INFORMATYKA EKONOMICZNA BUSINESS INFORMATICS 3(57) • 2020
ISSN 1507-3858
e-ISSN 2450-0003

Jerzy Surma
Warsaw School of Economics
e-mail: jerzy.surma@sgh.waw.pl
ORCID: 0000-0002-5544-2573

ATTACK VECTORS ON SUPERVISED MACHINE LEARNING SYSTEMS IN BUSINESS APPLICATIONS

WEKTORY ATAKÓW NA NADZOROWANE SYSTEMY UCZĄCE SIĘ W ZASTOSOWANIACH BIZNESOWYCH

DOI: 10.15611/ie.2020.3.05
JEL Classification: C890

Quote as: Surma, J. (2020). Attack vectors on supervised machine learning systems in business applications. *Informatyka Ekonomiczna. Business Informatics*, (3).

Abstract: Machine learning systems have become incredibly popular and now have practical applications in many fields. An area of business applications has been developing particularly well, starting from the prediction of customers' purchase preferences and up to the automation of critical business processes. In this context, the security of such systems in a situation of a threat of intentional attacks carried by organized crime is extremely important. A theoretical framework of attacks on supervised machine learning systems, which are the most popular in business applications, is set out in this article. The possible attack vectors are widely discussed. The main contribution of this article is to recognize that the black box type attack scenario is the most probable, therefore the scenario of this kind of attacks was described extensively.

Keywords: adversarial machine learning, supervised machine learning, security of machine learning systems.

Streszczenie: Systemy uczące się stają się coraz bardziej popularne i mają wiele praktycznych zastosowań. Szczególnie istotny i szybko rozwijający się jest obszar zastosowań biznesowych. W tym kontekście bezpieczeństwo informacyjne takich systemów jest niezwykle ważne, zwłaszcza przy dużej aktywności zorganizowanych grup cyberprzestępców. W artykule przedstawiono taksonomię intencjonalnych ataków na systemy uczące się pod nadzorem, które to są obecnie najpopularniejsze w zastosowaniach biznesowych. Omówiono także potencjalne wektory ataków. Wskazano ataki typu „czarna skrzynka" jako najbardziej prawdopodobne scenariusze ataków i omówiono je bardziej szczegółowo.

Słowa kluczowe: antagonistyczne maszynowe uczenie się, nadzorowane maszynowe uczenie się, bezpieczeństwo systemów uczących się.

1. Introduction

Historically, the term ‘hacking’ is connected with an activity which aims to break into information systems, and in this context it is identified with actions made for malicious purposes with unethical intentions. This is a particularly important issue in a situation of the currently marginal awareness of such real threats on machine learning systems, as confirmed by the research conducted by Shankar (2020) in 28 companies which use and develop learning systems independently to a considerable maturity. The researched employees (leaders of programming teams and managers responsible for cyber security) in their vast majority expressed opinions on the futuristic nature of attacks on learning systems and declared a lack of resources to analyse this type of threats.

In the first part of the article, related works are described. In the next section, a classification task in supervised machine learning systems is defined, and three possible objectives of the attacker are presented. Further on, the theoretical framework of attacks is introduced and the black box type attack is explained in detail. This article ends with final conclusions.

2. Related work

The problem of intentional attacks on learning systems was set out for the first time in a comprehensive manner in 2004 during the conference Knowledge Discovery in Databases (Dalvi, Domingos, P., Sumit, M., and Verma, 2004), where the method of data manipulation applied in the learning process was studied in order to increase the number of false positive errors. This new research topic was continued in the work by Barreno (Barreno, Nelson, Sears, Joseph, and Tygar, 2006) and in Nelson’s PhD thesis (2010). The summary of the research results from the first phase of the research and indication of the development directions was discussed in the Machine Learning Journal in the special issue “Machine Learning in Adversarial Environments” (Laskov and Lippmann, 2010). In the article by Barreno et al. (2010), a systematic approach to the classification of potential threats on learning systems was shown for the first time, and a theoretical model of interaction between the attacker and defender was proposed. A fully developed concept of the taxonomy of attacks was also presented in the article by Huang (Huang, Joseph, Nelson, Rubinstein, and Tygar, 2011).

