UNIVERSITY OF SCIENCE AND TECHNOLOGY OF HANOI

UNDERGRADUATE SCHOOL

INTERSHIP AND DEVELOPMENT

BACHELOR THESIS

By

Dang Vu Lam

Information and Communication Technology

Title

Extreme Learning Machine

Supervisors

Dr.Sebastián Basterrech

Hanoi, July 2017

Abstract

To be done later, it is a short text describing the thesis, around 500 characters (between 5 and 10 sentences).

Acknowledgement

List of Abbreviations

- BP: Backpropagation
- ELM: Extreme Learning Machine
- MNIST: Mixed National Institute of Standards and Technology
- PSO: Particle Swarm Optimization
- USTH: University of Science and Technology of Hanoi

Mathematical notation

In order to add clarity to the paper, we decided to introduce the following formalism:

- $F(X)$: The result of applying activation function f(x) to all element of the matrix X
- H : The output matrix of the hidden layer (before apply the output weight)
- $H^{\dagger}$: The Moore-Penrose psuedoinverse of the matrix H
- X : The input matrix, which include all datapoint in consideration
- y : The expected output matrix
- $\hat{y}$: The approximated output matrix
- w : The weights matrixes, including
  - w_i : Input weights matrix
  - w_o : Output weights matrix
- η : The predefined learning rate
- ε : Random error

Contents

1 Introduction

2 Background

2.1 Linear Regression

2.1.1 Assumption of the model

The Linear Model assume the linear relationship between the predictor X and respond Y:

$$Y = F(X) + \varepsilon = \beta_0 + \beta_1 * X_1 + ... + \beta_n X_n + \varepsilon$$

where Y is the respond, X is the predictor [1]. The vector $[\beta_0,..,\beta_n]$ is called a weights vector, and must be tuned in order for the model to yield accurate result.

2.1.2 Gradient Descent

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. [2]

Another view on the algorithm is to consider it as an optimization process [3]. In this view, Gradient Descent is considered as a constrained minimization problem of the model's error function.

Bishop [2, 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 η is the pre defined learning rate and $E(w^{(\tau)})$ is the error function, which 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

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

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

2.1.3 Least Square

2.2 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 [3].

2.2.1 Algorithm

Backpropagation algorithm can be considered layered version of Gradient Descent [2]. 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: [4, 161]

- 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.

- 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

2.3 Extreme Learning Machine

Extreme Learning Machine (ELM) [5] is a simple yet powerful neural network algorithm. A single layer perceptron, its model can be formalize as simple matrix operations, which in turn can be massively parallelized for GPGPU operations

2.3.1 Mathematic Model

As

2.4 Particle Swarm Optimization

3 Implementation

3.1 Simple dataset: Salary dataset

3.1.1 Description

The salary dataset present a classical regression problem. Given a professor's details, including ... we should be able to predict their expected salary.

3.1.2 Extreme learning machine applied to regression problem

3.1.3 Particle Swarm Optimized Extreme Learning Machine

Using PSO, we were able to lower the neuron count for similar accuracy by 1/10th.

3.2 MNIST

- Data Description
- ELM Regression
- ELM Classification
- PSO Optimized ELM

4 Discussion

4.1 Performance of ELM

- Prediction Performance:

  Evidently, ELM systems are lack the prediction performance of traditional algorithm such as Backpropagation for similar number of neurons. Result from MNIST compared to state of the art networks listed on MNIST website are vastly underperformed. However there are works have been done to improve such accuracy issue. Thanks to the speed that ELM able to learn, multiple

References

[1] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning*, volume 103 of *Springer Texts in Statistics*. Springer New York, 2013.

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

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

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

[5] Guang-Bin Huang, Qin-Yu Zhu, and Chee-kheong Siew. Extreme learning machine: Theory and applications. 2016-05-16.