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 discusses a conceptual and technical framework to measure risks which is called Risk-Measurement-Framework (RMF). The RMF will build a conceptual and technical framework upon approaches by Jakub Breier et al. [2] and Paul Schwerdtner et al. [10].

In Section 2 the related work explain and discuss the basis and necessary terms. Section 3 is one of the main parts of the thesis. The section discuss and describes the conceptual framework and gives the basis for the in Section 4 explained technical framework.

1.1 Motivation for this thesis

1.2 Goals of this present thesis

The goals of this thesis are formulated in the following research questions:

- Which ISO 27004 measurement metrics are useful to measure the risks of poisoning attacks?
- How can the size of a dataset be used to measure the risks of poisoning attacks?
- What are risk indicators of poisoning attacks?
- Which risk indicators can be used for the ML model apart from the dataset?
- How can the effort of an attack be measured?
- Which measurement requirements of ISO 27004 can be used to measure the effort of an attack in ML security?
- Which risk indicators from the poisoning attacks and the attackers effort are useful to evaluate the risks with the RMF?
- What are possible methods in the RMF to measure the effort of an attacker?
- Which backdoor attacks must execute an attacker and objective properties must be fulfilled by the attacker to find how much damage an attacker wants to do with his attack?

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 [6] 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 [5]. Barabanov et al. [1] and Tarnes [12] 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.

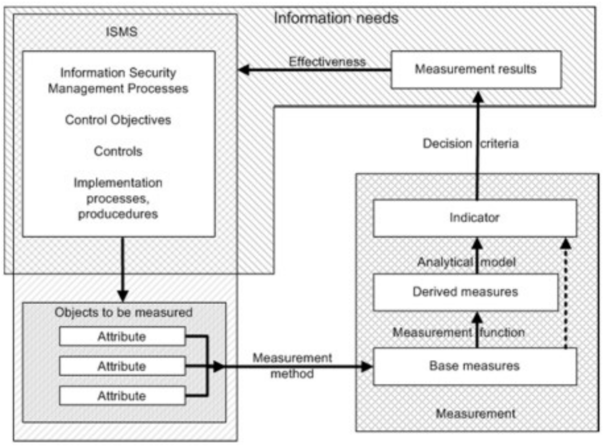


Figure 1: The information security measurement model [12]

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 that will be discussed in Subsection 3.3. The attributes in Figure 1 are the properties in the SMF.

2.2 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. These attacks are used in this thesis as properties to classify attacks. These different properties

are attack specificity, attack time and attacker's knowledge. Attack time is split in training time and deployment time. Training time is the attack time when the model gets manipulated while it trains. Deployment time is the attack time when the hacker attacks a ML model after its release. Attacker's knowledge is the amount of information the hacker has available. Attackers specificity is the amount an attacker needs to manipulate the output of a ML model. These three properties may serve as a basis for further properties useful for risk measurement.
Paul Schwerdtner et al. [10] is the second approach of this thesis. Schwerdtner et al. show a technical framework to evaluate the risks for ML models. Schwerdtner et al. give an evaluation whether it is secure to deploy a ML model or not. The ML model in Schwerdtner et al. must be a fully developed ML model that is trained and tested. Schwerdtner et al. concentrate on inference data when the ML model is executed. This thesis discuss this paper as an approach to estimate where the RMF could be used for.

2.3 Security risks in context of Machine Learning

Xiao et al. [14] 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.

2.4 Risk assessment in context of Machine Learning

In addition to ISO 27004, the paper by Sendi et al. [11] shows at which point in IT security management risk measurement takes place for the thesis and how it is carried out. In their paper, Sendi et al. evaluated 125 works published between 1995 and 2014. They developed categories and the last category is risk measurement. This category is the last step of risk assessment. To evaluate risks by measuring them, there are different properties which have an impact for risk measurement. Sendi et al. explain that the type of the attack, the dependency severity between resources and the type of defined permissions between resources are needed to measure risks.

2.5 Adversarial-Robustness-Toolbox

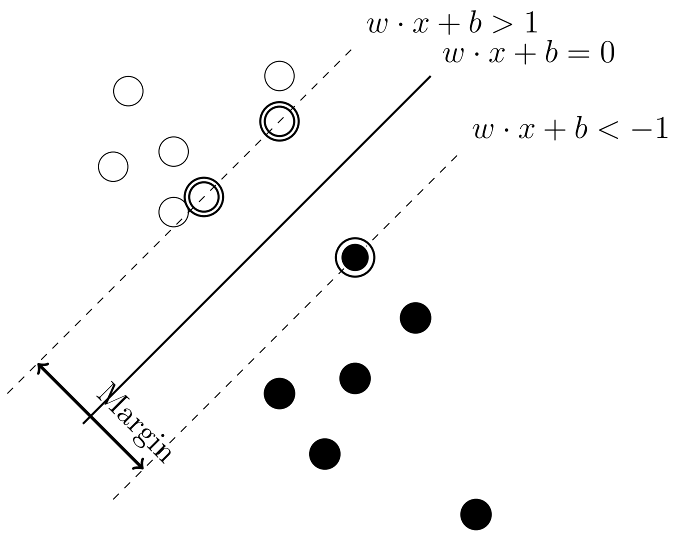
For this thesis the technical framework Adversarial-Robustness-Toolbox (ART) [8] is a main component. Nicolae et al. [7] evaluate in their work the technical framework ART. ART is a Python library that supports several ML frameworks for example TensorFlow and PyTorch to increase the defense of ML models. ART support 39 attacks and 29 defense' functions. This thesis only focuses on the attack functions for poisoning attacks which will be discussed in the following section more detailed. The backdoor attacks in the technical framework ART are introduced by Gu et al. [4].