Nowadays this research area is most often defined as adversarial machine learning. One can classify the research by Goodfellow et al. (2014) within generative adversarial networks (GAN), and the work on the black box type adversary attacks using models of deep learning (Papernot et al., 2017) into the most innovative research works in the recent period. Over the last few years, this research area has been developing extremely rapidly. A representative overview of the best out of the related studies can be found in the article by McDaniel (2016), the paper by Chakraborty (2018), and the book by Munoz-Gonzalez (2019).

3. Supervised learning

For the purpose of this article, the attacks on machine learning systems will be restricted to supervised learning within the classification task. This choice stems from the universality of the use of this approach in practical business applications.

The classification task is to find function Ψ, which will assign object x represented by a vector of features to class i from the set of class labels $M = \{1, 2, \ldots, M\}$[1]. Such a projection of features' space into the set of class labels is called a classification model (classifier) and formally has the following form (Kurzyński, 1997):

$$\Psi(x) = I, \text{ where } i \subset M.$$

In practical applications when a priori probabilities of classes and conditional probabilities of features in classes are unknown, a process of learning with the use of a learning set is applied. A learning set $D^{(\text{learning})}$ consisting of N elements is defined as follows:

$$D^{(\text{learning})} = \{(x_1, j_1), (x_2, j_2), \ldots, (x_N, j_N)\}, \text{ where } j \subset M.$$

It can be informally stated that the learning set is created thanks to the assistance of a 'teacher', which assigns every object x in the learning set to class j (it is for this reason that this approach is called supervised learning). A classifier (a model) constructed with the use of the learning set has the following form:

$$\Psi\,(x, D^{(\text{learning})}) = i, \text{ where } i \subset M.$$

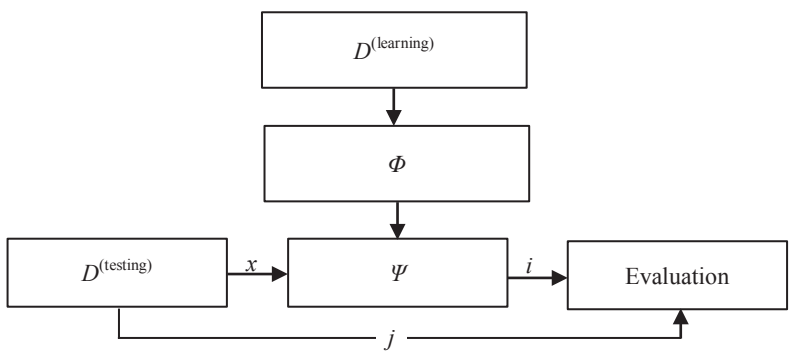

Fig. 1. Classifier construction

Source: based on (Surma, 2011).

The process of learning, aiming to build classifier Ψ, is implemented using a chosen algorithm of supervised learning Φ:

$$\Psi \leftarrow \Phi(D^{(\text{learning})}).$$

[1] In the case of the regression task, set M is a set of real numbers.

A process of testing, aiming to assess the quality of a classifier, is implemented using testing set $D^{(\text{testing})}$, which has the same structure as the learning set. The process of testing consists in a comparison of result $\Psi(x)$ to j for every element $(x, j) \subset D^{(\text{testing})}$. The process of learning, testing and classifier evaluation is given in Figure 1.

4. Attacker objectives

An intentional attack on supervised machine learning systems can have the following objectives:

1. **General misclassification** through generating errors in order to reduce classification accuracy, which implies a reduction of confidence in a system and even the termination of its use in extreme cases. This is due to the fact that wrong classifications generate real and potential (e.g. connected with a loss of reputation) costs.

