Semi-supervised Learning : Inference

Gyumin Lee

Yonsei University

August 29, 2023

Introduction

Brief summary

The purpose of the presentation is to introduce semi-supervised inference and related studies.

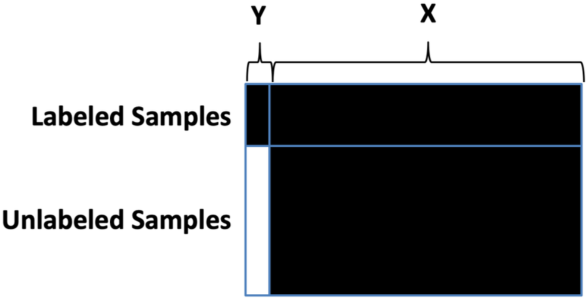
Papers

1. Doubly Robust Self-Training
    - Banghua Zhu, Mingyu Ding, Philip Jacobson, Ming Wu, Wei Zhan, Michael I. Jordan, and Jiantao Jiao(June 2023)
    - https://arxiv.org/abs/2306.00265
2. Prediction-Powered Inference
    - Anastasios N. Angelopoulos, Stephen Bates, Clara Fannjiang, Michael I. Jordan, and Tijana Zrnic.(February 2023)
    - https://arxiv.org/abs/2301.09633

Contents

Preliminaries: Semi-Supervised Inference

- Basic assumption:
  labeled data are more difficult/expensive to acquire than unlabeled data.
  $\rightarrow$ utilize unlabeled data!



- Examples
  - Survey sampling
  - Electronic health record
  - homeless consensus

  ⋮

Preliminaries: Semi-Supervised Inference

Basic Setting

- Labled data : $(X, Y) \in (\mathcal{X}, \mathcal{Y})^n$ where $X = (X_1, ..., X_n)$ and $Y = (Y_1, ..., Y_n)$.
- Unlabeled data : $(\tilde{X}, \tilde{Y}) \in (\mathcal{X}, \mathcal{Y})^m$ where $\tilde{Y}$ is not observed.
- Assume that (X, Y) and $(\tilde{X}, \tilde{Y})$ are i.i.d. samples from a common distribution $\mathbb{P}$.
- Prediction rule : $f : \mathcal{X} \to \mathcal{Y}$, independent of the observed data.(e.g. pretrained model)

Goal

Our estimand of interest is denoted as θ^* which could be $E[Y]$(mean estimation), $\min\{\theta : P(Y \leq \theta) \geq q\}$(quantile estimation), or $\arg\min_{\theta \in \mathbb{R}^d} E[l_\theta(X, Y)]$ for some loss function l_θ.

Preliminaries: Semi-Supervised Inference

Example

Let our goal of estimation be $\theta^* = E[Y]$.
We have several candidates of estimators as follows

- $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$
- $\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - f(X_i)) + \frac{1}{n+m}\sum_{i=1}^{n+m} f(X_i)$

Note that both are unbiased estimators. How about variance?

- $Var(\bar{Y}) = Var(\frac{1}{n}\sum_{i=1}^{n} Y_i) = \frac{1}{n}Var(Y)$
- $$\begin{aligned} Var(\hat{\theta}) &= Var(\frac{1}{n}\sum_{i=1}^{n}(Y_i - f(X_i)) + \frac{1}{n+m}\sum_{i=1}^{n+m} f(X_i)) \\ &= \frac{1}{n}(Var(Y) + \frac{m-n}{m}(Var(f(X)) - 2Cov(Y, f(X)))) \end{aligned}$$

$\rightarrow$ If $Var(f(X)) < 2Cov(Y, f(X))$, then $\hat{\theta}$ improves the original one!

Preliminaries: Semi-Supervised Inference

Example(cont.)

