If Influence Functions are the Answer, Then What is the Question?

Paper review

Alex Ali Maleknia

Contents

Setting

Consider a prediction task (regression problem) with:

- Input space $\mathcal{X}$;
- Output space $\mathcal{Y}$;
- Training set $\mathcal{D} = \{z_i\}_{\{i=1\}}^{n}$ where $z_i = (x_i, y_i)$ for all $i = 1, \ldots, n$, $X = (x_1, \ldots, x_n)$, $Y = (y_1, \ldots, y_n)$;
- Parameter $\theta \in \Theta := \mathbb{R}^d$;
- $f(\theta; x)$ estimator of $\mathcal{Y}|X$;
- $l : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ loss function (e.g., $l(y', y) \mapsto \|y' - y\|^2$).

We aim to minimize the training error:

$$L(\theta; \mathcal{D}) = \frac{1}{n} \sum_{i=1}^{n} l(f(\theta; x_i), y_i).$$

What happens if we change the importance of a training point $z = (x, y)$ of the dataset?

Call $\hat{\theta} = \arg\min_{\theta \in \Theta} L(\theta; \mathcal{D})$.
How different is it from

$$\hat{\theta}_{\varepsilon,-z} = \arg\min_{\theta \in \Theta} (L(\theta; \mathcal{D}) - \varepsilon l(f(\theta; x), y)) \quad ?$$

We can re-train the whole model on $\mathcal{D} \setminus \{z\}$ (Leave One Out method), or...

Influence functions

Definition

Given $\overline{z} \in \mathcal{D}$, the *influence function* relative to $\overline{z} = (\overline{x}, \overline{y})$ is:

$$\mathcal{I}(\overline{z}) = \frac{d}{d\varepsilon}\left[L(\theta; \mathcal{D}) - \varepsilon l(f(\theta; \overline{x}), \overline{y})\right]\Bigg|_{\varepsilon=\frac{1}{n}}$$

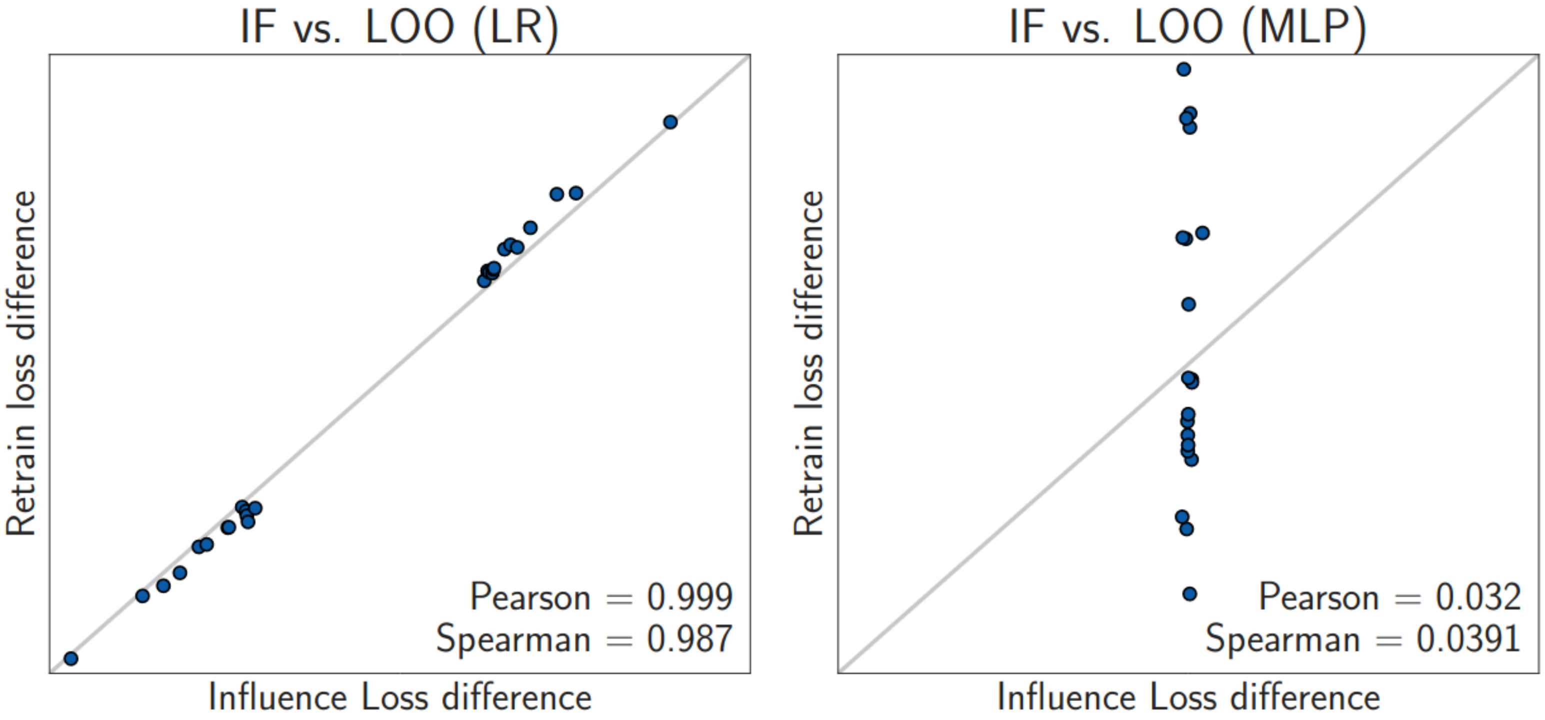
Classical interpretation: It indicates how much the training error changes when we remove a training data $\overline{z}$.

