keep coding

Machine Learning y Ciberseguridad

ML Adversarial Attacks

Índice

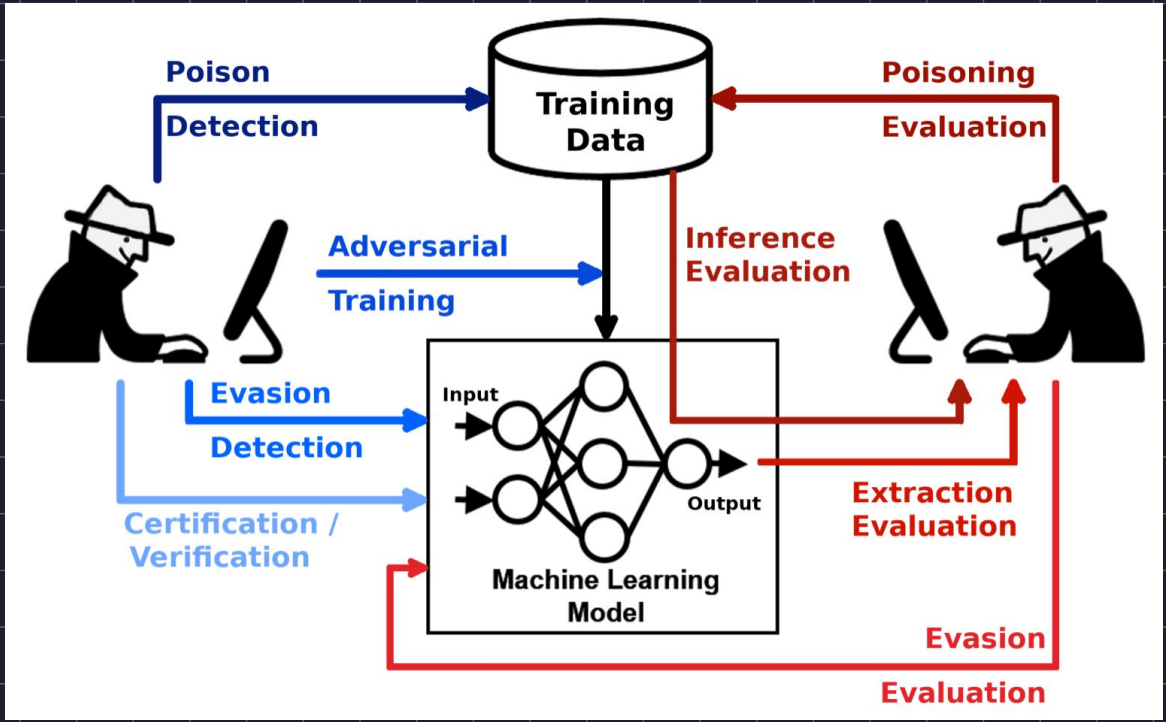
keep coding

¿Por Qué Ataques Adversarios en Cyber?

- Ataques adversarios manipulan inputs de ML para engañar modelos, común en cyber para evadir detección (e.g., malware disfrazado).
- Importancia en 2025: Con AI en firewalls y antivirus, hackers usan adversarios para breaches; ART ayuda a testear robustez.
- Beneficios de ART: Biblioteca gratuita para Red/Blue Teams, soporta múltiples frameworks.
- Especialización: Seguridad de ML, ethical hacking con AI.

keep coding

Blue/Red Team en Adversarial Attacks



¿Qué Son los Ataques Adversarios?

- Definición: Perturbaciones intencionales en datos de entrada para causar errores en modelos ML, sin cambiar su apariencia humana (un humano no puede detectar el cambio)
- Fases:
  - Exploratorio (ataque en producción) vs. Causativo (durante entrenamiento).
- En cyber: Engañar detectores de phishing o intrusiones

