La sécurité informatique à l'ère de l'intelligence artificielle

Pierre-François Gimenez
CentraleSupélec

ENS Rennes
13 octobre 2021

CentraleSupélec

Qu'y a-t-il dans cette présentation?

1. Cybersécurité
2. Intelligence artificielle et machine learning
3. La sécurité aidée du machine learning
4. Le machine learning attaqué
5. Conclusion

CentraleSupélec

Cybersécurité

Quelques attaques récentes. . .

Twitch Has Been Hit By A Major Hacking – And All Of Its Source Code Has Been Leaked

Armeen Khan October 8, 2021 0

SolarWinds Attack The 'Largest And Most Sophisticated' Ever, Microsoft CEO Says

MIcrosoft estimates that over 1,000 engineers worked on the massive SolarWinds supply chain compromise as details continue to emerge.

February 16, 2021 Zachary Comeau Leave a Comment

. . . et d'autres plus anciennes

LILY HAY NEWMAN SECURITY 05.12.2017 02:03 PM

The Ransomware Meltdown Experts Warned About Is Here

It's not just British hospitals. A nasty strain of ransomware is sweeping the world.

National Security

Stuxnet was work of U.S. and Israeli experts, officials say

By **Ellen Nakashima** and **Joby Warrick**
June 2, 2012

La sécurité aujourd'hui

Une menace grandissante

- Le nombre et l'importance de ces attaques augmentent
- De plus en plus d'attaques sont menées par des groupes organisés
- Toutes les couches informatiques peuvent devenir un vecteur d'attaque : matériel, systèmes d'exploitation, applications...

Tout est perdu ?

Non ! Un irréductible village de chercheurs et d'ingénieurs en sécurité travaillent à prévenir et limiter ces attaques.

Comment étudier la sécurité ?

Trois concepts clés

La sécurité informatique vise à assurer trois propriétés :

Confidentialité les données sensibles ne peuvent être lues que par les personnes autorisées

Intégrité les données sensibles ne peuvent être modifiées que par les personnes autorisées

Disponibilité le service doit être accessible aux utilisateurs avec un court temps de réponse

Exemples d'attaque

- Attaque de l'homme du milieu : Alice veut parler à Bob. Sauf qu'en fait, Alice parle à Eve qui répète à Bob !
- Attaque par déni de service : rendre un système inutilisable, généralement en le surchargeant de requêtes
- Phishing (hameçonnage) : tromper un usager pour qu'il donne des informations confidentielles (ses identifiants par exemple) ou qu'il installe un malware

CentraleSupélec

Des outils classiques

La cryptographie

- La cryptographie est la science de la protection des messages
- Une donnée chiffrée ne peut être lue que par le destinataire du message
- La cryptographie est aussi utilisée pour signer des messages

Le hachage

- Fonction mathématique injective difficile à inverser. Transforme son entrée en un "hash" (un grand nombre)
- Si deux fichiers ont le même hash, alors ce sont les mêmes. Mais impossible de retrouver le fichier à partir du hash
- Utile pour vérifier l'intégrité d'un message, pour stocker des mots de passe, pour fournir une preuve de travail

L'organisation de la sécurité

Les acteurs de la sécurité

- ANSSI (Agence nationale de la sécurité des systèmes d'information) est un organisme gouvernemental qui apporte son expertise, effectue des audits, certifie des solutions, etc.
- Dans une organisation : le RSSI est le responsable de la sécurité des systèmes d'informations. Il établit une politique de sécurité.

La politique de sécurité

Elle est établie en différentes étapes :

- Identifier la surface d'attaque, les ressources à protéger, la gravité en cas d'attaque
- Déterminer les objectifs de sécurité
- Élaborer les procédures de protection (la structure du réseau, badges d'accès, programmes de sauvegarde, protocole en cas de problèmes, etc.)

Une défense est aussi efficace que son élément le plus faible !

CentraleSupélec

Quelques conseils de sécurité. . .

