Easy guide for people who are not deeply familiar with MIA

Task: Membership inference attack
Implemented by: Smita Bhattacharya
Date: December, 16, 2024

1 Overview of membership inference attack

Protecting sensitive data is essential in today's data-driven environment because machine learning models are being trained with such data, resulting in these models being prone to memorize the details of such sensitive data. The attacker can unethically attack such models to gain that information, leading to privacy concerns. For instance, a membership inference attack attempts to determine if a specific data sample was in the model's training set. This document will provide the implementation details for such a membership inference attack.

2 Components of the Approach

Google Colab Notebook is used for implementation. Tesla T4 of Google Colab environment is used as GPU. The notebook contains step-by-step instructions for the implementation and further environmental details. The following sections contain details regarding the data preparation, models, results and a brief remark on how to enhance this implementation.

- **Data preparation :** For performing the MIA, `imdb` dataset is used, consisting of the movie reviews in English and the associated labels to denote the respective review as positive or negative. The training and the test dataset of `imdb` are first tokenized using `get_tokenizer` and then converted to `tensor`. Further, the data is passed through the `collate_fn` function to pad the sequences to a fixed length. These fixed length sequences are then loaded in `DataLoader` to use them in mini-batches for model training.

- **Target and shadow model:** The target model is whose behaviour the attacker is trying to infer. This model is typically trained on a dataset, and the attacker aims to extract sensitive information about it, such as whether a particular data sample was used in training. A shadow model replicates the target model that the attacker trains and is trained on a similar dataset as the target model. Its purpose is to mimic the behaviour of the target model.
  Here, for this sentiment classification task, a multilayer perceptron model is used as a target and a shadow model. These are trained on the same dataset. It consists of one embedding layer and two fully connected feed-forward layers. As optimizer `Adam` is used, as loss function `BCEWithLogitsLoss` is used. Necessary hyperparameters are initialized and the models are trained for 5 epochs, and the training loss is decreased with each epoch.

- **Preparation of the data for attack model:** In-sample data is a member of the training dataset and is used to train the model, and the model can be attacked to get information about the data. Out-sample data is non-training data; the target model has no information about it. The information about the membership of the data in a training set is required to train the attack model. For the data preparation of the attack model, the training data of the `imdb` dataset is prepared as in-sample data and test data is prepared as the out-sample data. The logit value of the prediction is inferred from the shadow models. The top 90% confident logit samples are filtered to train and evaluate the attack model. The confidence is calculated by taking the absolute value of the logit and is sorted to get the top samples. The training samples are annotated as 1(as training members), and the test samples are annotated as 0(as non-training members). While predicting, the shadow model will be more confident if they have seen the data before and will reflect in the logits. Finally, these logits are concatenated to get the whole dataset for the attack model.

- **Attack model and result:** The attack dataset is again split into training and testing datasets after shuffling. A `XGBoost` model is trained as an attack model.
  To evaluate the attack model's performance, evaluation metric such as accuracy and ROC-AUC are calculated. The accuracy of 0.561 and the ROC-AUC of 0.577 indicate that the membership inference attack model performs only slightly better than random guessing, suggesting limited effectiveness in reliably distinguishing between in-sample and out-sample data. This result highlights that the target model exhibits some resistance to membership inference attacks but may still have minor vulnerabilities.

3 Remark:

Here, a simple multilayer perceptron model is chosen as the target and the shadow model. In future, more advanced models can be opted for, and different methods of differential privacy can be further included to make them more robust against such attacks and vulnerability. Moreover, the attack can be evaluated with noisy or incomplete data to simulate a real-world scenario.