Some Applications of Machine Learning in Cryptography

Jollanda Shara
Dept.Mathematics&Computer Science
University “Eqrem Cabej”
Gjirokaster, Albania
e-mail jokrisha@yahoo.com

***Abstract*— In the 1940’s and 50’s the computer science made great progress relying on some theoretical developments of the 1930’s. The cryptography and machine learning, from the very beginning, were tightly related with this new technology. Cryptography, on the one hand, played an important part during the World War II, where some computers of that time were destined to accomplish cryptanalytic tasks. On the other hand, many authors, such as Turing, Samuel, etc. examined the possibility that computers could “learn” to perform tasks.**

Machine learning techniques have had a long list of applications in recent years. However, the use of machine learning in information and network security is not new. Machine learning and cryptography have many things in common. The most apparent is the processing of large amounts of data and large search spaces.

In its varying techniques, machine learning has been an interesting field of study with massive potential for application. In general, machine learning and cryptanalysis have more in common that machine learning and cryptography. This is due to that they share a common target; searching in large search spaces. A cryptanalyst’s target is to find the right key for decryption, while machine learning’s target is to find a suitable solution in a large space of possible solutions. In addition to cryptography and cryptanalysis, machine learning has a wide range of applications in relation to information and network security. In these notes, we underline some of them.

Keywords—cryptography; application; machine learning; data.

I. INTRODUCTION

Cryptography is the process that involves encryption and decryption of text using various mechanisms or algorithms. A cryptographic algorithm is a mathematical function that can be used in the process of encryption and decryption.

Cryptography is a technique used today hiding any confidential information from the attack of an intruder. Today data communication mainly depends upon digital data communication, where prior requirement is data security, so that data should reach to the intended user. The protection of multimedia data, sensitive information like credit cards, banking transactions and social security numbers is becoming very important. The protection of these confidential data from unauthorized access can be done with many encryption techniques. So, for providing data security many cryptography techniques are employed, such as symmetric and asymmetric techniques. [8] Cryptography enables the user to transmit confidential information across any insecure network so that it cannot be used by an intruder.

Machine Learning (ML) is a branch of AI and is closely related to (and often overlaps with) computational statistics, which also focuses on prediction making using computers. It has strong ties to mathematical optimization, which delivers methods, theory and application domains to the field. ML is occasionally conflated with data mining, but the latter subfield focuses more on exploratory data analysis and is known as unsupervised learning. ML can also be unsupervised and be used to learn and establish baseline behavioral profiles for various entities and then used to find meaningful anomalies. The pioneer of ML, Arthur Samuel, defined ML as a ``field of study that gives computers the ability to learn without being explicitly programmed." ML primarily focuses on classification and regression based on known features previously learned from the training data. [4]

As it is pointed out in [3], the area of Machine Learning deals with the design of programs that can learn rules from data, adapt to changes, and improve performance with experience. In addition to being one of the initial dreams of Computer Science, Machine Learning has become crucial as computers are expected to solve increasingly complex problems and become more integrated into our daily lives.

Machine Learning Theory, also known as Computational Learning Theory, aims to understand the fundamental principles of learning as a computational process. This field seeks to understand at a precise mathematical level what capabilities and information are fundamentally needed to learn different kinds of tasks successfully, and to understand the basic algorithmic principles involved in getting computers to learn from data and to improve performance with feedback. The goals of this theory are both to aid in the design of better automated learning methods and to understand fundamental issues in the learning process itself.

Machine Learning is a field of research that focuses on extracting information from datasets.

If the dataset is very large, it is also often referred to as Big Data or Data Mining. There are countless algorithms in Machine Learning with inputs ranging from numeric over categorical to text-based. The applications today seem endless: We have the first self-driving cars, which have learned to do this via Neural Networks, we have smartphone keyboards that predict the next word based on your individual writing style, researchers are working on algorithms that can predict illness from a set of measured attributes or even a persons genome, and many more. However, many of these application scenarios involve sensitive data : people do not feel safe sending e.g. their medical data to a service provider, because they either do not trust the provider or are worried about a data breach even if they do trust the provider. This has lead to Machine Learning being a popular topic in the context of privacy-preserving computations in general, and Fully Homomorphic Encryption in particular.

