OSSIR : *JSSI*
12 mars

Enseignant-Chercheur
R. ERRA — ESIEA

Machine Learning et Cybersécurité: *Un (Petit) Tour d'Horizon*

esiea
INGÉNIEUR·E·S D'UN NUMÉRIQUE UTILE

Mars 2024

Table des matières

1 Introduction

1 Introduction

1.1 Disclaimer

Disclaimer.

- Cette conférence s'adresse à des non spécialistes de l'Intelligence Artificielle (IA)
- L'IA va t'elle bouleverser la Cybersécurité ? Si oui, pourra-t-elle en même temps permettre d'inventer de nouvelles approches/défenses (vision optimiste) ?
- *Une franche réponse à cette question est : oui !*
- **Objectif Un** : Présenter (rapidement) quelques algorithmes de l'IA
- **Objectif Deux** : Préciser les problèmes à résoudre quand on a un projet
- **Objectif Trois** : présenter un exemple.

— **Alan TURING.** *If a machine is expected to be infaillible, it cannot also be intelligent.*

Il y a 5 . . .ans : MISC HS Numéro 18.

- Il suffisait de se promener dans les allées du FIC 2018 pour se rendre compte qu'un invité de marque était présent sur pas mal de stands : le *Machine Learning*, soit ce qu'on appelle l'apprentissage automatique ou mieux : **l'apprentissage machine (***Machine Learning***)** : la star actuelle de l'Intelligence Artificielle.
- On trouvait, jusqu'à la nausée, cette expression sur les flyers, les frontons des stands, les kakemonos, et dans quelques conférences.
- Vous faut-il une armée d'ordinateurs pour faire vos premiers pas ? Vous faut-il un doctorat en mathématiques pour calculer et utiliser votre premier réseau de neurones ?
- **Non !**

— **Alan TURING.** *If a machine is expected to be infaillible, it cannot also be intelligent.*

Machine Learning ?.

- Machine Learning (ML) ou *Apprentissage Automatique* ou encore *Apprentissage Machine* ? *Une "sous-catégorie" de l'intelligence artificielle (IA)*
- Idée ? *Laisser des algorithmes découvrir des "schémas/patterns"*
- Schéma ? *Motif(s) récurrent(s), dans un ensemble de données*
- Données ? *Cela peut être des nombres, des mots, des images, des fichiers pcap, des logs, des statistiques voire des "comportements"*

1 Introduction

1.2 Machine Learning

Machine Learning.

- Classée par la célèbre revue *Technology Review* du MIT comme une des 10 technologies de rupture il y a déjà . . .environ 10 ans. Est ce une mode ? Oui. Et non.
- Traitement d'images, de la vision, du traitement ou reconnaissance de la parole ou traduction automatique : *succès nombreux et impressionnants*
- Google a fait sensation avec AlphaGo, qui a donné bien des soucis aux meilleurs joueurs de cet ancestral jeu. Puis AlphaFold, puis
- *FunSearch : Making new discoveries in mathematical sciences using Large Language Models* (Google DeepMind)
- C'est en fait une *véritable révolution !*

2 Révolution(s) ?

Machine Learning : deux révolutions ?

— **1ère révolution :.**

- C'est évidemment le fait que la « puissance de calcul »
- et **la grande quantité de données nécessaire**s pour avoir des résultats utiles
- . . .sont disponibles
- À portée de bourse et de clavier pourrait-on dire.

Machine Learning : deux révolutions ?

— **2d révolution :.** Peut être la plus intéressante :

- Aujourd'hui le Machine Learning s'est démocratisé
- il est quasiment à portée de quiconque ayant un PC.
- Au pire, il suffit de s'acheter par exemple une carte GPU ou de dépenser quelques dizaines d'euros sur un Cloud.
- Mais même sans Cloud, des calculs inimaginables il y a encore 20 ans sont aujourd'hui tout à fait faisables chez vous, at home !

