PROTEGGERE LE APPLICAZIONI DELL'INTELLIGENZA ARTIFICIALE

Autore: *Aldo Pedico* – Enterprise Security & Privacy Architect

Contatto: pedicoaldo@gmail.com

1. INTRODUZIONE

Un passo verso la protezione delle applicazioni dell'Intelligenza Artificiale (AI), in particolare contro le manipolazioni contraddittorie del Machine Learning (ML), sviluppa una tassonomia e una terminologia dell'Adversarial Machine Learning (AML).

Sebbene l'intelligenza artificiale includa anche vari sistemi basati sulla conoscenza, l'approccio basato sui dati del ML introduce ulteriori sfide alla sicurezza nelle fasi di formazione e test (inferenza) delle operazioni di sistema.

L'AML si occupa della progettazione di algoritmi ML in grado di resistere le sfide alla sicurezza, lo studio delle capacità degli aggressori e la comprensione delle conseguenze degli attacchi.

I componenti ML di un sistema di IA includono i dati, il modello e i processi per l'addestramento, il test e la convalida.

Sebbene l'IA includa anche vari approcci basati sulla conoscenza, l'approccio basato sui dati del machine learning introduce ulteriori sfide alla sicurezza nelle fasi di formazione e test (inferenza) delle operazioni di ML.

Queste sfide alla sicurezza includono il potenziale per la manipolazione contraddittoria dei dati di addestramento e lo sfruttamento contraddittorio delle sensibilità del modello per influenzare negativamente le prestazioni della classificazione e della regressione ML.

L'AML si occupa della progettazione di algoritmi ML in grado di resistere alle sfide di sicurezza, dello studio delle capacità degli aggressori e della comprensione delle conseguenze degli attacchi.

Gli attacchi sono lanciati da avversari con intenzioni malevoli e la sicurezza del ML si riferisce alle difese intese a prevenire o mitigare le conseguenze di tali attacchi.

Nella sicurezza informatica sia la robustezza sia la resilienza sono misurate dal rischio, che è una misura della misura in cui un'entità (ad esempio un sistema) è minacciata da una circostanza o un evento potenziale (ad es. attacco).

NIST per la conduzione delle valutazioni dei rischi (la valutazione del rischio è una delle componenti fondamentali di un processo di gestione del rischio organizzativo) stabilisce la seguente definizione di SCOPO delle valutazioni:

> *Scopo delle valutazioni del rischio è informare i decisori e supportare le risposte al rischio identificando:*
>
> *(i) minacce rilevanti alle organizzazioni o minacce dirette attraverso organizzazioni contro altre organizzazioni;*
>
> *(ii) vulnerabilità sia interne che esterne alle organizzazioni;*
>
> *(iii) impatto (vale a dire, danno) alle organizzazioni che può verificarsi dato il potenziale di minacce che sfruttano le vulnerabilità; e*
>
> *(iv) probabilità che si verifichi un danno.*

Su tale base, un approccio basato sul rischio comincerebbe identificando le minacce, le vulnerabilità e gli impatti rilevanti.

Nel caso di AML:

1. *le* MINACCE *sono definite dai* **tipi di attacchi** *e dai* **contesti contraddittori** *in cui possono verificarsi gli attacchi;*
2. *le* VULNERABILITÀ *sono definite dai* **tipi di difese** *o dalla* **loro mancanza**, *per prevenire o mitigare gli attacchi;*
3. *gli* IMPATTI *sono definiti dalle conseguenze che derivano dagli attacchi e dalle difese associate contro tali attacchi.*

2. TASSONOMIA

L'intento è quello di introdurre una tassonomia dell'AML in modo che possa supportare gli sforzi per valutare e gestire i rischi operativi nelle applicazioni pratiche di ML.

I livelli più alti della tassonomia risultante includono vari aspetti di attacchi e difese, come illustrato dalla Figura 1 nel contesto delle fasi di addestramento e test (inferenza) della pipeline di apprendimento automatico.

