Testo

M.I.A e I.I.A

I modelli di machine learning possono presentare dei rischi per la privacy nel caso in cui facciano trasparire delle informazioni sensibili sui dati di training attraverso il loro output.
Si parla di Membership Inference Attack se è possibile con una certa accuratezza determinare se un certo dato fosse presente nei dati di training di un modello.
Si parla invece di Identity Inference Attack se è possibile determinare se nel training dataset fosse presente qualche dato corrispondente ad una certa persona avendo a disposizione un altro dato della stessa persona (ad esempio a partire da una foto di un individuo determinare se nel training dataset fosse presente un'altra foto dello stesso individuo).

Privacy nei modelli di machine learning

Dobbiamo definire cosa si intende per privacy nel machine learning:
Una definizione possibile è quella che viene chiamata "differential privacy" ovvero si richiede che un attaccante non possa usare il modello per risalire a informazioni sul training dataset che non avrebbe potuto dedurre da un altro modello

addestrato sulla stessa distribuzione del dataset.
Questo quindi non significa che non si possa dedurre alcuna informazione sulla distribuzione dei dati di training. Infatti sarebbe una richiesta troppo forte in quanto un modello per essere utile deve generalizzare il più possibile le informazioni apprese dal training dataset a tutta la distribuzione che ha generato il dataset.
Richiedere che dal modello non si possa dedurre alcuna informazione sulla distribuzione del dataset equivale a chiedere che il modello non funzioni.
Differential privacy quindi è la richiesta che dal modello non si possa risalire a informazioni specifiche dei dati della distribuzione che erano presenti nel training dataset.
Ad esempio un modello che identifica che una certa categoria di persone è predisposta ad una certa malattia non viola la differential privacy se:

1. È possibile costruire un altro modello tale che venendo addestrato sulla stessa distribuzione troverà la stessa relazione tra categoria di persone e malattia.
2. Non è possibile ricavare informazioni sulle persone presenti nel training dataset, quindi il modello deve essere resistente a M.I.A. e I.I.A.
   In questo esempio si nota che violare la differential privacy è particolarmente grave in quanto permette di accedere allo stato di salute di una specifica persona se presente nel dataset di addestramento.

Comprehensive_Privacy_Analysis_of_Deep_Learning_Passive_and_Active_White-box_Inference_Attacks_against_Centralized_and_Federated_Learning.pdf (qua evidenziato in rosa e anche nell'altro paper di Reza) TODO da mettere a posto

Black Box e White Box

Per gli inference attacks possiamo usare due paradigmi diversi.

- White box (scatola bianca) ovvero l'attaccante conosce informazioni sul modello da attaccare come, tipo di modello utilizzato, numero di layer, parametri.
- Black box (scatola nera) ovvero l'attaccante non ha alcuna informazione sul funzionamento interno del modello e può solo utilizzarlo osservando gli output corrispondenti agli input che gli fornisce.
  Il modello black box può essere visto come una API a cui è possibile fare delle richieste con degli input scelti dall'attaccante e ottenere le risposte da utilizzare per l'attacco.
  Ci concentreremo su questo secondo modello in quanto è più simile ad un caso reale in cui un attaccante effettui un M.I.A. su un modello a cui non ha accesso direttamente.

M.I.A. su classificatori

Chiamiamo $\mathcal{T}$ il modello target su cui vogliamo svolgere l'attacco.
Chiamiamo $\mathcal{D}_{train,\mathcal{T}}$ il dataset di training di $\mathcal{T}$.

Nei modelli di machine learning di classificazione, il modello determina a quale tra
k classi è più probabile appartenga l'input.
Il classificatore da in output un vettore lungo k dove ogni componente rappresenta la probabilità che l'input appartenga alla corrispondente classe. Ad esempio $\begin{bmatrix} cane \\ gatto \\ orso \\ volpe \end{bmatrix} = \begin{bmatrix} 0.6 \\ 0.1 \\ 0.1 \\ 0.2 \end{bmatrix}$
L'intuizione su cui ci basiamo è che il modello classificherà in maniera diversa input che erano già presenti nei dati di training ($\mathcal{D}_{train,\mathcal{T}}$), per esempio con una confidenza maggiore.
Ad esempio:
$\begin{bmatrix} cane \\ gatto \\ orso \\ volpe \end{bmatrix} = \begin{bmatrix} 0.9 \\ 0.025 \\ 0.025 \\ 0.05 \end{bmatrix}$
Come mostra il Paper di REZA(aggiungi link TODO) l'overfitting del modello da attaccare $\mathcal{T}$ rende maggiori le differenze nella confidenza della classificazione tra dati nuovi e dati già visti dal modello durante il training.
Quindi l'overfitting facilita attacchi di inferenza sul modello.

Possiamo quindi addestrare un nuovo modello di machine learning $M_{inference}$ (un classificatore binario) che a partire da

queste differenze nell'output tra dati in $\mathcal{D}_{train,\mathcal{T}}$ e in $\overline{\mathcal{D}_{train,\mathcal{T}}}$ (il complemento)
determini se l'input $\in \mathcal{D}_{train,\mathcal{T}}$ o no.

Per poter addestrare $\mathcal{M}_{inference}$ avremmo bisogno dei vettori di predizione (e.g. $\begin{bmatrix} cane \\ gatto \\ orso \\ volpe \end{bmatrix}$) con la corrispondente label `in` o `out` in base all'appartenenza a $\mathcal{D}_{train,\mathcal{T}}$.

Non abbiamo a disposizione questi dati per il modello $\mathcal{T}$ quindi creiamo una serie di "shadow models" $\mathcal{S}_i$ che andranno a imitare $\mathcal{T}$.
Siccome questi $\mathcal{S}_i$ sono creati da noi possiamo controllarne il training set $\mathcal{D}_{train,\mathcal{S}_i}$(creazione dataset shadow) e quindi abbiamo a disposizione dei vettori di predizione con la corrispondente label `in` e `out`.
Possiamo quindi addestrare il classificatore binario $\mathcal{M}_{inference}$ in modo che capisca se un certo input appartenga a $\mathcal{D}_{train,\mathcal{S}_i}$ oppure no in base al vettore di predizione corrispondente.
L'idea è che se gli shadow models $\mathcal{S}_i$ si comportano in maniera abbastanza simile a $\mathcal{T}$ la capacità di $\mathcal{M}_{inference}$ di discriminare in e out di $\mathcal{D}_{train,\mathcal{S}_i}$ si tradurrà nella capacità di discriminare in e out in $\mathcal{D}_{train,\mathcal{T}}$.
In questo caso avere informazioni aggiuntive su $\mathcal{T}$ (ad esempio il tipo del modello) permette di creare dei $\mathcal{S}_i$ più simili aumentando l'accuratezza del nostro attacco.

Per l'addestramento dei $\mathcal{S}_i$ avremo che $\mathcal{D}_{train,\mathcal{S}_i}$ contiene tutti dati con una determinata classe i (ad esempio $\mathcal{D}_{train,\mathcal{S}_i}$ ha tutti cani,$\mathcal{D}_{train,\mathcal{S}_2}$ ha tutti gatti ecc...), questo perché la distribuzione del vettore di predizione può dipendere dalla classe dell'input (ad esempio se è più facile per un modello classificare certi animali rispetto ad altri).
Per il modello $\mathcal{M}_{inference}$ si può utilizzare qualsiasi modello che permetta la classificazione binaria.

---

creazione dataset shadow

come creiamo i dataset shadow TODO se serve

M.I.A su GAN