Suppose we take an ideal prediction function $f(X) = E(Y|X)$
Since

$$\begin{aligned} Cov(Y, E(Y|X)) &= E[(Y - E[Y])(E[Y|X] - E[Y])] \\ &= E[E[(Y - E[Y])(E[Y|X] - E[Y])|X]] \\ &= E[(E[Y|X] - E[Y])^2] \\ &= Var(E(Y|X)) \end{aligned}$$

, we obtain the result assuming $n << m$ as

$$\begin{aligned} Var(\hat{\theta}) &= \frac{1}{n}E(Var(Y|X)) + \frac{1}{m}Var(E(Y|X)) \\ &\leq \frac{1}{n}E(Var(Y|X)) + \frac{1}{n}Var(E(Y|X)) = \frac{1}{n}Var(Y) = Var(\bar{Y}) \end{aligned}$$

Hence, $\hat{\theta}$ has smaller variance than the sample mean $\bar{Y}$.

Related Works

Semi-supervised inference: general theory and estimation of means(Anru Zhang, Lawrence D. Brown and T. Tony Cai)

For the ideal semi-supervised inference, where $m = \infty$, the proposed estimator is 'least square estimator', which is defined as

$$\hat{\theta}_{LS} = \boldsymbol{\mu}'\hat{\beta} = \hat{\beta}_1 + \mu'\hat{\beta}_{(2)} = \bar{Y} - \hat{\beta}'_{(2)}(\bar{X} - \mu).$$

Here, $\boldsymbol{\mu} = (1, \mu')' = \mathbb{E}X$ is known and $\hat{\beta} = [\hat{\beta}_1 \hat{\beta}'_{(2)}]' = (X'X)^{-1}X'Y$, where X is $n \times (p+1)$ matrix including intercept column of ones.

On the other hand, for the ordinary semi-supervised inference, where $m < \infty$, the proposed estimator is 'semi-supervised least squared estimator', which is defined as

$$\hat{\theta}_{SSLS} = \hat{\boldsymbol{\mu}}'\hat{\beta} = \bar{Y} - \hat{\beta}'_{(2)}(\bar{X} - \hat{\mu}).$$

Here, $\hat{\boldsymbol{\mu}} = (1, \hat{\mu}')' = \frac{1}{n+m}\sum_{k=1}^{n+m} X_k$.

Related Works

Semi-supervised inference: general theory and estimation of means(Anru Zhang, Lawrence D. Brown and T. Tony Cai)

- Results about l_2 risks

$$
\begin{aligned}
nE(\bar{Y}-\theta)^2 &= \tau^2 + \beta_{(2)}^\top \Sigma_n \beta_{(2)} \\
nE(\hat{\theta}_{LS}^1-\theta)^2 &= \tau^2 + O(\frac{p^2}{n}) \\
nE\left(\hat{\theta}_{\mathrm{SSLS}}^1-\theta\right)^2 &= \tau_n^2 + \frac{n}{n+m}\beta_{(2)}^\top \Sigma_n \beta_{(2)} + O(\frac{p^2}{n}) \\
&\approx \frac{n}{n+m}E(\bar{Y}-\theta)^2 + \frac{m}{n+m}E(\hat{\theta}_{LS}^1-\theta)^2
\end{aligned}
$$

.

Related Works

Semi-supervised inference: general theory and estimation of means(Anru Zhang, Lawrence D. Brown and T. Tony Cai)

Let $(Y_1, X_1), \cdots, (Y_n, X_n)$ be i.i.d. copies from P, and assume that $[Y, X]$ has finite second moments, $\vec{\Xi}$ is non-singular and $\tau^2 > 0$. Then, under the setting that P is fixed and $n \to \infty$

$$\frac{\hat{\theta}_{\mathrm{LS}} - \theta}{\tau/\sqrt{n}} \xrightarrow{d} N(0,1),$$

and

$$MSE/\tau^2 \xrightarrow{d} 1, \text{ where } MSE := \frac{\sum_{i=1}^{n} \left(Y_i - X_i^\top \hat{\beta}\right)^2}{n-p-1}, \tau^2 = E(Y_i - \vec{X}_i^\top \beta)^2$$

.

