Iterative Approximate Cross-Validation

Yuetian Luo

University of Chicago

Joint work with Zhimei Ren and Rina Foygel Barber

Background

- Cross-validation (CV) is a popular tool for assessing and selecting predictive models.

Background

- Cross-validation (CV) is a popular tool for assessing and selecting predictive models.
- Leave-one-out CV or n-fold CV

  ★ Has high accuracy for out-of-sample error estimation
  [Arlot and Celisse, 2010].

Background

- Cross-validation (CV) is a popular tool for assessing and selecting predictive models.
- Leave-one-out CV or n-fold CV

  ★ Has high accuracy for out-of-sample error estimation [Arlot and Celisse, 2010].
- Computing leave-one-out CV can be expensive as the model needs to be fitted n times.

Background

- Cross-validation (CV) is a popular tool for assessing and selecting predictive models.
- Leave-one-out CV or *n*-fold CV

  ★ Has high accuracy for out-of-sample error estimation [Arlot and Celisse, 2010].
- Computing leave-one-out CV can be expensive as the model needs to be fitted *n* times.
- Can we find efficient approximations for leave-one-out CV?

Background

- Cross-validation (CV) is a popular tool for assessing and selecting predictive models.
- Leave-one-out CV or n-fold CV

  ★ Has high accuracy for out-of-sample error estimation [Arlot and Celisse, 2010].
- Computing leave-one-out CV can be expensive as the model needs to be fitted n times.
- Can we find efficient approximations for leave-one-out CV?

Much progress has been made to speed up Leave-one-out CV under the ERM framework [Beirami et al., 2017, Giordano et al., 2019, Wang et al., 2018, Wilson et al., 2020, Rad and Maleki, 2020, Stephenson and Broderick, 2020].

Prediction Error Estimation in ERM via CV

- Empirical Risk Minimization:

$$\widehat{\theta} = \arg\min_{\theta \in \mathbb{R}^p} F(\mathcal{Z}; \theta) := \sum_{j=1}^{n} \ell(Z_i; \theta) + \lambda \pi(\theta),$$

$\mathcal{Z} = \{Z_i\}_{i=1}^n$ is the data, $\ell(Z_i, \theta)$ is the loss on data Z_i with parameter θ.

Prediction Error Estimation in ERM via CV

- Empirical Risk Minimization:

$$\widehat{\theta} = \arg\min_{\theta \in \mathbb{R}^p} F(\mathcal{Z}; \theta) := \sum_{j=1}^{n} \ell(Z_i; \theta) + \lambda \pi(\theta),$$

$\mathcal{Z} = \{Z_i\}_{i=1}^n$ is the data, $\ell(Z_i, \theta)$ is the loss on data Z_i with parameter θ.

- Leave-one-out CV estimation for the prediction error of $\widehat{\theta}$:

$$\mathrm{CV}(\{\widehat{\theta}_{-i}\}_{i=1}^n) = \sum_{i=1}^{n} \ell(Z_i; \widehat{\theta}_{-i}),$$

where

$$\widehat{\theta}_{-i} = \arg\min_{\theta \in \mathbb{R}^p} F(\mathcal{Z}_{-i}; \theta) := \sum_{j=1, j \neq i}^{n} \ell(Z_j; \theta) + \lambda \pi(\theta)$$

Computing $\widehat{\theta}_{-i}$ for $i = 1, \ldots, n$ can be expensive.

Existing Approaches for Approximating $\widehat{\theta}_{-i}$

- $F(\cdot, \theta)$ is twice continuously differentiable in θ:

Existing Approaches for Approximating $\widehat{\theta}_{-i}$

- $F(\cdot, \theta)$ is twice continuously differentiable in θ:
  - One Newton-Step (NS) estimator [Beirami et al., 2017]:

$$\widetilde{\theta}_{-i}^{\mathrm{NS}} = \widehat{\theta} - \left(\nabla_\theta^2 F(\mathcal{Z}_{-i}; \widehat{\theta})\right)^{-1} \nabla_\theta F(\mathcal{Z}_{-i}; \widehat{\theta}),$$

