Machine Unlearning Summary

November 15, 2023

Summarizer: Xingjian Zhang

Contribution | Takeaway

All of the following arguments are based on linear model and convex optimization assumptions.

Certified Removal

- The author gives the definition of ε**-certified removal and** (ε, δ)**-certified removal**, which are (the first to be) inspired by **differential privacy**.
- The intuition of certified removal is to measure how hard **a model unlearned by some algorithm** can be distinguished from **a model retrained by a dataset removing some data point (gold standard reference)**.

Newton Updated Removal

- The author prove how well is the **newton updated removal method** for single data point, batch removal (multiple data points removal), and multiple removal (multiple iterations removal) by **evaluating the gradient residual**. The gradient residual should be close 0.
- The author provides data independent (loose and cheap) and data dependent bound (tight and expensive) for them.

Algorithm: Loss perturbation

- The author proposes an empirical method to **test whether the certified removal can be established** by carefully **introducing noise into the loss function**.

Strength

- The definition of certified removal is **well-motivated under a stochastic learning framework**.
- The author provides **strong theoretical supports** for most arguments.
- It is the first paper that **extend differential privacy to machine unlearning**.

Weakness

The weakness are provided by Weijie and Pingbang.

- The paper is based on the **strong assumption of linear model and convex optimization**.
  - Unique global optimal may be the key for such first-order method to work.
- There is **no demo on how the model change after the removal**. Potential ways may include:
  - Inspecting one specific data point to see prediction changes.
  - Testing the uncertainty of data points that are removed.
  - Testing resilience against a membership inference attack.
- There is **no demo on the effect of loss perturbation**.
- There is a **potential risk of data leakage in the upstream (non-linear) encoder** in a non-pre-trained scenario.
- The proposed method has **high compute complexity** due to inverse Hessian matrix.

Future directions

- Non linear models and non convex optimization.
- Dynamic system where data are constantly evolving.