Certified Data Removal from Machine Learning Models

Challenge

Pingbang Hu

University of Illinois Urbana-Champaign

November 15, 2023

Table of Content

Reactions

Contribution:

- Define *certified removal*, a novel notion for "the right to be forgotten".
- Design certified removal algorithm with *provable guarantee*.

Strength:

- Well-motivated definition.
  - Fills the gap between *retraining* and *differential privacy*.
- Approach the problem from multiple angles.
  - *Vanilla uniform* v.s. *data-dependent* bound.
  - Extend to *iterative* removal and *batch* removal.

Challenge

Analysis Weakness:

- *Convex*: Unique optimal is the key for such first-order approximation method to work. In some sense, authors didn't go anywhere beyond this current first-order approximation paradigm.
- *Linear*: Linear model, huh.
- Being *nit-picking*:
  - Similar analysis is already available [Gio+].
  - For a broader class of functions and setup, e.g., M-estimator, ERM.

Experiment Weakness:

- *Actual Comparison*: No demo on how in reality the model changes *before* and *after* the removal?
- *Algorithmic design*: No demo on the effect of *loss perturbation*, the heart of the algorithm.

Table of Content

References

[Gio+] Ryan Giordano et al. *A Swiss Army Infinitesimal Jackknife*. DOI: 10.48550/arXiv.1806.00550. arXiv: 1806.00550 [stat]. URL: http://arxiv.org/abs/1806.00550. preprint.

[Guo+] Chuan Guo et al. *Certified Data Removal from Machine Learning Models*. arXiv: 1911.03030 [cs, stat]. URL: http://arxiv.org/abs/1911.03030. preprint.