UNIVERSITÀ
DEGLI STUDI
DI BRESCIA

Sicurezza Informatica
A.A. 2021-2022
Proff. Federico Cerutti
Fabio Bresciani

Regolarizzazione come difesa dagli attacchi membership inference

Alessandro Gaudenzi

Indice

1 Introduzione

Negli ultimi anni si è visto un incremento dell'utilizzo di tecniche di machine learning in molti ambiti, sia commerciali che non, ad esempio allo scopo di ricerca oppure di migliorare servizi esistenti. Conseguenza di ciò è l'utilizzo di dati, anche sensibili, appartenenti agli utenti finali.

Questa situazione ha portato ad una discussione sulla privacy degli utenti, anche in luce della recente normativa europea del GDPR: il risultato è stato riuscire a dimostrare l'esistenza di tecniche che possono circuire l'apparente struttura black box dei modelli di machine learning e la possibilità di recuperare informazioni sui dati utilizzati da per l'addestramento degli stessi da parte degli attaccanti.

Una di queste tecniche, la membership inference attack, consiste nell'addestrare un modello di attacco che impara a riconoscere le istanze appartenenti o meno al training set del modello attaccato sulla base della distribuzione di probabilità in uscita da esso per ogni input.

Andando poi ad indagare sulle cause della fuoriuscita di informazioni, si è scoperto che una di esse è l'overfitting del modello sul training set e che di conseguenza le tecniche atte a ridurlo possono costituire una buona linea di difesa, tra queste vi è la regolarizzazione.

Avendo visto che in letteratura si è posta enfasi soprattutto sulla regolarizzazione L2, in questo progetto si è cercato di fare un confronto tra essa e tra le altre due principali forme di regolarizzazione, L1 ed Elastic Net, cercando di capire se anche esse possono essere utilizzate come difesa dagli attacchi di membership inference.

2 Background

2.1 Privacy nei modelli di machine learning

Idealmente un modello di machine learning, quando interrogato, non dovrebbe rivelare più informazioni sull'input di quante si possano ricavare esaminando l'input stesso. Applicare letteralmente questa definizione porta ad un modello poco utile, perchè almeno la classe dell'input deve essere rivelata per avere una qualche utilità. Inoltre se il modello è basato su fatti statistici relativi agli attributi degli oggetti di una popolazione, permette di avere informazioni e quindi violare questa proprietà sull'intera popolazione.

Una definizione di privacy ristretta sugli elementi del training set è più significativa nel mondo reale, siccome nel caso di utilizzo di dati sensibili (per esempio in campo medico o finanziario) è essenziale evitare che un malintenzionato inferisca informazioni sulle istanze del training set. Il livello di privacy raggiunto dal modello in questo caso viene considerato come quantità di informazione che il modello ha inferito che differenzia gli elementi del training set con gli altri elementi della popolazione con la stessa distribuzione statistica. Se questa informazione viene fatta trapelare dall'output del modello, allora un attaccante può riconoscere con una certa probabilità la presenza o meno di un determinato elemento nel training set.

Per fare ciò l'attaccante deve montare un Membership Inference Attack (MIA): consiste nell'addestrare molteplici modelli ombra che simulano il comportamento del modello bersaglio per poter costruire, attraverso l'output di quest'ultimi, un dataset binario in cui ogni istanza, viene classificata come appartenente o non appartenente al training set. Successivamente si può usare il dataset così ottenuto per addestrare un modello attaccante che classificherà le istanze come presenti o assenti. Per l'addestramento dei modelli ombra si possono stimare le caratteristiche del modello bersaglio e utilizzare per l'addestramento delle istanze di una popolazione con una simile distribuzione di probabilità di quelle utilizzate nel training set del modello bersaglio, perchè anche in presenza di rumore importante nei modelli e nei dati proxy il grado di accuratezza resta comunque buono: non essendo quindi necessario un accesso white-box al modello bersaglio, questa tipologia di attacca risulta relativamente facile da montare.

Le cause principali di vulnerabilità di un modello a questi attacchi sono: l'overfitting del modello sul training set, la presenza di molte classi, la predisposizione di alcuni modelli rispetto ad altri e l'eccessiva ingegnerizzazione del modello stesso (per esempio nel caso di una rete neurale con troppi layer rispetto al necessario)

2.2 Overfitting

L'overfitting si verifica nel momento in cui il modello si adatta troppo alle istanze del training set e fallisce nel generalizzare bene a tutta la popolazione di cui il training set fa parte. È facilmente riconoscibile confrontando la differenza di performance del modello sul training set e sul validation set rispetto a quelle ottenute sul test set formato da istanze non ancora

conosciute dal modello. Se la differenza è elevata significa che il modello si comporta in modo diverso in base alla conoscenza pregressa o meno delle istanze.

Questa differenza di comportamento è anche il motivo per cui il livello di vulnerabilità del modello è correlato all'overfitting: l'attacco di membership inference si basa soprattutto sull'individuazione di pattern differenti nell'elaborazione di oggetti appartenenti al training set rispetto ai non appartenenti.

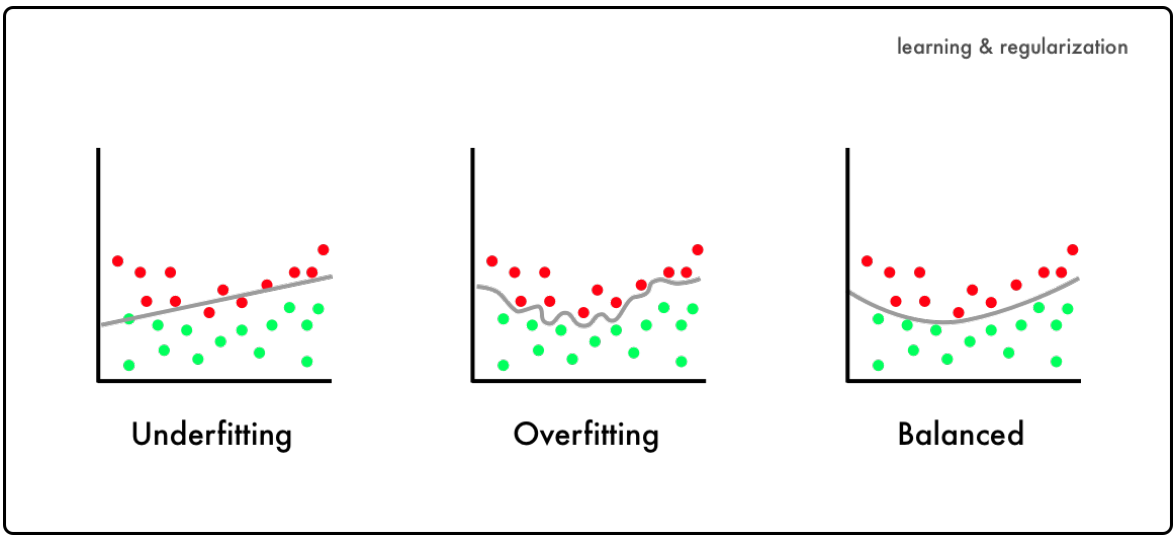

Figura 1: Esempio di overfitting e underfitting

2.3 Regolarizzazione

La regolarizzazione è delle tecniche impiegate per ridurre l'overfitting di un modello (e di conseguenza la sua vulnerabilità agli attacchi membership inference). Consiste nel aggiungere un'ulteriore penalità alla funzione di errore del modello, in modo che risulti più rilassato impedendo ai coefficienti di assumere valori estremi.