Related Works

Semi-supervised inference: general theory and estimation of means(Anru Zhang, Lawrence D. Brown and T. Tony Cai)

On the other hand, Let $(Y_1, X_1), \cdots, (Y_n, X_n)$ be i.i.d. labeled samples from P, and let $X_{n+1}, \cdots, X_{n+m}$ be m additional unlabeled independent samples from P_X. Suppose $\vec{\Xi}$ is nonsingular and $\tau^2 > 0$. If P is fixed and $n \to \infty$, then

$$\frac{\sqrt{n}\left(\hat{\theta}_{\mathrm{SSLS}} - \theta\right)}{\nu} \xrightarrow{d} N(0,1),$$

and

$$\frac{\hat{\nu}^2}{\nu^2} \xrightarrow{d} 1$$

where $\hat{\nu}^2 = \frac{m}{m+n} MSE + \frac{n}{m+n} \hat{\sigma}_Y^2$ with $MSE = \frac{1}{n-p-1} \sum_{k=1}^n \left(Y_i - \vec{X}_k^\top \hat{\beta}\right)^2$, $\nu^2 = \tau^2 + \frac{n}{n+m} \beta_{(2)}^\top \Sigma_n \beta_{(2)}$ and $\hat{\sigma}_Y^2 = \frac{1}{n-1} \sum_{k=1}^n \left(Y_i - \overline{\mathbf{Y}}\right)^2$.

Related Works

- Semi-supervised inference: general theory and estimation of means(Anru Zhang, Lawrence D. Brown and T. Tony Cai, The Annals of Statistics) - mean estimation on regression setting
- "Semisupervised inference for explained variance in high dimensional linear regression and its applications(T. Tony Cai and Zijian Guo,JRSS)" - Variance estimation
- "High-dimensional Semi-supervised Learning: in Search for Optimal Inference of the Mean(Yuqian Zhang and Jelena Bradic, Biometrika)" - high-dimensional setting
- "The correlation-assisted missing data estimator(Timothy I. Cannings and Yingying Fan, JMLR)" - missing data approach, extension to U-statistics
- "Methods for correcting inference based on outcomes predicted by machine learning(Siruo Wang, Tyler H. McCormick, and Jeffrey T. Leek, Proceedings of the National Academy of Sciences, 2020)" - different approach using bootstrap
- "Valid inference after prediction(Keshav Motwani and Daniela Witten)" - showing the above method is a bad approach

Doubly Robust Self-Training

Motivation

- Given a teacher model, a large unlabeled dataset and a small labeled dataset, how can we design a principled learning process that ensures consistent and sample-efficient learning of the true model?

Self-Training

- involves using a teacher model to generate pseudo-labels for all unlabeled data, and then training a new model on a mixture of both pseudo-labeled and labeled data.
- can lead to overreliance on the teacher model and can miss important information provided by the labeled data.

$\rightarrow$ highly sensitive to the accuracy of the teacher model

Doubly Robust Self-Training

Problem setting

- Given
  Unlabeled samples $\mathcal{D}_1 = \{X_1, ..., X_m\}$ drawn from $\mathbb{P}_X$,
  Labeled samples $\mathcal{D}_2 = \{(X_{m+1}, Y_{m+1}), ..., (X_{m+n}, Y_{m+n})\}$ drawn from $\mathbb{P}_X \times \mathbb{P}_{Y|X}$
  Pre-trained model $\hat{f} : \mathcal{X} \to \mathcal{Y}$
- Goal: find θ^* such that $\theta^* = argminE[l_\theta(X, Y)]$ for some loss function l_θ

Main result

$$\mathcal{L}^{DR}_{\mathcal{D}_1,\mathcal{D}_2} = \frac{1}{m+n}\sum_{i=1}^{m+n} l_\theta(X_i, \hat{f}(X_i)) - \frac{1}{n}\sum_{i=m+1}^{m+n} l_\theta(X_i, \hat{f}(X_i)) + \frac{1}{n}\sum_{i=m+1}^{m+n} l_\theta(X_i, Y_i)$$

is doubly robust!

