

HUMBOLDT-UNIVERSITÄT ZU BERLIN
MATHEMATISCH-NATURWISSENSCHAFTLICHE FAKULTÄT
INSTITUT FÜR INFORMATIK



Risk assessment in Machine Learning security - a framework for risk measuring

Masterthesis

for the attainment of the academic degree
Master of Science (M. Sc.)

submitted by: Jan Schröder

born on: 03.03.1996

born in: Lemgo

Surveyor: Martin Schneider

Prof. Dr. Holger Schlingloff

submitted on:

defended on:

Contents

| | | |
|----------|--|----------|
| 1 | Introduction | 2 |
| 1.1 | Motivation | 2 |
| 2 | Related Work | 3 |
| 2.1 | ISO 27001 | 3 |
| 2.2 | Security risks in context of Machine Learning | 3 |
| 2.3 | Risk assessment in context of Machine Learning | 3 |
| 2.4 | Adversarial-Robustness-Toolbox | 3 |
| 2.5 | Approaches | 3 |

Abstract

Acknowledgements

1 Introduction

Machine Learning (ML) is a constantly growing field and is essential for many modern applications such as highly-automated and autonomous driving. Resulting from this, there is an increased need to maintain security. This thesis concentrates on risk measuring in context of ISO 27001 which will be discussed in 2. Risk measuring is a part of risk assessment to help where investments are needed to defend a system against attackers.

This thesis explains and discuss' a conceptual and technical framework to measure risks which is called Security-Measuring-Framework (SMF).

1.1 Motivation

2 Related Work

This chapter presents the relevant background knowledge and show the approaches from other scientific paper.

2.1 ISO 27001

2.2 Security risks in context of Machine Learning

Xiao et. al [4] evaluate the security risks in deep learning for common frameworks i.e. TensorFlow. Xiao et. al uses the framework sample applications along the frameworks. One statement of Xiao et. al is that the named frameworks TensorFlow, Caffe and Torch are implemented with many lines of code which make them vulnerable for many security vulnerabilities i.e. heap overflow or integer overflow.

2.3 Risk assessment in context of Machine Learning

2.4 Adversarial-Robustness-Toolbox

For this present thesis the technical framework Adversarial-Robustness-Toolbox (ART) [2] is a main component.

2.5 Approaches

Jakub Breier et. al [1] propose in their paper different proposals to measure risks with different aspects.

Paul Schwerdtner et. al [3] present in their work a framework to evaluate ML model by input corrupted data. This thesis discuss this paper as an approach to estimate where the SMF could be used for.

References

- [1] Jakub Breier, Adrian Baldwin, Helen Balinsky, and Yang Liu. Risk management framework for machine learning security. *CoRR*, abs/2012.04884, 2020.
- [2] Maria-Irina Nicolae, Mathieu Sinn, Minh Ngoc Tran, Beat Buesser, Ambrish Rawat, Martin Wistuba, Valentina Zantedeschi, Nathalie Baracaldo, Bryant Chen, Heiko Ludwig, Ian Molloy, and Ben Edwards. Adversarial robustness toolbox v1.2.0. *CoRR*, 1807.01069, 2018.
- [3] Paul Schwerdtner, Florens Greßner, Nikhil Kapoor, Felix Assion, René Sass, Wiebke Günther, Fabian Hüger, and Peter Schlicht. Risk assessment for machine learning models. *CoRR*, abs/2011.04328, 2020.
- [4] Qixue Xiao, Kang Li, Deyue Zhang, and Weilin Xu. Security risks in deep learning implementations. In *2018 IEEE Security and Privacy Workshops, SP Workshops 2018, San Francisco, CA, USA, May 24, 2018*, pages 123–128. IEEE Computer Society, 2018.

Selbständigkeitserklärung

Ich erkläre hiermit, dass ich die vorliegende Arbeit selbständig verfasst und noch nicht für andere Prüfungen eingereicht habe. Sämtliche Quellen einschließlich Internetquellen, die unverändert oder abgewandelt wiedergegeben werden, insbesondere Quellen für Texte, Grafiken, Tabellen und Bilder, sind als solche kenntlich gemacht. Mir ist bekannt, dass bei Verstößen gegen diese Grundsätze ein Verfahren wegen Täuschungsversuchs bzw. Täuschung eingeleitet wird.

Berlin, den November 27, 2021

.....