L1: La penalità è il valore assoluto dei coefficienti, stima la mediana dei dati, viene usata per la riduzione dimensionale e per la selezione di feature, siccome ne azzera effettivamente alcune

L2: La penalità è il quadrato dei coefficienti, stima la media dei dati, risulta utile in caso di feature collineari

Elastic net: Combina le precendenti, eliminando prima alcune feature e successivamente applicando la regolarizzazione L2

3 Metodologia

È importante valutare il membership risk di modelli sviluppati che trattano dati sensibili e individuare soluzione per minimizzarlo. In letteratura si è già definito che una causa (sufficiente ma non necessaria) del leak di informazioni è la presenza di overfitting: i risultati sperimentali confermano questa ipotesi mostrando una correlazione tra la differenza tra accuratezza in validation set e test set (principale indicatore di overfitting) e la vulnerabilità del modello ad un attacco di membership inference. L'attenzione dev'essere quindi rivolta a tutte quelle tecniche che mirano a ridurre l'overfitting, cercando di mantenere e/o migliorare le performance del modello allo stesso tempo.

Dalla letteratura sappiamo che l'utilizzo della regolarizzazione come tecnica di riduzione dell'overfitting permette di ridurre la vulnerabilità del modello. Si trovano principalmente esperimenti con la regolarizzazione L2, che è anche la più usata. È di interesse confrontarla con altre due tecniche di regolarizzazione, L1 e Elastic Net, al fine di verificare in che misura quest'ultime permettono di ridurre la vulnerabilità del modello.

3.1 Ipotesi

Tramite il confronto tra le regolarizzazioni si può stabilire se anche le altre tecniche di regolarizzazione meno usate siano efficaci nell'aumentare la privacy di un modello, stabilendo inoltre ulteriormente la correlazione tra overfitting e vulnerabilità, essendo più di una soluzione ad esso correlata all'aumento della privacy del modello.

Inoltre un altro obiettivo è valutare se queste tecniche di regolarizzazione vadano a ledere le performance del modello in maniera eccessiva: infatti se generalmente sono utilizzate per ridurre l'overfitting eccessivo e migliorare le prestazioni del modello, non è scontato che su un modello che presenta poco overfitting (ma comunque sufficiente a minarne la privacy) non possa avvenire il contrario.

3.2 Esperimento

È stata utilizzato come riferimento un modello basato su rete neurale convoluzionale, formata da: internamente due layer convoluzionali e un layer Dense con funzione di attivazione ReLU, in uscita un Layer Dense con funzione lineare.

Sono stati scelti due dataset di classificazioni immagini, il Cifar10 e il fashion-mnist, perchè sono tra gli standard utilizzati in letteratura sia per la valutazione della privacy sia per le performance dei modelli in generale. Inoltre la presenza di dieci classi comporta una vulnerabilità maggiore agli attacchi di membership inference rispetto ai dataset binari o con poche classi.

Altri dataset considerati per l'esperimento sono stati il dataset UCI seeds e il dataset kaggle Customer, entrambi scartati perchè utilizzando come modello una rete neurale comparabile a quella utilizzata definitivamente non presentavano abbastanza overfitting tale da minarne la privacy: si poteva gonfiare artificialmente l'overfitting andando a peggiorare il

