Overview of Statistical Machine Learning

Doug Aberdeen

November 4, 2004

Assignment 1: Stochastic gradient descent

Two alternative, and non-optimal, sets of parameters of hypothesis h are given by $\mathbf{v} = [v_1, \ldots, v_P]$, and $\mathbf{u} = [u_1, \ldots, u_P]$. The hypothesis could be for pattern classification, regression, whatever. In general, there is an error function $E(\mathbf{w}) : \mathbb{R}^P \to \mathbb{R}$ associated with an arbitrary set of the hypothesis parameters $\mathbf{w}$. The aim is to approximately locate a local minima in E in a single step, given only *noisy* gradient estimates

$$\widehat{\nabla_{\mathbf{v}}} = \widehat{\nabla} E(\mathbf{w})\Big|_{\mathbf{w}=\mathbf{v}} \quad \text{and} \quad \widehat{\nabla_{\mathbf{u}}} = \widehat{\nabla} E(\mathbf{w})\Big|_{\mathbf{w}=\mathbf{u}},$$

where

$$\nabla E(\mathbf{w}) = [\nabla_{w1}, \ldots, \nabla_{wP}] = \left[\frac{\partial E(\mathbf{w})}{\partial w_1}, \ldots, \frac{\partial E(\mathbf{w})}{\partial w_P}\right].$$

Assume that the two noisy gradient estimates are well conditioned in the sense that they do not point away from eachother. The approximation should be exact if the error surface is quadratic and the gradient estimates are exact. You do not need higher order derivatives. No solution is possible for $P = 1$, but your method should be valid for any $P > 1$.

- Provide an analytic solution for the approximate local minima $\mathbf{w} = [w_1, \ldots, w_P]$ in terms of $\mathbf{v}$, $\mathbf{u}$, $\widehat{\nabla_{\mathbf{v}}}$, and $\widehat{\nabla_{\mathbf{u}}}$.
- Provide a short paragraph sketching how more than two gradient estimates might be combined.
- Provide a short paragraph stating what problems might arise using this scheme in practice.
- Why is no solution possible for $P = 1$?

Hints and Clarifications

- Doc updated to reflect the use of $\mathbf{w}$ for parameters instead of $\mathbf{x}$.