**Vietnam General Confederation of Labor**

**TON DUC THANG UNIVERSITY**

**FACULTY OF INFORMATION TECHNOLOGY**



**MIDTERM REPORT**

**INTRODUCTION TO**

**MACHINE LEARNING**

*Instructor*: **Ph.D LE ANH CUONG**

*Student*: **Ho Huu An – 521H0489**

**Tran Nhut Anh – 521H0491**

**Do Minh Quan – 521HH0290**

*Class* **: 21H50301**

*Year* **: 25**

**HO CHI MINH CITY, 2023**

Vietnam General Confederation of Labor

**TON DUC THANG UNIVERSITY**

**FACULTY OF INFORMATION TECHNOLOGY**



**MIDTERM REPORT**

**INTRODUCTION TO**

**MACHINE LEARNING**

*Instructor*: **Ph.D LE ANH CUONG**

*Student*: **Ho Huu An – 521H0489**

**Tran Nhut Anh – 521H0491**

**Do Minh Quan – 521HH0290**

*Class* **: 21H50301**

*Year* **: 25**

**HO CHI MINH CITY, 2023**

ACKNOWLEDGEMENT

We would like to express my sincere appreciation to Le Anh Cuong for his exceptional teaching on machine learning. His expertise and guidance have provided us with a comprehensive understanding of various models, metrics, and practical applications. His dedication to our growth, patient approach, and real-world examples have been truly inspiring. I am grateful for his unwavering support and belief in our abilities. Thank you, Le Anh Cuong, for being an outstanding educator and empowering us in the field of machine learning..

*Ho Chi Minh city, 14th October, 2023*

*Author*

*(Sign and write full name)*

**THIS PROJECT WAS COMPLETED AT**

**TON DUC THANG UNIVERSIY**

We fully declare that this is our own project and is guided by Mr. Le Anh Cuong; The research contents and results in this topic are honest and have not been published in any form before. The data in the tables for analysis, comments and evaluation are collected by the author himself from different sources, clearly stated in the reference section.

Besides that, the project also uses a number of comments, assessments as well as data from other authors, other agencies and organizations, with citations and source annotations.

**Should any frauds were found, I will take full responsibility for the content of my report.** Ton Duc Thang University is not related to copyright and copyright violations caused by me during the implementation process (if any).

*Ho Chi Minh city, 14th January, 2022*

*Author*

*(Sign and write full name)*

CONFIRMATION AND ASSESSMENT SECTION

**Instructor confirmation section**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

*Ho Chi Minh January, 2022*

*(Sign and write full name)*

**Evaluation section for grading instructor**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

*Ho Chi Minh January 2022*

*(Sign and write full name)*

SUMMARY

This is a report on Introduction to Machine Learning by Faculty of Information Technology of Ton Duc Thang University.

INDEX

[ACKNOWLEDGEMENT i](#_Toc148906577)

[CONFIRMATION AND ASSESSMENT SECTION iii](#_Toc148906578)

[SUMMARY iv](#_Toc148906579)

[INDEX 1](#_Toc148906580)

[CHAPTER 1 – INTRODUCTION 1](#_Toc148906581)

[Introduce the overfitting problem 1](#_Toc148906582)

[Overfitting definition 2](#_Toc148906583)

[Cause of overfitting 2](#_Toc148906584)

[CHAPTER 2 – Reducing Overfitting 3](#_Toc148906585)

[Cross-validation 3](#_Toc148906586)

[Regularization 4](#_Toc148906587)

[Dropout 4](#_Toc148906588)

[Early stopping 5](#_Toc148906589)

[Data augmentation 6](#_Toc148906590)

CHAPTER 1 – INTRODUCTION

Introduce the overfitting problem

Overfitting is a fundamental problem in supervised learning , which prevents generalizing the models to fit well observed data on training data, as well as unseen data on testing set. Because of the presence of noise, the limited size of training set, and the complexity of classifiers, overfitting happens. This paper is going to talk about overfitting from the perspectives of causes and solutions. This document is going to talk about overfitting from the perspectives of causes and solutions.

Overfitting definition

Overfitting refers to a model which does not generalize well from observed data to unseen data. On the other hand the model performs perfectly on the training dataset , while fitting poorly on the testing dataset.

More in-depth explanation is that the model has difficulty predicting the testing dataset which is different from the training dataset. In other words , over-fitted models tend to remember all the data, including the noise on the training dataset , rather than learning the pattern of the data

In the Picture below . The over-fitted Model (represent as blue line) trying the fit all data point rather learning the pattern like the True function (represent as orange line) from the data

Cause of overfitting

* **Noise learning on the training set:** when the training set is too small in size or has less representative data or too many noises. This situation makes the noises have great chances to be learned, and later act as a basis of predictions.
* **Highly complex machine learning model:** the complex model (especially neural networks) tends to learn the datapoint to carefully leading to overfitting due to their extreme flexibility and capacity
* **hypothesis complexity**: the trade-off in complexity, a key concept in statistic and machining learning, is a compromise between Variance and Bias. It refers to a balance between accuracy and consistency.

A graph with blue and orange lines

Description automatically generated

(Source : [Example of overfitting and underfitting in machine learning (keeeto.github.io)](https://keeeto.github.io/blog/bias_variance/) )

CHAPTER 2 – Reducing Overfitting

Cross-validation

Cross validation is a technique for evaluating machine learning models on a limited data sample. It aims to estimate how well a model will generalize to new data.

Cross-validation is a method in which we train models by splitting the dataset into training and validating sub-data . After training the model and validating the data , we continue to split the dataset in which the new training sub-dataset and the validating sub-dataset is not the same as the previous until there is no sub-dataset to be validated . Notice after a validation step will provide the error for evaluating the cross-validation score.After the cross-validation we will choose the model which have the least error

A green rectangular boxes with white text

Description automatically generated

Source : [GeeksforGeeks | A computer science portal for geeks](https://www.geeksforgeeks.org/)

Regularization

Dropout

Dropout is a regularization technique that randomly sets a fraction of neuron outputs to zero during training. For example, with a dropout rate of 0.2, 20% of neurons will be "dropped out" in each training iteration.

A diagram of a network

Description automatically generated

(source: [What is Dropout Regularization Technique? (xenonstack.com)](https://www.xenonstack.com/glossary/dropout-regularization) )

Early stopping

The idea of this method is we stop the training process at certain points where the models performance decline based on evaluating the model on the validation dataset

A graph of a graph with a triangle and a warning sign

Description automatically generated with medium confidence

* Horizontal axis: epochs ( or number of iterations)
* Vertical axis: error
* blue: training error
* red: validation error

In this figure, as the model reaches a certain point, the training error declines, but the validation error increases. So early stopping will shut down the training process before this happens.

Data augmentation

The idea of data augmentation is to increase the amount of training data by generating modified versions of existing dataset . This help the model to learn better as one datapoint provide more variant versions of it

A diagram of a butterfly

Description automatically generated