Optimization

Aciditeam

September 18, 2015

While being self-sufficient, those notes focus on the recent results in optimization and don't pretend to be an exhaustive review of optimization methods.

Chapter 1

Error function

1.1 Euclidean norms

1.2 Cross-entropy

Cross-entropy is a measure between two distribution probability P and Q defined on the same set of events X. It measures the average number of bits needed to be able to identify an event drawn from the set if we use a coding scheme optimized for the distribution Q, but the data are drawn from the distribution p.

The optimal length of the coding message for each event $x_i \in X$ is given by $l_i = -log_2(q(x_i)$ if we suppose that Q models the distribution of X. Hence, H(p,q) is given by

$$H(p,q) = \mathbb{E}_p\left[l_i\right] = \mathbb{E}_p\left[-log(q)\right] \tag{1.1}$$

An other interesting writing of this formula is

$$H(p,q) = H(p) + D_{KL}(p||q) \tag{1.2}$$

which highlight the fact that the minimum of the cross-entropy function is given by the constant term of this formula $H(p)$, since the Kullback-Leibler divergence is non-negative.

However, the "true" distribution P is often unknown. For instance, when defining a cost function in order to train a neural network by gradient descent, P is the distribution we are trying to model. For a training set T,, an approximation of the cross-entropy can be obtained by

$$\hat{H}(T,q) = -\sum_{x_i \in T} \frac{1}{|T|} \log_2 q(x_i) \tag{1.3}$$

Chapter 2

Parameter update

2.1 Stochastic gradient descent (SGD)

2.1.1 Momentum

2.2 Hessian-free optimization

2.2.1 AdaGrad

2.2.2 AdaDelta

2.3