SOFIA UNIVERSITY
St. KLIMENT OHRIDSKI

LARP: Learner-Agnostic Robust Data Prefiltering

Kristian Minchev, Dimitar I. Dimitrov, Nikola Konstantinov

INSAIT, Sofia University "St. Kliment Ohridski"

Growing Importance of Public Datasets

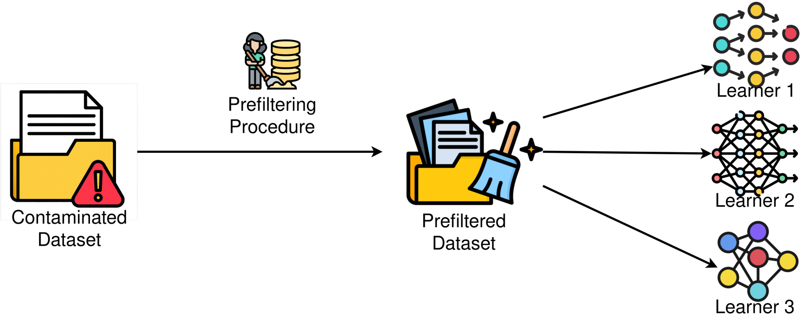
- Public datasets are becoming increasingly important for machine learning (ML) development.
- Increasing ML pipeline fragmentation brings up the role of public dataset curators
- **Challenges:**
  - Ensuring good data quality
  - Compatibility with various downstream models
  - Trustworthy release

Motivation

- Need a framework for provably robust prefiltering procedures
- Data prefiltering should be principled and transparent [2, 4]
- Procedures should be **learner-agnostic**
- In this work, we study the problem of **L**earner-**A**gnostic **R**obust **P**refiltering (LARP)

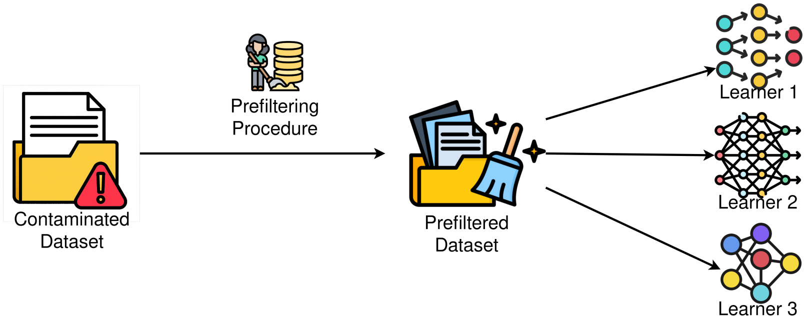
Framework

- Prefiltering procedure – function F from dataset S to $S' \subseteq S$



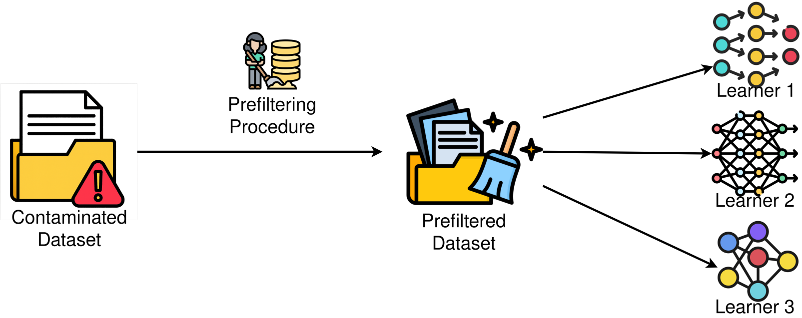
Framework

- Prefiltering procedure – function F from dataset S to $S' \subseteq S$
- Each downstream learner l takes prefiltered dataset S' and returns a hypothesis $l(S')$



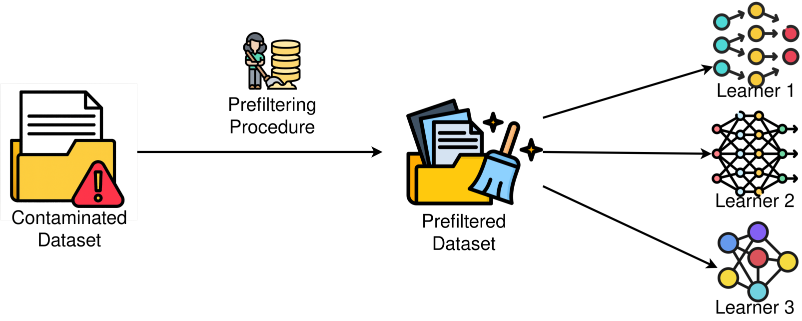
Framework

- Prefiltering procedure – function F from dataset S to $S' \subseteq S$
- Each downstream learner l takes prefiltered dataset S' and returns a hypothesis $l(S')$
  - Mean estimation: $l(S') \in \mathbb{R}$