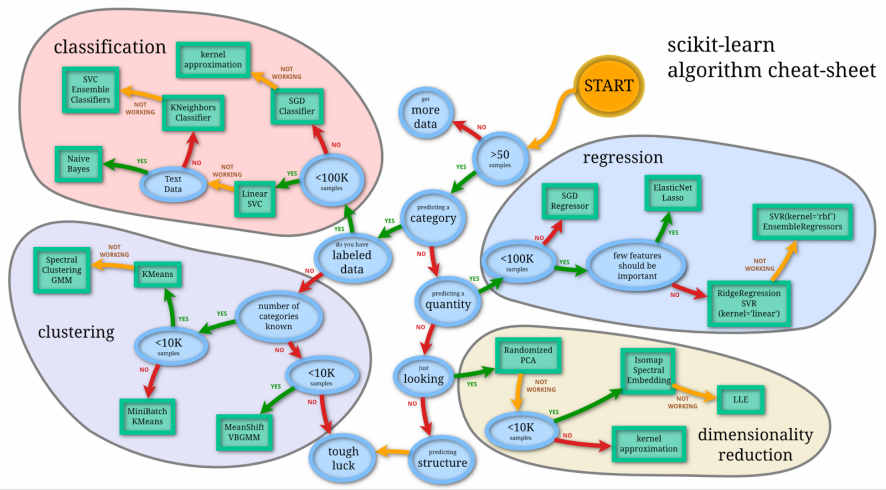
3 Non Disputemus Sed Calculemus : algorithmes ?

Quelques problématiques et algorithmes du ML.

1. Clusterisation
2. Classification
3. Régression linéaire et généralisation
4. Réduction de dimension

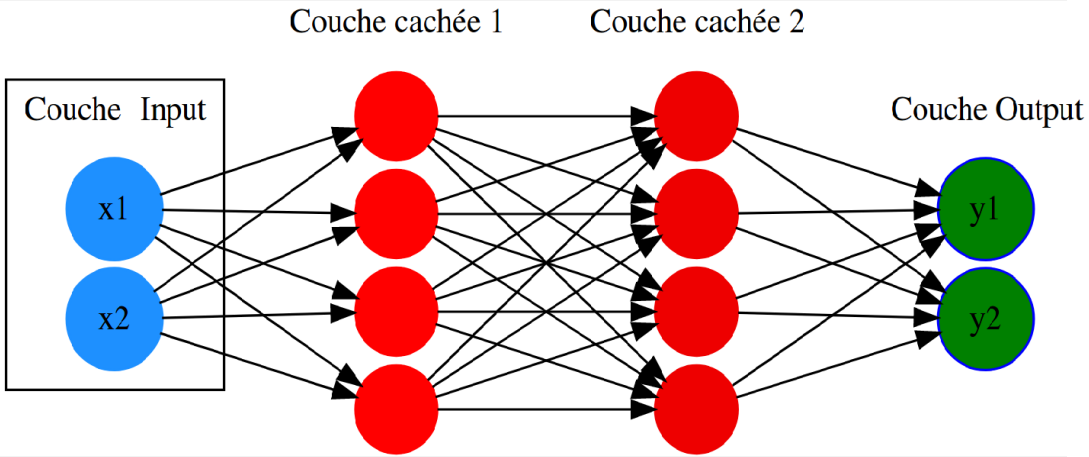
Théoriquement les algorithmes du ML vous aident à *apprendre des connaissances sur un jeu de données*. Comme le verbe *to learn* peut se traduire par *apprendre* mais aussi par *étudier* , on peut faire remarquer que tout projet de ML vous oblige donc aussi à *comprendre vos données*.

Principaux Algorithmes de SciKit-Learn.