L'humain est une porte d'accès pour les attaquants : il est donc important d'éduquer

- Utilisez des mots de passe différents pour chaque site
- Faîtes des sauvegardes régulières
- N'ouvrez pas de fichiers (exécutables, PDF, documents Office) de source inconnue ou peu fiable
- Gardez votre système à jour
- Ne devenez pas parano :)

CentraleSupélec

Intelligence artificielle et machine learning

CentraleSupélec

Pourquoi tant de hype ?

L'intelligence artificielle, le *machine learning* et les réseaux de neurones ne sont pas nouveaux (50's) mais ont souffert de cycle d'hype, de grandes promesses et de désillusion

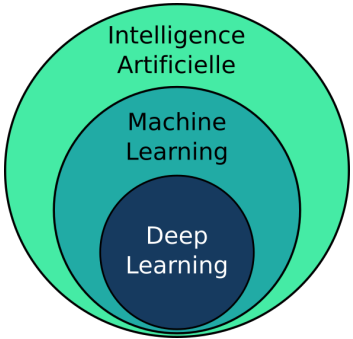
Pourquoi un nouvel effet de mode depuis 10 ans ?

- Stockage de données bon marché, explosion des données disponibles grâce à la numérisation de nos vies
- Explosion de la puissance de calcul via les GPU et le calcul distribué
- Avancées théoriques en modèles et algorithmes d'apprentissage

Quelques domaines révolutionné par le *deep learning*

- Traitement de l'image, du son, de la vidéo
- Traduction automatique
- Jeux (Go, StarCraft II)
- Repliement des protéines

Machine learning ? Intelligence artificielle ? Deep learning ?



Intelligence artificielle (IA)

Vise à résoudre des problèmes complexes (par exemple : planification avec contraintes, recherche de chemin, représentation des connaissances, etc.)

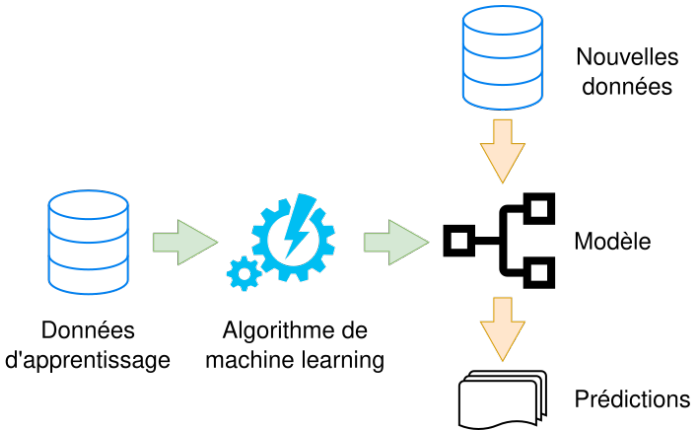
Machine learning (ML) = apprentissage automatique

Construction d'un "modèle" à partir de données, dans l'objectif de réaliser une tâche après (jouer au go, détecter des intrusions, etc.)

Deep learning (DL) = apprentissage profond

Techniques de machine learning basées sur des réseaux de neurones avec plusieurs couches. Nécessite beaucoup de données et de puissance de calcul, mais obtient généralement les meilleures performances

Le principe du machine learning



Deux étapes principales en machine learning

- Dans la phase d'apprentissage, un algorithme d'apprentissage transforme des données d'apprentissage en un modèle
- Dans la phase d'inférence, le modèle fait des prédictions sur de nouvelles données

Les données pour le machine learning

Quelle entrée pour l'apprentissage ?

