MODULO 02

Introduzione all'AI per la sicurezza

I fondamenti di machine learning e deep learning necessari a valutare criticamente le soluzioni di sicurezza basate su AI, fino agli LLM e agli agenti autonomi.

ML / DL

LLM & AGENTI

METRICHE E LIMITI

Corso magistrale · Cyber Security · Digital Forensics & Incident Response

SCHEDA DEL MODULO

Abstract e obiettivi formativi

Abstract

Il modulo fornisce i fondamenti di machine learning e deep learning necessari a valutare criticamente le soluzioni di sicurezza basate su AI: pipeline di addestramento, metriche di valutazione e limiti statistici della detection, fino ai modelli linguistici di grandi dimensioni e agli agenti autonomi.

Obiettivi formativi (Bloom)

- **Comprendere**: distinguere apprendimento supervisionato, non supervisionato e per rinforzo
- **Analizzare**: interpretare precision, recall, F1 e ROC/AUC nel contesto della detection
- **Valutare**: discutere i limiti statistici dei classificatori su eventi rari

Traccia core

01 02 07

Detection & Response

03 04 05 06

o

Digital Forensics

08 09 10 11

o

Sicurezza dell'AI

06 12 13 14

Traccia core

15 16

SCHEMA · TEORIA

Machine learning e LLM per la sicurezza

TEORIA · PARADIGMI DI APPRENDIMENTO

Paradigmi di apprendimento e usi in sicurezza

SUPERVISIONATO

Etichette note

Classificazione di malware e phishing a partire da esempi etichettati; richiede dataset curati e aggiornati

NON SUPERVISIONATO

Struttura nei dati

Clustering e anomaly detection: rileva ciò che devia dalla baseline, senza esempi d'attacco

PER RINFORZO

Politiche da ricompense

Agenti che apprendono strategie di risposta; oggi soprattutto in ricerca e simulazione

La pipeline di apprendimento

Dati, feature, addestramento, validazione, rilascio, monitoraggio. L'overfitting riduce la capacità di generalizzare; il **concept drift** degrada il modello in esercizio, per cui la verifica delle prestazioni prosegue dopo il rilascio.

Addestramento e impiego degli LLM

Tecniche principali

- **Pre-training**: conoscenza generale appresa dal corpus
- **Fine-tuning**: adattamento dei pesi a un dominio
- **RAG**: conoscenza esterna recuperata al momento della risposta
- **Agenti**: pianificazione in più passi e invocazione di strumenti

Rilevanza per il SOC

- Il RAG basa le risposte su playbook e telemetria dell'organizzazione
- Con l'uso di strumenti il modello compie azioni: servono limiti di privilegio e registrazione
- Spiegabilità (XAI): requisito per usare l'output in contesti critici

TEORIA · LIMITI STATISTICI

Limiti statistici della detection

99%

di accuratezza, con un attacco ogni 100.000 eventi, lascia circa un falso positivo ogni 100 eventi, cioè migliaia di falsi allarmi al giorno anche con volumi modesti

Conseguenze operative

- Con classi sbilanciate si usano **precision, recall e F1** al posto dell'accuratezza
- La soglia di allarme è una scelta di costo tra falsi positivi e falsi negativi
- Lo sbilanciamento delle classi va trattato sia in addestramento sia in valutazione

Sommer e Paxson (IEEE S&P 2010) spiegano lo scarso impiego operativo dell'anomaly detection basata su ML: scarsa capacità di riconoscere attività assenti dai dati di addestramento, costo elevato degli errori, distanza tra anomalia e attacco, variabilità del traffico, difficoltà di valutazione.

SCHEMA · APPROFONDIMENTI

Valutare e progettare un sistema di detection

APPROFONDIMENTO 1

Precision, recall e soglia di allarme

La **precision** è la quota di veri positivi tra gli elementi che il sistema segnala e misura quanto spesso un alert è fondato. Il **recall** è la quota di veri positivi tra gli attacchi effettivamente presenti e misura quanti attacchi il sistema intercetta. Al variare della soglia di allarme le due metriche tendono a muoversi in direzioni opposte: alzandola il recall non può aumentare e la precision di norma cresce, anche se non in modo monotono. La scelta della soglia è quindi una decisione di costo, che confronta il danno atteso di un attacco non rilevato con il tempo che gli analisti dedicano ai falsi allarmi.

