Dataset Inference: Ownership Resolution in Machine Learning

Pratyush Maini†‡, Mohammad Yaghini‡, Nicolas Papernot‡
† Indian Institute of Technology Delhi | ‡ University of Toronto and Vector Institute

Key Points: 1. Dataset Inference (DI) resolves model ownership without the need for retraining; and does not have a trade-off with task accuracy.
2. We prove that the success of Membership Inference decreases as overfitting reduces, whereas DI is independent of the same.
3. We introduce a new method for black-box ownership resolution that requires less than 50 private training points from the victim's dataset.

Motivation

- Developing high performing ML models is becoming expensive.
- Adversaries may steal ML models by copying their predictions with little cost.

Threat Models

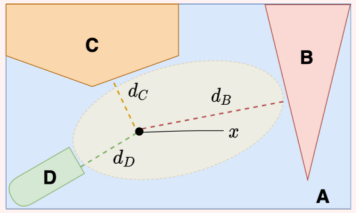
An adversary may have varying degrees of access to your *knowledge:*
Data Access, A_D: Insider access or release of dataset by victim for academic purposes.
Query Access, A_Q: Such as in MLaaS.
Model Access, A_M: Release of complete model by the victim for academic purposes.

Intuition

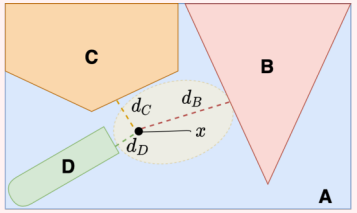
- All *knowledge* of an ML model can be linked back to the dataset it was trained on.
- There exists a distinction in the model response to any point in its training set as compared to an unseen set.

Train-Test Prediction Margin

The optimization step pushes boundaries of incorrect classes further away from points in the train set, but not for points in the test set.



If x was in Train set



If x was in Test set

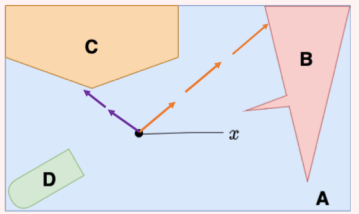
The expected prediction margin for inputs in train set is larger than that for unseen inputs.

Blind Walk

Black-box approximation of the prediction margin by creating an embedding,

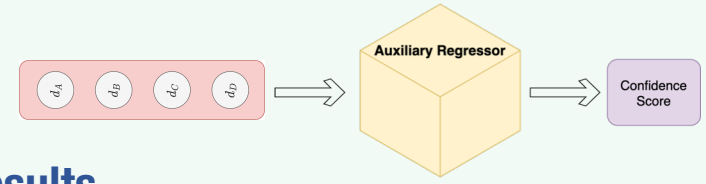
$$\text{emb}^i_{(x,y)}(f) = \min_{k \in \mathbb{N}} d(x, x + k\delta_i)$$

$$s.t.\ \ \delta_i \sim \mathcal{N}(0, \sigma^2 I);$$
$$f(x + k\delta_i) \neq y$$



Dataset Inference

1. Sample data points from the train & test set.
2. Generate embeddings indicating prediction margin.
3. Pass embeddings through a confidence regressor.
4. One sided t-Test --- H_0: $\mu_{test} > \mu_{train}$.



Results

Dataset Inference resolves ownership with fewer than 60 private samples revealed, at a FPR < 1% on CIFAR10, SVHN, CIFAR100 and ImageNet.