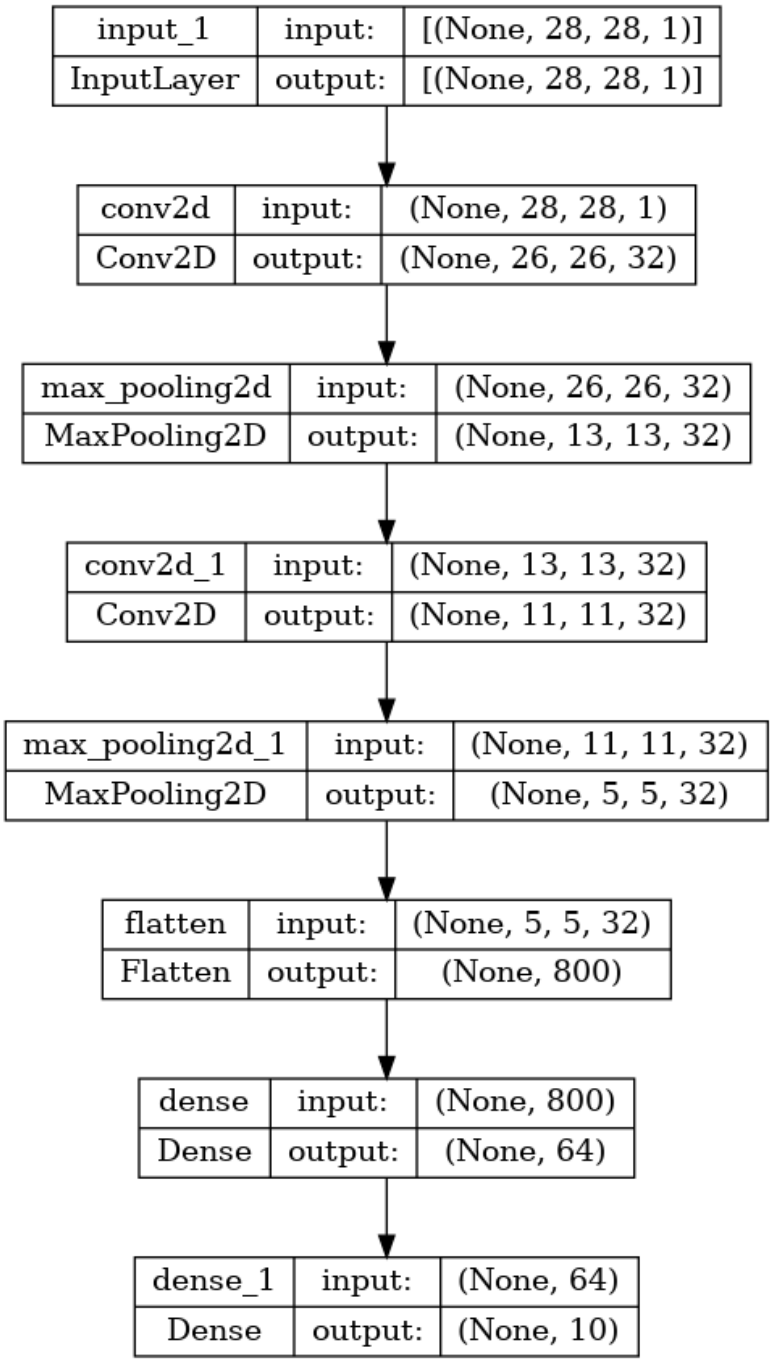

Figura 2: Schema della rete neurale utilizzata per il dataset mnist

modello apposta, ma non sarebbe stata una scelta realistica. Il dataset sound-mnist invece è stato scartato perchè dovendo convertire le tracce sonore in immagini spettrografiche sarebbe stato analogo ai due dataset di immagini già considerati.

Ciascuno dei due dataset è stato utilizzato per addestrare il modello, prima senza regolarizzazione e poi con ciascuna delle tre regolarizzazioni, utilizzando un lambda di 0.01 e 0.001.

3.3 Valutazione vulnerabilità

Per valutare la privacy del modello è stata utilizzata la libreria tensorflow-privacy che fornisce un membership inference attack standard e può essere chiamata ad ogni epoch dell'addestramento. Il report in uscita della libreria fornisce l'AUC del modello attaccante e il vantaggio che possiede l'attaccante per ogni epoch di addestramento della rete. È possible utilizzare due attacchi, uno di tipo threshold e l'altro basato su regressione logistica, inoltre fornisce direttamente il rapporto tra accuratezza sul validation set e privacy, in modo da poter da avere un indicazione per cercare un compromesso tra privacy e performance del modello.

La libreria mette a disposizione sei tipologie di attacco divise in due gruppi:

Trained attacks viene addestrato un modello ombra che viene utilizzato per generare il dataset del modello attaccante.

Thresholded attack data una certa quantità di threshold, si calcola quante istanze hanno una probabilità di essere membri maggiore della threshold.

Per questo esperimento è stato considerato un attacco per tipo, il primo basato su regressione logistica e il secondo un attacco threshold standard.

4 Risultati

4.1 Dataset fashion-mnist

I grafici mostrano che per quanto riguarda il dataset fashion-mnist tutte le forme di regolarizzazione si sono dimostrate efficaci nel diminuire la vulnerabilità del modello, in entrambe le tipologie di attacco e con entrambi i valori di lambda considerati, non solo come valore assoluto ma anche come andamento della curva con l'aumentare delle epoch nell'addestramento del modello.