Problems

Assume L is strongly convex. Evaluating influence functions requires heavy computations:

$$\mathcal{I}(z) = \frac{1}{n} H_{\hat{\theta}}^{-1} \nabla l\Big(f\big(\hat{\theta}; x\big), y\Big),$$

where $H_{\hat{\theta}}$ is the Hessian of L, which can be difficult to compute.



The strong convexity assumption is essential!

Solution?

Solution 1. For iHVP, there are good approximations that only require $O(nd)$ flops instead of $O(n^3)$.

Solution 2. Change point of view: Influence functions are not approximators of LOO retraining, but instead of the proximal Bregman response function (PBRF).

We define the response function in the general setting as:

$$\hat{r}_z(\varepsilon) = \arg\min_{\theta\in\Theta}(L(\theta; \mathcal{D}) - \varepsilon l(f(\theta; x), y)).$$

Note that $\hat{r}_z(\varepsilon) = \hat{\theta}_{\varepsilon,-z}$ and $\hat{r}_z(0) = \hat{\theta}$. Since (by IFT) $\hat{r}$ is differentiable at 0, we can define the influence functions as first order approximant of $\hat{r}$. In fact, expanding with Taylor near 0 we get:

$$\hat{r}_{z,\mathrm{lin}}\left(\frac{1}{n}\right) = \hat{r}_z(0) + \left.\frac{d\hat{r}_z}{d\varepsilon}\right|_{\varepsilon=0}(\varepsilon - 0) = \hat{\theta} + \frac{1}{n}H_{\hat{\theta}}^{-1}\nabla l\left(f\left(\hat{\theta}; x\right), y\right).$$

We need H_θ to be positive definite in order to invert it, so θ must be a minimum point.

In order for the influence functions to be computable in the MLP case, we need to address the hessian inverse. This can be done by approximating $H_{\hat{\theta}}$ with the Gauss-Newton Hessian (GNH) and adding a damping term to ensure GNH is invertible:

$$\mathcal{J}^{\dagger}(z) = \frac{1}{n}\left(J_{y\hat{\theta}}^{\top} H_{\hat{\theta}} J_{y\hat{\theta}} + \lambda \boldsymbol{I}\right)^{-1} \nabla l\left(f\left(\hat{\theta}; x\right), y\right),$$

where $J_{y\hat{\theta}}$ is the Jacobian of $F(\theta) = (f(\theta; x_1), \ldots, f(\theta; x_n))$ in $\hat{\theta}$.
We can get the previous formula by linearizing near 0:

$$\begin{aligned} \hat{r}_{z,\mathrm{damp}}(\varepsilon) &= \arg\min_{\theta \in \Theta} L(\theta; \mathcal{D}) - \varepsilon l(f(\theta; x), y)) + \frac{\lambda}{2} \|\theta - \hat{\theta}\|^2, \\ \hat{r}_{z,\mathrm{damp,lin}}(1/n) &\approx \hat{\theta} + \mathcal{J}^{\dagger}. \end{aligned}$$

What is PBRF?