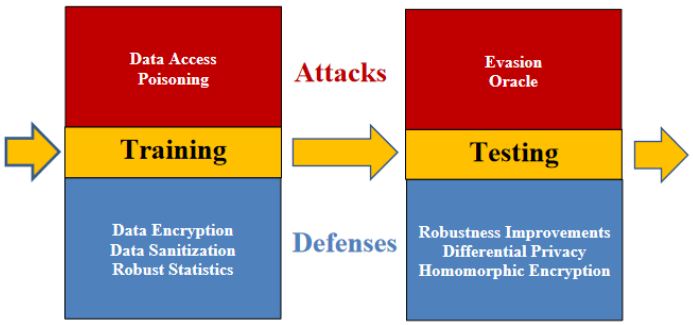

Figure 1. An illustration of example Attacks and Defenses in the Machine Learning Pipeline.

2.1 – ATTACCHI

2.1.1 – OBIETTIVI DEGLI ATTACCHI

Gli obiettivi degli attacchi sono definiti dalle fasi della pipeline ML, tra cui il Dominio Fisico dei sensori di input, la Rappresentazione Digitale per la pre-elaborazione, il modello di Apprendimento Automatico (ML) stesso o il Dominio Fisico delle azioni di output.

I tipi di metodi che generano un modello di apprendimento automatico (MACHINE LEARNING MODEL) *includono l'apprendimento supervisionato* (SUPERVISED LEARNING), *l'apprendimento non supervisionato* (UNSUPERVISED LEARNING) *e l'apprendimento per rinforzo* (REINFORCEMENT LEARNING).

1°. *Nel* **SUPERVISED LEARNING**, *i dati di addestramento sono forniti sotto forma di input etichettati con output corrispondenti e il modello apprende una mappatura tra input e output. L'attività di apprendimento è definita* ***classificazione*** quando gli output assumono valori categoriali *e* ***regressione*** quando gli output assumono valori numerici.

2°. *In* **SUPERVISED LEARNING**, *i dati di addestramento sono input senza etichetta e il modello apprende una struttura sottostante dei dati. Ad esempio, il modello può eseguire il*

raggruppamento di input in base a una metrica di somiglianza o una riduzione della dimensionalità per proiettare i dati in sottospazi dimensionali inferiori.

3°. *In* **REINFORCEMENT LEARNING**, *una politica basata sulla ricompensa (*REWARD-BASED*) per agire in un ambiente è appresa dai dati di addestramento rappresentati come sequenze di azioni, osservazioni e ricompense. In alcune applicazioni,* REINFORCEMENT LEARNING *può essere combinato con l'apprendimento supervisionato e l'apprendimento non supervisionato.*

2.1.2 – TECNICHE DI ATTACCO

*Le tecniche contraddittorie (*ADVERSARIAL TECHNIQUES*) utilizzate per lanciare attacchi contro obiettivi possono applicarsi alle fasi di addestramento o test (*INFERENZA*) del funzionamento del sistema. Gli attacchi nella fase di addestramento tentano di acquisire o influenzare i dati di addestramento o il modello stesso.*

*Negli attacchi di accesso ai dati, è possibile accedere ad alcuni o tutti i dati di addestramento e possono essere utilizzati per creare un modello sostitutivo. Questo modello sostitutivo può quindi essere utilizzato per testare l'efficacia di potenziali input prima di inviarli come Attacchi nella fase operativa di Test (*INFERENZA*).*

*Nell'avvelenamento (*POISONING*), noto anche come attacchi causali (*CAUSATIVE ATTACKS*), i dati o il modello vengono alterati indirettamente o direttamente.*

Nell'avvelenamento indiretto, gli avversari senza accesso ai dati preelaborati utilizzati dal modello di destinazione devono invece avvelenare i dati prima della preelaborazione.

Nell'avvelenamento diretto, i dati vengono alterati da DATA INJECTION *o* DATA MANIPULATION*, oppure il modello viene alterato direttamente da* LOGIC CORRUPTION*.*

In DATA INJECTION*, gli input contraddittori sono inseriti nei dati di training originali, modificando in tal modo la distribuzione dei dati sottostanti senza modificare le caratteristiche o le etichette dei dati di training originali.*

I campioni contraddittori iniettati possono essere ottimizzati mediante metodi di programmazione lineare che spostano il confine decisionale di un modello centrale (in UNSUPERVISED LEARNING*) o mediante ascesa del gradiente sull'errore di test del modello per degradare l'accuratezza della classificazione (in* SUPERVISED LEARNING*).*

*La manipolazione dei dati comporta la modifica in contraddittorio delle etichette di output (*LABEL MANIPULATION*) e dei dati di input (*INPUT MANIPULATION*) dei dati di addestramento originali.*