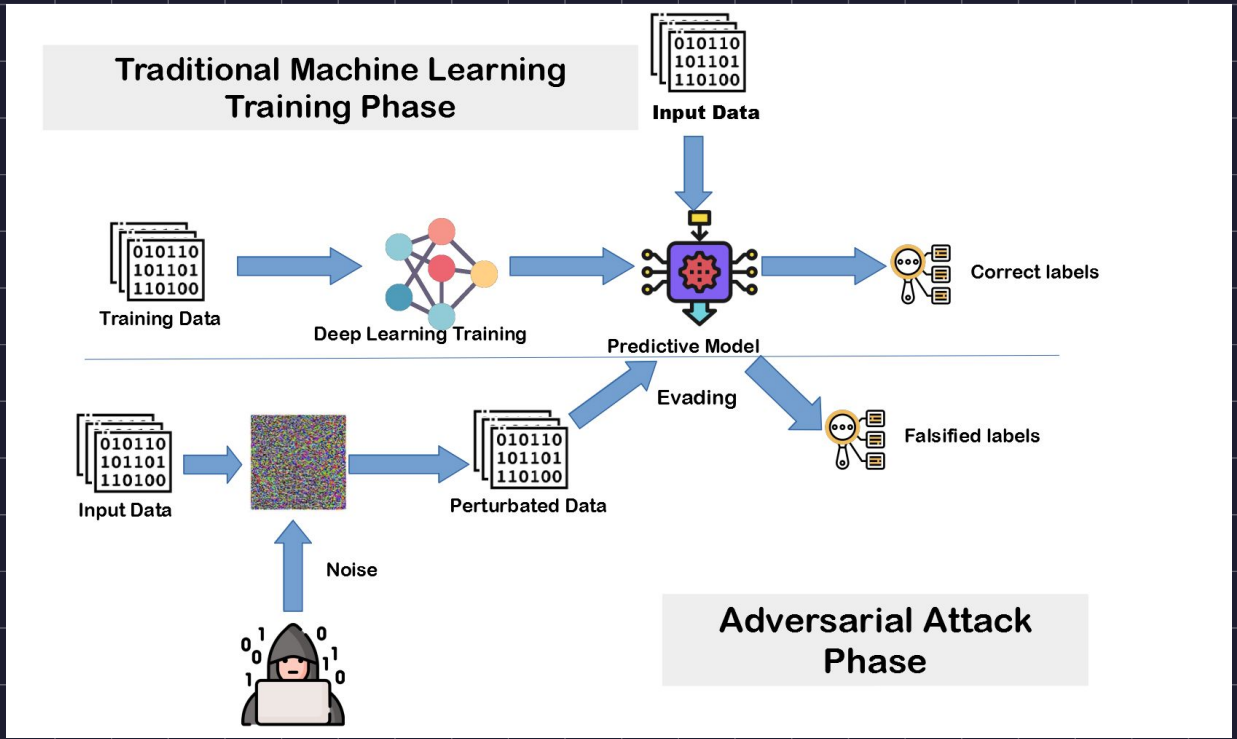
keep coding

Tipos de Ataques Adversarios - Evasión

- Evasión: Alterar inputs en inferencia para evadir clasificación (e.g., agregar ruido a malware para que parezca benigno).
- Ejemplo en cyber: Modificar payloads en ataques DDoS para burlar SIEM basados en ML.
- Blanco vs. Negro: Conocimiento del modelo (blanco: completo; negro: limitado).
- Especialización: Red teaming en cyber-ML.