Existing Approaches for Approximating $\widehat{\theta}_{-i}$

- $F(\cdot, \theta)$ is twice continuously differentiable in θ:
  - One Newton-Step (NS) estimator [Beirami et al., 2017]:

$$\widetilde{\theta}_{-i}^{\mathrm{NS}} = \widehat{\theta} - \left(\nabla_\theta^2 F(\mathcal{Z}_{-i}; \widehat{\theta})\right)^{-1} \nabla_\theta F(\mathcal{Z}_{-i}; \widehat{\theta}),$$

  - Infinitesimal jackknife (IJ) estimator [Giordano et al., 2019] :

$$\widetilde{\theta}_{-i}^{\mathrm{IJ}} = \widehat{\theta} - \left(\nabla_\theta^2 F(\mathcal{Z}; \widehat{\theta})\right)^{-1} \nabla_\theta F(\mathcal{Z}_{-i}; \widehat{\theta}).$$

Existing Approaches for Approximating $\widehat{\theta}_{-i}$

- $F(\cdot, \theta)$ is twice continuously differentiable in θ:
  - One Newton-Step (NS) estimator [Beirami et al., 2017]:

$$\widetilde{\theta}_{-i}^{\mathrm{NS}} = \widehat{\theta} - \left(\nabla_\theta^2 F(\mathcal{Z}_{-i}; \widehat{\theta})\right)^{-1} \nabla_\theta F(\mathcal{Z}_{-i}; \widehat{\theta}),$$

  - Infinitesimal jackknife (IJ) estimator [Giordano et al., 2019] :

$$\widetilde{\theta}_{-i}^{\mathrm{IJ}} = \widehat{\theta} - \left(\nabla_\theta^2 F(\mathcal{Z}; \widehat{\theta})\right)^{-1} \nabla_\theta F(\mathcal{Z}_{-i}; \widehat{\theta}).$$

- These two methods rely on the assumption $\widehat{\theta}$ can be exactly obtained.

Guarantees for Existing Methods and Limitations

This assumption can be restrictive in a couple of scenarios:

- large-scale problems with limited computational budget
- algorithm has a slow rate of convergence such as SGD
- stop early to avoid overfitting

Guarantees for Existing Methods and Limitations

This assumption can be restrictive in a couple of scenarios:

- large-scale problems with limited computational budget
- algorithm has a slow rate of convergence such as SGD
- stop early to avoid overfitting

What if $\widehat{\theta}$ is unknown?

New solution: Iterative Approximate Cross-Validation (IACV).

General Setup

$$F(\mathcal{Z};\theta) = g(\mathcal{Z};\theta) + h(\theta),$$

where $g(\mathcal{Z};\theta)$ is twice-differentiable in θ while $h(\theta)$ may be nondifferentiable.

- Iterative solver:

$$\widehat{\theta}^{(t)} = \operatorname*{argmin}_{\theta} \left\{ \frac{1}{2\alpha_t} \|\theta - \theta'\|_2^2 + h(\theta) \right\},$$

  where $\theta' = \widehat{\theta}^{(t-1)} - \alpha_t \nabla_\theta g(\mathcal{Z}_{S_t}; \widehat{\theta}^{(t-1)})$.
- $S_t \subseteq [n]$: subset of indices, $\mathcal{Z}_{S_t} := \{Z_i : i \in S_t\}$, and $\alpha_t > 0$: learning rate
- Examples: GD, proxGD and SGD ...

Iterative Approximate CV (IACV)

Recall, for $i = 1, \ldots, n$:

$$\widehat{\theta}_{-i}^{(t)} = \operatorname*{argmin}_{\theta} \left\{ \frac{1}{2\alpha_t} \|\theta - \theta'\|_2^2 + h(\theta) \right\}, \tag{1}$$

where $\theta' = \widehat{\theta}_{-i}^{(t-1)} - \alpha_t \nabla_\theta g(\mathcal{Z}_{S_t}; \widehat{\theta}_{-i}^{(t-1)})$.

Iterative Approximate CV (IACV)

Recall, for $i = 1, \ldots, n$:

$$\widehat{\theta}_{-i}^{(t)} = \operatorname*{argmin}_{\theta} \left\{ \frac{1}{2\alpha_t} \|\theta - \theta'\|_2^2 + h(\theta) \right\}, \tag{1}$$

where $\theta' = \widehat{\theta}_{-i}^{(t-1)} - \alpha_t \nabla_\theta g(\mathcal{Z}_{S_t}; \widehat{\theta}_{-i}^{(t-1)})$.

