Soft Weighted Machine Unlearning

Paper review by A.A.M.

Xinbao Qiao, Ningning Ding, Yushi Cheng, Meng Zhang

https://arxiv.org/abs/2505.18783

Contents

Setting

Consider a prediction task (regression problem) with:

- Input space $\mathcal{X}$;
- Output space $\mathcal{Y}$;
- Training set $\mathcal{D}^n = \{z_i = (\boldsymbol{x_i}, y_i)\}_{\{i=1\}}^n \subset \mathcal{Z} = \mathcal{X} \times \mathcal{Y}$;
- Parameter $\theta \in \Theta := \mathbb{R}^d$;
- $f(\theta; x)$ estimator of $y|\boldsymbol{x}$;
- $\ell : \mathcal{Z} \times \Theta \to \mathbb{R}$ loss function (e.g., $\ell(z, \theta) \mapsto \|y - f(\theta; \boldsymbol{x})\|^2$).

We seek the empirical risk minimizer:

$$\hat{\theta} = \arg\min_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} \ell(z_i, \theta).$$

Influence Functions

Given $z_j \in \mathcal{D}^n$, the *influence function* relative to z_j is:

$$\mathcal{I}(z_j) = \frac{d}{d\varepsilon_j}\left[\frac{1}{n}\sum_{i=1}^{n}\ell\left(z_i, \hat{\theta}\right) + \varepsilon_j \ell\left(z_j, \hat{\theta}\right)\right]\Bigg|_{\varepsilon_j=-1}$$

Interpretation: It indicates how much the training error changes when we remove a training data z_j.

Remark

We can rewrite:

$$\mathcal{I}\left(z_j\right) = -H_{\hat{\theta}}^{-1} \nabla_\theta \ell\left(z_j, \hat{\theta}\right)$$

where H is the Hessian matrix of the empirical risk at $\hat{\theta}$.