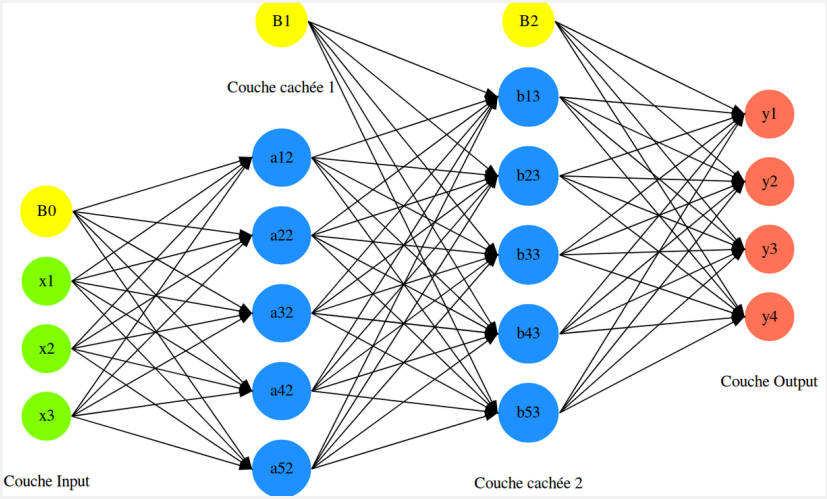


4 Au début étaient les Réseaux de Neurones Artificiels

Un Exemple.


Couche cachée 1
Couche cachée 2
Couche Input
Couche Output
x1
x2
y1
y2

Un Exemple de RNA avec "Biais".



Un Exemple d'Autoencoder.

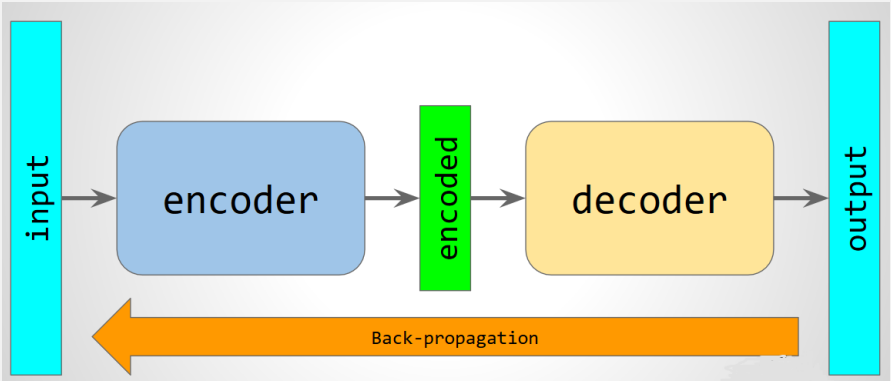


Figure – Un Autoencoder "classique"

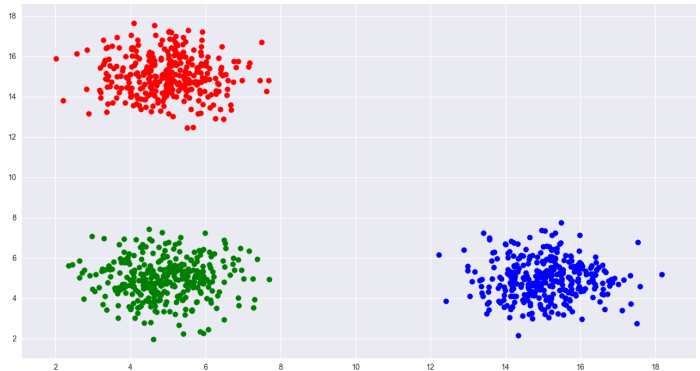
5 Un projet d'IA en Cybersécurité ?

Une stratégie classique.

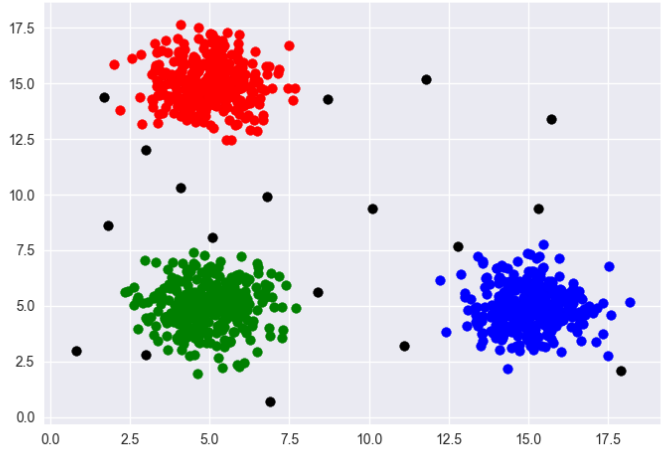
1. Définir et calculer un Vecteur de Caractéristiques (VC/FV) pour chaque donnée
2. Cela vous donne l'ensemble des *Vecteurs de Caractéristiques* (EVC)
3. Définir une distance (calculable) entre deux vecteurs (numériques) de caractéristiques, cela vous donne *la similarité* (ou une quasi-similarité) entre deux VCs.
4. Exécuter votre ou vos algorithme(s) sur l'*Ensemble des Vecteurs de Caractéristiques*
5. Essayer de comprendre/interpréter les résultats.

