# Week 4 Report: Backpropagation review

Lam Dang

2017-04-27

# 1 Gradient Descent

Gradient Descent is a popular method in Data Mining to obtain the most suitable model. Gradient Descent usually used in tandem with other algorithm in order to create a model for prediction or classification, such as Linear Regression, Neural Network etc.

The concept of Gradient Descent is to "go down" the gradient "well" at a predefined rate. After each "step" the algorithm re-evaluate for new weight vector. These steps are repeated until a pre-defined number of iterations or until the global minima have been reach. [1]

Another view on the algorithm is to consider it as an optimization process [2]. In this view, Gradient Descent is considered as a constrained minimization problem of the model's error function.

## 1.1 Mathematical Model

Bishop [1, p240] provide a simple formular to update weights vector base on Gradient information

$$w^{(\tau+1)} = w^{(\tau)} - \eta \nabla E(w^{(\tau)}) \tag{1}$$

whereas $\eta$ is the pre defined learning rate and $E(w^{(\tau)})$ is the error function defined in the following section.

## 1.2 Error Function

Bishop [1, p242] also describe an evaluation of weight vector base on derivative of error function. The error function is given as

$$E(w) = \frac{1}{2} \sum_{n=1}^{N} ||y(x_n, w) - t_n||^2 \tag{2}$$

From $E(w)$ the derivative $\nabla E(w)$ is given by

$$\frac{\partial E}{\partial a_k} = y_k - t_k \tag{3}$$

Different method can be used for Gradient Descent. For Stochastic method, one data point is considered at one time. On the other hand, for batch method, multiple data point are considered and their error functions' derivatives are sumed to form the total error function for the iteration.

### 1.3 Learning Rate

An important aspect of Gradient Descent is its Learning Rate $\eta$. One of the major drawback of the method is the learning rate must be carefully considered. If $\eta$ is too small, the algorithm might take very long time to finish or will not reach the minima before the exit condition is met. Meanwhile a $\eta$ too large will risk overshooting the desired result.

## 2 Backpropagation

Backpropagation is one of the most popular method used in Data Mining and Artificial Intelligent field for obtaining a model for Artificial Neural Network. Discovered and re-discovered many time in the $20^{th}$ century, the method is popularized and refined by multiple independent party, one of the most prominion are Werbos and Le Cun [2]

### 2.1 Algorithm

Backpropagation algorithm can be considered layered version of Gradient Descent [1]. It use indefinitely differentiable activation function for its neurons. We denote these function as $F$ and their derivative as $F'$. The algorithm then went through 2 steps: [3, 161]

#### 2.1.1 Forward Propagation

The input $X_i$ is feed into the network. The functions $F(X_i)$ are calculated and propagate forward into next layers. The derivative functions $F'(X_i)$ are stored.

#### 2.1.2 Back Propagation

The constant 1 is assigned to the output unit and feed into the network in reverse. The incomming values to each node is added and multiply with value previously stored in those nodes. The result is then transfered upward to the input layer as derivative of the network function with respect to $X_i$

### 2.2 Weight update

The weight of a connection (vertice) is adjusted proportionally to an error signal $\delta$

### 2.3 Learning Rate

## 3 Current Result

### 3.1 Linear Regression

### 3.2 Extreme Learning Machine

### 3.3 Backpropagation

## References


[1] Christopher M. Bishop. *Pattern recognition and machine learning*. Information science and statistics. Springer, 2006.

[2] Yann Lecun. A theoretical framework for back-propagation. In *Artificial neural networks*. IEEE Computer Society Press, 1992.

[3] Ral Rojas. *Neural Networks*. Springer Berlin Heidelberg, 1996. DOI: 10.1007/978-3-642-61068-4.