VIETNAM GENERAL CONFEDERATION OF LABOUR

**TON DUC THANG UNIVERSITY**

**FACULTY OF INFORMATION TECHNOLOGY**



**NGUYỄN MINH QUÂN - 521H0473**

**NGUYỄN ĐÌNH VIỆT HOÀNG - 522H0120**

**FINAL REPORT**

**INTRODUCTION TO MACHINE LEARNING**

**HO CHI MINH CITY, YEAR 2024**

VIETNAM GENERAL CONFEDERATION OF LABOUR

**TON DUC THANG UNIVERSITY**

**FACULTY OF INFORMATION TECHNOLOGY**



**NGUYỄN MINH QUÂN - 521H0473**

**NGUYỄN ĐÌNH VIỆT HOÀNG - 522H0120**

**FINAL REPORT**

**INTRODUCTION TO MACHINE LEARNING**

Instructor

**Mr. Lê Anh Cường**

**HO CHI MINH CITY, YEAR 2024**

**ACKNOWLEDGEMENT**

We sincerely thank Mr. Lê Anh Cường for teaching us the Introduction to Machine Learning course with great enthusiasm. We want to express our deep appreciation for the dedication and professional knowledge that you shared with us. Through your classes, we gained a better understanding of the fundamental aspects of the Introduction to Machine Learning, thanks to your detailed explanations and practical applications. You helped us grasp the knowledge and apply it effectively. Finally, we extend our heartfelt gratitude to Mr. Lê Anh Cường for your commitment and invaluable support throughout our learning journey in this course. The skills and knowledge we acquired will continue to impact our future development. We sincerely thank you and wish your health, success, and happiness.

*Ho Chi Minh City, May 20, 2024*

*Authors:*

*Nguyễn Minh Quân*

*Nguyễn Đình Việt Hoàng*

**DECLARATION OF AUTHORSHIP**

Our group assures that this is our own report and was guided by Mr. Lê Anh Cường. The research content and results in this report are honest and have not been published in any form before. The figures in the tables used for analysis, comments, and evaluations were collected by the authors from various sources clearly stated in the reference section.

Additionally, the report includes some comments, evaluations, and data from other authors and organizations, all of which are cited and noted for their origin.

**If any fraud is detected, we fully take responsibility for the content of our midterm report for the second semester of the 2023-2024 academic year.** Ton Duc Thang University is not involved in any copyright or intellectual property violations that we may cause during the process (if any).

*Ho Chi Minh City, May 20, 2024*

*Authors:*

*Nguyễn Minh Quân*

*Nguyễn Đình Việt Hoàng*

**INSTRUCTOR RUBRIC**

Supervisor’s Name: ……………………………………………………………..............

Comments: ……………………………………………………………………………...

Total Score Based on Rubric Evaluation: ………………………………………………

*Ho Chi Minh City, date … month … year …*

*Supervisor*

*(sign and write your full name)*

**TABLE OF CONTENTS**

# LIST OF FIGURES

# LIST OF TABLES

# LIST OF ABBREVIATIONS

|  |  |
| --- | --- |
| SGD | Stochastic Gradient Descent |
| NAG | Nesterov Accelerated Gradient |
| Adagrad | Adaptive Gradient Algorithm |
| RMSprop | Root Mean Square Propagation |
| Adam | Adaptive Moment Estimation |
| Nadam | Nesterov-accelerated Adaptive Moment Estimation |

QUESTION 1:

**a) Discuss different optimizers used in training Neural Networks. Describe the advantages and disadvantages of these methods. Compare these methods for the same problem (using the same dataset).**

1. *Stochastic Gradient Descent (SGD):*

+ SGD updates the weights of the network incrementally for each training sample.

* Advantages:
* Simplicity: Easy to understand and implement.
* Efficiency: Suitable for large datasets as it processes one sample at a time.
* Disadvantages:
* Convergence: May have slow convergence and get stuck in local minima.
* Noise: High variance in updates can lead to noisy convergence.

1. *Mini-Batch Gradient Descent:*

+ Mini-Batch Gradient Descent is a compromise between SGD and Batch Gradient Descent. It updates the weights based on a small random subset of the data (mini-batch).

* Advantages:
* Speed: Faster convergence than Batch Gradient Descent.
* Stability: Reduces the variance in updates compared to SGD.
* Disadvantages:
* Complexity: Requires careful tuning of the mini-batch size.
* Memory: Requires more memory than SGD.