Generally, Machine Learning can be divided into two categories: supervised and unsuper-
vised learning.

II. SOME APPLICATIONS OF MACHINE LEARNING

As we mentioned above, Machine Learning can be of great help in producing useful information from extremely large amounts of data.

Classification is one of the most widely used applications of machine learning… (E. Alpaydin, 2014). A common example of classification is the classification that banks use for loans; low-risk, and high-risk.

Another example of machine learning applications is regression. Regression is the type of problem that produces a number based on multiple inputs…The output would be a specific number driven from the inputs. However, there has to be training that would enable the system to be more accurate gradually and to learn the impact of change in each of the input elements.

Learning associations is also one of the applications of machine learning. For example, analysis of shopping baskets data can produce useful information to supermarkets that they can use to improve their sales.

Unsupervised learning can also be used in machine learning. When there is no reference output to compare to, the learning is done through input data only. This type of learning or training is called unsupervised.

Reinforcement learning can also be used in machine learning applications. In certain applications, the output of the system is a sequence of actions. A single action by itself is not important, but of importance is the sequence of the correct action. [2]

Machine Learning happens when we need a machine(computer) to learn to solve a problem based on, usually large amounts of, data previously fed into the machine. Machine Learning can be a good solution finder for problems for which we do not have a clear algorithm to solve. For example, when we want the computer to be able to detect spam emails, there is not clear algorithm that is 100% accurate in finding spam. Hence, Machine Learning can be a closer to optimum solution when we feed the machine hundreds or thousands of spam and non-spam examples. Gradually, the machine will learn to us more and more accurate in detecting spam emails. The larger the data used for learning becomes, the more accurate the classification becomes. (E. Alpaydin, 2014)

With the increasingly in-depth integration of the Internet and social life, the Internet is changing how people learn and work, but it exposes us to increasingly serious security threats, as well. A key issue, which must be solved immediately, is how to identify various network attacks, particularly not previously seen attacks.

Cybersecurity is a set of technologies and processes designed to protect computers, networks, programs and data from attacks and unauthorized access, alteration, or destruction… (S. Aftergood, 2017). A network security system consists of a network security system and a computer security system. Each of these systems includes firewalls, antivirus software, and intrusion detection systems (IDS). IDSs help discover, determine and identify unauthorized system behavior such as use, copying, modification and destruction [9].

In the intervening forty years, the field of computer and network security has come to encompass an enormous range of threats and domains: intrusion detection, web application security, malware analysis, social network security, advanced persistent threats, and applied cryptography, and these are only a few of them.. But even today spam remains a major focus for those in the email or messaging space, and for the general public spam is probably the aspect of computer security that most directly touches their own lives. [6]

Machine learning was not invented by spam fighters, but it was quickly adopted by statistically inclined technologists who saw its potential in dealing with a constantly evolving source of abuse. Email providers and Internet service providers (ISPs) have access to a wealth of email content, metadata, and user behavior. Leveraging email data, content-based models can be built to create a generalizable approach to recognize spam. Metadata and entity reputations can be extracted from email to predict the likelihood that an email is spam without even looking at its content. By instantiating a user behavior feedback loop, the system can build a collective intelligence and improve over time with the help of its users.

Email filters have thus gradually evolved to deal with the growing diversity of circumvention methods that spammers have thrown at them. Even though 86% of all emails sent today are spam (according to one study, see [10]) the best spam filters today block more than 99.9% of all spam, (see [11]) and it is a rarity for users of major email services to see unfiltered and undetected spam in their inboxes. These results demonstrate an enormous advance over the simplistic spam filtering techniques developed in the early days of the Internet, which made use of simple word filtering and email metadata reputation to achieve modest results.(see [12])

Computer systems and web services have become increasingly centralized, and many applications have evolved to serve millions or even billions of users. Entities that become arbiters of information are bigger targets for exploitation, but are also in the perfect position to make use of the data and their user base to achieve better security. Coupled with the advent of powerful data crunching hardware, and the development of more powerful data analysis and Machine Learning algorithms, there has never been a better time for exploiting the potential of Machine Learning in security.(see [6])