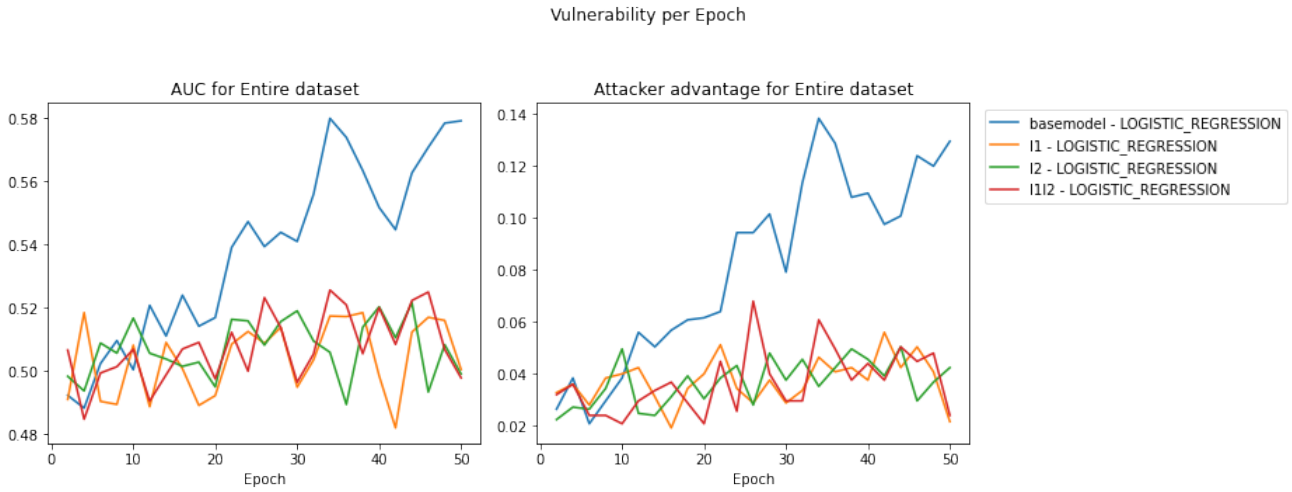

Figura 3: Attacco logreg, lambda=0.01

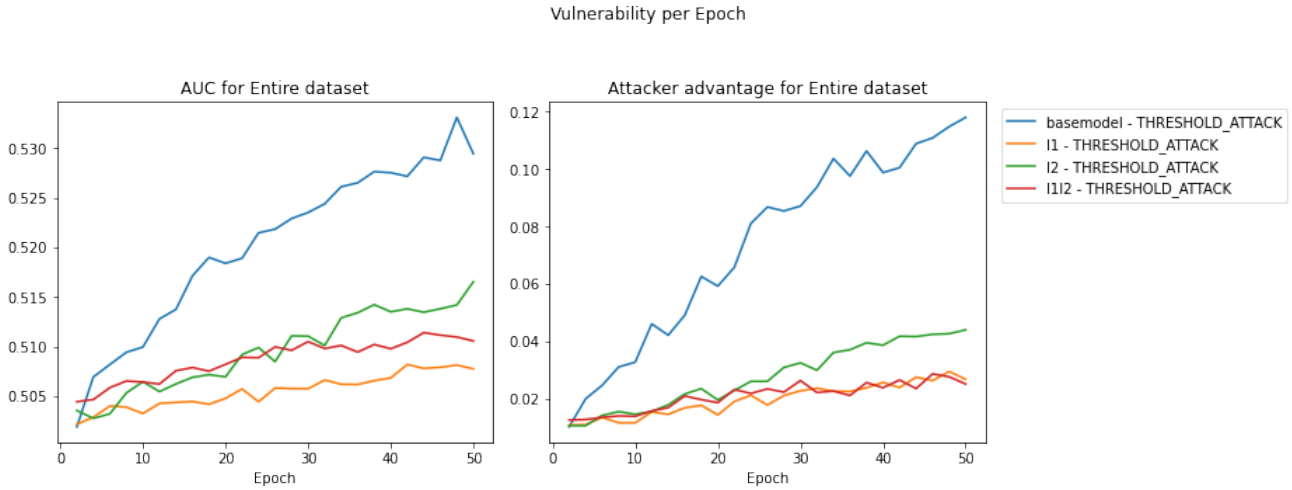

Figura 4: Attacco threshold, lambda=0.01

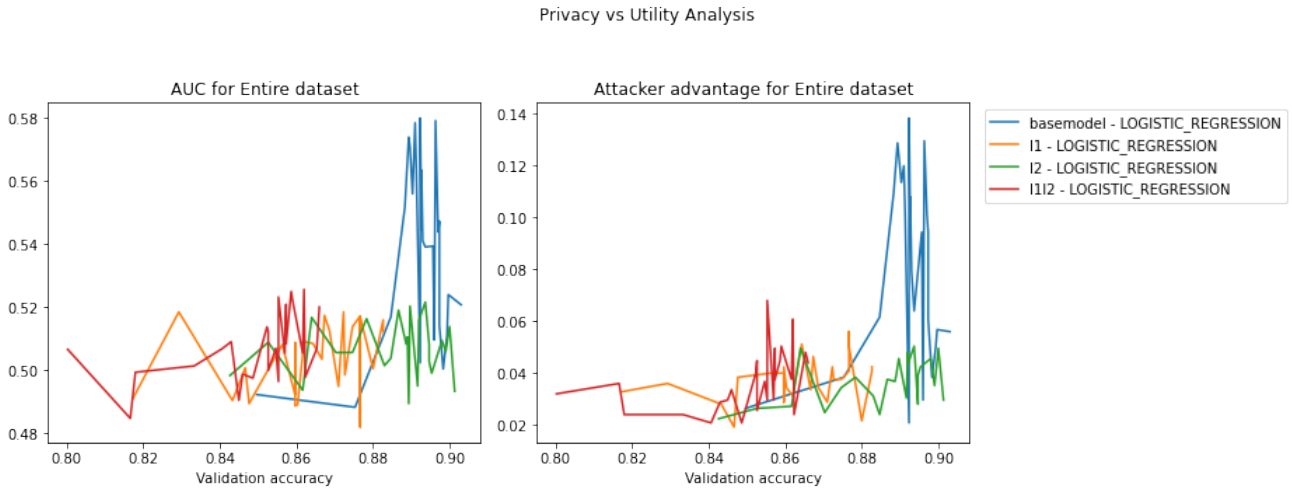

Figura 5: Rapporto accuratezza attacco logreg, lambda=0.01

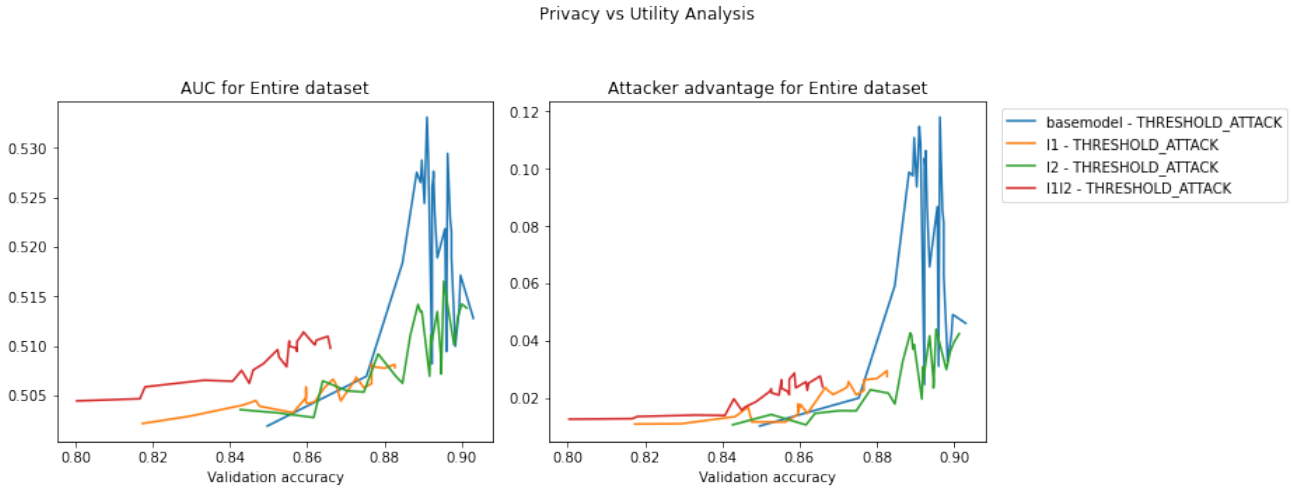

Figura 6: Rapporto accuratezza attacco threshold, lambda=0.01

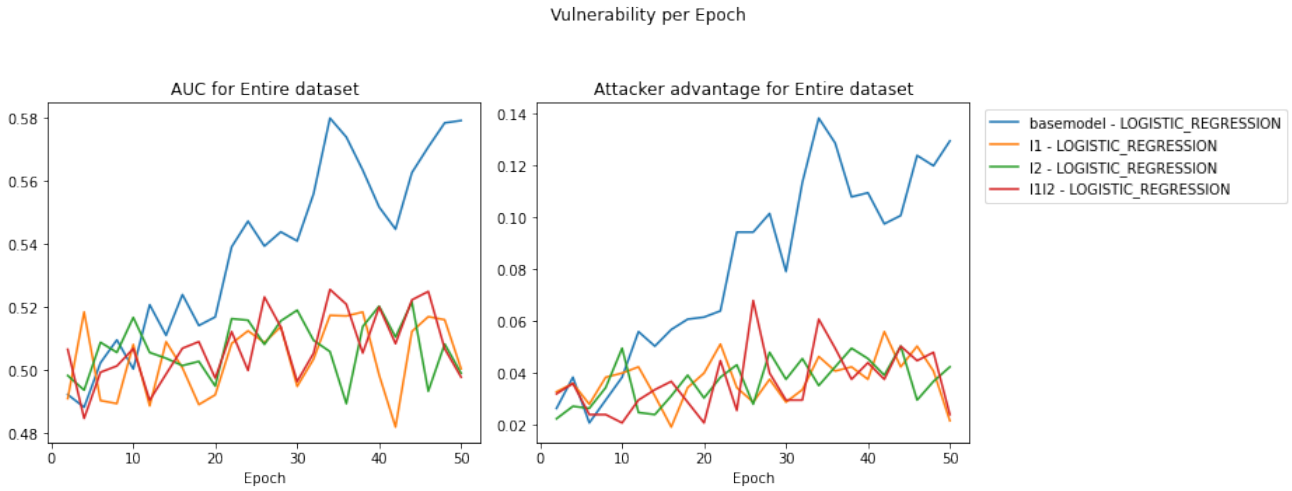

Figura 7: Attacco logreg, lambda=0.001

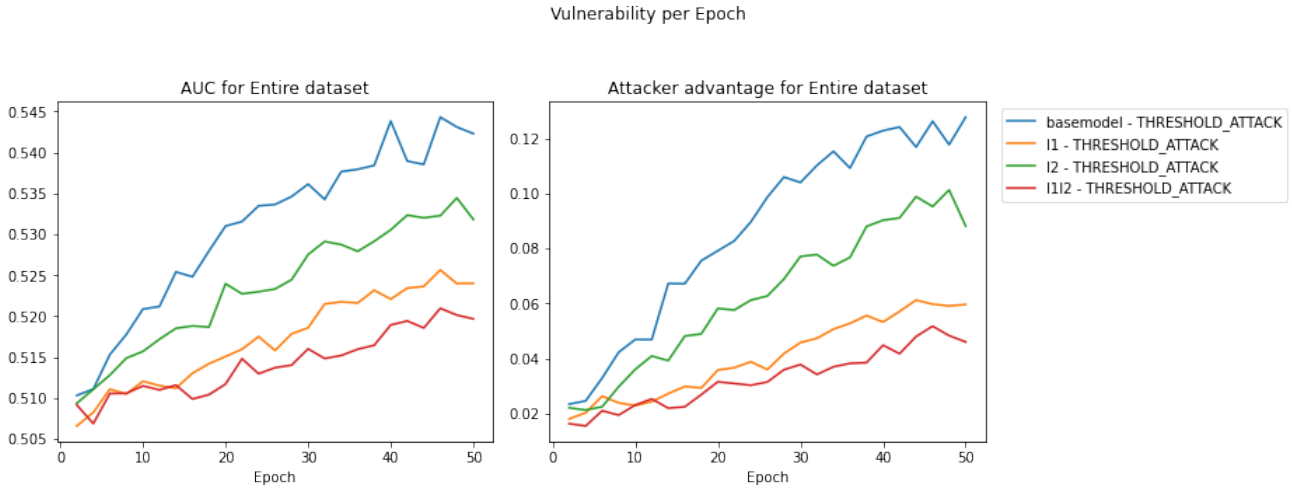