ça évidemment c'est la stragégie théorique !

Mais en pratique il vous aussi faut des tactiques. Combien de clusters voyez vous ?

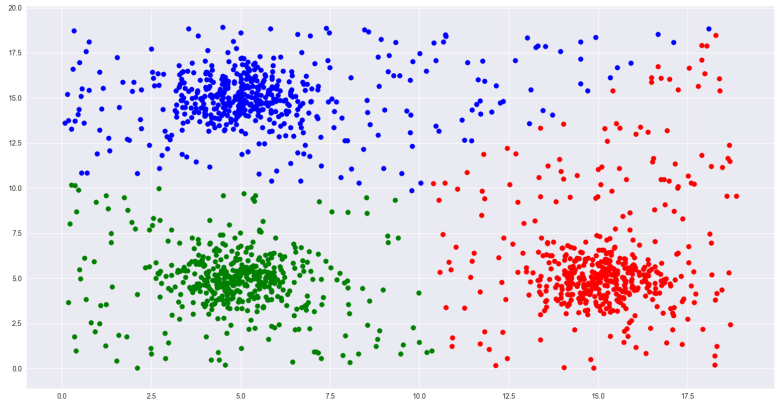


Voyez vous toujours 3 "clusters" ?.

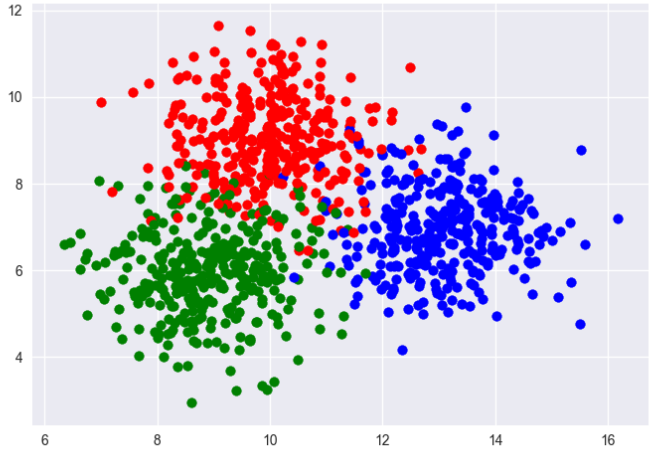


Stratégies & Tactiques :3/4 (Kmeans)

Mais il vous faut des tactiques en pratique. Voyez vous toujours 3 "clusters" ?



Ce que vous aurez probablement !.



6 Clusterisation de Malware à Grande Échelle

6 Clusterisation de Malware à Grande Échelle

6.1 Machine Learning in a Nutshell

Théoriquement :. Les algorithmes de ML vous aident, et même vous "forcent", à mieux comprendre vos données

Pratiquement :. . . .si vous disposez d'assez de puissance de calcul et d'une connaissance *a priori* minimale de votre ensemble de données et d'algorithmes rapides ils vous aident à apprendre, mais (seulement) à partir d'un très grand ensemble de données

Deux grandes familles d'algorithmes :.

- Les algorithmes *supervisés* : nous avons des informations, des étiquettes, sur chaque élément de l'ensemble de données
- Les algorithmes *non supervisés* : nous n'avons pas d'étiquettes.

Clusterization :.

- Classification automatique (en Français et en Anglais "clusterization") : regrouper en classes des objets qui sont les plus similaires
- et à séparer ceux qui sont différents ou dissimilaires.
- On désire donc partitionner en classes !
- Si possible, ces classes doivent être distinctes, chaque classe contenant des objets qui se ressemblent, on utilisera dans la suite l'anglicisme clusterisation.
- Le critère de regroupement est réalisé grâce à une *fonction de similarité* qui mesure la ressemblance entre deux programmes.