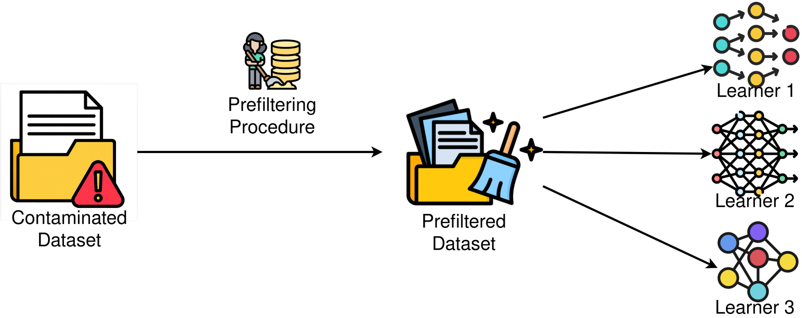
Framework

- Prefiltering procedure – function F from dataset S to $S' \subseteq S$
- Each downstream learner l takes prefiltered dataset S' and returns a hypothesis $l(S')$
  - Mean estimation: $l(S') \in \mathbb{R}$
  - Classification: $l(S') : \mathcal{X} \to \{0, \ldots, c-1\}$



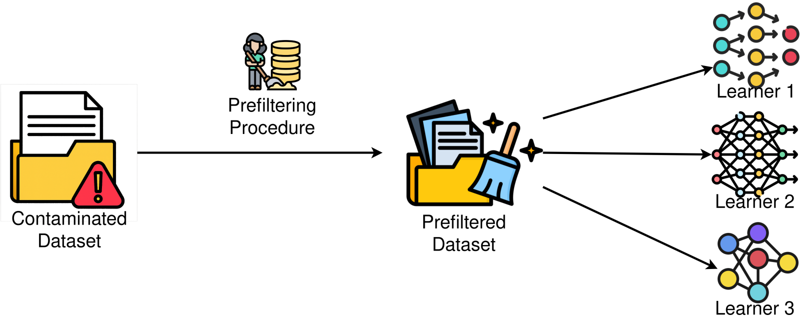
Framework

- Prefiltering procedure – function F from dataset S to $S' \subseteq S$
- Each downstream learner l takes prefiltered dataset S' and returns a hypothesis $l(S')$
- Learner performance is measured by a risk function $R_l(S')$



Framework

- Prefiltering procedure – function F from dataset S to $S' \subseteq S$
- Each downstream learner l takes prefiltered dataset S' and returns a hypothesis $l(S')$
- Learner performance is measured by a risk function $R_l(S')$
- Minimize worst-case loss across the prespecified learner set $\mathcal{L}$
- Learner-agnostic risk: $R_{\mathrm{agn}}(F) := \max_{l \in \mathcal{L}} R_l(S')$



LARP: Theory for Gaussian Scalar Mean Estimation

- Setting: Gaussian scalar mean estimation with Huber contamination [3]

LARP: Theory for Gaussian Scalar Mean Estimation

- Setting: Gaussian scalar mean estimation with Huber contamination [3]
- Learner set - Huber M-estimators

$$l_\delta(S') = \arg\min_{\mu \in \mathbb{R}} \sum_{X \in S'} \rho_\delta(X - \mu)$$

with various parameters $\delta \in \Delta$

LARP: Theory for Gaussian Scalar Mean Estimation

- Setting: Gaussian scalar mean estimation with Huber contamination [3]
- Learner set - Huber M-estimators

$$l_\delta(S') = \arg\min_{\mu \in \mathbb{R}} \sum_{X \in S'} \rho_\delta(X - \mu)$$

  with various parameters $\delta \in \Delta$
- Learner risk is measured as $R_l(S') := (l(S') - \theta)^2$.