Figura 8: Attacco threshold, lambda=0.001

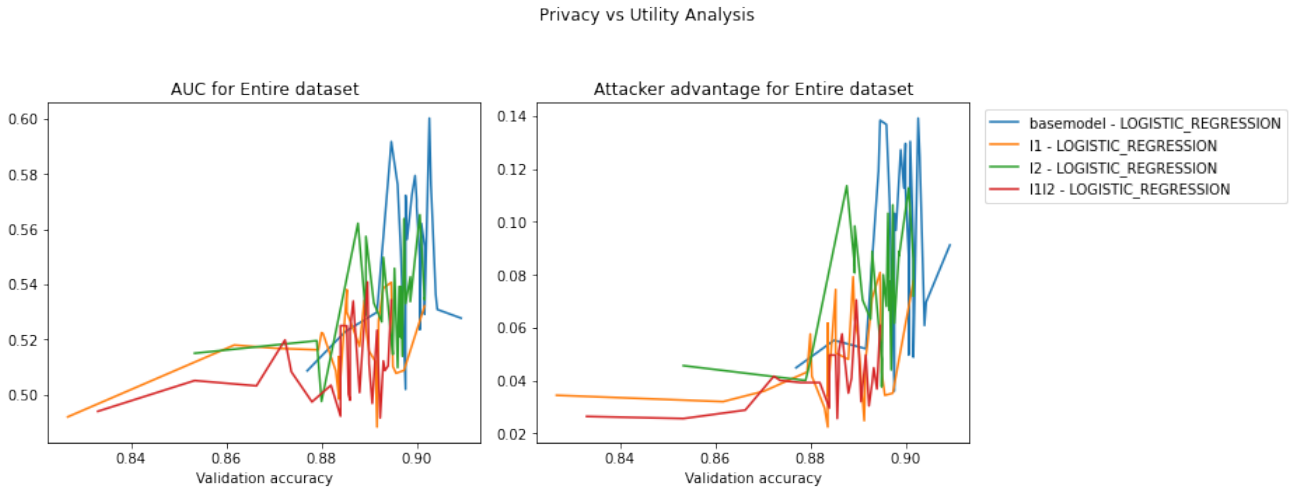

Figura 9: Rapporto accuratezza attacco logreg, lambda=0.001

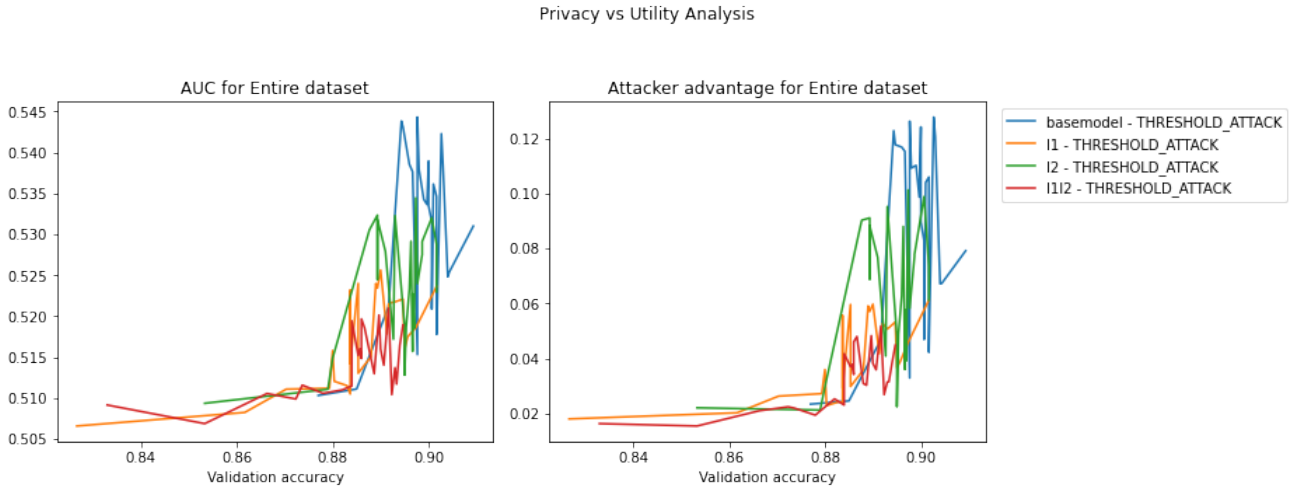

Figura 10: Rapporto accuratezza attacco threshold, lambda=0.001

Nella tabella 1 viene mostrata l'accuratezza di ciascun modello, si vede che la differenza di accuratezza tra validation set e test set è sempre minore quando applicata una forma di regolarizzazione, segno che l'overfitting è diminuito.

