].

References

[1] X. Boyen and D. Koller. Tractable inference for complex stochastic processes. In *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence*, pages 33–42, 1998.

[2] T. Minka. *A family of algorithms for approximate Bayesian inference*. PhD thesis, MIT Media Lab, MIT, 2001.