2. **Targeted misclassification** through obtaining an erroneous classification for specified objects. In such situation a classifier incorrectly classifies a specific object or a set of objects in accordance with the intention of an attacker. In this approach, the attacker is interested in the quality of the classifier being on the proper high level and thereby inspiring confidence in users.

3. **Availability limitation** i.e. obtaining an unacceptably long system reaction time to input data, and in extreme cases the termination of the working classifier.

5. Attack on a supervised learning system

5.1. Theoretical framework

From the formal point of view, the problem of learning systems security can be described as a game between the attacker (a criminal) and the defender (a user) of the system. This type of a game can be formalised through (Huang et al., 2011):

- Φ – an algorithm of machine learning,
- $A^{(\text{learning})}$ – a procedure of the attacker violating the integrity of the learning set,
- $A^{(\text{testing})}$ – a procedure of the attacker violating the integrity of the testing set.

For such defined variables, the game is as follows:

1. The defender (the user) chooses Φ in order to construct the classifier Ψ with the use of the known data: $D^{(\text{learning})}$ and $D^{(\text{testing})}$.

2. The attacker (the criminal) chooses an attack scenario (potentially with the knowledge on Φ): $A^{(\text{learning})}$ or $A^{(\text{testing})}$.

3. The learning process:

a. Using $A^{(\text{learning})}$ generate an 'infected' learning set $D^{*(\text{learning})}$,

b. Build the classifier: $\Psi^* \leftarrow \Phi\,(D^{*\,(\text{learning})})$.

4. The testing process:
a. Using $A^{(\text{testing})}$ generate 'infected' testing set $D^{*\,(\text{testing})}$,
b. Compare the result $\Psi^{*}(x)$ to j for every element of the set $(x, j) \subset D^{*\,(\text{testing})}$.

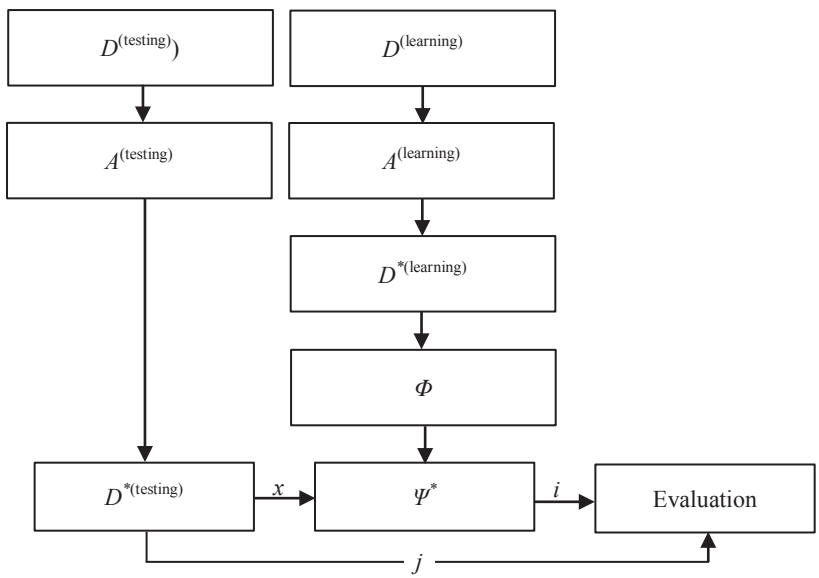

Fig. 2. Attack scenarios on classifier

Source: based on (Huang et al., 2011).

In this type of game (see Figure 2), the defender chooses Φ so as to obtain the best quality of classification for the known $D^{(\text{learning})}$ and $D^{(\text{testing})}$ and simultaneously without having knowledge on $A^{(\text{learning})}$ and $A^{(\text{testing})}$. However, the attacker tries to reduce classification accuracy through the proper fitting of $A^{(\text{learning})}$ or $A^{(\text{testing})}$.