Doubly Robust Self-Training

Main result

- Traditional:

$$\mathcal{L}^{\mathrm{SL}}_{\mathcal{D}_1,\mathcal{D}_2}(\theta) = \frac{1}{m+n}\left(\sum_{i=1}^{m} \ell_\theta\left(X_i, \hat{f}\left(X_i\right)\right) + \sum_{i=m+1}^{m+n} \ell_\theta\left(X_i, Y_i\right)\right)$$

$$= \frac{1}{m+n}\sum_{i=1}^{m+n} \ell_\theta\left(X_i, \hat{f}\left(X_i\right)\right) - \frac{1}{m+n}\sum_{i=m+1}^{m+n} \ell_\theta\left(X_i, \hat{f}\left(X_i\right)\right) + \frac{1}{m+n}\sum_{i=m+1}^{m+n} \ell_\theta\left(X_i, Y_i\right)$$

- Doubly Robust:

$$\mathcal{L}^{DR}_{\mathcal{D}_1,\mathcal{D}_2} = \frac{1}{m+n}\sum_{i=1}^{m+n} l_\theta(X_i, \hat{f}(X_i)) - \frac{1}{n}\sum_{i=m+1}^{m+n} l_\theta(X_i, \hat{f}(X_i)) + \frac{1}{n}\sum_{i=m+1}^{m+n} l_\theta(X_i, Y_i)$$

Doubly Robust Self-Training

Mean estimation

- Goal: find θ^* such that $\theta^* = argminE[(\theta - Y)^2]$
- Loss only from labeled data
  : $\mathcal{L}^{TL}_{\mathcal{D}_1,\mathcal{D}_2} = \frac{1}{n}\sum_{i=m+1}^{m+n}(\theta - Y_i)^2$
  $\rightarrow \hat{\theta}_{TL} = \frac{1}{n}\sum_{i=m+1}^{m+n} Y_i$
- Loss from self-training
  : $\mathcal{L}^{SL}_{\mathcal{D}_1,\mathcal{D}_2} = \frac{1}{m+n}\left(\sum_{i=1}^{m}(\theta - \hat{f}(X_i))^2 + \sum_{i=m+1}^{m+n}(\theta - Y_i)^2\right)$
  $\rightarrow \hat{\theta}_{SL} = \frac{1}{m+n}\left(\sum_{i=1}^{m}\hat{f}(X_i) + \sum_{i=m+1}^{m+n} Y_i\right)$
- Doubly robust loss: $\mathcal{L}^{DR}_{\mathcal{D}_1,\mathcal{D}_2} =$
  $\frac{1}{m+n}\sum_{i=1}^{m+n} l_\theta(X_i,\hat{f}(X_i)) - \frac{1}{n}\sum_{i=m+1}^{m+n} l_\theta(X_i,\hat{f}(X_i)) + \frac{1}{n}\sum_{i=m+1}^{m+n} l_\theta(X_i, Y_i)$
  $\rightarrow \hat{\theta}_{DR} = \frac{1}{m+n}\sum_{i=1}^{m+n}\hat{f}(X_i) - \frac{1}{n}\sum_{i=m+1}^{m+n}\hat{f}(X_i) - Y_i$

Doubly Robust Self-Training

Mean estimation(cont.)

- $E[(\theta^* - \hat{\theta}_{TL})^2] = \frac{1}{n}Var(Y)$
- $E[(\theta^* - \hat{\theta}_{SL})^2] \leq \frac{2m^2}{(m+n)^2}E[(\hat{f}(X) - Y)^2] + \frac{2m}{(m+n)^2}Var(\hat{f}(X) - Y) + \frac{2n}{(m+n)^2}Var(Y)$
- $E[(\theta^* - \hat{\theta}_{DR})^2] \leq$
  $2min\left(\frac{1}{n}Var(Y) + \frac{m+2n}{(m+n)n}Var(\hat{f}(X)), \frac{m+2n}{(m+n)n}Var(\hat{f}(X) - Y) + \frac{1}{m+n}Var(Y)\right)$

$\rightarrow$ $E[(\theta^* - \hat{\theta}_{DR})^2] \leq \frac{4}{n}(\mathrm{Var}[Y] + \mathrm{Var}[\hat{f}(X)])$ no matter how poor the estimator $\hat{f}(X)$ is.

