INSTITUTO POLITÉCNICO NACIONAL

CENTRO DE INVESTIGACIÓN Y DESARROLLO DE TECNOLOGÍA DIGITAL

MAESTRÍA EN CIENCIAS EN SISTEMAS DIGITALES

THESIS

MACHINE LEARNING LIBRARY TO SUPPORT APPLICATIONS WITH EMBEDDED SYSTEMS AND PARALLEL COMPUTING

TO OBTAIN THE DEGREE IN

MAESTRÍA EN CIENCIAS EN SISTEMAS DIGITALES

PRESENTS

CÉSAR MIRANDA MEZA

UNDER THE DIRECTION OF

DR. JUAN JOSÉ TAPIA ARMENTA

MARCH 2022 TIJUANA, B.C., MEXICO

MACHINE LEARNING LIBRARY TO SUPPORT APPLICATIONS WITH EMBEDDED SYSTEMS AND PARALLEL COMPUTING

Abstract

The currently available machine learning libraries have strongly addressed deep learning and parallel computing, but have neglected traditional machine learning methods and support for embedded systems in comparison. Therefore, in this thesis, a new machine learning library with a total of 53 functions has been developed and contributes with 6 new traditional machine learning methods, while supporting parallel computing and embedded systems. For a reference against which to validate and benchmark the developed library, the Dlib, PyTorch, scikit-learn and TensorFlow libraries were chosen as the main comparators. During the testing and validation process, several algorithms were developed in sequential mode: 6 for statistical methods; 6 for feature scaling methods; 9 for evaluation metric methods; 12 for regression methods; 12 for classification methods; and 2 for deep learning methods. Furthermore, additional algorithms were developed in order to parallelize the deep learning algorithms in CPU, single GPU and multiple GPU. However, only 36 of the sequential algorithms were compared with equivalent algorithms and the results indicate that about 83.33% of the library functions of this thesis were faster; 8.33% were equally fast; and another 8.33% were slower. Finally, some implementations were made on an Arduino UNO and STM32F446RE microprocessor to test the support of embedded systems with the developed library.

Keywords. Machine learning, library, embedded systems, parallel computing.

Contents

Resumen **i**

Abstract **ii**

List of Figures **vii**

List of Tables **xi**

List of Algorithms **xiii**

1 Introduction **1**

1.1 Problem statement 3

1.2 Research questions 3

1.3 Justification 4

1.4 Hypothesis 5

1.5 General objective 5

1.5.1 Specific objectives 5

1.6 Organization of the thesis 6

2 Theoretical framework **7**

2.1 Artificial intelligence 7

2.2 Machine learning 8

2.2.1 Machine learning algorithms 9

2.2.2 Machine learning libraries 10

2.2.3 General pipeline of machine learning applications 14

2.3 Parallel computing . . . 15
2.3.1 Performance analysis for parallel computing . . . 15
2.3.2 Parallel programming libraries . . . 16

3 Mathematical formulations **18**
3.1 Statistical equations . . . 20
3.1.1 The mean of a sample set . . . 20
3.1.2 The median (second quartile) of a sample set . . . 20
3.1.3 The variance of a sample set . . . 21
3.1.4 The standard deviation of a sample set . . . 21
3.1.5 The mode of a sample set . . . 22
3.1.6 Mean intervals . . . 22
3.2 Feature scaling for machine learning . . . 23
3.2.1 Min max normalization . . . 23
3.2.2 L2 normalization . . . 24
3.2.3 Z score normalization (standardization) . . . 24
3.3 Formulations of machine learning algorithms . . . 25
3.3.1 Simple linear regression . . . 25
3.3.2 Multiple linear regression . . . 30
3.3.3 Polynomial regression . . . 33
3.3.4 Multiple polynomial regression (without interaction terms) . . . 35
3.3.5 Logistic regression . . . 37
3.3.6 Gaussian regression . . . 40
3.3.7 Linear logistic classification . . . 42
3.3.8 Simple linear machine classification . . . 44
3.3.9 Kernel machine classification . . . 47
3.3.10 Single neuron in Deep Neural Network . . . 52
3.3.11 Deep Neural Network with a single output . . . 64

4 Methodology **78**
4.1 Methods . . . 78

4.2 Materials . . . 82
4.2.1 Hardware used . . . 82
4.2.2 External libraries and packages used . . . 83
4.3 Evaluation metrics for regression problems . . . 83
4.3.1 Mean squared error . . . 84
4.3.2 Coefficient of determination . . . 84
4.3.3 Adjusted coefficient of determination . . . 85
4.4 Evaluation metrics for classification problems . . . 85
4.4.1 Cross entropy error function . . . 85
4.4.2 Confusion matrix . . . 86
4.4.3 Accuracy . . . 86
4.4.4 Precision . . . 86
4.4.5 Recall . . . 87
4.4.6 F1 score . . . 87
4.5 Evaluation of underfitting and overfitting in machine learning models . . . 87

5 Results **89**
5.1 Statistical functions . . . 90
5.2 Feature scaling functions for machine learning . . . 93
5.3 Evaluation metric functions for machine learning algorithms . . . 95
5.4 Machine learning functions . . . 99
5.4.1 Regression algorithms . . . 99
5.4.2 Classification algorithms . . . 103
5.4.3 Deep learning algorithms . . . 108
5.4.4 Parallel computing in deep learning algorithms . . . 110
5.5 Results obtained in embedded systems . . . 114

6 Discussions **116**
6.1 Statistical functions . . . 116
6.2 Feature scaling functions for machine learning . . . 117
6.3 Evaluation metric functions for machine learning algorithms . . . 118

6.4 Machine learning functions 118
6.4.1 Regression algorithms 119
6.4.2 Classification algorithms 120
6.4.3 Deep learning algorithms 120
6.4.4 Parallel computing in deep learning algorithms 121
6.5 Results obtained in embedded systems 123

7 Conclusions **124**

8 Annexes **128**
8.1 Detailed specifications of the computer system used 128

References **133**