Esempio

Su 100 attacchi presenti il modello ne segnala 90 (recall 0,90) e produce anche 60 falsi allarmi: la precision è 90/150 = 0,60. L'F1, media armonica delle due metriche, vale 0,72 e riassume il bilanciamento in un solo valore.

Metodo

Al variare della soglia si ottiene la curva ROC, che riporta il tasso di veri positivi in funzione di quello di falsi positivi; l'area sottesa (AUC) misura la capacità discriminante indipendentemente dalla soglia e serve a confrontare modelli. Con classi molto sbilanciate è più informativa la curva precision-recall.

APPROFONDIMENTO 2

Concept drift naturale e avversario

Un modello apprende la distribuzione dei dati disponibile al momento dell'addestramento, e in sicurezza questa distribuzione cambia rapidamente. Il **drift naturale** nasce da nuove applicazioni, migrazioni verso il cloud e cambi di abitudini, che spostano la normalità appresa dal modello. Il **drift avversario** è prodotto dagli attaccanti, che osservano ciò che viene rilevato e modificano le proprie tecniche di conseguenza; questa componente distingue il machine learning applicato alla sicurezza dalla maggior parte degli altri domini applicativi.

Esempio

Un classificatore di phishing addestrato prima della diffusione dei modelli linguistici generativi si basa in parte su errori grammaticali e costrutti innaturali. Nei messaggi generati da un LLM queste caratteristiche scompaiono e le prestazioni del classificatore calano.

Contromisure

Si monitorano la distribuzione degli input e le prestazioni in esercizio, si riaddestra periodicamente il modello con dati recenti e si attiva un allarme quando il tasso di segnalazione si discosta dalla serie storica.

APPROFONDIMENTO 3

Agenti basati su LLM

Un LLM usato da solo trasforma testo in testo, e l'errore possibile è una risposta sbagliata. Un **agente** è un LLM inserito in un ciclo in cui osserva, pianifica, invoca strumenti e valuta il risultato prima del passo successivo. Gli strumenti hanno effetti sui sistemi: interrogare un SIEM, aprire un ticket, bloccare un account. L'errore può quindi tradursi in un'azione sbagliata, e la sequenza di passi lo amplifica, perché una premessa errata al primo passo orienta tutti i successivi.

Scenario didattico

Un agente di triage legge un alert, interroga il SIEM sull'host coinvolto, consulta la threat intelligence sull'indirizzo IP e propone di chiudere il caso come falso positivo. Le tre chiamate agli strumenti e la conclusione vanno registrate perché siano verificabili.

Cautele

Ogni strumento concesso all'agente è una capacità operativa: si applicano il minimo privilegio, l'accesso in sola lettura come impostazione predefinita e l'approvazione umana per le operazioni di scrittura. I Moduli 6, 13 e 14 sviluppano questi principi.

APPROFONDIMENTO 4

Scelta del paradigma di apprendimento

La scelta del paradigma dipende in primo luogo dalla disponibilità di etichette. L'**apprendimento supervisionato** richiede esempi etichettati come malevoli o benigni e ne generalizza le regolarità: si presta alla classificazione di phishing e malware, per cui esistono ampie raccolte etichettate. L'**apprendimento non supervisionato** non richiede etichette, modella la struttura dei dati e segnala le deviazioni; è adatto a rilevare comportamenti mai osservati, al prezzo di molti falsi positivi. L'apprendimento per rinforzo ottimizza una politica in base a una ricompensa ed è studiato per la risposta automatica e per honeypot adattivi, ma è poco diffuso in produzione perché l'apprendimento per tentativi su sistemi in esercizio ha costi difficili da accettare.

Esempio

Per individuare account compromessi, un modello supervisionato addestrato sugli incidenti passati riconosce schemi già visti, mentre un modello non supervisionato basato su orari, volumi e destinazioni degli account segnala l'account che si discosta dal proprio storico anche in presenza di tecniche nuove. Se i due modelli si usano in sequenza, quello non supervisionato produce le segnalazioni e quello supervisionato aiuta a ordinarle.

Limiti

Le etichette dei dati di sicurezza sono sbilanciate e imperfette: gli attacchi etichettati sono quelli scoperti, mentre i dati considerati benigni possono contenere attacchi passati inosservati. Un modello supervisionato addestrato su questi dati eredita le lacune della detection precedente.

