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

Fraunhofer
FOKUS

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

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

Abstract

Acknowledgements

1 Introduction

Machine Learning (ML) is a constantly growing field and is essential for many innovative 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-Measurement-Framework (SMF).

1.1 Motivation

2 Related Work

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

2.1 ISO/IEC 27004:2009

This present thesis based the requirements of Risk measurement of ISO 27004, among other things. ISO 27004 is a international security standard from the ISO 27000 [4] family which guides on continious basis evaluation methods. The present ISO can be related with ISO 27001 or used as a standalone standard. In ISO 27001 it is declared as a requirement where the effectiveness must be measured of a Information Security Management System [1]. The ISO 27004 standard specifies what to be measured, when the measurement is needed and types of measurement [3]. Barabanov et al. [1] and Tarnes [8] describe in their works the different properties of ISO/IEC 27004:2009 for Risk measurement. Tarnes shows the information security measurement model which is shown in Figure 1.

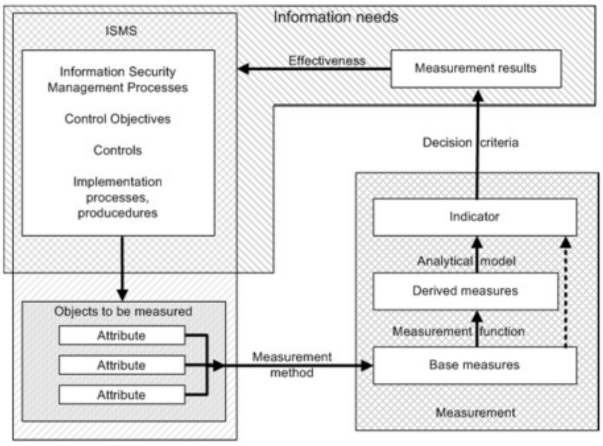

Figure 1: The information security measurement model [8]

For this thesis the objects to be measured and the measurement are the important parts of the information security measurement model. The measurement method is the SMF which measure based on different properties that are derived from risk indicators 3.1. The attributes in Figure 1 are the properties in the SMF.

2.2 Security risks in context of Machine Learning

Xiao et al. [9] evaluate the security risks in deep learning for common frameworks for example 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 for example heap overflow or integer overflow. Xiao et. al work is only in context of deep learning e.g. only for neural networks.

2.3 Risk assessment in context of Machine Learning

Paul Schwerdtner et al. [7] 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.

2.4 Adversarial-Robustness-Toolbox

For this thesis the technical framework Adversarial-Robustness-Toolbox (ART) [6] is a main component. Nicolae et al. [5] evaluate in their work the ART framework. ART is a Python library supports several ML frameworks for example TensorFlow and PyTorch. ART support 39 attacks and 29 defense functions. This thesis only focus on the attack functions for poisoning attacks especially inserting backdoor.

2.5 Approaches from Jakub Breier et. al and Paul Schwerdtner et. al

This present thesis is divided into two approaches. Jakub Breier et al. [2] propose in their paper different proposals to measure risks with different aspects.
Paul Schwerdtner et al. [7] is the second approach of this thesis.

3 The conceptual framework

3.1 Risk indicators

3.2 Poisoning attacks especially backdoor attacks

3.3 Finding attackers effort by objective properties

4 The technical framework

5 Evaluation

Test

References

[1] Rostyslav Barabanov, Stewart Kowalski, and Louise Yngström. Information security metrics: State of the art: State of the art. 2011.

[2] Jakub Breier, Adrian Baldwin, Helen Balinsky, and Yang Liu. Risk management framework for machine learning security. *CoRR*, abs/2012.04884, 2020.

[3] Kristoffer Lundholm, Jonas Hallberg, and Helena Granlund. Design and use of information security metrics. *FOI, Swedish Def. Res. Agency, p. ISSN*, pages 1650–1942, 2011.

[4] Ines Meriah and Latifa Ben Arfa Rabai. Comparative study of ontologies based ISO 27000 series security standards. In Elhadi M. Shakshuki, Ansar-Ul-Haque Yasar, and Haroon Malik, editors, *The 10th International Conference on Emerging Ubiquitous Systems and Pervasive Networks (EUSPN 2019) / The 9th International Conference on Current and Future Trends of Information and Communication Technologies in Healthcare (ICTH-2019) / Affiliated Workshops, Coimbra, Portugal, November 4-7, 2019*, volume 160 of *Procedia Computer Science*, pages 85–92. Elsevier, 2019.

[5] Maria-Irina Nicolae, Mathieu Sinn, Tran Ngoc Minh, Ambrish Rawat, Martin Wistuba, Valentina Zantedeschi, Ian M. Molloy, and Benjamin Edwards. Adversarial robustness toolbox v1.0.0. *CoRR*, abs/1807.01069, 2019.

[6] 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.

[7] 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.

[8] Marte Tarnes. Information security metrics: An empirical study of current practice. *Specialization Project, Trondheim, 17th December*, 2012.

[9] 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 December 11, 2021 ....