- Les données doivent être structurées en vecteurs (appelés instances)
- Chaque vecteur a un *nombre fixé* d'attributs
- Chaque attribut peut être
  - discret : un nombre fini de valeurs arbitraires (catégories)
  - continu : une quantité qu'on peut manipuler avec des opérateurs comme $+$ et $>$

« Garbage in, garbage out »

- La sortie du machine learning ne peut être qu'aussi bonne que ses entrées
- Toujours des erreurs humaines et techniques dans les données collectées
- Spécialement critique pour l'aide à la décision, car des précédents décisions biaisés vont aboutir à des recommandations biaisées

CentraleSupélec

Un moyen d'organiser les techniques de ML

Une grosse boîte à outils

- Les techniques de machine learning forment une boîte à outils pour divers problèmes
- Mais c'est une grosse boîte à outils... c'est difficile de trouver ce qui est pertinent !
- Cette boîte à outils est organisée en "types de problème"

Types d'apprentissage

- Ces "types de problème" sont appelés *types d'apprentissage*
- Un type d'apprentissage spécifie quel est l'objectif de la technique de ML et sur quelles données elle s'appuie
- La plupart des méthodes sont adaptées à un type d'apprentissage, donc il est crucial d'identifier le type d'apprentissage avant de chercher quel outil utiliser !

$\Rightarrow$ on va voir ensemble les trois principaux types d'apprentissage

Apprentissage supervisé

Apprentissage supervisé

Dans les données d'apprentissage, il y a une *étiquette* pour chaque instance. Cette étiquette est ce qu'on cherche à prédire

- Si l'étiquette est discrète, c'est une tâche de *classification*. Par exemple : prédire la famille d'un malware, détecter un spam
- Si l'étiquette est continue, c'est une tâche de *régression*. Par exemple : prédire le nombre de connexions à un service

Apprentissage non-supervisé

Apprentissage non-supervisé

Il n'y a pas d'étiquette dans les données d'apprentissage

- *Détection de valeurs aberrantes* : identifier les instances qui ne ressemblent pas aux autres
- *Clustering* (partitionnement) : regrouper les instances similaires en différents paquets
- *Génération de données* : générer de nouvelles instances qui se rapprochent des données d'apprentissage

Apprentissage semi-supervisé

Apprentissage semi-supervisé

Une catégorie fourre-tout où il y a seulement certaines étiquettes

- *Classification à une seule classe*, où toutes les données proviennent d'une seule classe

Ça reste abstrait ? On va voir à quoi ressemblent vraiment les modèles de machine learning !

k-NN (apprentissage supervisé)

Principe

Pour prédire la classe d'une instance, on cherche ses k plus proches voisins (k nearest neighbors, d'où k-NN) dans les données d'apprentissage et on prédit la classe majoritaire parmi eux

Remarques

- De bonnes performances et une prédiction interprétable
- Généralement utilisé avec des attributs continus
- Sensible aux attributs redondants et non-pertinents
- Variation : LOF (basé sur la densité local) pour la détection de valeurs aberrantes

Arbre de décision (apprentissage supervisé)

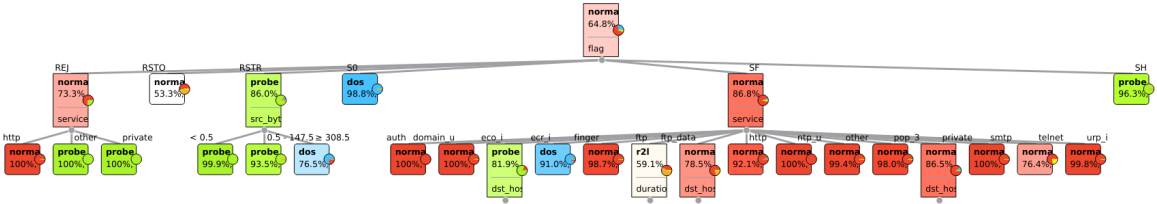
Principe

- Un arbre de décision est un ensemble de règles sous la forme d'un arbre
- Chaque nœud contient une règle simple sur un attribut, par exemple "A=val1" ou "B>0.5"

Remarques

- Très simple à interpréter, mais pas les meilleures performances
- Fonctionne avec peu de données
- Robuste aux attributs non-pertinents ou redondant
- Peut être utilisé comme base pour des modèles plus complexes

Exemple d'arbre de décision



Exemple de détection d'attaques dans les réseaux

Réseaux de neurones (apprentissage supervisé)