Backdoor Attacks

Due to the rising amount of training data, human supervision to check trustworthiness is less possible. That exposes vulnerabilites in training datasets like backdoors. Backdoor attacks can cause far- reaching consequences for example bypass critical authentication. In [9] Salem et al. introduces dynamic backdoors to trigger (a secret pattern of neighboring pixels) random patterns and locations to reduce the efficacy on identifying backdoors. Salem et al. discuss in their work three backdoors, Random Backdoor, Backdoor Generating Network and Conditional Backdoor Generating Network. Gu et al. show in their paper different backdoor attacks and do a case study with a traffic sign detection attack. The evaluated backdoors are a single pixel backdoor and a pattern backdoor. The single pixel backdoor changes a pixel to a bright pixel and the pattern backdoor adds a pattern of bright pixels in an image. The implemented attacks from Gu et al. are single target attack and an all-to-all attack.

2.6 Support-Vector-Machine

Support-Vector-Machine (SVM) is a supervised ML algorithm which classifies a set of objects (splitted in two groups) between a hyperplane in an *N-dimensional* coordinate system. The goal is to find the maximum distance between the objects in both classes. As the name SVM says, this ML algorithm uses Support Vectors. That are the objects close to the hyperplane. The most maximized margin between the sets of objects is the best hyperplane. When the set of objects are more complex the SVM needs a higher dimensional hyperplane. The following example shows a two dimensional hyperplane. If linear separation is not possible a so called kernel realizes the non-linear to a feature space.



Hyperplane

The hyperplane is in a SVM a linear line between a set of objects (one set of object is called a class on one side of a hyperplane). The line differentiate the set of objects for classification. The hyperplane is used for two-dimensional coordinate systems.

Support Vector

Support Vectors are the minimum margin on both sides of the hyperplane. The maximum margin is the nearest object to the hyperplane in both classes.

SVM optimization

The kernel trick

The kernel trick is used if the positions of the sets of objects is not redundant to classify them with a hyperplane. Kernel trick is also used if there are more than two classes to classify. If there are more than two classes the SVM do a multi-class classification. The idea of multi-class classification is separating the classes in a binary classification [13].

3 The conceptual framework

In contrast to Schwerdtner et al., the framework of this thesis concentrates on training, especially Risk Measurement before and during training of the ML model. The conceptual framework discusses and explains the RMF. The RMF is a conceptual and technical framework which measures risks of backdoor attacks and measures the attacker effort. The attacker effort is measured by objective properties. These objective properties are the base of the risk indicators for the attacker effort explained in the following subsection. Objective properties

3.1 Finding the attacker's effort

Attacker characteristics found by the threat model

In their paper, Doynikova et al. [3] show a formal attacker model with input data for experiments, the data handling process and describe the experiment that was executed. Doynikova et al. explain that the attacker models can be split into high-level and low-level. These models contain attributes which used in this thesis as properties. High-level properties are...

3.2 Characteristics of backdoor attacks

3.3 Risk indicators

4 The technical framework

4.1 Using ART as the basis for the technical framework

4.2 Implementation of the logging function

4.3 Implementation of the visualization

4.4 Build in the risk indicatiors

5 Evaluation

5.1 Case Study: Developing a SVM fo traffic sign detection

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] Elena Doynikova, Evgenia Novikova, Diana Gaifulina, and Igor V. Kotenko. Towards attacker attribution for risk analysis. In Joaquín García-Alfaro, Jean Leneutre, Nora Cuppens, and Reda Yaich, editors, *Risks and Security of Internet and Systems - 15th International Conference, CRiSIS 2020, Paris, France, November 4-6, 2020, Revised Selected Papers*, volume 12528 of *Lecture Notes in Computer Science*, pages 347–353. Springer, 2020.

[4] Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg. Badnets: Identifying vulnerabilities in the machine learning model supply chain. *CoRR*, abs/1708.06733, 2017.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] Ahmed Salem, Rui Wen, Michael Backes, Shiqing Ma, and Yang Zhang. Dynamic backdoor attacks against machine learning models. *CoRR*, abs/2003.03675, 2020.

[10] 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.

[11] Alireza Shameli Sendi, Rouzbeh Aghababaei-Barzegar, and Mohamed Cheriet. Taxonomy of information security risk assessment (ISRA). *Comput. Secur.*, 57:14–30, 2016.

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

[13] Angelos Tzotsos and Demetre Argialas. Support vector machine classification for object-based image analysis. In *Object-Based Image Analysis*, pages 663–677. Springer, 2008.

[14] 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 21, 2021 ....