Classification après clusterization ?.

- But : placer un programme malveillant dans un groupe/cluster connu afin que celui-ci soit "plus proche" d'eux que ceux situés à l'extérieur du groupe.
- En clair : on veut labelliser un programme inconnu mais considéré comme malveillant en l'associant à un ensemble de programmes pré-labellisés et connus.
- *Exemple le plus simple de classification est la classification binaire des programmes en bénins (goodware) ou malveillants (malware). Une fois la classe choisie (le label) d'une variante, on peut « élire » un ou plusieurs représentants et comparer plus finement cette variante aux représentants, ceci permet de classifier (en général) une nouvelle variante*
- *Ce qui permet éventuellement de comprendre les mécanismes d'infection, d'attaque et de propagation de cette nouvelle variante*

6 Clusterisation de Malware à Grande Échelle

6.2 Quelques problèmes à résoudre

Vous aurez à résoudre certains problèmes :.

PB0/ **Maths** : Vous devrez probablement rafraîchir certaines connaissances sur l'entropie, les statistiques, etc. pour comprendre votre jeu de données. Mais aussi : algèbre linéaire numérique, algorithmes d'optimisation, etc.

PB1/ **Vecteurs de Caractéristiques** (VCs) : Vous avez d'abord besoin de bonnes informations statiques. Malheureusement, il n'existe pas de représentation universelle !

PB2/ **Parseur** : Vous aurez besoin d'un "très bon" parseur pour calculer ces VCs.

Vous aurez à résoudre certains problèmes :.

PB3/ **Distance** : Vous aurez également besoin d'une métrique de distance ou d'une fonction de similarité (au moins une) sur vos VCs pour définir la similarité entre paires de malwares mais tout algorithme en $O(n^2)$ est inutile.

PB4/ **Algorithme(s)** : Quel est le *"Meilleur Algorithme de Clustering"* ? Malheureusement, il n'existe pas !

- Donc vous essayerez, par exemple, K-means et DBSCAN, qui sont de bons candidats, et vous déciderez probablement, comme nous, de les utiliser parfois ensemble.

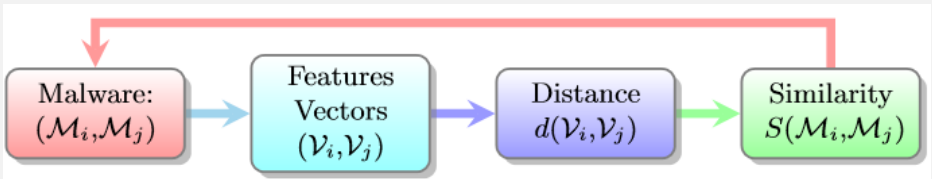
Vous aurez à résoudre certains problèmes :.

PB5/ **Apprentissage Profond [ML]** : Vous devrez essayer probablement plusieurs algorithme de ML avant d'un trouver un qui vous satisfasse

OB6/ **Graphes** : Vous aurez probablement besoin/envie d'utiliser des algorithmes de graphes : **Louvain** est très bien adapté pour calculer les communautés (clusters !) dans un très grand graphe "creux" (mais vous aurez besoin de calculer un très grand graphe. Eh bien, cela signifie que vous devez comprendre (par exmple) ce qu'est un graphe t-spanner ...)

Notre (première) Chaîne de Traitement.

Distance ≠ Similarité.