APPROFONDIMENTO 5

Feature engineering sui log di sicurezza

Un modello di machine learning opera su vettori numerici; il **feature engineering** traduce i log in questa forma. Da un flusso di autenticazioni si possono estrarre, per esempio, il numero di accessi falliti per finestra temporale, il rapporto tra successi e fallimenti, il numero di host distinti contattati, l'ora del giorno codificata come variabile ciclica e la novità della coppia utente-dispositivo. Un modello semplice con feature che descrivono il comportamento d'attacco ottiene spesso risultati migliori di un modello complesso con feature povere. Ogni feature è una scelta progettuale, e documentarla fa parte della spiegabilità (Modulo 8).

Esempio

Nel password spraying il conteggio degli accessi falliti per account non evidenzia anomalie, perché ogni account riceve uno o due tentativi. L'attacco emerge aggregando per origine: il numero di account distinti con accessi falliti dallo stesso indirizzo IP in una finestra temporale; se i tentativi arrivano da molti indirizzi, si aggrega sull'intero tenant.

Metodo

Il deep learning può apprendere rappresentazioni direttamente da sequenze di eventi (embedding), con minore spiegabilità e maggiore dipendenza dalla quantità di dati. Quando il risultato va difeso, come in ambito forense o disciplinare, le feature esplicite e documentate sono la scelta prudente.

APPROFONDIMENTO 6

Base rate fallacy e falsi positivi

Gli attacchi sono rari rispetto agli eventi legittimi e questa asimmetria altera l'interpretazione delle metriche. Si consideri un detector con recall del 99% e tasso di falsi positivi dell'1%, applicato a 10 milioni di eventi al giorno di cui 100 malevoli: produce circa 99 veri positivi e circa 100.000 falsi positivi. La probabilità che un alert corrisponda a un attacco è circa 99 su 100.100, cioè lo 0,1%. Trascurare questa frequenza di base e leggere il recall del 99% come affidabilità dell'alert è la **base rate fallacy**: con una prevalenza molto bassa, un tasso di falsi positivi piccolo in percentuale genera grandi numeri assoluti.

Esempio

Un fornitore dichiara di rilevare il 99,9% degli attacchi, un altro di produrre un falso positivo ogni 100.000 eventi. Le due affermazioni misurano grandezze diverse, il recall e il tasso di falsi positivi, e nessuna indica da sola quanti alert al giorno riceverà il SOC: in una prova sul campo occorre misurare il numero di alert giornalieri sul traffico dell'organizzazione e la quota di quelli fondati.

Metodo

Un singolo detector difficilmente sostiene questi volumi. Si progettano quindi architetture a cascata, con un primo stadio ad alto recall e stadi successivi di arricchimento, correlazione o classificazione con un secondo modello, che filtrano gli alert prima dell'analista. L'architettura è ripresa nei Moduli 3 e 5.

LABORATORIO · ESERCITAZIONE

Classificatore su traffico di rete

Svolgimento (Python / scikit-learn)

1. Caricamento di CIC-IDS2017, nella versione corretta pubblicata da Engelen et al., e analisi esplorativa
2. Addestramento di un classificatore su traffico benigno e malevolo
3. Valutazione: precision, recall, F1, curva ROC e curva precision-recall
4. Simulazione di una prevalenza realistica degli attacchi e misura dell'effetto sui falsi positivi

Consegna

- Notebook con metriche commentate
- Stima dei falsi positivi al giorno a prevalenza realistica
- Una pagina di valutazione sull'impiego in un SOC

CHIUSURA DEL MODULO

Letture e autovalutazione

Letture obbligatorie

- Goodfellow, Bengio, Courville, *Deep Learning*, MIT Press 2016, capp. 1 e 5
- Sommer & Paxson, *Outside the Closed World*, IEEE S&P 2010
- Axelsson, *The Base-Rate Fallacy and the Difficulty of Intrusion Detection*, ACM TISSEC 2000

Domande di autovalutazione

1. Perché un'accuratezza del 99% può essere insufficiente nella detection?
2. Cosa distingue fine-tuning e RAG?
3. Cos'è il concept drift e perché degrada i modelli di sicurezza?
4. Quali metriche privilegiare con classi fortemente sbilanciate?
5. Cosa si intende per agente autonomo basato su LLM?

FINE DELLA LEZIONE · MODULO 02

Grazie

Introduzione all'AI per la sicurezza · Corso «Incident Response potenziata dall'AI»

Domande e materiali: ricevimento del docente. Profilo e contatti: vincenzocalabro.it

Vincenzo Calabrò

CYBER SECURITY · DIGITAL FORENSICS

VINCENZOCALABRO.IT