Security for A.I.

Parte III

Indice

Capitolo 1

STRIDE-AI

Con *Security for A.I.* si intende fare un *threat modeling* per modelli di intelligenza artificiale; in questa lezione vediamo l'approccio STRIDE-AI.
STRIDE-AI vuole fornire una modalità strutturata per capire quali sono gli asset critici, per poi identificare quali sono i rispettivi failure mode e relativi threat.

Step		Description
1	Objectives Identification	*States the security properties the system should have.*
2	Survey	*Determines the system's assets, their interconnections and connections to outside systems.*
3	Decomposition	*Selects the assets that are relevant for the security analysis.*
4	Threat Identification	*Enumerates threats to the system's components and assets that may cause it to fail to achieve the security objectives.*
5	Vulnerabilities Identifications	*Examines identified threats and determines if known attacks show that the overall system is vulnerable to them.*

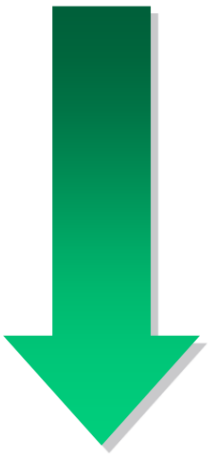
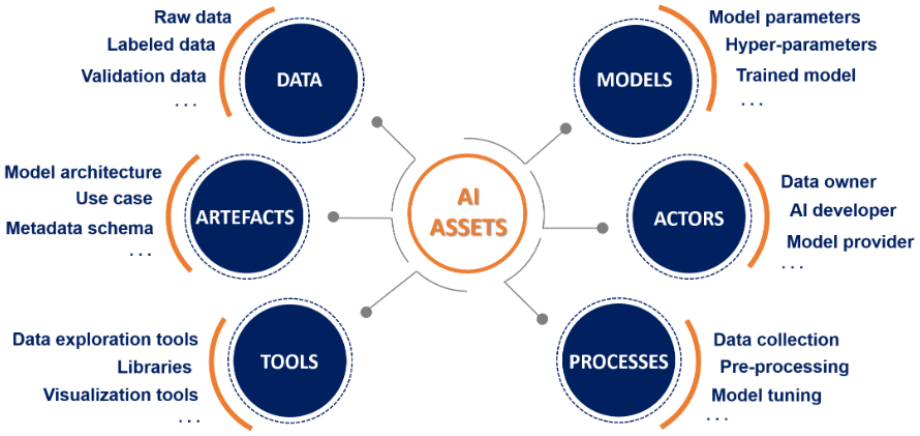


Figura 1.1: Processo di threat modeling



STRIDE-AI prende i threat classici della metodologia STRIDE e li mappa in delle proprietà specifiche per modelli di intelligenza artificiale:

STRIDE THREATS IN A NUTSHELL.

Threat	Description
Spoofing Identity	*A user takes on the identity of another. For example, an attacker takes on the identity of an administrator.*
Tampering with Data	*Information in the system is modified by an attacker. For example, an attacker changes a data item.*
Repudiation	*Information about a transaction is deleted in order to deny it ever took place. For example, an attacker deletes a login transaction to deny he ever accessed an asset.*
Information Disclosure	*Sensitive information is stolen and sold for profit. For example, information on user behavior is stolen and sold to a competitor.*
Denial of Service (DoS)	*This involves exhausting resources required to offer services. For example, in a DoS against a data flow the attacker consumes network resources.*
Elevation of Privilege (EoP)	*This is a threat similar to spoofing, but instead of taking on the ID of another, the attacker elevates his own security level to an administrator.*

ML-SPECIFIC $CIA^3 - R$ HEXAGON.

Property	ML-specific definition
Authenticity	*The output value delivered by a model has been verifiably generated by it.*
Integrity	*Information used or generated throughout a model's life-cycle cannot be changed or added to by unauthorized third parties.*
Non-repudiation	*There is no way to deny that a model's output has been generated by it.*
Confidentiality	*Using a model to perform an inference exposes no information but the model's input and output.*
Availability	*When presented with inputs, the model computes useful outputs, clearly distinguishable from random noise.*
Authorization	*Only authorized parties can present inputs to the model and receive the corresponding outputs.*

Dunque, il processo che segue è:

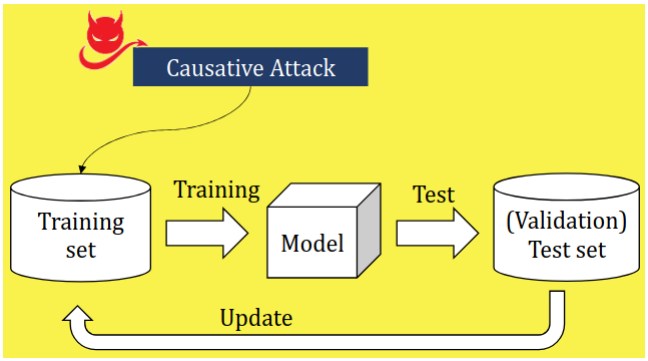
- fare un'analisi dell'architettura del sistema
- identificare quali sono gli asset
- identificare quali sono i failure mode per ciascun asset
- dai failure mode risalire a quali sono le proprietà di sicurezza che vengono negate

Ad ogni proprietà sono associati dei threat noti, e si possono applicare delle misure di sicurezza.

Capitolo 2

Sicurezza di modelli ML contro attacchi di poisoning

In questa lezione vediamo un framework di difesa contro attacchi di poisoning a modelli di machine learning.
Una delle assunzioni chiavi del machine learning è che un modello addestrato sui dati di training avrà una buona accuratezza e performerà bene dato che il training set è ben rappresentativo dei dati che il modello vedrà in fase di deploy... un training malevolo, dovuto ad esempio ad attacchi di poisoning, può mettere fuori uso un modello di machine learning.



Alcune strategie di difesa contro attacchi ai dati di training sono:

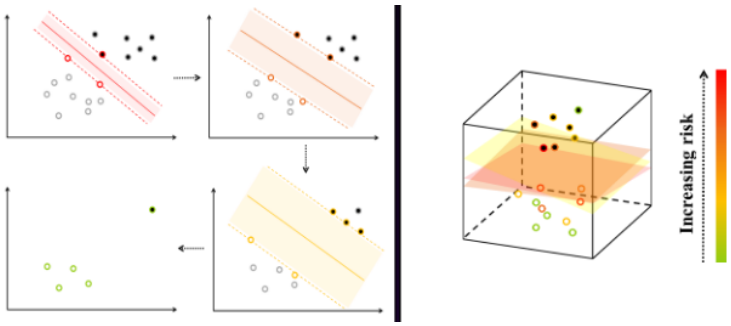
- **meccanismi di rilevamento** che mirano ad identificare i dati avvelenati e sanitificarli, oppure escludere i dati sospetti (come gli outlier)
- **meccanismi di *miglioramento*** che agiscono nella fase di addestramento cercando di impedire che possa avvenire l'avvelenamento (ad esempio vengono introdotti dati avvelenati per ridurre la sensibilità ad un attacco)

- **model composition** che prevede di ridurre l'interferenza dei punti avvelenati tramite una suddivisione del training set

Bisogna sempre considerare un trade-off tra accuratezza e robustezza di un modello di machine learning.

Analisi del rischio dei data assets di un modello di ML

- Viene usato un classificatore *Support Vector Machine* (ovvero un tipo di classificatore lineare) per stimare il rischio associato ai dati di addestramento
- L'idea è quella di mettere in relazione l'indice di rischio con la vicinanza dei punti ai SVM, ovvero lo superfici di separazione delle classi
- L'attaccante usa una sua SVM per stilare quali sono secondo lui i punti più a rischio, e decidere in quale direzione fare una modifica
- Il risultato è ottenere delle partizioni del training set che vengono utilizzate per addestrare dei sottomodelli, che saranno poi combinati tra loro per ottenre il risultato finale



La strategia prevede di usare l'*anti-clustering* per ottenere delle partizioni con una massima diversità interna e una minima diversità tra i gruppi; in particolare, si prevede che punti vicini con alti livelli di rischio siano distribuiti in partizioni diverse, in modo da ridurre il rischio di compromettere più sottomodelli.
Possiamo quindi dire che:

- suddividendo oppurtanamente i dati e combinando le predizioni dei modelli, **gli errori sui dati risultano indipendenti tra loro**, garantendo così robustezza
- segue che un **numero sufficiente di sottomodelli intatti** garantisce che il risultato finale non sia corrotto.