$\rightarrow$ when $\mathrm{Var}[\hat{f}(X) - Y]$ is small, $E[(\theta^* - \hat{\theta}_{DR})^2] \leq \frac{2}{m+n}\mathrm{Var}[Y]$.

$\rightarrow$ $E[(\theta^* - \hat{\theta}_{SL})^2]$ always has a non-vanishing term, $\frac{2m^2}{(m+n)^2}\mathbb{E}[(\hat{f}(X) - Y)]^2$ unless the predictor is accurate.

Doubly Robust Self-Training

General loss

With probability at least $1-\delta$,

$$\left\|\nabla_\theta \mathcal{L}^{\mathrm{DR}}_{\mathcal{D}_1,\mathcal{D}_2}(\theta^\star)\right\|_2 \leq C \min\left(\left\|\Sigma^{\hat{f}}_{\theta^\star}\right\|_2 \sqrt{\frac{d}{(m+n)\delta}} + \left\|\Sigma^{Y-\hat{f}}_{\theta^\star}\right\|_2 \sqrt{\frac{d}{n\delta}}, \left\|\Sigma^{\hat{f}}_{\theta^\star}\right\|_2 \left(\sqrt{\frac{d}{(m+n)\delta}} + \sqrt{\frac{d}{n\delta}}\right) + \left\|\Sigma^{Y}_{\theta^\star}\right\|_2 \sqrt{\frac{d}{n\delta}}\right),$$

where C is a universal constant, and we denote $\Sigma^{Y-\hat{f}}_\theta = \mathrm{Cov}\left[\nabla_\theta \ell_\theta(X,\hat{f}(X)) - \nabla_\theta \ell_\theta(X,Y)\right]$ and let $\Sigma^{\hat{f}}_\theta = \mathrm{Cov}\left[\nabla_\theta \ell_\theta(X,\hat{f}(X))\right]$, $\Sigma^{Y}_\theta = \mathrm{Cov}\left[\nabla_\theta \ell_\theta(X,Y)\right]$.

- $\left\|\nabla_\theta \mathcal{L}^{\mathrm{SL}}_{\mathcal{D}_1,\mathcal{D}_2}(\theta^\star)\right\|_2 \geq C$ for some positive constant C
- When $\hat{f}$ is a perfect predictor, one has $\mathcal{L}^{\mathrm{DR}}_{\mathcal{D}_1,\mathcal{D}_2}(\theta^\star) = \frac{1}{m+n}\sum_{i=1}^{m+n} \ell_\theta(X_i, Y_i)$.

Doubly Robust Self-Training

The case of Distribution mismatch

$$\mathcal{L}^{\mathrm{DR2}}_{\mathcal{D}_1,\mathcal{D}_2}(\theta) = \frac{1}{m}\sum_{i=1}^{m}\ell_\theta\left(X_i, \hat{f}\left(X_i\right)\right) - \frac{1}{n}\sum_{i=m+1}^{m+n}\frac{1}{\pi\left(X_i\right)}\ell_\theta\left(X_i, \hat{f}\left(X_i\right)\right) + \frac{1}{n}\sum_{i=m+1}^{m+n}\frac{1}{\pi\left(X_i\right)}\ell_\theta\left(X_i, Y_i\right)$$

$\mathbb{E}\left[\mathcal{L}^{\mathrm{DR2}}_{\mathcal{D}_1,\mathcal{D}_2}(\theta)\right] = \mathbb{E}_{\mathbb{P}_{X,Y}}\left[\ell_\theta(X, Y)\right]$ as long as one of the two assumptions hold:

- For any $x, \pi(x) = \frac{\mathbb{P}_X(x)}{\mathbb{Q}_X(x)}$.
- For any $x, \ell_\theta(x, \hat{f}(x)) = \mathbb{E}_{Y\sim\mathbb{P}_{Y|X=x}}\left[\ell_\theta(x, Y)\right]$.