5.2. Attack on a classifier construction

Within the attack on the process of classifier construction (learning, update and testing), it is possible to violate the integrity of both the learning set, testing set and the process of learning/update and testing. Therefore the following taxonomy of attacks was proposed by Surma (2020):

1. **Poisoning attack** – an attack on the learning/update process based on an intrusion in the learning set:

- Using $A^{(\text{learning})}$ generate 'infected' learning set $D^{*(\text{learning})}$,
- Build the classifier: $\Psi^{*} \leftarrow \Phi\,(D^{*\,(\text{learning})})$.

The procedure of the attacker violating the integrity of the learning set $A^{(\text{learning})}$ can be implemented through:

- Data injection – carried out through adding false examples to the learning set and also through data modification or removing existing elements from the learning set.
- Data manipulation – carried out through affecting the structure of the learning set both through adding, modifying or removing a feature of vector x and also label modification of class j.

2. **Invasion attack** – an attack on the process of testing, which consists in intrusion into the testing set:

- Using $A^{(\text{testing})}$ generate 'infected' testing set $D^{*\,(\text{testing})}$,
- Compare the result $\Psi(x)$[2] to j for every element of the set $(x, j) \subset D^{*\,(\text{testing})}$.

The procedure of the attacker violating the integrity of the testing set $A^{(\text{testing})}$ can be implemented through data injection as was described in the case of a poisoning attack.

- **Model logic corruption** – an attack carried out directly on model Ψ so as to obtain a 'fake' version of Ψ^*. This kind of an attack can occur in a situation when the user unknowingly uses machine learning algorithm Φ, which has been downloaded from infected programming environments (Chakraborty, Alam, Dey, Chattopadhyay, and Mukhopadhyay, 2018).

5.3. Attack on a working classifier

The classifier after construction can also be subject to attacks. Taking into account this range of information, the knowledge of the attacker can be divided into three groups: complete knowledge (a white box attack), partial knowledge (a grey box) and lack of knowledge (a black box). In each of these cases, even when there is a complete lack of knowledge, effective actions on the part of the attacker are possible. In the case of acquiring complete knowledge, one talk about a model extraction attack, which means gathering information covering:

- object x together with its specification of features;
- set of class labels M;
- classifier Ψ;
- algorithm Φ with its parameters;
- learning set $D^{(\text{learning})}$ and update methods applied;
- testing set $D^{(\text{testing})}$ and testing methods applied;
- libraries used and the programming environments;
- the context of the use of the system: intentions, purpose of using the system, work organisation, committed employees, clients, etc.

In the case of acquiring complete or partial knowledge, the attacker has the opportunity for reconstructing the classifier's model. For instance, acquiring knowledge of the learning set and of the programming environment used can enable an independent construction of a proper model and an analysis of its vulnerabilities.

[2] If the poisoning attack is followed by an invasion attack, then $\Psi(x) = \Psi^*(x)$.

6. Black box attack

In a situation of the lack of any knowledge, the system is treated as a black box. In this case the classifier is the subject of experiments aiming to examine what will be the its reaction to input data x. From the formal point of view, this is an identification task which aims to construct a model of the system on the basis of experimental research, i.e. measurement data collected from an input and output of the identified system (Bubnicki, 1974). The collection of an appropriate number of input-output pairs enables obtaining learning set $D^{(\text{learning})}$ and the construction of a substitute of real classifier Ψ by the attacker. In general, the attacker can carry out an active experiment, which is the result of the free choice of the experimenter. Otherwise, one talk about a passive experiment, in which input-output data constitute the result of object measurements in a normal operation mode. In the case of the active experiment, this requires access to the targeted system of the attacker. For example, an attack on the object detection system in Tesla cars was carried out using a real car with an autopilot function which was openly bought by hackers (Brewster, 2019). In practice, the passive experiment is more real when the attacker can observe the targeted system and collect data. Naturally, in such an event the collection of a proper learning set will be significantly more laborious and the obtained database will be typically of a lower quality. After creating the substitute of the classifier, the attacker can develop adversarial examples that will be used in the attack on an actual system.