III. ML AND CRYPTOGRAPHY

"…Machine Learning and cryptanalysis can be viewed as "sister fields", since they share many of the same notions and concerns. In a typical cryptanalytic situation, the cryptanalyst wishes to "break" some cryptosystem. Typically this means he wishes to find the secret key used by the users of the cryptosystem, where the general system is already known. The decryption function thus comes from a known family of such functions (indexed by the key), and the goal of the cryptanalyst is to exactly identify which such function is being used. He may typically have available a large quantity of ciphertext and plaintext to use in his analysis. This problem can also be described as the problem of "learning an unknown function" (that is, the decryption function) from examples of its input/output behavior and prior knowledge about the class of possible functions…" [1]

1. "…The notion of "secret key" in cryptography corresponds to the notion of "target function" in machine learning theory, and more generally, the notion of "key space" in cryptography corresponds to the notion of the "class of possible target functions".
2. "…A critical aspect of any cryptanalytic or learning scenario is the specification of how the cryptanalyst (learner) may gather information about the unknown target function… Even if information is gathered from random examples, cryptanalytic/learning scenarios may also vary in the prior knowledge available to the attacker/learner about the distribution of those examples…" [1]

Prior work at the intersection of machine learning and cryptography has focused on the generation and establishment of cryptographic keys (Ruttor, 2006; Kinzel & Kanter, 2002), and on corresponding attacks (Klimov et al., 2002). [5]

Machine Learning Theory also has a number of fundamental connections to other disciplines. In cryptography, one of the key goals is to enable users to communicate so that an eavesdropper cannot acquire any information about what is being said. Machine Learning can be viewed in this setting as developing algorithms for the eavesdropper. In particular, provably good cryptosystems can be converted to problems one cannot hope to learn, and hard learning problems can be converted into proposed cryptosystems. Moreover at the technical level, there are strong connections between important techniques in Machine Learning and techniques developed in Cryptography. For example, Boosting, a Machine Learning method designed to extract as much power as possible out of a given learning algorithm, has close connections to methods for amplifying cryptosystems developed in cryptography. [3]

Machine Learning and Cryptography have many things in common: the amount of data to be handled and large search spaces for instance. The application of Machine Learning in Cryptography is not new, but with over 3 quintillion bytes of data being generated every day, it is now more relevant to apply Machine Learning techniques in cryptography than ever before. Machine Learning generally automates analytical model building to continuously learn and adapt to the large amount of data being fed as input. Machine Learning techniques can be used to indicate the relationship between the input and output data created by cryptosystems. Machine Learning techniques such as Boosting and Mutual Learning can be used to create the private cryptographic key over the public and insecure channel. Methods such as Naive Bayesian, support vector machine, and AdaBoost, which come under the category of classification, can be used to classify the encrypted traffic and objects into steganograms used in steganography. Besides the application in cryptography, which is an art of creating secure systems for encrypting/decrypting confidential data, the Machine Learning techniques can also be applied in cryptanalysis, which is an art of breaking cryptosystems to perform certain side-channel attacks.

Another arena in which cryptography and machine learning relate is that of data compression. It has been shown by Blumer et al. that pac-learning and data compression are essentially equivalent notions. Furthermore, the security of an encryption scheme is often enhanced by compressing the message before encrypting it. Learning theory may conceivably aid cryptographers by enabling ever more effective compression algorithms. (see [1])

IV. CRYPTOGRAPHY AND NEURAL NETWORKS

Cryptanalysis has been an area of great research interest in the past decade owing to advancements in

Machine Learning algorithms, particularly in neural networks. The process of

discovering the plaintext from a ciphertext without knowing any information about the system or the key that was used to encrypt the plaintext is called cryptanalysis. Any mode of communication is secure only as long as the cryptographic system that encrypts the messages between the sender and the receiver is strong. Once a third party listening in on the communication channel is able to decipher the encrypted texts, the cipher system is said to have flaws and to be broken. All ciphers are vulnerable to brute-force attacks in that the attackers try to break the cipher system by exploring its key space. Though this takes a lot of time and computational power, it is possible to break the system.