Doubly Robust Self-Training

Experiment: ImageNet

Minimize the curriculum-based loss in epoch

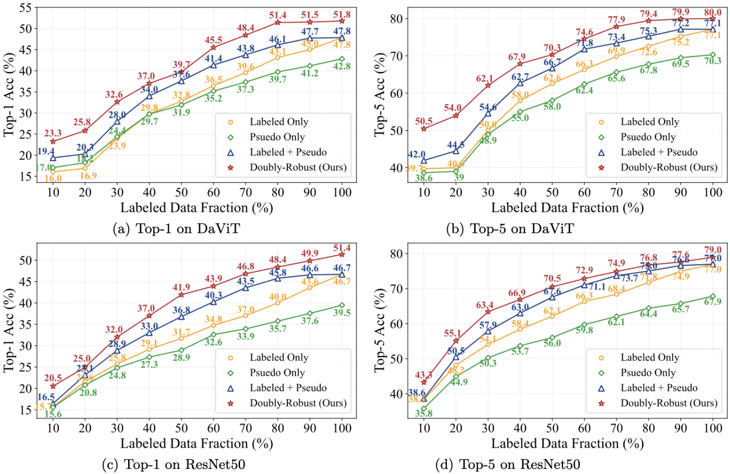
$$\mathcal{L}^{\mathrm{DR},t}_{\mathcal{D}_1,\mathcal{D}_2}(\theta) = \frac{1}{m+n}\sum_{i=1}^{m+n} \ell_\theta\left(X_i, \hat{f}(X_i)\right) - \alpha_t \cdot \left(\frac{1}{n}\sum_{i=m+1}^{m+n} \ell_\theta\left(X_i, \hat{f}(X_i)\right) - \frac{1}{n}\sum_{i=m+1}^{m+n} \ell_\theta(X_i, Y_i)\right).$$

Table 1: Comparisons on mini-ImageNet100, all models trained for 100 epochs.

Labeled Data Percent	Labeled Only		Pseudo Only		Labeled + Pseudo		Doubly robust Loss	
	top1	top5	top1	top5	top1	top5	top1	top5
1	2.72	9.18	**2.81**	9.57	2.73	9.55	2.75	**9.73**
5	3.92	13.34	4.27	13.66	4.27	14.4	**4.89**	**16.38**
10	6.76	20.84	7.27	21.64	7.65	22.48	**8.01**	**21.90**
20	12.3	31.3	13.46	30.79	**13.94**	**32.63**	13.50	32.17
50	20.69	46.86	20.92	45.2	24.9	50.77	**25.31**	**51.61**
80	27.37	55.57	25.57	50.85	30.63	58.85	**30.75**	**59.41**
100	31.07	60.62	28.95	55.35	**34.33**	62.78	34.01	**63.04**

Doubly Robust Self-Training

Experiment: ImageNet



(a) Top-1 on DaViT (b) Top-5 on DaViT

(c) Top-1 on ResNet50 (d) Top-5 on ResNet50

Figure 1: Comparisons on ImageNet100 using two different network architectures. Both Top-1 and Top-5 accuracies are reported. All models are trained for 20 epochs.

Doubly Robust Self-Training

Experiment: nuScenes

$$\mathcal{L}_{obj}^{\mathrm{DR}}(\theta) = \frac{1}{M+N_{ps}} \sum_{i=1}^{M+N_{ps}} \ell_\theta\left(X_i, f\left(X_i\right)\right) - \frac{1}{N_{ps}} \sum_{i=M+1}^{M+N_{ps}} \ell_\theta\left(X_i', f\left(X_i'\right)\right) + \frac{1}{N} \sum_{i=M+1}^{M+N} \ell_\theta\left(X_i, Y_i\right),$$

Table 2: Performance comparison on nuScenes *val* set.