7. Conclusion

Supervised machine learning systems are currently the most popular in business applications (Surma, 2011). It is reasonable to say that the most cost-effective hacking strategy on that kind of system is a black box attack on a working classifier. This approach is based on an attempt to construct a substitute of an actual model with the use of the passive experiment (see Section 6). This results from the fact that the poisoning attack or invasion attack (see Section 5) are very difficult to implement in real-life situations and can generate disproportionately high costs compared to the potential benefits.

An analysis of profitability of the attack, namely comparing the costs of carrying out an attack with the subsequent profits is an important managerial issue. This will enable a risk analysis, and finally a determination of the most probable attack vectors. Obviously, this approach is adequate for organized crime groups governed by the profit motivation, which is not the motivation for state actors.

In conclusion, it must be clearly underlined that black box attacks on machine learning systems are currently the most probable situation. Based on this knowledge, appropriate defence strategies must be developed in companies in order to mitigate this kind of reputational and operational risk. In this business context, research on secure and robust machine learning systems is one of today's most significant challenges.

References

Chakraborty, A., Alam, M., Dey, V., Chattopadhyay, A. and Mukhopadhyay, D. (2018). *Adversarial attacks and defences: a survey*. Retrieved from arXiv:1810.00069

Barreno, M., Nelson, B., Sears, R., Joseph, A., Tygar, J. (2006). *Can machine learning be secure*? In ASIACCS'06, 16-25.

Barreno, M. Nelson, B., Joseph, A., and Tygar, J. (2010). The security of machine learning. *Machine Learning*, *81*(2), 121-148.

Brewster, T. (2019). Hackers use little stickers to trick tesla autopilot into the wrong lane. *Forbes Magazine*, April 1.

Bubnicki, Z. (1974). *Identyfikacja obiektów sterowania*. Warszawa: PWN.

Dalvi, N., Domingos, P., Sumit, M., and Verma, D. (2004). *Adversarial classification* (Proceedings of the tenth ACM SIGKDD international conference on Knowledge Discovery and Data Mining (KDD'04), pp. 99-108) ACM Press.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). *Generative Adversarial Networks*, arXiv:1406.2661

Huang, L., Joseph, A., Nelson, B., Rubinstein, B., Tygar, J. (2011). *Adversarial machine learning* (Proceedings of the 4th ACM workshop on Security and Artificial Intelligence (AISec '11), pp. 43-58), ACM Press.

Kurzyński, M. (1997). *Rozpoznawanie obrazów. Metody statystyczne*. Wrocław: Oficyna Wydawnicza Politechniki Wrocławskiej.

Laskov, P., and Lippmann, R. (2010). Machine learning in adversarial environments. *Machine Learning*, *81*(2), 115-119.

Muńoz-González, L. (2019). The security of machine learning systems. In L.F. Sikos (Ed.), *AI in cybersecurity* (pp. 47-79). Springer.

Nelson, B. (2010). *Behavior of machine learning algorithms in adversarial environments* (Technical Report No. UCB/EECS-2010-140. Electrical Engineering and Computer Sciences). University of California at Berkeley.

McDaniel, P., Papernot, N., and Celik, Z. (2016). Machine learning in adversarial settings. *IEEE Security & Privacy*, May/June, 68-72.

Papernot, N, McDaniel, P., Goodfellow, I., Jha, S., Celik, Z., and Swami, A. (2017). *Practical Black--Box Attacks against Machine Learning* (Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security (ASIA CCS '17), pp. 506-519). ACM: New York, NY, USA.

Shankar, R., Nyström, M., Lambert, J., Marshall, A., Goertzel, M., Comissoneru, A., Swann, M., and Xia, S. (2020). *Adversarial machine learning – industry perspectives*. Retrieved from arXiv:2002.05646

Surma, J. (2011). *Business intelligence: making decisions through data analytics.* New York: Business Expert Press.

Surma, J. (2020). *Hacking machine learning: towards the comprehensive taxonomy of attacks against machine learning systems* (ICIAI 2020: ACM Proceedings of the 2020 the 4th International Conference on Innovation in Artificial Intelligence, pp. 1-4).