Gli **ATTACCHI NELLA FASE DI TEST (INFERENZA)**, *noti anche come attacchi esplorativi (*EXPLORATORY ATTACKS*), non alterano il modello di destinazione o i dati utilizzati nell'addestramento.*

Invece questi attacchi generano esempi contraddittori come input che sono in grado di eludere la corretta classificazione dell'output da parte del modello, in EVASION ATTACKS*, o raccogliere e dedurre informazioni sul modello o dati di addestramento, in* ORACLE ATTACKS*.*

In **EVASION ATTACKS**, *l'avversario risolve un problema di ottimizzazione vincolata per trovare una piccola perturbazione dell'input che provoca un grande cambiamento nella funzione di perdita e risulta in una classificazione errata dell'output.*

Questo in genere coinvolge algoritmi di ricerca basati su gradiente come Limitedmemory Broyden–Fletcher–Goldfarb–Shanno (L-BFGS), Fast Gradient Sign Method (FGSM) o Saliency Map Attack (JSMA) basato su Jacobian.

L-BFGS è stato il primo algoritmo utilizzato per generare classificazioni errate da un modello di sistema di visione artificiale utilizzando perturbazioni di input impercettibili per gli osservatori umani.

FGSM migliora l'efficienza computazionale della salita del gradiente, in un approccio Single Step che elimina le iterazioni necessarie per ottenere una perturbazione che causerà un grande cambiamento nella funzione di perdita.

Rispetto a FGSM, JSMA è un algoritmo iterativo che fornisce un controllo più dettagliato delle caratteristiche perturbate e quindi può generare esempi contraddittori più convincenti, anche se a un costo computazionale maggiore.

Questi e altri algoritmi per gli attacchi di evasione richiedono la conoscenza del modello, o un modello sostitutivo, al fine di calcolare i gradienti nella funzione di perdita attraverso gli accoppiamenti input-output.

In **ORACLE ATTACKS**, *un avversario utilizza un'interfaccia di programmazione dell'applicazione per presentare il modello con gli input e per osservare gli output del modello.*

Anche quando l'avversario non ha una conoscenza diretta del modello stesso, gli accoppiamenti input-output ottenuti da ORACLE ATTACKS *possono essere utilizzati per addestrare un modello sostitutivo che opera in modo molto simile al modello di destinazione, a causa della proprietà di trasferibilità esibita da molte architetture di modello.*

Questo modello sostitutivo, a sua volta, può quindi essere utilizzato per generare esempi contraddittori da utilizzare negli attacchi di evasione contro il modello bersaglio.

Gli **ORACLE ATTACK** *includono attacchi di:*

1. ESTRAZIONE;
2. INVERSIONE;
3. INFERENZA *dell'appartenenza.*

In **EXTRACTION ATTACKS**, *un avversario estrae i parametri o la struttura del modello dalle osservazioni delle previsioni del modello, includendo tipicamente le probabilità restituite per ogni classe.*

Nel caso di **INVERSION ATTACK**, *le caratteristiche dedotte possono consentire all'avversario di ricostruire i dati utilizzati per addestrare il modello, comprese le informazioni personali che violano la privacy di un individuo.*

In un **MEMBERSHIP INFERENCE ATTACK**, *l'avversario utilizza i ritorni delle query del modello di destinazione per determinare se punti dati specifici appartengono alla stessa distribuzione del set di dati di addestramento, sfruttando le differenze nella confidenza del modello sui punti che sono stati o non sono stati visti durante l'addestramento.*

2.1.3 – SAPERE

Oltre alle tecniche utilizzate per lanciare gli attacchi contro i bersagli, le minacce ai componenti di Machine Learning dipendono anche dalla conoscenza dell'avversario del modello di destinazione.

BLACK BOX ATTACKS, *l'avversario non ha alcuna conoscenza del modello eccetto campioni input-output di dati di addestramento o accoppiamenti input-output ottenuti utilizzando il modello di destinazione come un* ORACLE.