Regolarizzazione	Validation accuracy	Test accuracy	Delta accuracy
Nessuna	0.9949	0.8922	0.1027
L1 (0.01)	0.8945	0.8762	0.0183
L2 (0.01)	0.9273	0.8866	0.0407
L1L2 (0.01)	0.8798	0.8582	0.0216
L1 (0.001)	0.9427	0.8862	0.0565
L2 (0.001)	0.9781	0.8926	0.0855
L1L2 (0.001)	0.9309	0.8774	0.0535

Tabella 1: Accuratezza per il dataset fashion-mnist

4.2 Dataset cifar10

I grafici mostrano che anche per quanto riguarda il dataset cifar10 tutte le forme di regolarizzazione si sono dimostrate efficaci nel diminuire la vulnerabilità del modello, in entrambe le tipologie di attacco e con entrambi i valori di lambda considerati, non solo come valore assoluto ma anche come andamento della curva con l'aumentare delle epoch nell'addestramento del modello.

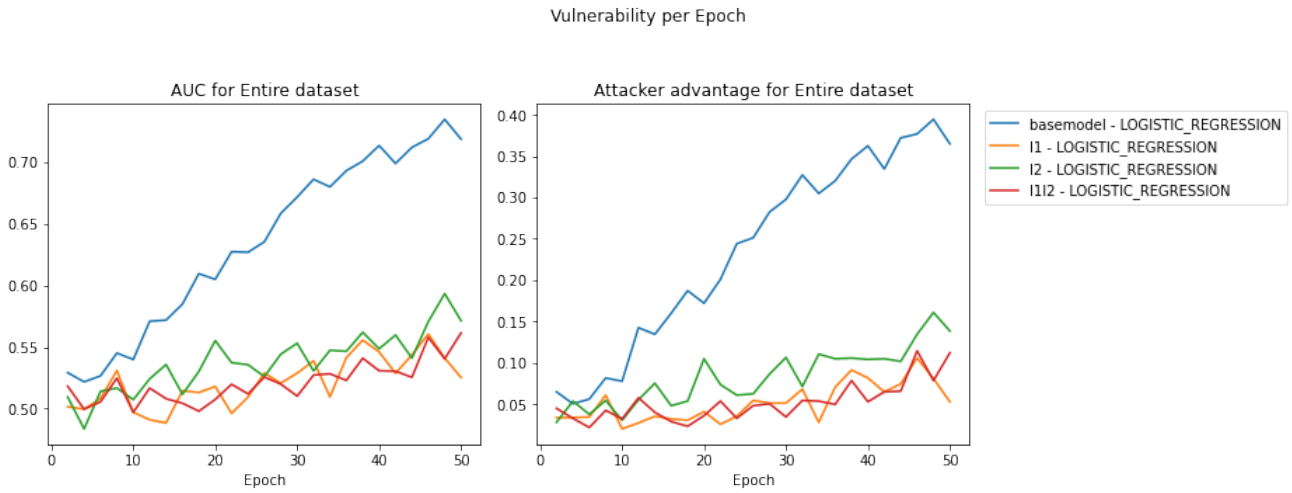

Figura 11: Attacco logreg, lambda=0.01

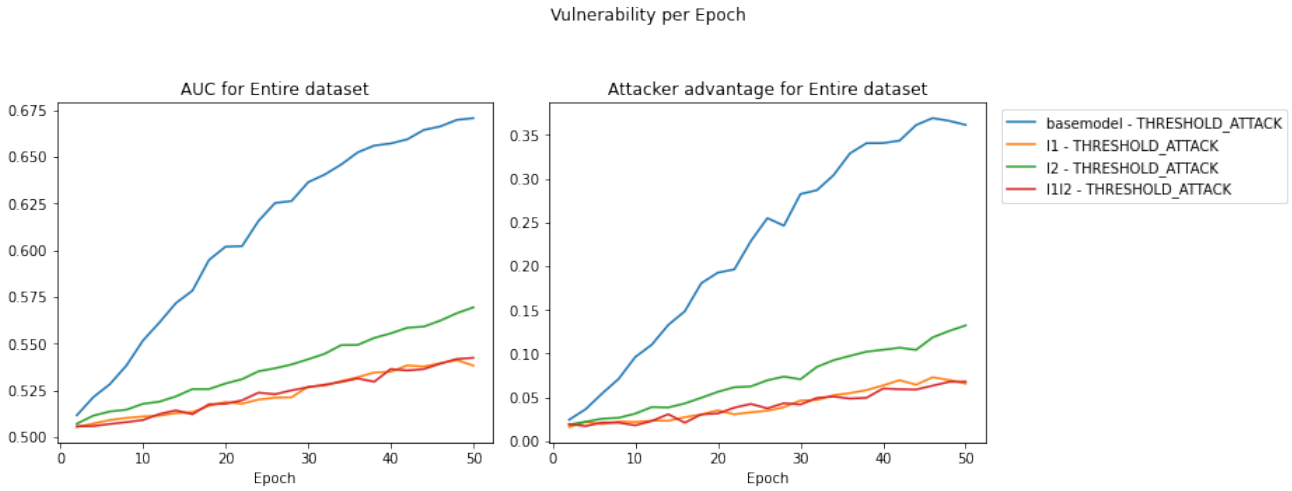

Figura 12: Attacco threshold, lambda=0.01

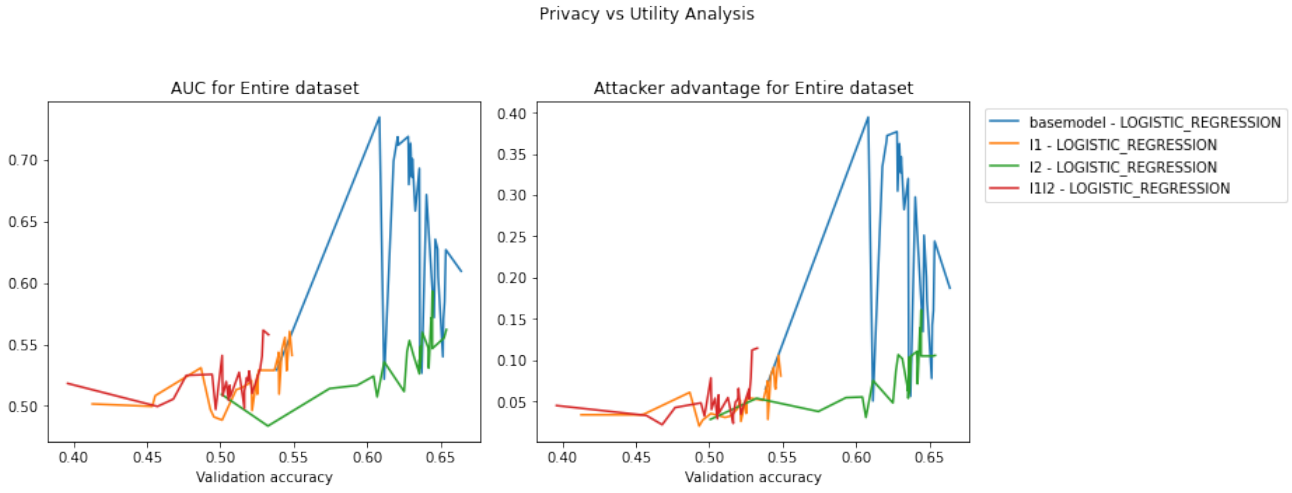

Figura 13: Rapporto accuratezza attacco logreg, lambda=0.01

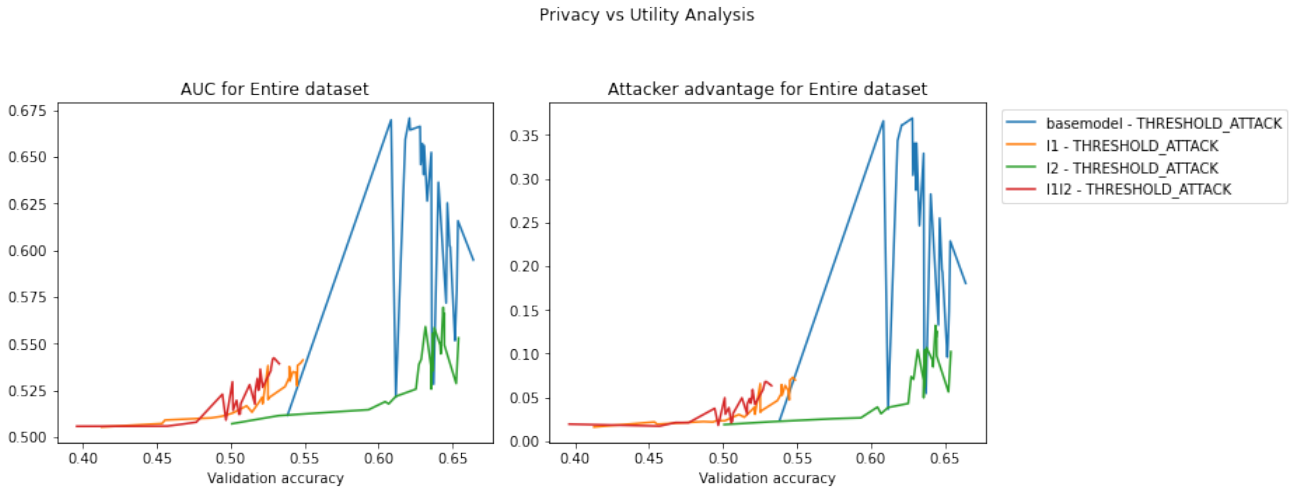

Figura 14: Rapporto accuratezza attacco threshold, lambda=0.01

Figura 15: Attacco logreg, lambda=0.001

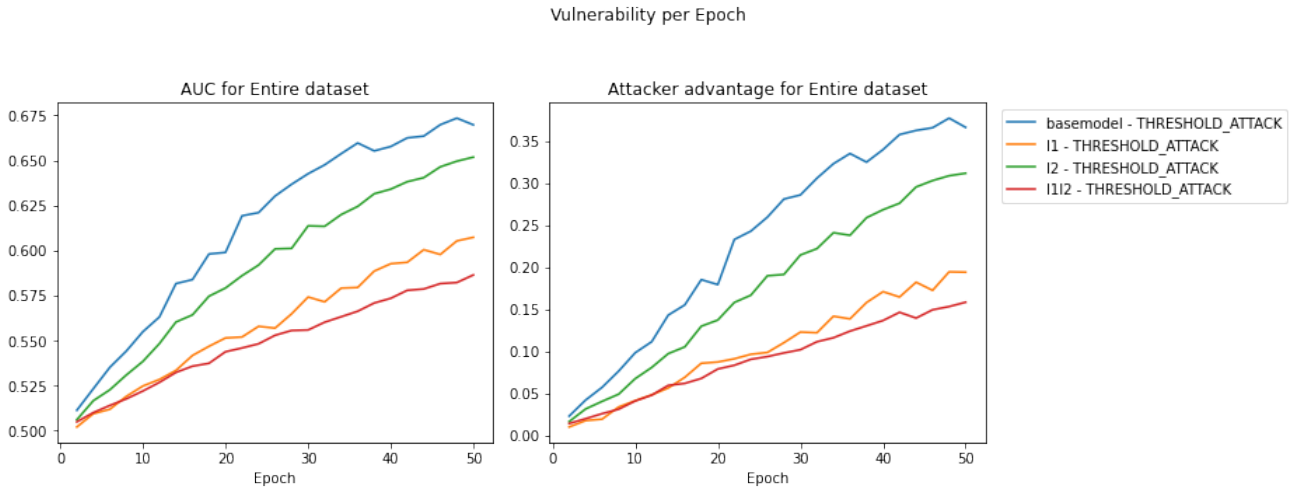

Figura 16: Attacco threshold, lambda=0.001

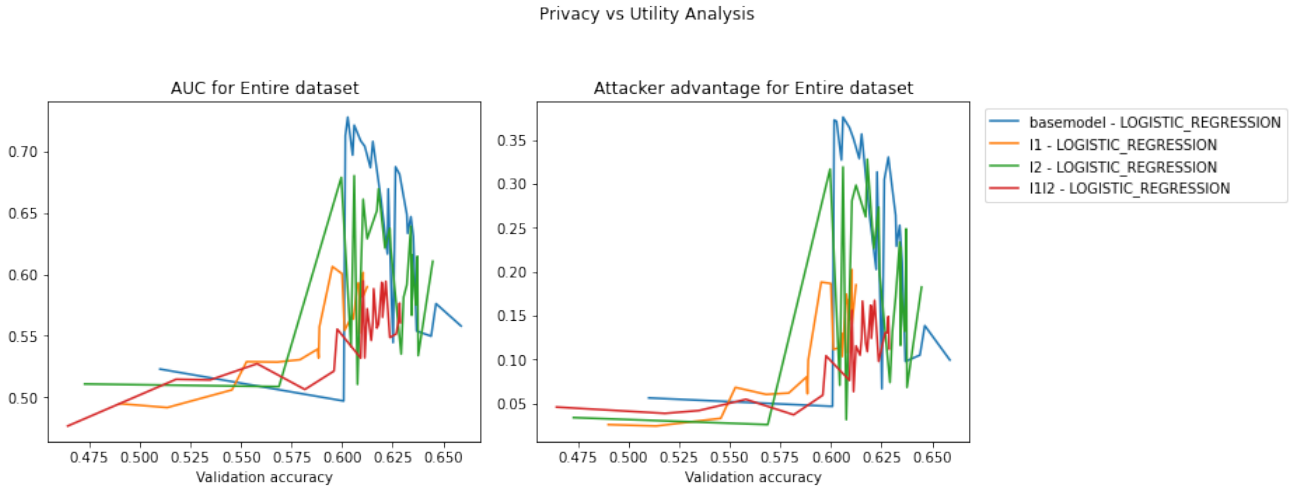

Figura 17: Rapporto accuratezza attacco logreg, lambda=0.001

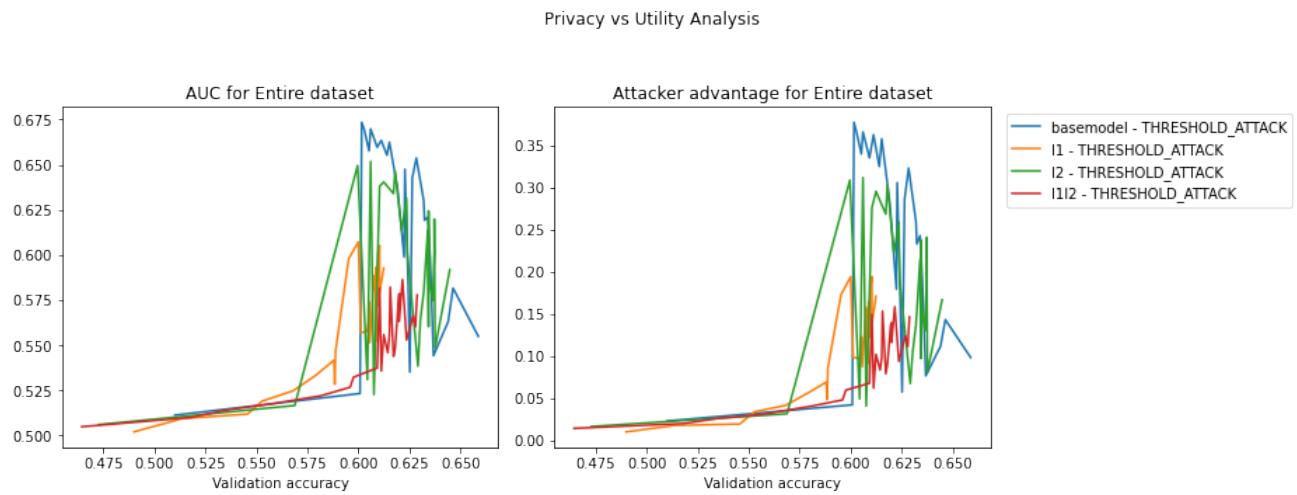