Principe

Chaque neurone fait un calcul simple. Les neurones sont arrangés en couche, où la sortie de la couche N est l'entrée de la couche N+1. Le réseau sort des nombres continus (tâche de régression) ou une distribution de probabilité (tâche de classification)

Remarques

- Les meilleurs performances... quand c'est bien paramétré
- Peut traiter des Go de données mais nécessite de la puissance de calcul
- Impossible à interpréter sans un expert
- Variations :
  - Auto-encodeur pour la classification à une classe
  - Auto-encodeur variationnel et réseaux génératifs adversariaux (GAN) pour la génération de données

Ce qu'on peut faire avec les GAN

`https://thispersondoesnotexist.com/`

C'est saisissant de réalisme ! Mais des artefacts restent visibles

CentraleSupélec

La sécurité aidée du machine learning

Systèmes de détection d'intrusions

Objectif

Détecter des attaques dans les systèmes et les réseaux

Deux principales sources de données

- Des données réseaux récoltées à un point d'intérêt, comme un pare-feu à la périphérie du réseau à protéger
- Des données issues d'une machine en particulier, où sont surveillés la consommation des ressources, les appels systèmes, les logs systèmes, etc.

Deux grandes approches

- L'approche *par règle* qui reconnaît les attaques via des signatures
- L'approche *comportementale* qui modélise le comportement normal et vérifie s'il n'y a pas de déviations

Systèmes de détection d'intrusions

Approche par règle

- Historiquement, les règles sont construites par des experts qui connaissent les signes à surveiller
- Depuis 25 ans, on utilise aussi du machine learning
- Il s'agit d'un cas d'apprentissage supervisé où l'étiquette à prédire est "bénin / attaque", voire de quelle attaque plus précisément il s'agit
- On arrive à atteindre de très bonnes performances ($>$95% d'exactitude)
- Par contre, on est démuni face aux nouvelles attaques

Systèmes de détection d'intrusions

Approche comportementale

- Héritage d'une approche par spécification : on savait ce qui pouvait arriver ou non
- Les systèmes actuels sont beaucoup trop complexes pour ça...
- Il s'agit d'un cas d'apprentissage semi-supervisé (apprentissage à une seule classe) où on a uniquement des données du comportement normal
- Tout évènement qui sort du comportement normal appris est une anomalie !
- On peut détecter des attaques nouvelles, mais cette approche amène souvent plus de fausses alertes

Systèmes de gestion d'alertes

La gestion des alertes

- De nos jours, on dispose d'un grand nombre de système de détection hétérogène
- On peut avoir 10 000 alertes par jour dans une grande entreprise !
- Un SIEM (Security Information and Event Management) est un outil qui va traiter ces alertes pour les :
  - normaliser
  - agréger
  - corréler
  - prioriser
- Le machine learning peut intervenir à plusieurs de ces étapes

Analyse de malware

Deux grandes approches

- Analyse statique : on étudie le binaire sans l'exécuter. C'est rapide, mais le fichier peut être chiffré
- Analyse dynamique : on exécute le binaire. C'est lent, mais c'est plus difficile pour un malware de se cacher

Le machine learning peut aider...

- C'est encore de l'apprentissage supervisé !
- Analyse statique : l'entrée est le fichier binaire (ou une représentation)
- Analyse dynamique : l'entrée est une séquence d'appels systèmes

CentraleSupélec

À quand la révolution par le machine learning ?

Pas (encore) de révolution

- Une recherche forte avec de constants progrès
- Contrairement à d'autres domaines, les solutions commerciales (antivirus, détection d'intrusion) en sécurité n'intègrent pas beaucoup de machine learning
- La différence fondamentale avec les autres domaines ? En sécurité, les systèmes de machine learning se font attaquer !


LILY HAY NEWMAN SECURITY 07.16.2021 06:00 PM

Hackers Got Past Windows Hello by Tricking a Webcam

The security researchers used infrared photos and third-party hardware to best Microsoft's facial-recognition tech.