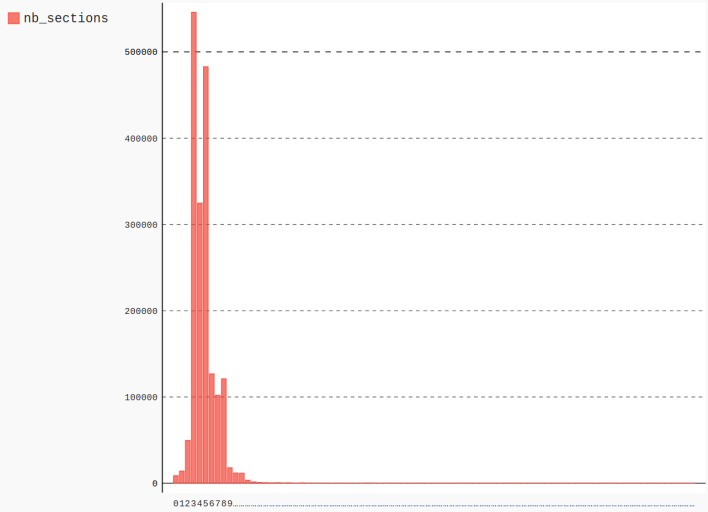
Mais le monde réel gagne toujours : Vous verrez rapidement, par exemple que :.

1. . . .que certains parseurs n'apprécient pas les sections avec des noms très étranges.
2. Vous déciderez donc rapidement d'écrire votre propre parseur (ou de modifier un parseur existant)
3. Nous (en fait Sébastien LARINIER) en avons développé deux !
4. . . .dont un est Open Source : *PEtoJson*

Quel Vecteur de Caractéristiques pour comparer des Malware :. Sur environ $2,3 \times 10^6$ malware (en 2019) :

1. Number of sections : $8543138 \approx 8,5 \times 10^6$
2. Number of different sections : : $2793782 \approx 2,7 \times 10^6$
3. Number of different section names : 97263
4. Minimal number of sections/malware : 0
5. Maximal number of sections/malware : 90
6. Minimal length of a section : 0
7. Maximal length of a section : 4294966784 (probably a joke because $4294966784 = 2^{32} - 2^9$)

Nombre de sections par malware : un histogramme.



7 Une conclusion (en forme de problèmes (et des références) !)

Vous aurez à résoudre certains problèmes :.

Con1 Vous devrez récupérer des données

Con2 Vous devrez les nettoyer (enlever ou gérer les données incomplètes, enlever les données redondantes etc.), par exemple vous pouvez choisir de supprimer les différentes copies d'un malware qui ont juste un nom différent, si vous en avez.

Con3 Vous devrez normaliser vos données, par exemple pour un jeu de données de malware, si vous calculez le nombre de sections et l'entropie de chaque section, si vous avez un malware à plus de 50 sections, cette valeur risque de « dévorer » les autres.

Con4 Vous devrez aussi « explorer » vos données, *i.e.* commencer à les comprendre. On pourrait presque dire qu'il vous faudra les apprivoiser

Vous aurez à résoudre certains problèmes :.

Con5 Vous devrez choisir un vecteur de caractéristiques (très délicat), la qualité et la pertinence de ce vecteur vont fortement influencer la qualité de vos résultats. Pour un malware, on peut commencer avec : sa taille (en octets), le nombre de sections et leur taille, l'entropie du fichier, l'entropie de chaque section, le nombre de symboles de la table d'import, le nombre de symboles de la table d'export, le nombre de ressources, etc

- Vous devrez maîtriser le « passage à l'échelle » (scalability) : c'est à dire maîtriser la complexité pratique des algorithmes

Con6 Vous devrez choisir/calculer les hyperparamètres d'un algorithme (paramètres non optimisées par l'algorithme : soit choisis par l'utilisateur soit calculés de manière différente). Très complexe, et très coûteux. Par exemple, pour un réseau de neurones : combien de couches cachées faut-il choisir ? Et pour chaque couche, quelle taille ?

Vous aurez à résoudre certains problèmes :.

Con7 Vous devrez probablement aussi réviser (un peu ou beaucoup) vos maths : statistiques et probabilités, algèbre linéaire, calcul différentiel, algorithmes numériques de gradient pour la minimisation, etc.

Con8 Et, last but not least, vous devrez choisir vos algorithmes, en fonction de vos données, de vos attentes, de votre puissance CPU etc. Bon là, c'est à vous !

Con9 Ah oui : évidemment vous devrez avoir une petite idée de ce que vous voulez mettre en évidence, il n'y a pas de magie.

Con10 Et n'oubliez pas : le ML est essentiellement un processus itératif, vous devrez revenir sur vos choix en fonctions des résultats préliminaires, refaire des tests, revenir sur vos hyperparamètres etc.

esiea

Merci pour votre attention !