1. *Momentum:*

+ Momentum accelerates SGD by adding a fraction of the previous update vector to the current update vector.

* Advantages:
* Speed: Faster convergence and helps escape local minima.
* Stability: Smoother updates compared to vanilla SGD.
* Disadvantages:
* Parameters: Requires tuning of the momentum parameter.

1. *Nesterov Accelerated Gradient (NAG):*

+ NAG is an extension of momentum that anticipates the future position of the parameters and corrects them before computing the gradient.

* Advantages:
* Speed: Faster convergence than standard momentum.
* Precision: More accurate updates.
* Disadvantages:
* Complexity: More complex to implement and requires additional parameter tuning.

1. *Adagrad (Adaptive Gradient Algorithm):*

+ Adagrad adapts the learning rate for each parameter based on the historical gradients.

* Advantages:
* Adaptability: Suitable for sparse data and handles varying learning rates automatically.
* Parameter-Free: Reduces the need to manually tune the learning rate.
* Disadvantages:
* Convergence: Learning rate can become very small, causing the training to stop prematurely.

1. *RMSprop (Root Mean Square Propagation):*

+ RMSprop is a modification of Adagrad that deals with the diminishing learning rate problem by using a moving average of squared gradients.

* Advantages:
* Stability: Maintains a good learning rate throughout the training.
* Convergence: Faster and more reliable convergence compared to Adagrad.
* Disadvantages:
* Parameters: Requires tuning of the decay parameter.

1. *Adam (Adaptive Moment Estimation):*

+ Adam combines the advantages of both RMSprop and momentum by maintaining a moving average of both the gradients and the squared gradients.

* Advantages:
* Efficiency: Suitable for large datasets and high-dimensional parameter spaces.
* Adaptability: Automatically adjusts the learning rate.
* Convergence: Generally fast and reliable convergence.
* Disadvantages:
* Parameters: Requires tuning of several hyperparameters (learning rate, beta1, beta2).

1. *Nadam (Nesterov-accelerated Adaptive Moment Estimation):*

+ Nadam is an extension of Adam that incorporates Nesterov momentum.

* Advantages:
* Speed: Faster convergence than Adam in some cases.
* Precision: Combines the benefits of Nesterov momentum and Adam.
* Disadvantages:
* Complexity: More complex and requires tuning of additional parameters.

1. *AdaMax:*

+ AdaMax is a variant of Adam that uses the infinity norm instead of the L2 norm.

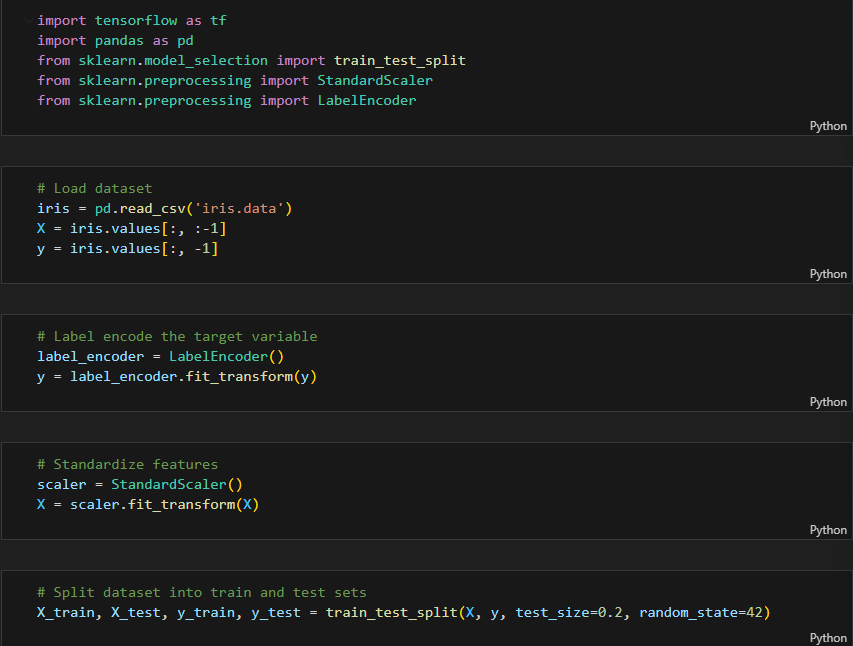
* Advantages:
* Stability: More stable in some cases compared to Adam.
* Robustness: Handles large gradients better.
* Disadvantages:
* Specificity: May not always outperform Adam.

