DEPENDABLE AND SECURE AI-ML (AI60006)

ASSIGNMENT 2

Database Reconstruction Attack

You are required to perform a database reconstruction attack on a given dataset and apply differential private training as a mitigation strategy Follow the steps below and provide a report detailing your findings.

Steps

1. **Load Data**
   - Load the dataset of interest.
   - Split the dataset into training and test subsets.
2. **Train Model**
   - Choose a machine learning model (e.g., from scikit-learn).
   - Train the model on the training data.
3. **Launch Attack**
   - Select a row from the training dataset to remove.
   - Use the trained model to perform the database reconstruction attack.
4. **Evaluate Attack**
   - Calculate metrics such as RMSE or accuracy of the inferred values compared to the true values.
5. **Mitigate Attack with Differential Privacy**
   - Train the model with differential privacy guarantees.
   - Repeat the attack and evaluate its effectiveness.

Perform the following Tasks

Task 1: Logistic Regression on Iris Dataset:

a. Using scikit-learn, train a logistic regression model on the Iris dataset.
b. Remove all samples one at a time from the dataset and attempt to reconstruct the removed samples with correct labels.
c. Report the number of samples successfully reconstructed out of the total size of the dataset.

Task 2: Logistic Regression on Two-Class Classification Dataset:

a. Train a logistic regression model on a two-class classification dataset of your choice.
b. Attempt to reconstruct the data after removing samples.
c. If reconstruction fails, provide reasons for the failure.

Task 3: Logistic regression and Gaussian Naive Bayes on Four-Class Classification Dataset:

a. Generate a synthetic four-class classification dataset using `make_classification` from `sklearn.datasets`.(Check references below)
b. Train both on the generated dataset.make sure training accuracy is at least 60 percent
c. Perform a database reconstruction attack by removing sample and attempting to reconstruct them.
d. For the above case,Iterate over 100 parallel databases(two databases are parallel if they differ in a single row) and report the number of samples successfully reconstructed using both logistic regression and Gaussian naive Bayes classifiers.

Task 4: Apply Differential Privacy as Mitigation wherever applicable:

a. Wherever the database reconstruction attack succeeds in the previous three cases, apply differential privacy using the IBM differential privacy library.https://github.com/IBM/differential-privacy-library
b. Evaluate the effectiveness of differential privacy in mitigating the reconstruction attack for different values of epsilon and also plot the graph of accuracy vs epsilon similar to
https://github.com/IBM/differential-privacy-library/blob/main/notebooks/logistic_regression.ipynb

Analysis and Discussion

Analyze the results obtained from each task, including the success rate of database reconstruction and mitigation effectiveness or insights gained. Write your findings properly using proper headers in markdown in colab or kaggle or jupyter notebook

Model Inversion Attack

Intro

The goal of the model inversion attack is to reconstruct sensitive information about the input data, such as images of faces, from the output of a trained machine learning model. By exploiting the model's predictions, an adversary attempts to infer details about the original data that was used to train the model.

Tasks to be performed

In this attack, we implemented the model inversion technique on the AT and T, which was discussed in class and provided in the course materials. Now you need to implement the below exactly similar to (https://github.com/Trusted-AI/adversarial-robustness-toolbox/blob/main/notebooks/model_inversion_attacks_mnist.ipynb):

- For the AT and T dataset (https://git-disl.github.io/GTDLBench/datasets/att_face_dataset/), we performed the model inversion attack in the class with this notebook (https://colab.research.google.com/drive/1d1evCuWzHG7wb9Tbi6r535IHtVXcOuhB?usp=sharing) but you have to implement using MI face algorithm by ART library similar to (https://github.com/Trusted-AI/adversarial-robustness-toolbox/blob/main/notebooks/model_inversion_attacks_mnist.ipynb):
- For the CIFAR-10 dataset, utilize the MI Face algorithm for reconstructing samples similar to (https://github.com/Trusted-AI/adversarial-robustness-toolbox/blob/main/notebooks/model_inversion_attacks_mnist.ipynb):
- The algorithm implementation can be found (https://github.com/Trusted-AI/adversarial-robustness-toolbox/blob/main/art/attacks/inference/model_inversion/mi_face.py).

Results

Describe the results and observations of the model inversion attack on the AT and T either w.r.t to a custom model defined by you or as discussed in class notebook (https://colab.research.google.com/drive/1d1evCuWzHG7wb9Tbi6r535IHtVXcOuhB?usp=sharing)and for CIFAR-10 dataset, you can use the tensor flow model as defined here for CIFAR-10 dataset (https://www.kaggle.com/code/viratkothari/image-classification-of-cifar-10-using-tensorflow)

References

1. **Apply Differential Privacy:** Follow the instructions provided in the notebook to apply the mitigation strategy (https://github.com/IBM/differential-privacy-library/blob/main/notebooks/logistic_regression.ipynb) and (https://github.com/Trusted-AI/adversarial-robustness-toolbox/blob/main/notebooks/attack_database_reconstruction.ipynb), this includes both attack and mitigation

2. Use this link (https://scikit-learn.org/stable/modules/generated/sklearn.datasets.make_classification.html) to generate synthetic datasets for 4 class classification problem mentioned in above tasks

3. You can find the database reconstruction attack in the (https://colab.research.google.com/drive/1wwWoPv4WNGdVxc_klMZakX9XaUg8Vbfh?usp=sharing) and you can also find the full attack and mitigation of a naive Bayes classifier here (https://github.com/Trusted-AI/adversarial-robustness-toolbox/blob/main/notebooks/attack_database_reconstruction.ipynb)

Submission Guidelines

For the submission, please ensure the following:

1. Prepare three Jupyter Notebook files(ipynb), each addressing a specific task:
   - Database reconstruction
   - Model inversion for AT and T dataset
   - Model inversion for CIFAR-10 dataset

2. Name the ZIP file according to the following format: **yourname_rollnumber.zip**

Note: Ensure that all files are correctly named and organized within the ZIP file before submission.