keep coding

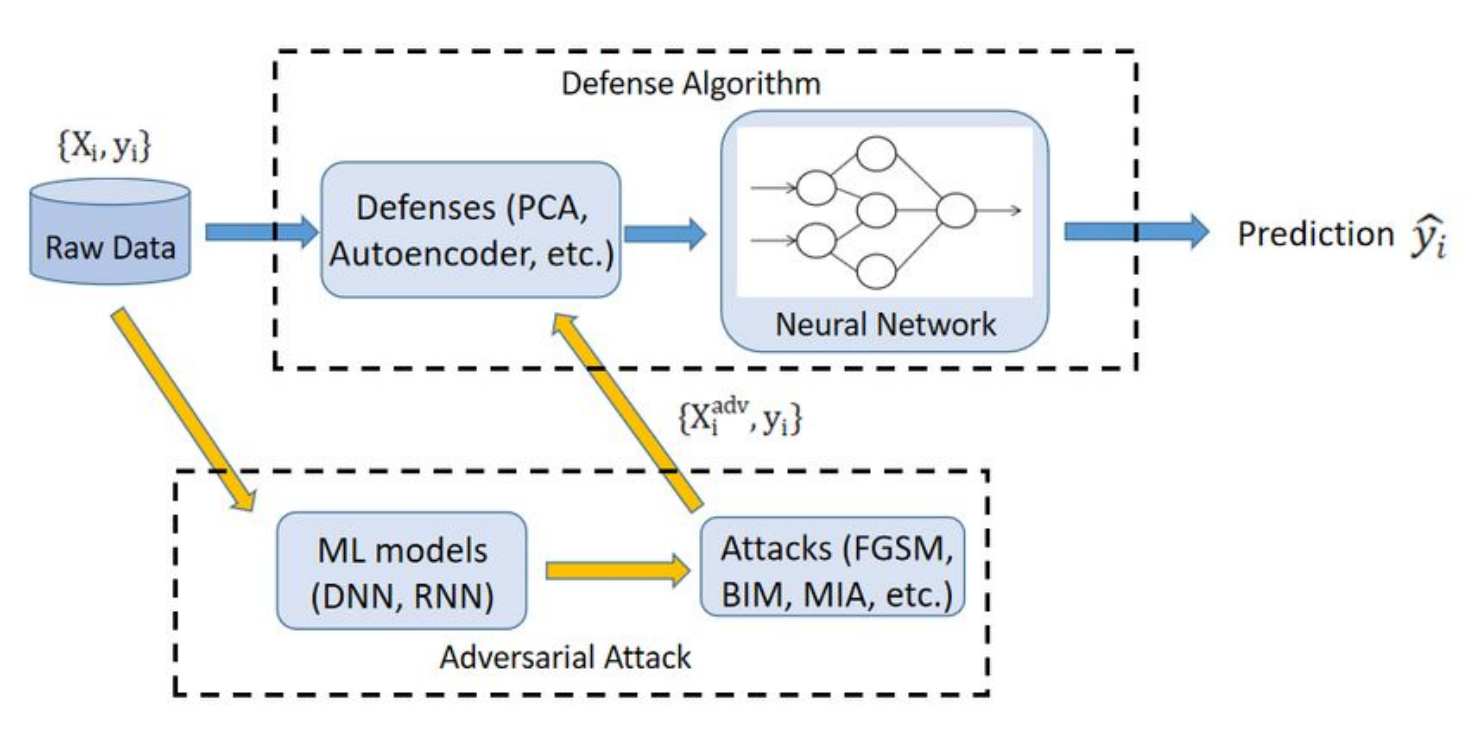
Adversarial Attacks - Evasión



Tipos de Ataques Adversarios - Envenenamiento

- Envenenamiento: Corromper datos de entrenamiento para sesgar el modelo (e.g., inyectar samples falsos en datasets de threats).
- Ejemplo en cyber: Alterar logs de entrenamiento para que un detector ignore ciertos ataques. Introducir datos de mi negocio para aumentar SEO.
- Backdoor: Insertar triggers ocultos.
- Especialización: Blue Team, Data poisoning defense.

Adversarial Attacks - Envenenamiento



Tipos de Ataques Adversarios - Extracción

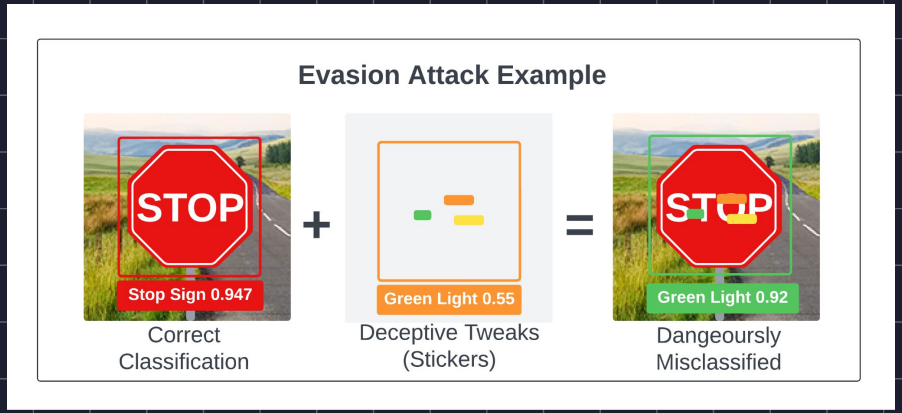
- Extracción: Robar el modelo mediante queries (e.g., inferir parámetros de un clasificador de malware).
- Ejemplo en cyber: Extraer un modelo de detección de un servicio cloud para evadirlo.
- Model stealing: Recrear funcionalidad.
- Especialización: Model security auditing.