Figura 18: Rapporto accuratezza attacco threshold, lambda=0.001

Nella tabella 2 viene mostrata l'accuratezza di ciascun modello, si vede che la differenza di accuratezza tra validation set e test set è sempre minore quando applicata una forma di regolarizzazione, segno che l'overfitting è diminuito.

Regolarizzazione	Validation accuracy	Test accuracy	Delta accuracy
Nessuna	0.9682	0.6242	0.344
L1 (0.01)	0.6215	0.538	0.0835
L2 (0.01)	0.7525	0.6324	0.1201
L1L2 (0.01)	0.5817	0.5186	0.0631
L1 (0.001)	0.7889	0.5996	0.1893
L2 (0.001)	0.9107	0.614	0.2967
L1L2 (0.001)	0.7567	0.614	0.1427

Tabella 2: Accuratezza per il dataset fashion-mnist

5 Conclusioni e lavori futuri

I risultati empirici mostrano come l'efficacia delle tre regolarizzazioni considerate nell'aumentare il livello di privacy del modello è tangibile e tra di loro paragonabile, così com'è paragonabile il grado di accuratezza raggiunto dai modelli modificati. Si può anche affermare che un lambda maggiore porta a migliori risultati in termini di privacy.

Oltre a confermare che anche la regolarizzazione L1 e l'Elastic Net sono efficaci, questi risultati forniscono un'ulteriore prova nell'individuare l'overfitting come causa di vulnerabilità del modello.