Nous cherchons des stages et/ou contrats d'alternance
pour nos élèves du BACHELOR CYBERSÉCURITÉ
Stage : Juin 2024—Août 2024 :

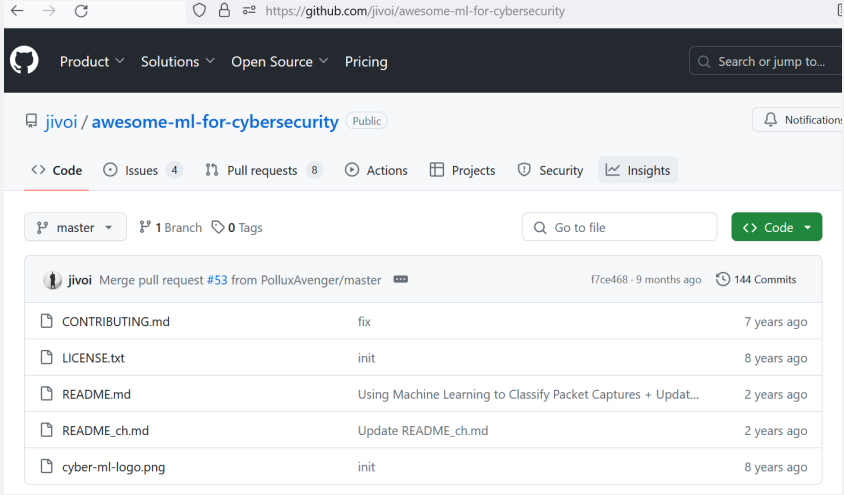
Alternance : Septembre 2024—Août 2025 :
robert.erra@esiea.fr

Si vous voulez commencer :.

1. Un jeu de données récent ? **SOREL-20M (Sophos/ReversingLabs-20 Million)** : *it is a large-scale dataset consisting of nearly 20 million files with pre-extracted features and metadat . . .malware sample [https : // paperswithcode. com/ dataset/ sorel-20m]*
2. Des algorithmes et du code ? **Catch'em all : Classification of Rare, Prominent, and Novel Malware Families** : où vous trouvez des rééférences et du code . . .
3. Si vous voulez vraiment jouer (en Python) : *https : // github. com/ lanl/ T-ELF*

https ://github.com/jivoi/awesome-ml-for-cybersecurity

Une page pleine de ressources.



Une page pleine de ressources.

Awesome Machine Learning for Cyber Security

A curated list of amazingly awesome tools and resources related to the use of machine learning for cyber security.

Table of Contents

Lectures complémentaires

1/2

- Scikit-Llearn : `http://scikit-learn.org`
- Anaconda : `https://www.anaconda.com/`
- THE MNIST DATABASE : `http://yann.lecun.com/exdb/mnist/`
- Y. LeCun, Y. Bengio & G. Hinton, Deep learning, Nature, vol. 521, 28 MAY 2015 , disponible à : `https://www.cs.toronto.edu/~hinton/absps/NatureDeepReview.pdf`
- A. Géron : Machine Learning avec Scikit-Learn, Mise en oeuvre et cas concrets, Dunod, 2017
- A. Géron : Deep Learning avec TensorFlow , Mise en oeuvre et cas concrets, Dunod, 2017.
- F. Chollet, Deep learning with Python, Manning Publications Co., 2018.

2/2

- Ember Malware Dataset : https ://arxiv.org/pdf/1804.04637.pdf
- theZoo Malware Dataset : https ://github.com/ytisf/theZoo
- A Python Notebook (by S. Larinier aka @Sebdraven) : Python and Machine Learning : How to clusterize a malware dataset ? https ://github.com/sebdraven/hack_lu_2017
- PeToJson : https ://github.com/sebdraven/petojson
- *Malware Data Science Attack Detection and Attribution* : by Joshua Saxe with Hillary Sanders : https ://nostarch.com/malwaredatascience
- MISC HS Numéro 18 : "Machine Learning et Sécurité" : https ://boutique.ed-diamond.com/en-kiosque/1363-misc-hs-18.html