Paper implementation
Membership Inference Attacks Against Machine Learning Models

Adrien BENAMIRA

May 12, 2019

1 Introduction

This report presents the results of my implementation of [2] and explains the experimental figures. I was asked to implement the most basic attack which assumes the adversary has the data which comes from the same distribution as the target model's the training dataset. I choose to evaluate on MNIST, CIFAR10 and CIFAR100. I used the framework pytorch for the target and the shadow models and scikit-learn for the attack model. The code is available on Github: XXX

2 General configuration

2.1 Architecture

We use Pytorch and we train a standard convolutional neural network (CNN) with two convolution and max pooling layers plus a fully connected layer of size 128 and a SoftMax layer. We use tanh as the activation function as in the publication. For the attacker model, we use the ligth gradient boosting [1]. We set the number of estimator to 10000 and a reg ambada to 10.

2.2 General parameters

We set the learning rate to 0.001, the learning rate decay to $1e07$ervery epoch, and the number epochs of training to 100. The size of the target train dataset is 2500 and of the test dataset is 1000. There are 25 shadow models. The batch size is 64, the momentum is 0.5

3 Results

Table 3 are the general results for the attack model.

3.1 CIFAR 10

Figure 1 are the results for the target model. Figure 2 are the results for the attack model.

Dataset	**Accuracy**	**Precision**	**Recall**
CIFAR10	76.23 %	71.78%	64.13 %
CIFAR100		0	0
MNIST	94.40 %	93.69 %	92.45 %

Table 1: General results on Dataset of the attack model

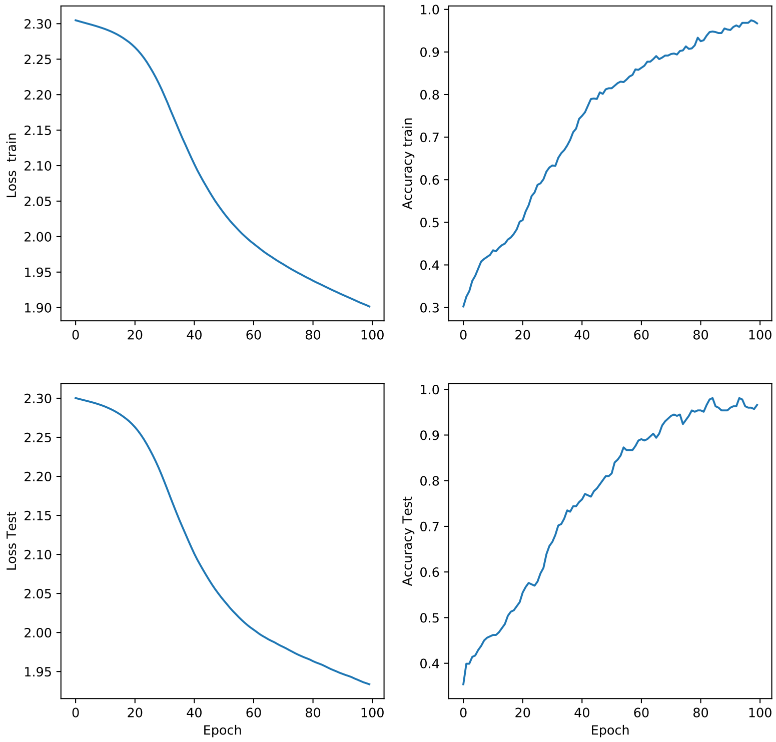

Figure 1: Plot of the learning curves (loss and accuracy) during training for the target model on CIFAR10

3.2 MNIST

Figure 3 are the results for the target model. Figure 4 are the results for the attack model.