Prefiltering with provable upper bounds

- Trimming-based prefiltering F_p^q gives theoretical guarantees

Theorem

Assume that the target distribution is $\mathcal{D}_\theta = \mathcal{N}(\theta, \sigma^2)$*, and that* $\epsilon < 2/7$*. Let* F_p^q *be the quantile prefiltering procedure with* $p \in (0, 1/2)$*. Then, if* $n \geq \Omega(\log(1/\delta_0))$*, with probability* $1 - \delta_0$ *the downstream Huber learners with parameter set* Δ *produce mean estimates* $\hat{\theta}_\delta$ *such that*

$$R_{agn}(F_p^q) \leq \mathcal{O}\left(\left(\epsilon^2 + \frac{\log(1/\delta_0)}{n}\right)\sigma^2 + \max_{\delta \in \Delta} \delta^2\right).$$

Prefiltering with provable upper bounds

- Trimming-based prefiltering F_p^q gives theoretical guarantees

Theorem

Assume that the target distribution is $\mathcal{D}_\theta = \mathcal{N}(\theta, \sigma^2)$, and that $\epsilon < 2/7$. Let F_p^q be the quantile prefiltering procedure with any $p \in (0, 1/2)$. Then, if $n \geq \Omega(\log(1/\delta_0))$, with probability $1 - \delta_0$ the downstream Huber learners with parameter set Δ produce mean estimates $\hat{\theta}_\delta$ such that

$$R_{agn}(F_p^q) \leq \mathcal{O}\left(\left(\epsilon^2 + \frac{\log(1/\delta_0)}{n}\right)\sigma^2 + \max_{\delta \in \Delta} \delta^2\right).$$

- Risk bound depends on contamination rate, sample size, and heterogeneity of learner set

Prefiltering with provable upper bounds

- Trimming-based prefiltering F_p^q gives theoretical guarantees

Theorem

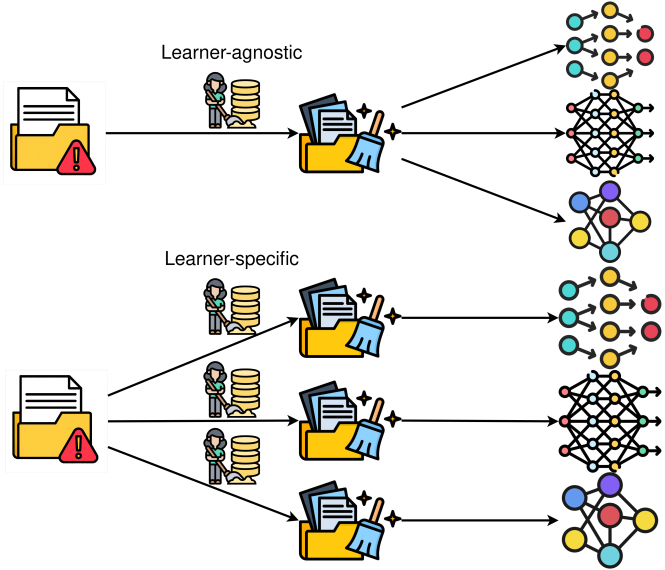
Assume that the target distribution is $\mathcal{D}_\theta = \mathcal{N}(\theta, \sigma^2)$, and that $\epsilon < 2/7$. Let F_p^q be the quantile prefiltering procedure with any $p \in (0, 1/2)$. Then, if $n \geq \Omega(\log(1/\delta_0))$, with probability $1 - \delta_0$ the downstream Huber learners with parameter set Δ produce mean estimates $\hat{\theta}_\delta$ such that

$$R_{agn}(F_p^q) \leq \mathcal{O}\left(\left(\epsilon^2 + \frac{\log(1/\delta_0)}{n}\right)\sigma^2 + \max_{\delta \in \Delta} \delta^2\right).$$

- Risk bound depends on contamination rate, sample size, and heterogeneity of learner set
- First two bounds are close to robust statistics results [1], last term increases with heterogeneity

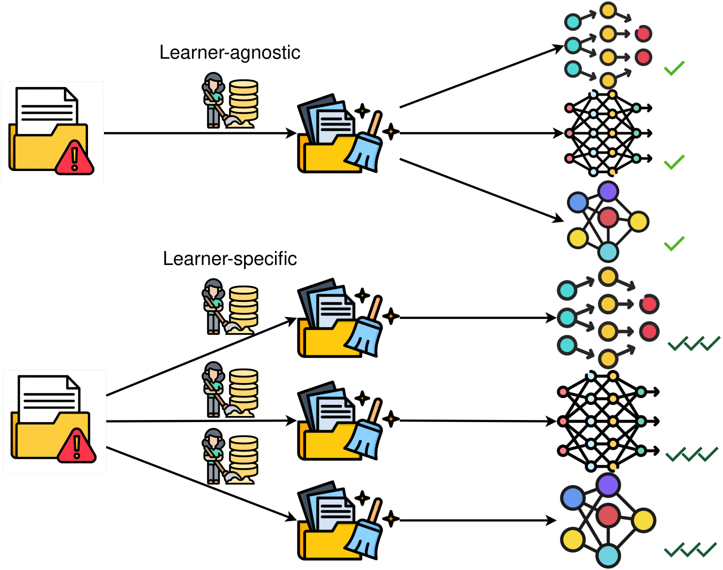
The Price of Learner-Agnostic Prefiltering

