CPSC 550: Machine Learning II **2008/9 Term 2**

Lecture 15 — Mar 17, 2009

Lecturer: Nando de Freitas *Scribe: Matt Hoffman*

15.1 Previous lectures

Previous lectures have presented Younes' *maximum likelihood* (ML) algorithm and *contrastive divergence* (CD) for classification using restricted Boltzmann machines (RBMs). The basic idea is that CD stays "close" to the data, while ML can simulate "everywhere" over the parameter space, so CD *might* be more efficient, although this is somewhat uncertain.

It is easier to understand the convergence properties of Younes' algorithm by viewing it as a stochastic approximation method. This lecture will introduce some of the basic underpinnings of stochastic approximation. Subsequent lectures will show how stochastic approximation applies to RBMs.

15.2 Examples of stochastic approximation

Stochastic approximation has a wide variety of uses:

1. Stochastic boosting and mirrored averaging (c.f. J.Friedman, A.Tsybakov)
2. Control and reinforcement learning:
   - Bellman's operator in dynamic programming
   - Stochastic policy gradient optimization
3. Sensor networks
4. Experimental design
5. Online expectation maximization (EM) algorithms

15.3 Fixed-point iterations

The idea behind stochastic approximation is to set up a fixed-point equation, $\mathbb{E}_{p(y|x)}\big[g(y,x)\big] = x$, whose solution, x, corresponds to the desired optimum. This can be solved by writing

$$\begin{aligned} x &= \mathbb{E}_{p(y|x)}\big[g(y,x)\big] = \mathbb{E}[g] \\ \gamma x &= \gamma \mathbb{E}[g] \\ x + \gamma x &= x + \gamma \mathbb{E}[g] \\ x &= (1-\gamma)x + \gamma \mathbb{E}[g]. \end{aligned}$$

If it is possible to sample from $p(y|x)$, then given some initial x_0, x can be solved for iteratively as

$$x_{t+1} = (1-\gamma_t)x_t + \gamma_t \frac{1}{N}\sum_{i=1}^{N} g\big(\widetilde{y}_t^{(i)}, x_t\big)$$

where $\{\widetilde{y}_t^{(i)}\}$ is a set of N samples taken from $p(y|x_t)$. The use of multiple samples can prove wasteful from a computational standpoint. Only a single sample is needed. The recursion can then be written as

$$x_{t+1} = (1-\gamma_t)x_t + \gamma_t g(\widetilde{y}_t, x_t)$$

where $\widetilde{y}_t$ is one sample from $p(y|x_t)$.

Example. Consider some cost function $G(x, y)$ parameterized by x and some probability density $p(y)$. The minimum expected cost is

$$\min_x \int G(x,y)\, p(y)\, dy.$$

The minimum expected cost can be computed through stochastic approximation.

Assuming differentiability, etc., the gradient of the cost function can be written as $g(x, y) = \nabla G(x, y)$ and the fixed-point equation can be written as $\mathbb{E}_{p(y)}[g(x, y)] = 0$. This can then be plugged into the recursion above to obtain

$$x_{t+1} = x_t + \gamma_t\, g(x_t, \widetilde{y}_t).$$

15.4 Robbins-Monro Form

One form of the stochastic approximation update equations that is particularly useful for analysis is known as the *Robbins-Monro* stochastic approximation algorithms. The update equations can be written in Robbins-Monro form as follows

$$\begin{aligned} x_{t+1} &= (1-\gamma_t)x_t + \big(\gamma_t\, \mathbb{E}_{p(y|x_t)}\big[g(y,x_t)\big] - \gamma_t\, \mathbb{E}_{p(y|x_t)}\big[g(y,x_t)\big]\big) + \gamma_t\, g(\widetilde{y}_t, x_t) \\ &= (1-\gamma_t)x_t + \gamma_t\, \mathbb{E}_{p(y|x_t)}\big[g(y,x_t)\big] + \gamma_t\, w_t, \end{aligned}$$

where the w_t term can be thought of as "stochastic noise", and is given by

$$w_t = g(\widetilde{y}_t, x_t) - \mathbb{E}_{p(y|x_t)}\big[g(y,x_t)\big].$$

Example. Assume, for the sake of argument, that $w_t \sim \mathcal{N}(0, \sigma^2)$ and $\gamma_t = \gamma \in [0, 1]$ is some constant (this form of w_t does not hold in general). Then x will "oscillate" around a region with variance $\gamma^2\sigma^2$.