- ✓ **GREY BOX ATTACKS**, *l'avversario ha informazioni parziali sul modello, che possono includere l'architettura del modello, i valori dei parametri, il metodo di addestramento (funzione di perdita) o i dati di addestramento.*
- ✓ **WHITE BOX ATTACKS**, *l'avversario ha una conoscenza completa del modello, inclusi architettura, parametri, metodi e dati. Anche quando un avversario non ha la conoscenza completa necessaria per un attacco* WHITE BOX, *gli attacchi* DATA ACCESS *o* ORACLE *che producono accoppiamenti input-output possono essere utilizzati per addestrare un modello sostitutivo, che funziona in modo molto simile al modello reale a causa della proprietà di trasferibilità esibita da molte architetture modello. Questo modello sostitutivo può quindi essere utilizzato come* WHITE BOX *per generare esempi contraddittori da utilizzare negli attacchi di evasione.*

2.2 - DIFESE

Le difese possono essere caratterizzate dal fatto che si applichino agli attacchi lanciati contro le fasi di ADDESTRAMENTO *o di* TEST (INFERENZA) *del funzionamento del sistema.*

In entrambi i casi, i metodi difensivi spesso possono comportare un sovraccarico delle prestazioni e avere un effetto dannoso sull'accuratezza del modello.

Le **DEFENSES AGAINST TRAINING ATTACKS** *che coinvolgono l'accesso ai dati includono tradizionali misure di controllo degli accessi come la crittografia dei dati.*

Le difese contro gli **POISONING ATTACKS** *includono la sanificazione dei dati e statistiche robuste.*

In **DATA SANITIZATION**, *gli esempi contraddittori sono identificati testando l'impatto degli esempi sulle prestazioni di classificazione.*

Gli esempi che causano alti tassi di errore nella classificazione vengono quindi rimossi dal training set, in un approccio noto come REJECT ON NEGATIVE IMPACT.

Piuttosto che tentare di rilevare dati avvelenati, le statistiche robuste utilizzano vincoli e tecniche di regolarizzazione per ridurre le potenziali distorsioni del modello di apprendimento causate da dati avvelenati.

Le difese contro gli ATTACCHI *di* TEST (INFERENZA) *includono vari miglioramenti della robustezza del modello, tra cui:*

- ✓ ADVERSARIAL TRAINING;
- ✓ GRADIENT MASKING;
- ✓ DEFENSIVE DISTILLATION;
- ✓ ENSEMBLE METHODS;
- ✓ FEATURE SQUEEZING;
- ✓ REFORMERS/AUTOENCODER.

Sebbene utilizzate come Difese contro gli Attacchi effettuati nella fase di TEST *(Inferenza), queste Difese sono schierate dal difensore nella fase di* ADDESTRAMENTO *che precede il* TEST *(Inferenza).*

In **ADVERSARIAL TRAINING**, *gli input contenenti perturbazioni contraddittorie ma con etichette di output corrette sono iniettati nei dati di training al fine di ridurre al minimo gli errori di classificazione causati da esempi contraddittori.*

Il **GRADIENT MASKING** *riduce la sensibilità del modello a piccole perturbazioni negli input calcolando le derivate del primo ordine del modello rispetto ai suoi input e riducendo al minimo queste derivate durante la fase di apprendimento.*

Un'idea simile motiva **DEFENSIVE DISTILLATION**, *in cui un modello di destinazione è utilizzato per addestrare un modello più piccolo che presenta una superficie di output più liscia e* **METODI ENSEMBLE**, *in cui più classificatori sono addestrati insieme e combinati per migliorare la robustezza.*

Allo stesso modo, **FEATURE SQUEEZING**, *utilizza trasformazioni di livellamento delle funzionalità di input nel tentativo di annullare le perturbazioni contraddittorie.*

I riformatori prendono un dato input e lo spingono verso l'esempio più vicino nel set di addestramento, in genere utilizzando reti neurali chiamate **AUTOENCODER**, *per contrastare le perturbazioni contraddittorie.*

È importante riconoscere che l'avversario può sconfiggere vari ROBUSTNESS IMPROVEMENT DEFENSES *lanciando* DATA ACCESS *o* ORACLE ATTACKS *per ottenere accoppiamenti input-output.*

Questi abbinamenti possono essere successivamente utilizzati per addestrare un modello sostitutivo che non maschera i gradienti o uscite uniformi come il modello di destinazione.