- Comparison of learner-specific vs. learner-agnostic prefiltering



The Price of Learner-Agnostic Prefiltering

- Learner-agnostic prefiltering $\Rightarrow$ inherently worse guarantees



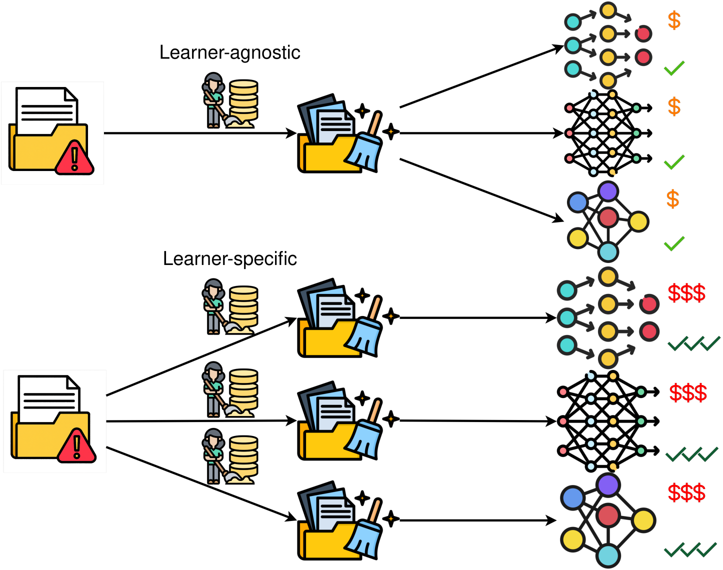
The Price of Learner-Agnostic Prefiltering

- Learner-agnostic prefiltering $\Rightarrow$ inherently worse guarantees
- This leads to reduced downstream utility, which can be captured through a function $\mathcal{U}_{red}\left(R_{worse}, R_{better}\right)$.
- Price of LARP defined via average utility drop across learners, compared to learner-specific optima:

$$P\left(F\right) := \frac{1}{|\mathcal{L}|}\sum_{l\in\mathcal{L}}\mathcal{U}_{red}\left(R_l(F), R_l(F_l^*)\right).$$

Tradeoff in agnostic vs specific prefiltering

- We propose a model where shared costs offset price of LARP



Game-Theoretic Justification

- Learners choose between shared vs. individual prefiltering

Game-Theoretic Justification

- Learners choose between shared vs. individual prefiltering
- Following result proves the benefit of LARP for large datasets:

Game-Theoretic Justification

- Learners choose between shared vs. individual prefiltering
- Following result proves the benefit of LARP for large datasets:

Lemma

If the total cost of prefiltering a dataset is Cn^{α} and the dataset size n satisfies

$$n > \left[\frac{|\mathcal{L}|}{C\left(|\mathcal{L}|-1\right)}P(F)\right]^{1/\alpha},$$

then there is a payment scheme $(p_l)_{l\in\mathcal{L}}$ such that no learner is incentivized to opt out of the learner-agnostic prefiltering scheme.

The Price of LARP: Mean Estimation Case

- We explore $P(F)$ in the univariate Gaussian setup

The Price of LARP: Mean Estimation Case

- We explore $P(F)$ in the univariate Gaussian setup
- We use the upper bound

$$R_{agn}(F_p^q) \leq \mathcal{O}\left(\left(\epsilon^2 + \frac{\log(1/\delta_0)}{n}\right)\sigma^2 + \max_{\delta \in \Delta} \delta^2\right).$$

as proxy for R_{agn}

The Price of LARP: Mean Estimation Case

- We explore $P(F)$ in the univariate Gaussian setup
- We use the upper bound

$$R_{agn}(F_p^q) \leq \mathcal{O}\left(\left(\epsilon^2 + \frac{\log(1/\delta_0)}{n}\right)\sigma^2 + \max_{\delta \in \Delta} \delta^2\right).$$

  as proxy for R_{agn}
- Utility drop function is $\mathcal{U}_{red}(R_{worse}, R_{better}) := R_{worse} - R_{better}$

The Price of LARP: Mean Estimation Case

- We then rephrase the lemma as follows:

Lemma

If in addition to the assumption to $n \geq \Omega\left(\log\left(1/\delta_0\right)\right)$, *the dataset size* n *satisfies*

$$n \geq \left(\frac{|\Delta|}{C(|\Delta|-1)}\left(\max_{\delta\in\Delta}\delta^2 - \min_{\delta\in\Delta}\delta^2\right)\right)^{1/\alpha},$$

then, there is a payment scheme $(p_i)_N^{i=1}$ *such that, with probability* $1-\delta_0$ *over the randomness of the noisy sample, we have* $U_{agn}^{(i)} \geq U_{spec}^{(i)}$ *for all* $i = 1, \ldots, N$.