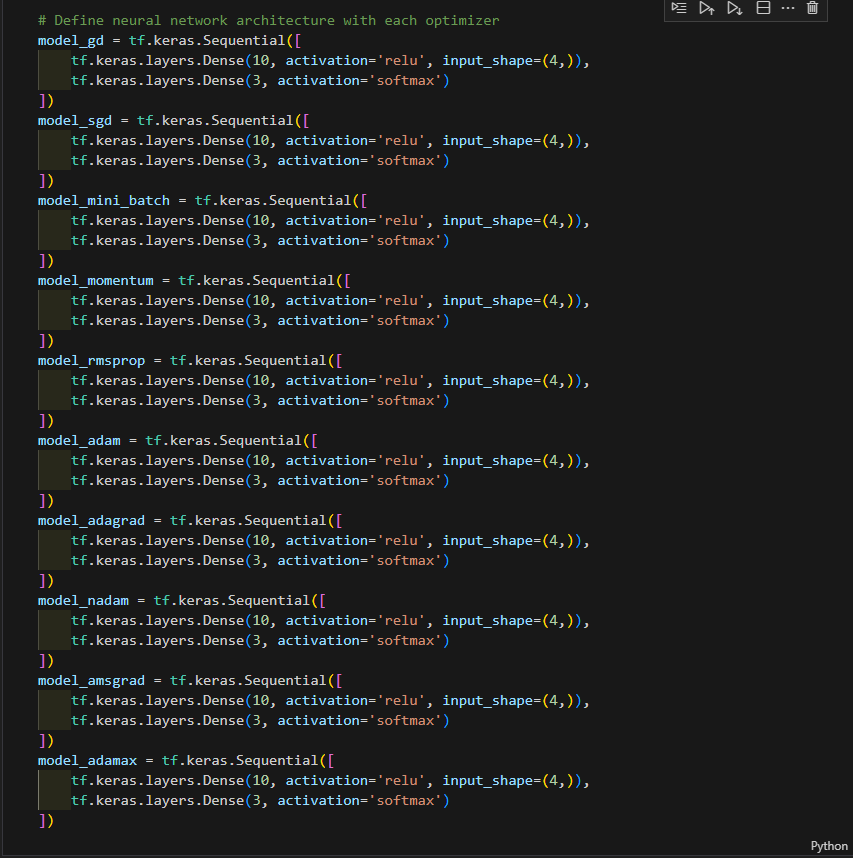
1. *AMSGrad:*

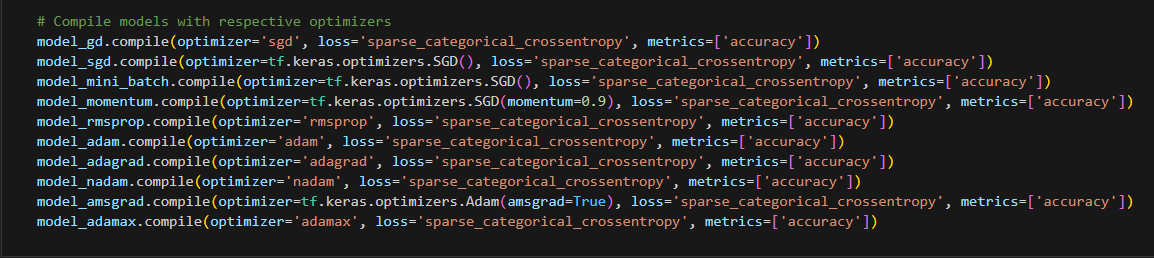
+AMSGrad is a variant of Adam that fixes some issues with the convergence of Adam by using the maximum of past squared gradients.