3.3 CIFAR 100

XXX

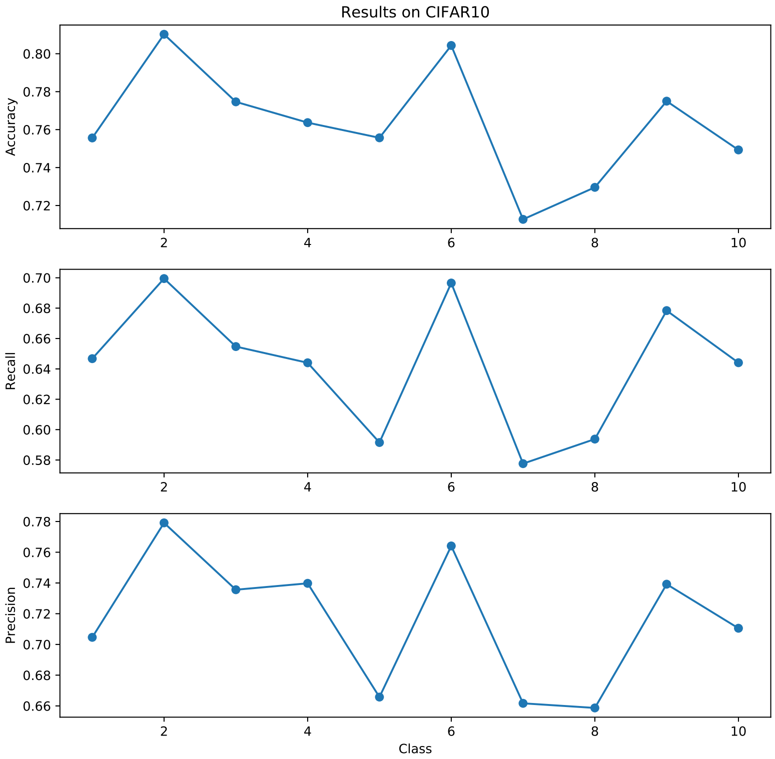

Figure 2: Plot of the accuracy, the recall and the precision value with classes on CIFAR10

References

[1] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems*, pages 3146–3154, 2017.

[2] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 3–18. IEEE, 2017.

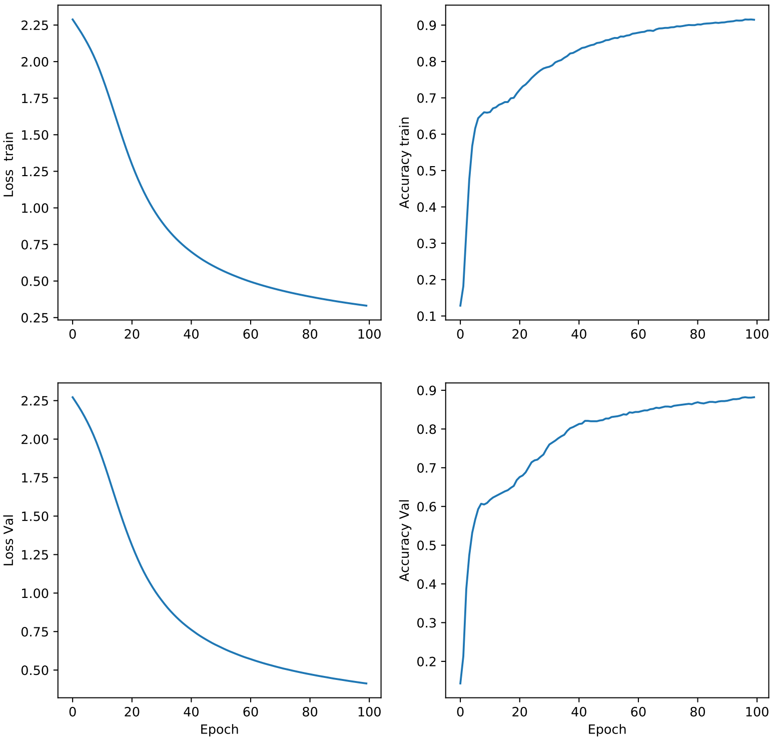

Figure 3: Plot of the learning curves (loss and accuracy) during training for the target model on MNIST

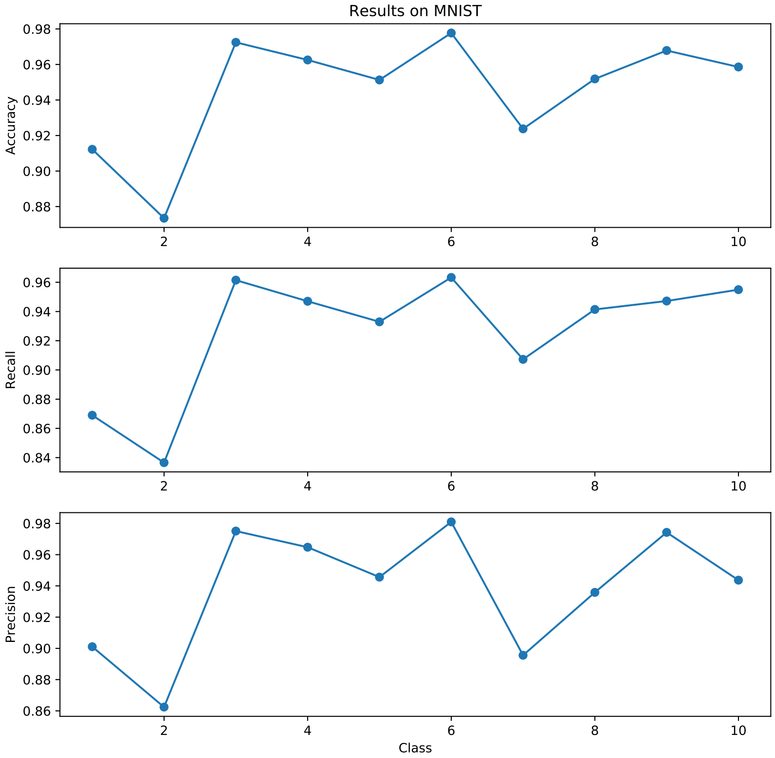

Figure 4: Plot of the accuracy, the recall and the precision value with classes on MNIST