- Holds for sufficiently large n
- Can be computed in advance

Experimental Setup

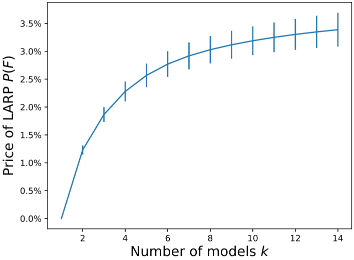
We measure the price of LARP for various classification setups:

- Datasets: CIFAR-10 (images), Adult (tabular)
- Corruption types: Label noise, Shortcuts
- Diverse learner sets: CNNs, SVMs, Boosting, and others
- Learner heterogeneity: different algorithms and/or hyperparameters
- Risk: Error rate on test set

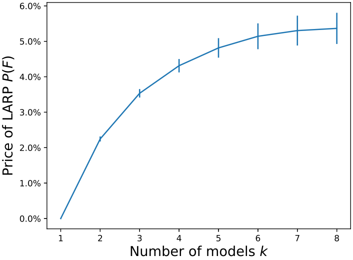
Empirical Findings

- Price of LARP is statistically significant
- Price increases with learner heterogeneity

Adult (Label)



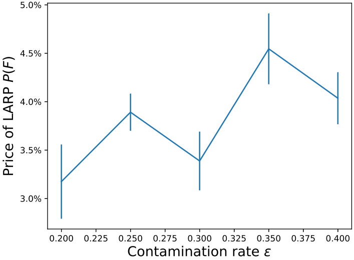
CIFAR-10 (Shortcut)



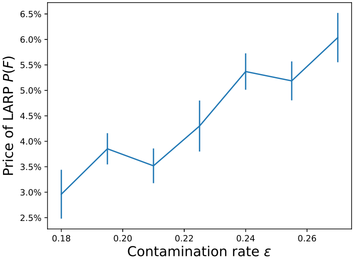
Empirical Findings

- Price increases with contamination ratio

Adult (Label)



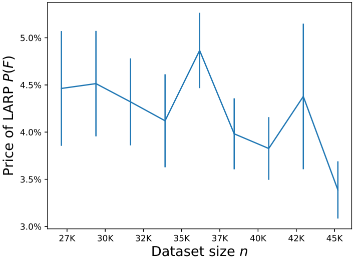
CIFAR-10 (Shortcut)



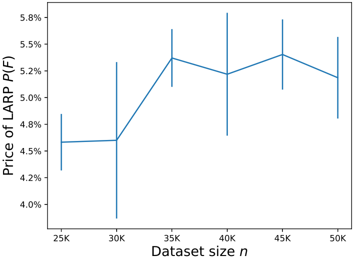
Empirical Findings

- No strong dependence on dataset size
- Evidence that Lemma might hold for large datasets

Adult (Label)



CIFAR-10 (Shortcut)



Summary & Takeaways

- Introduced a framework for robust learner-agnostic prefiltering
- Showed feasibility of LARP : Theoretical guarantees + empirical evidence
- Analyzed the trade-off between price of LARP and reduced prefiltering costs
- Ideas for future work:
  - Regression, classification, R^d mean estimation
  - Prefiltering with practical guarantees on price of LARP
  - Data curation beyond prefiltering

References

[1] Ilias Diakonikolas and Daniel M Kane. *Algorithmic high-dimensional robust statistics*. Cambridge university press, 2023.

[2] Timnit Gebru et al. "Datasheets for datasets". In: *Communications of the ACM* 64.12 (2021), pp. 86–92.

[3] Peter J Huber. "Robust statistics". In: *International encyclopedia of statistical science*. Springer, 2011, pp. 1248–1251.

[4] Weixin Liang et al. "Advances, challenges and opportunities in creating data for trustworthy AI". In: *Nature Machine Intelligence* 4.8 (2022), pp. 669–677.

Thank You!

Kristian Minchev

Dimitar I. Dimitrov

Nikola Konstantinov

Contact: kristian.minchev@insait.ai

More details in the paper: