Applying artificial intelligence in cyber security :
Malware detection using machine learning based approach

Introduction :

What the Security industry needs in term of algorithms is huge. Algorithms that can grow and keep on the road with the tsunami of cyberattacks its facing, just as the number of the devices hooked up to the internet is exponentially growing .At the same time there is the problem of skills of cyber workers that simply quit or switch to an other field because of the continuous stress and challenge .

All that makes the AI the new hope of Cyber Security to continue to exist by using the intelligent algorithms that proved to give better performance in different fields and even beat the world's leading expert. Those machine learning and deep learning algorithms can help automate better than any old traditional software-driven, since it approaches the whole process of cyber security from threat detection and identification to the threat to the threat deletion and evacuation.

Before talking about how AI and its statistical approaches can be used in the field of cyber security. I will first introduce those key words, and how they emerge as saviors in so many other fields as the best alternative to traditional ways. Then I will introduce some real world use cases along with cyber security projects from around the world and I will finish with a brief conclusion.

1- An overview of Artificial Intelligence and machine learning :

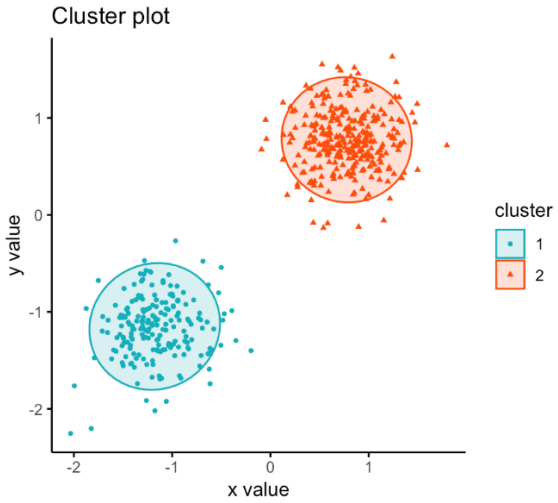
According to Encyclopedia Britannica , Artificial Intelligence (AI) is the ability of a digital computer or computer-controlled robot to perform tasks commonly associated with intelligent beings that can adapt to changing circumstances in their environment which means giving machines the ability to learn from experience , adjust to new inputs and perform human-like tasks .
from the 50s to the 80s Symbolic AI took the lead of that new field in computer science ,symbolic AI is based on system and symbolic representation of knowledge or "facts" which simply means that all is symbols and variables in the memory .and then after the 80s connectionist AI took the lead after the rise of neural networks and the high performance it shows in performing tasks without all that hardcoding that symbolic AI required, since it relies on coding all the universal facts so that the machine can learn how to do human-like reasoning to infer and generate new knowledge .
Connectionist AI relies on detecting patterns in huge amounts of data using statistical and probabilistic methods and technics.

Machine learning:

According to the classic definition given by AI pioneer Arthur Samuel , Machine learning is a set of methods that gives "computers the ability to learn without being explicitly programmed" .In other

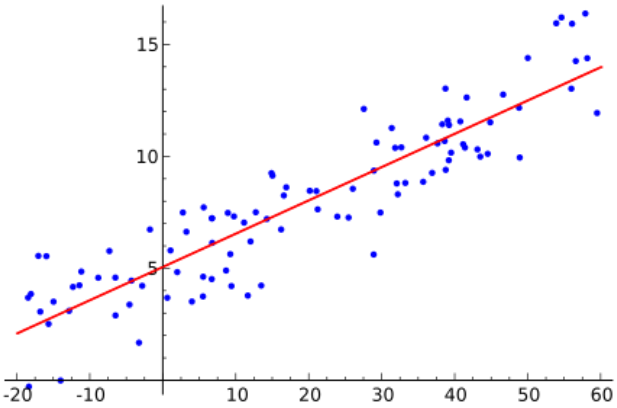
words , a machine learning algorithm discovers and formalizes the principals that underlie the data it sees .With this knowledge , the algorithm can reason the properties of previously unseen samples ,for example in malware detection an unseen sample could be a new file and its hidden properties could be malware or benign .A mathematically formalized set of principles underlaying data properties is called the model .
Machine learning has a variety of approaches from neural networks, statistics, operation research and physics in order to find or in better words “predict” the suitable solution to the problem.

One machine learning approach is **unsupervised learning** .In this setting we are given the data without the right answers for the tasks, the purpose of this approach is then to discover the structure of the data or the law behind the data generation, in other words the patterns that describes this set of data .An example of unsupervised algorithms is clustering that means splitting the data into groups of similar objects (communities that share common things) .
In cyber world large datasets are available and the cost of doing manual labeling by experts is very high, this makes unsupervised valuable for labeling new samples using Clustering for example.



Clustering the data in two clusters

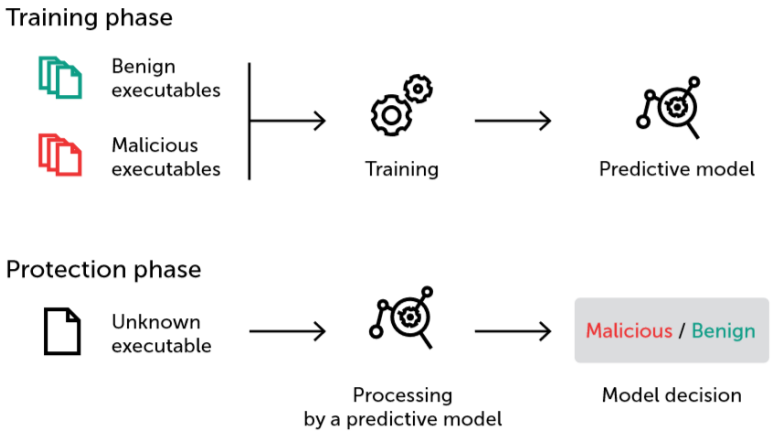
Another approach of machine learning is **supervised learning** that relies on labeled data wish means both data and the right answers for each object are available and the purpose is to find the model that will produce the right answers for new objects and samples.
This approach consists of two important steps :
1- training a model and fitting it to the available data , a simple basic example is linear regression and training means finding the coefficients of the line that passes from almost all our training objects .



A linear regression model.

2-Applying the trained model to new samples and obtaining predictions.

In more easy words we are given a set of objects, each one is represented with feature set X and mapped to right answer as Y. The training step is used to find the best model by selecting some family of models ,for example, neural networks or decision trees and see which one of those models predict with more accuracy the correct label Y for previously unseen objects given the feature X .
And in the case of malware detection , X could be some feature of file content or behavior .Label Y could be “malware” or “benign” or even a more fine-gained classification, such as virus, Trojan or adware .



Machine learning training and detection (applying) lifecycle

2- Machine Learning application specifics in cyber security:

Applying machine learning in the field of cyber security, for malware detection for example, has some specifics. Because in order to have the right model that will fit to the data you have and make the best prediction each time.

- we need to have a large representative dataset that will enable the model during the training step to extract the relevant features for prediction. Having small datasets may lead to predict that all files more than 10MB for example are malwares which is not true and when the model will be applied to real world data it will many false positives. To prevent that we need to add more benign files with larger sizes to the training step.
- Our model must be interpretable and that means we should use families of models that can easily lead us to conclude wish of the features causes or contributed to the decision of telling that the input file is a malware or benign. The families of models like deep neural networks or random forest are called Black Box Models because they produce the label Y of a given input X through a complex sequence of operations that can hardly be interpreted by humans .This could present a problem in real-life applications because we need to know what made the model decide that such file is a malware or benign se we can easily scale our machine learning driven malware detection.
- False positive rates must be low. False positive happen when a model mistakes a malicious label for a benign file . In real-life applications false positive rate must be very low Because one false positive in a million benign can create serious consequences for user. And this is very complicated because there are millions of new clean files that are been generated every day by users. To prevent that our model must be flexible and allow to fix false-positive on the fly.
- The model must allow a quick adaptation to malwares changes and evolution. Most of machine learning algorithms regularly work under the assumption of fixed data distribution which means that the model predict the label Y of any input data X based on the training set and that supposed to be fixed .For malware case it is wrong and a bad mistake to do because malware writers constantly work on avoiding detections and releasing new versions of malwares files that differ significantly from those that have been seen during the training step .And also thousands of new benign files are produced daily that the model didn't see during the training step too .To avoid those problems we should retrain our model regularly so it can be able to recognize new malware and benign files .

3- Implementing a machine learning model to detect malwares:

For the machine learning to implement a model that will be able to detect malware and benign files it needs to go along the following steps :

- First of all , we need to collect malware and benign samples ,and as we have seen previously we need to have lots of examples to prevent any false positive predictions .Thanks to open source community this step is quite easy , malwares dataset can be found here : https://zeltser.com/malware-sample-sources/ or buy one from Virus Total.
- The second important step in the workflow is featured extractions .This step depends on the file format you are using and what kind you can get from it .For PE i.e. Portable Executable file format files for example, the features you can extract can be :
    - Virtual Address
    - Os version
    - Resources size
    - Number of sections

    - Linker version
    - Size of stack reserve
    - Major image version
- Then comes the machine learning application. After preparing a dataset containing both types of files (malicious and clean) we split it into two parts one for training (generally 70% of the data) and an other (the rest of examples) as a testing dataset that we will use later after training the model to test the performance of our model and how able it is to recognize malicious files .
  We can use one of the classification algorithms (Random forests, SVM ,Logistic regression , neural networks,…etc) and for this kind of classification problems its recommended to use Random Forests because they are quite performing. For further details about the performance of classification algorithms please refer to this paper.
- After getting the model we can test its performance with the training dataset and see how many malicious files it can detect and measure the same performance as well. A performance that we can enhance by adding more features or even sometimes reduce them or change the classier (algorithm used for the classification task).

Conclusion:

I kept the article more general and I didn't dig more in the coding aspect of the issue nor real implementation, because the purpose was to introduce the use of AI and Machine learning based method in cyber security and in malware detection. For more details about coding this task, please do check the references where you can find some interesting projects and links.

Bibliography :

Kaspersky Lab Whitepaper : Machine Learning for malware detection

References:

[1] https://resources.infosecinstitute.com/machine-learning-malware-detection/

[2] https://www.technologyreview.com/s/611860/ai-for-cybersecurity-is-a-hot-new-thing-and-a-dangerous-gamble/

[3] https://eugene.kaspersky.com/2016/09/26/laziness-cybersecurity-and-machine-learning/

[4] https://www.datasciencecentral.com/profiles/blogs/artificial-intelligence-vs-machine-learning-vs-deep-learning

[5] https://www.forbes.com/sites/bernardmarr/2018/02/14/the-key-definitions-of-artificial-intelligence-ai-that-explain-its-importance/#7159b1044f5d

[6] https://resources.infosecinstitute.com/2-malware-researchers-handbook-demystifying-pe-file/

[7] https://github.com/obarrera/Machine-Learning-Malware-Detection/blob/master/Machine%20Learning%20Malware%20Detection%20/Machine%20Learning%20Malware%20Detection%20.md

[8] https://github.com/akash14204/Machine-learning-Malware-Detection-Final-year-Project

About the Author:

Zakaria EL BAZI

A computer sciences student at the National Institute of Statistics and applied economy in Rabat, Morocco. Artificial intelligence enthusiast ,Pentester and Graphic Designer. Interested in #algorithms, #data_structures, #computer_security, #new_technologies, and #artificial_intelligence.
website : https://elbazi.me Portfolio : http://www.elbazi.tk