Goal: generate approximations $\widetilde{\theta}_{-i}^{(t)} \approx \widehat{\theta}_{-i}^{(t)}$, at each iteration $t \geq 1$ and for each $i \in [n]$.

Iterative Approximate CV (IACV)

Recall, for $i = 1, \ldots, n$:

$$\widehat{\theta}_{-i}^{(t)} = \operatorname*{argmin}_{\theta} \left\{ \frac{1}{2\alpha_t} \|\theta - \theta'\|_2^2 + h(\theta) \right\}, \tag{1}$$

where $\theta' = \widehat{\theta}_{-i}^{(t-1)} - \alpha_t \nabla_\theta g(\mathcal{Z}_{S_t}; \widehat{\theta}_{-i}^{(t-1)})$.

Goal: generate approximations $\widetilde{\theta}_{-i}^{(t)} \approx \widehat{\theta}_{-i}^{(t)}$, at each iteration $t \geq 1$ and for each $i \in [n]$.

- (approx the previous iterate) $\widetilde{\theta}_{-i}^{(t-1)} \approx \widehat{\theta}_{-i}^{(t-1)}$
- (approx the gradient) taylor expansion at $\widehat{\theta}^{(t-1)}$

Iterative Approximate CV (IACV)

Recall, for $i = 1, \ldots, n$:

$$\widehat{\theta}_{-i}^{(t)} = \operatorname{argmin}_{\theta} \left\{ \frac{1}{2\alpha_t} \|\theta - \theta'\|_2^2 + h(\theta) \right\}, \quad (1)$$

where $\theta' = \widehat{\theta}_{-i}^{(t-1)} - \alpha_t \nabla_\theta g(\mathcal{Z}_{S_t}; \widehat{\theta}_{-i}^{(t-1)})$.

Goal: generate approximations $\widetilde{\theta}_{-i}^{(t)} \approx \widehat{\theta}_{-i}^{(t)}$, at each iteration $t \geq 1$ and for each $i \in [n]$.

- (approx the previous iterate) $\widetilde{\theta}_{-i}^{(t-1)} \approx \widehat{\theta}_{-i}^{(t-1)}$
- (approx the gradient) taylor expansion at $\widehat{\theta}^{(t-1)}$

$$\text{IACV}: \widetilde{\theta}_{-i}^{(t)} = \arg\min_{\theta} \left\{ \frac{1}{2\alpha_t} \|\theta - \theta'\|_2^2 + h(\theta) \right\},$$

where $\theta' = \widetilde{\theta}_{-i}^{(t-1)} - \alpha_t \left(\nabla_\theta g(\mathcal{Z}_{S_t \setminus i}; \widehat{\theta}^{(t-1)}) + \nabla_\theta^2 g(\mathcal{Z}_{S_t \setminus i}; \widehat{\theta}^{(t-1)}) [\widetilde{\theta}_{-i}^{(t-1)} - \widehat{\theta}^{(t-1)}] \right)$.

Advantages and Guarantees

- Smaller computation complexity
- Guaranteed per-iteration error control
- Recover the existing one-step Newton method in the limit
- Numerically performs well

Advantages and Guarantees

- Smaller computation complexity
- Guaranteed per-iteration error control
- Recover the existing one-step Newton method in the limit
- Numerically performs well

Iterative Approximate Cross-Validation

Exhibit Hall 1 https://arxiv.org/abs/2303.02732

Thank you!

Bibliography

Arlot, S. and Celisse, A. (2010).
A survey of cross-validation procedures for model selection.
Statistics surveys, 4:40–79.

Beirami, A., Razaviyayn, M., Shahrampour, S., and Tarokh, V. (2017).
On optimal generalizability in parametric learning.
Advances in Neural Information Processing Systems, 30.

Giordano, R., Stephenson, W., Liu, R., Jordan, M., and Broderick, T. (2019).
A swiss army infinitesimal jackknife.
In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1139–1147. PMLR.

Rad, K. R. and Maleki, A. (2020).
A scalable estimate of the out-of-sample prediction error via approximate leave-one-out cross-validation.