Il modello sostitutivo può quindi essere utilizzato come **WHITE BOX** *per creare esempi contraddittori, sfruttando la proprietà di trasferibilità dei* ML-TRAINED *models, quindi può essere difficile difendersi dagli di* EVASION ATTACKS *da parte di un avversario in grado di creare un modello sostitutivo.*

Oltre ai ROBUSTNESS IMPROVEMENTS *sopra menzionati, le Defenses Against Testing (inferenza) includono anche meccanismi di randomizzazione applicati ai dati di addestramento o agli output del modello per fornire garanzie di* DIFFERENTIAL PRIVACY.

La privacy differenziale formula la privacy come una proprietà soddisfatta da un meccanismo di randomizzazione su coppie di dataset adiacenti.

In definitiva, la proprietà DIFFERENTIAL PRIVACY *garantisce che gli output del modello non rivelino alcuna informazione aggiuntiva su un singolo record incluso nei dati di addestramento.*

Un approccio alternativo è la CRITTOGRAFIA OMOMORFICA *(***HOMOMORPHIC ENCRYPTION***), che crittografa i dati in una forma che una rete neurale può elaborare senza decifrare i dati. Ciò protegge la privacy di ogni singolo input ma introduce un sovraccarico delle prestazioni computazionali e limita l'insieme delle operazioni aritmetiche a quelle supportate dalla* **HOMOMORPHIC ENCRYPTION**.

2.3 – RIPERCUSSIONI

Le conseguenze degli attacchi contro gli obiettivi dipendono dalle difese implementate. Per una data combinazione di Attacco (inclusi OBIETTIVO, TECNICA *e* CONOSCENZA*) e Difesa, le conseguenze possono essere classificate categoricamente come:*

- ✓ VIOLATIONS OF INTEGRITY;
- ✓ AVAILABILITY;
- ✓ CONFIDENTIALITY O PRIVACY.

All'interno di ciascuna categoria, possono essere utilizzati anche diversi livelli di gravità per misurare la violazione della sicurezza.

In **INTEGRITY VIOLATIONS**, *il processo di inferenza è compromesso, con conseguente* **CONFIDENCE REDUCTION** *della fiducia o* **MISCLASSIFICATION** *a qualsiasi classe diversa dalla classe originale.*

Errori di classificazione più specifici includono l'errata classificazione mirata degli input a una specifica classe di output target e l'errata classificazione SOURCE-TARGET *di uno specifico input a una specifica classe di output target.*

Nel **UNSUPERVISED LEARNING**, *una* INTEGRITY VIOLATION *può produrre una rappresentazione priva di significato dell'input in un estrattore di funzionalità non supervisionato.*

In **REINFORCEMENT LEARNING**, *una* INTEGRITY VIOLATION *può far sì che l'agente di apprendimento agisca in modo poco intelligente o con prestazioni degradate nel suo ambiente.*

Le **AVAILABILITY VIOLATIONS** *inducono riduzioni della qualità (come la velocità di inferenza) o dell'accesso (*DENIAL OF SERVICE*) al punto da rendere il componente ML non disponibile per gli utenti.*

Sebbene le violazioni della disponibilità possano comportare riduzioni dell'affidabilità o classificazioni errate simili a quelle delle violazioni dell'integrità, la differenza è che le violazioni della disponibilità comportano comportamenti come velocità inaccettabile o negazione dell'accesso che rendono inutilizzabile l'output o l'azione di un modello.

VIOLAZIONI

CONFIDENTIALITY VIOLATIONS *si verificano quando un avversario estrae o deduce informazioni utilizzabili sul modello e sui dati.*

Gli attacchi alle informazioni riservate sul modello includono un EXTRACTION ATTACK *che rivela l'architettura o i parametri del modello o un* ORACLE ATTACK *che consente all'avversario di costruire un modello sostitutivo.*

Gli attacchi che rivelano informazioni riservate sui dati includono un attacco di inversione in cui un avversario sfrutta il modello di destinazione per recuperare i dati mancanti utilizzando input parzialmente noti o un attacco di inferenza di appartenenza in cui un avversario esegue un test di appartenenza per determinare se un individuo è stato incluso nel set di dati utilizzato per addestrare il modello di destinazione.

PRIVACY VIOLATIONS *sono una classe specifica di violazione della riservatezza in cui l'avversario ottiene informazioni personali su uno o più input del modello individuali e legittimi, inclusi o meno nei dati di formazione.*

Un esempio potrebbe essere quando un avversario acquisisce o estrae le cartelle cliniche di un individuo in violazione delle politiche sulla privacy.