Tipos de Ataques Adversarios - Inferencia

- Inferencia: Revelar datos privados del entrenamiento (e.g., membership inference en datasets de user behavior).
- Ejemplo en cyber: Inferir datos sensibles de un modelo UBA.
- Privacy attacks: Membership, attribute inference.
- Especialización: Privacy-preserving ML.

Impacto en Ciberseguridad

- Riesgos: Evasión de antivirus ML, falsos negativos en threat detection.
- Estadísticas: Hasta 90% de modelos vulnerables (basado en studies 2025).
- Herramientas como ART para testing.



keep coding

Introducción a ART

- Adversarial Robustness Toolbox
- Qué es: Biblioteca Python para ML security, desarrollada por IBM/LF AI & Data
- Propósito: Evaluar y defender contra evasión, poisoning, extraction, inference.

https://github.com/Trusted-AI/adversarial-robustness-toolbox

Características de ART

- Soporte: Frameworks como TensorFlow, PyTorch, scikit-learn; datos (imágenes, audio); tareas (clasificación, detección).
- Ataques/Defensas: Docenas implementadas, modulares.
- Métricas: Robustez, certificación.
- Útil para datasets tabulares en intrusiones.

www.keepcoding.io cursos@keepcoding.io (+34) 916 33 1779

keep coding

Cuerpo de texto

Título sección

Texto extra

keep coding

Título sección

Texto extra

Cuerpo de texto

Título solo

Cuerpo de texto

Título combinado

Cuerpo de texto

keep coding

Título Largo

keep coding

TÍTULO APARTADO ESPECÍFICO

TÍTULO APARTADO ESPECÍFICO

Ejemplo Bootcamp Web

Proyecto Final Web

Ejemplo Bootcamp Ciber

Proyecto Final Ciber

Cuerpo de texto (si aplica)

keep coding

Título elementos

Elemento 1

Cuerpo

Elemento 2

Cuerpo

Elemento 3

Cuerpo

Título elementos V2

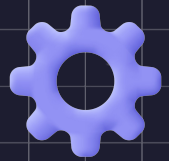
Elemento 1	Elemento 2	Elemento 3
Cuerpo	Cuerpo	Cuerpo

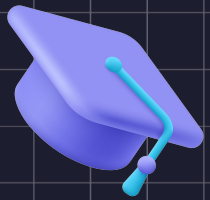
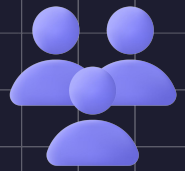
keep coding





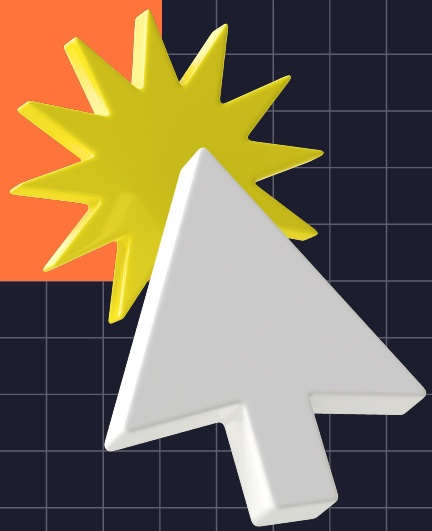


*Slide de ejemplo de aplicación del emoji

Módulo Fundamentos



keep coding

Cuerpo de texto

Título sección

Texto extra