CentraleSupélec

Le machine learning attaqué

Les attaques adversariales

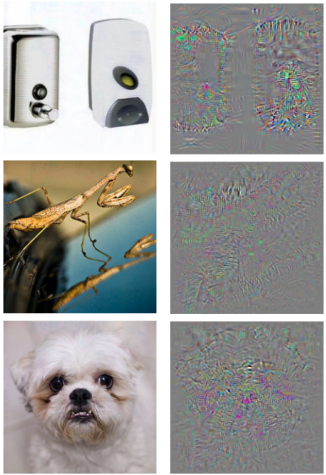
En 2014, une découverte étonnante. . . [1]

- Le modèle prédit bien les bonnes classes

[1] Intriguing properties of neural networks, *Szegedy et al.*

Les attaques adversariales

En 2014, une découverte étonnante... [1]

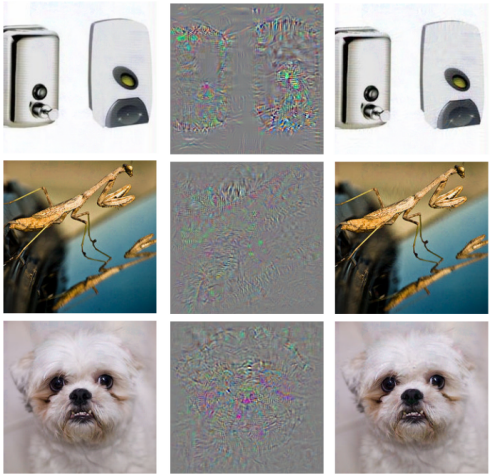


- Le modèle prédit bien les bonnes classes
- Mais si on ajoute une petite perturbation... (ici amplifiée pour être visible)

[1] Intriguing properties of neural networks, *Szegedy et al.*

Les attaques adversariales

En 2014, une découverte étonnante... [1]



- Le modèle prédit bien les bonnes classes
- Mais si on ajoute une petite perturbation... (ici amplifiée pour être visible)
- ... maintenant le modèle identifie tout comme des autruches !

[1] Intriguing properties of neural networks, *Szegedy et al.*

CentraleSupélec

Des modèles bien fragiles

Les modèles de ML sont attaquables

- Même si nos outils de détection fonctionnent bien avec des données réelles, de petites modifications bien choisies peuvent complètement les perturber !
- Ce sont principalement les réseaux de neurones qui sont attaquables
- Cette attaque fonctionne même si on ne sait pas comment marche le modèle

D'autres attaques

- Récupération de texte servant à l'apprentissage de modèles de langages (par exemple GPT-2 utilisé par AI Dungeon)
- Empoisonnement de données d'apprentissage : en manipulant les données d'apprentissage, un attaquant peut créer un moyen simple mais secret de modifier la prédiction
- ... et d'autres encore

CentraleSupélec

Des solutions ?

Des pistes actuellement recherchées

- Améliorer l'explicabilité des prédictions pour mieux détecter ces attaques
- Bruiter les données pour rendre le modèle plus robuste
- Modifier le modèle avec des contraintes mathématiques pour s'assurer que de petites perturbations en entrée ne produisent pas de grosses perturbations en sortie
- Utiliser du chiffrement homomorphique pour que les données d'apprentissage ne soient pas inférables du modèle

$\Rightarrow$ ces méthodes ont tendance à faire baisser l'efficacité des modèles ou à beaucoup augmenter le temps de calcul...

La recherche est très active dans ce domaine !

CentraleSupélec

Conclusion

CentraleSupélec

Conclusion

Ce qu'il faut retenir

- La menace informatique est bien réelle et continue de grandir
- Les chercheurs et les ingénieurs travaillent à la sécurité des systèmes
- On a développé des procédures, des outils informatiques et mathématiques pour protéger les systèmes d'information
- Le machine learning est à la fois :
  - une opportunité : pour la détection d'intrusion, la corrélation d'alertes et l'analyse de malware notamment
  - un risque : au mieux le modèle est inefficace, au pire il devient lui-même un vecteur d'attaque !
- Bilan : la révolution de la sécurité par le machine learning va devoir attendre !
- En attendant, restez vigilants :)

Merci de votre attention !