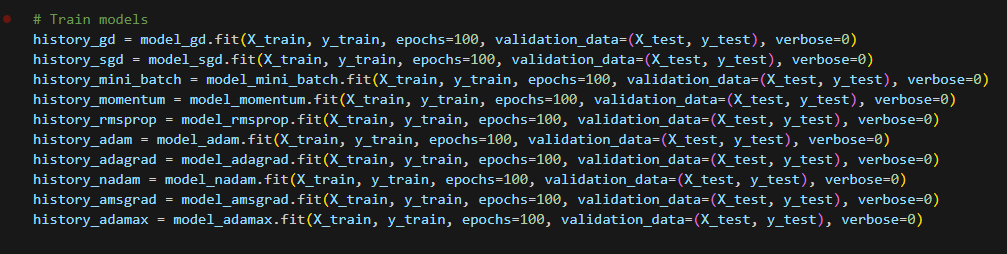
* Advantages:
* Convergence: Improved theoretical guarantees of convergence.
* Disadvantages:
* Complexity: Slightly more complex to implement than Adam.

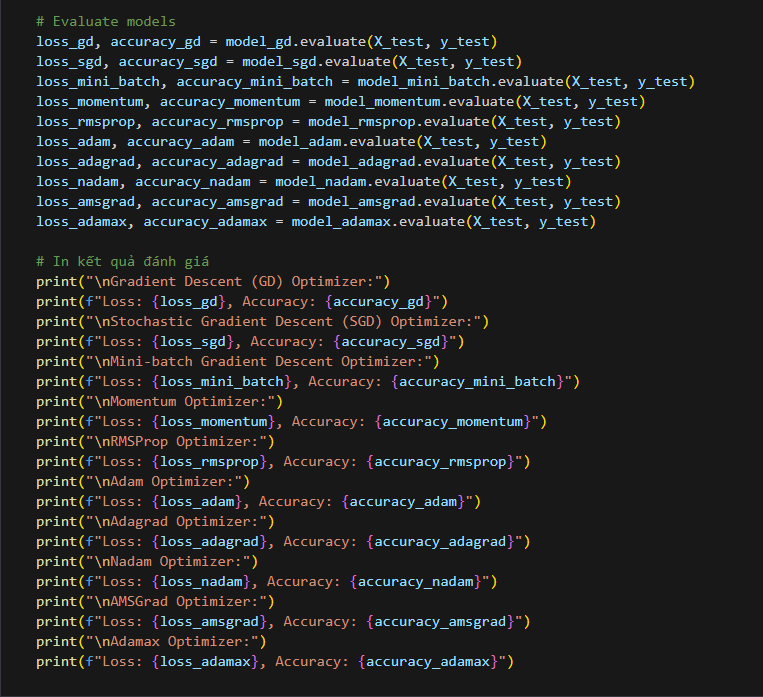
Compare these methods for the same problem (using the same dataset):

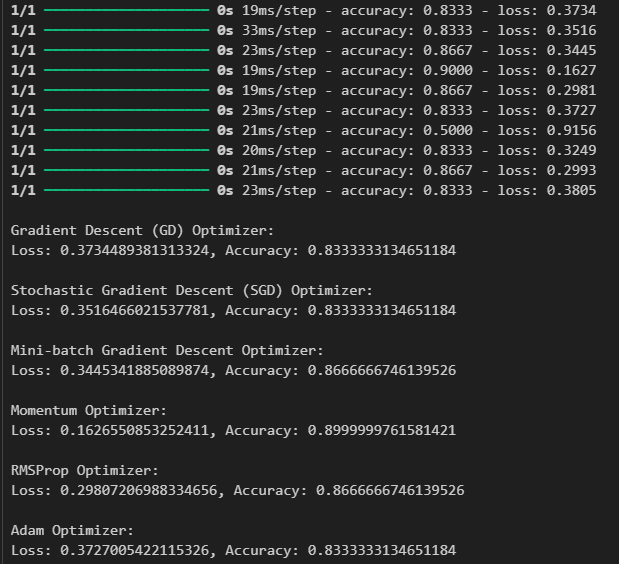


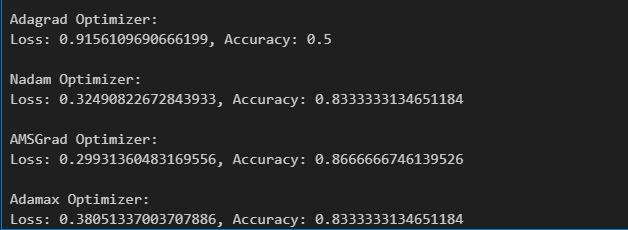










*Figure 1.a) Different optimizers used in training Neural Networks*

**b) Explain the backpropagation algorithm for learning parameters in Neural Networks and provide illustrative code.**

# REFERENCES

**Vietnamese**

**English**