B. Chandra and P. P. Varghese, (2007), used neural networks to classify the ciphertext based on the algorithm that was used to encrypt it. They had used Cascade Correlation Neural Network and Back Propagation Network to identify the cipher systems. For training they had used ciphertexts obtained from Enhanced RC6, a block cipher, and from SEAL, a stream cipher. They had used different types of datasets with same keys, different keys, same sets of plaintexts, different sets of plaintexts etc. and concluded that cascade correlation worked better than the back propagation method.

Another author,(Alani MM, 2012), had come up with an idea to break Data Encryption Standard (DES) cipher using neural network. The author had used the known-plaintext attack to arrive at the plaintext. The algorithm used by the author does not seem to attempt to find the key, but rather tries to directly find the plaintext. Though this approach is not considered to be a cryptographic attack, the work of the author is commendable as the author had designed a neural network for the process of identifying the plaintext using the same plaintext and ciphertext of the same key.

In the paper written by Albassal and Wahdan, (2004), the authors have described how they were able to use neural networks to break a hypothetical Feistel cipher, called HypCipher. The round function for the HypCipher had been chosen from the Advanced Encryption Standard (AES). The back propagation technique has been used showing success with 2 and 3 rounds of the cipher. An additional hidden layer had been added for 4 rounds. The model was successful in that it used a simple neural network with a simple activation function like the sigmoid function. The authors have proposed to use a distributed system to attack ciphers with more rounds.[7]

Let us consider a neural network with several components, and suppose that we wish to guarantee that one of the components does not rely on some aspect of the input data, perhaps because of concerns about privacy or discrimination. Neural networks are notoriously difficult to explain, so it may be hard to characterize how the component functions. A simple solution is to treat the component as an adversary, and to apply encryption so that it does not have access to the information that it should not use. Classical cryptography may be able to support some applications along these lines. In particular, homomorphic encryption enables inference on encrypted data (Xie et al., 2014; Gilad-Bachrach et al., 2016). On the other hand, classical cryptographic functions are generally not differentiable, so they are at odds with training by stochastic gradient descent (SGD), the main optimization technique for deep neural networks. Therefore, we would have trouble learning what to encrypt, even if we know how to encrypt. Integrating classical cryptographic functions—and, more generally, integrating other known functions and relations (e.g., (Neelakantan et al., 2015))—into neural networks remains a fascinating problem.[5]

V.CONCLUSIONS

In this paper, we have described, briefly, the relationship of ML and Cryptography. It is known that there exists a wide range of applications of ML in Cryptography and this range is becoming more and more larger in our times. We have separated here some of them trying to launch a beam of light to the throng of these applications.

ACKNOWLEDGMENT *(Heading 5)*

I want to express my gratitude to many authors such as [1,2,3] etc. for their excellent help and inspiration they offer to me in this research.

REFERENCES

[1] Ronald L.Rivest, “Cryptography and Machine Learning”.

[2] Mohammed M. Alani, “Applications of Machine Learning in Cryptography: A Survey”, 2019.

[3] Avrim Blum, Machine Learning Theory, Carnegie Mellon University, Department of Computer Science.

[4] Yang Xin, Lingshuang Kong, Zhi Liu , Yuling Chen, Yanmiao Li, Hongliang Zhu, Mingcheng Gao, Haixia Hou, Chunhua Wang, “Machine Learning and Deep Learning Methods for Cybersecurity”, 2018.

[5] Martin Abadi and David G. Andersen, “Learning to Protect Communications with Adversarial Neural Cryptography”.

[6] Clarence Chio and David Freeman, “Machine Learning and Security”, 2017.

[7] Kowsic Jayachandiran, “A Machine Learning Approach for Cryptanalysis”, RIT Computer Science.

[8] Ravi K. Sheth, Sarika P. Patel, “Analysis of Cryptography Techniques”, International Journal of Research in Advance Engineering, Volume-1, Issue-2, 2015.

[9] A. Milenkoski, M. Vieira, S. Kounev, A. Avritzer and B. D. Payne, “Evaluating computer intrusion detection systems: A survey of common practices”, ACM Comput. Surv., vol. 48, no. 1, pp. 1-41, 2015.

[10] https://www.bloomberg.com/ news/articles/2016-01-19/e-mail-spam-goesartisanal

[11] https://www.wired.com/2015/07/ google-says-ai-catches-99-9-percentgmail-spam/

[12] http://www.paulgraham.com/spam.html