In practice, θ is not a minimum point for L.
However, we can consider another training error for which the early arrested parameter θ^s is optimal:

$$\mathcal{L}(\theta; \theta^s, \mathcal{D}) = \frac{1}{n} \sum_{i=1}^{n} D_{l^{(i)}}(f(\theta; x_i), f(\theta^s; x_i)),$$

where $D_{l^{(i)}}(y, y') = l(y, y_i) - l(y', y_i) - \nabla_1 l(y', y_i)^\top (y - y')$ is called Bregman difference.
We can then define the PBRF as:

$$r^b_{z,\mathrm{damp}}(\varepsilon) = \arg\min_{\theta \in \Theta} \mathcal{L}(\theta; \theta^s, \mathcal{D}) - \varepsilon l(f(\theta, x), y) + \frac{\lambda}{2} \|\theta - \theta^s\|^2.$$

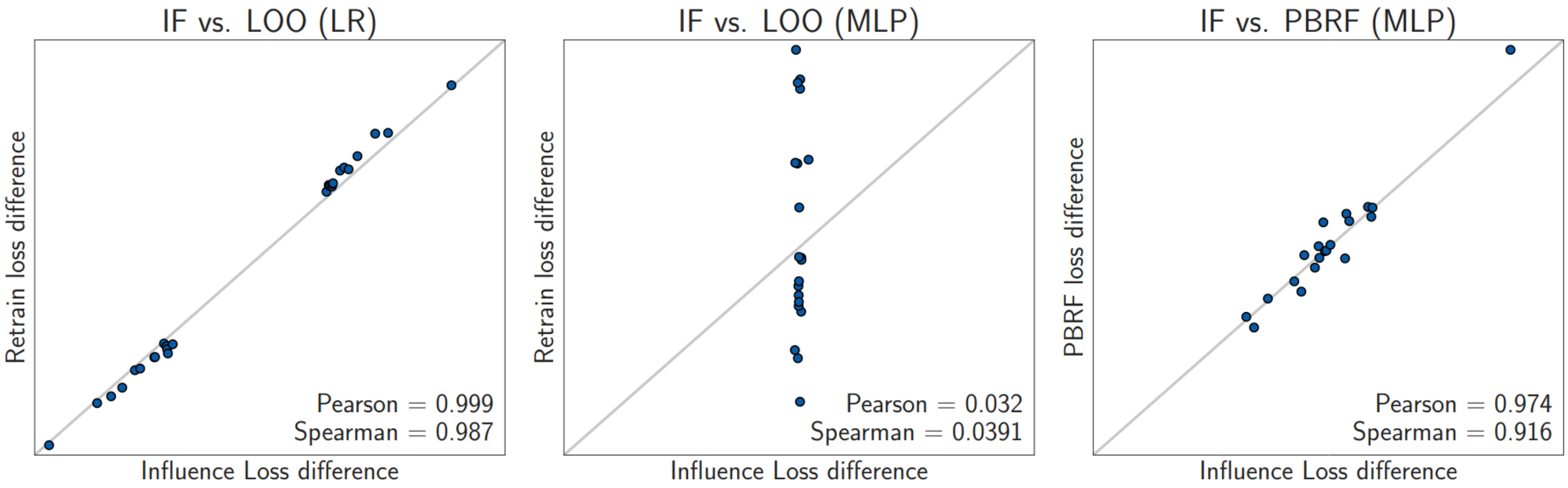
The optimal solution for the linearised PBRF is the same as the influence function estimation:

$$r^{b}_{z,\mathrm{damp},\mathrm{lin}}(1/n) = \theta^{s} + \mathcal{J}^{\dagger}(z).$$

Therefore, influence functions do NOT depict the retraining with LOO algorithm using L as training error.
Instead they estimate what happens after training from θ^{s} using as empirical risk:

$$\mathcal{L}(\theta;\theta^{s},\mathcal{D}) - \frac{1}{n} l(f(\theta,x),y) + \frac{\lambda}{2}\|\theta-\theta^{s}\|^{2}.$$

Example


IF vs. LOO (LR)
Retrain loss difference
Influence Loss difference
Pearson = 0.999
Spearman = 0.987
IF vs. LOO (MLP)
Retrain loss difference
Influence Loss difference
Pearson = 0.032
Spearman = 0.0391
IF vs. PBRF (MLP)
PBRF loss difference
Influence Loss difference
Pearson = 0.974
Spearman = 0.916