Appurato che tutte le tecniche considerate abbiano un tangibile livello di efficacia, un tecnico che vuole aumentare la privacy del proprio modello può essere più sicuro nel considerarne una piuttosto che l'altra, e far ricadere la propria scelta su altri fattori, quali le performance del modello o il tipo particolare di dati che rendono una regolarizzazione migliore di un'altra.

Si può pensare di ampliare la ricerca su questo campo andando a considerare dataset diversi (ricordando però che devono avere determinate caratteristiche per essere vulnerabili in primo luogo, da cui la scelta del cifar e del fashion-mnist) e lambda differenti per valutare anche se esistono casi in cui il rapporto tra privacy e lambda della regolarizzazione si evolve diversamente, nonchè provare con attacchi differenti da quelli presenti nella libreria tf-privacy.

Riferimenti bibliografici

[1] Shokri et al, *Membership Inference Attacks against Machine Learning Models*

[2] Song and Mittal, *Systematic Evaluation of Privacy Risks of Machine Learning Models*

[3] Truex et al, *Towards Demystifying Membership Inference Attacks*

[4] Yeom et al, *Privacy Risk in Machine Learning: analyzing the connection to overfitting*

A Usage Report

L'elaborato è stato prodotto usando i seguenti software e librerie:
Google Colab https://colab.research.google.com/
Python https://www.python.org/
Numpy https://numpy.org/
Tensorflow https://www.tensorflow.org/
Keras https://keras.io/
Tf-privacy https://github.com/tensorflow/privacy

CopyRight