Labeled Data Fraction	Labeled Only		Labeled + Pseudo		Doubly robust Loss	
	mAP↑	NDS↑	mAP↑	NDS↑	mAP↑	NDS↑
1/24	7.56	18.01	7.60	17.32	**8.18**	**18.33**
1/16	11.15	20.55	11.60	21.03	**12.30**	**22.10**
1/4	25.66	41.41	**28.36**	**43.88**	27.48	43.18

Table 3: Per-class mAP (%) comparison on nuScenes *val* set using 1/16 of total labels in training.

	Car	Ped	Truck	Bus	Trailer	Barrier	Traffic Cone
Labeled Only	48.6	30.6	8.5	6.2	4.0	6.8	4.4
Labeled + Pseudo	48.8	30.9	8.8	7.5	5.7	6.7	4.0
Improvement	+0.2	+0.3	+0.3	+1.3	**+1.7**	-0.1	-0.4
Doubly robust Loss	51.5	32.9	9.6	8.2	5.2	7.2	4.5
Improvement	**+2.9**	**+2.3**	**+1.1**	**+2.0**	+1.2	**+0.4**	**+0.1**

Prediction-Powered Inference

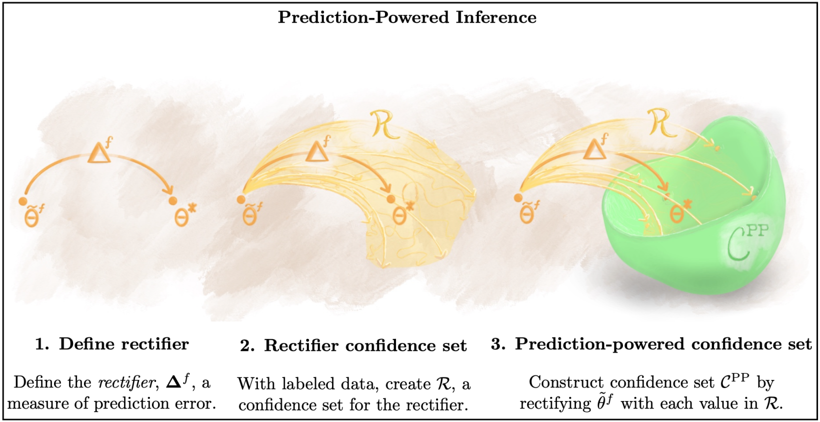
Motivation

How can we assess the role of prediction in terms of basic principles of statistical inference? Is it possible to exploit predictions from a machine-learning system while still providing guarantees of statistical validity?

- $\rightarrow$ suggest a framework for performing valid statistical inference when an experimental data set is supplemented with predictions.
- $\rightarrow$ use predictions from the model to perform inference, leverage the immense number of predictions to improve their confidence in a scientific conclusion.

Prediction-Powered Inference

General Framework



Prediction-Powered Inference

Mean estimation

Our estimtes

$$\hat{\theta}^{class} = \frac{1}{n}\sum_{i=1}^{n} Y_i \quad vs \quad \hat{\theta}^{\mathrm{PP}} = \underbrace{\frac{1}{N}\sum_{i=1}^{N} \tilde{f}_i}_{\tilde{\theta}^f} - \underbrace{\frac{1}{n}\sum_{i=1}^{n} (f_i - Y_i)}_{\hat{\boldsymbol{\Delta}}^f}.$$

and their corresponding confidence intervals

$$\underbrace{\hat{\theta}^{\text{class}} \pm 1.96\sqrt{\frac{\hat{\sigma}_Y^2}{n}}}_{\text{classical interval}} \quad \text{or} \quad \underbrace{\hat{\theta}^{\mathrm{PP}} \pm 1.96\sqrt{\frac{\hat{\sigma}_{f-Y}^2}{n} + \frac{\hat{\sigma}_{\tilde{f}}^2}{N}}}_{\text{prediction-powered interval}}$$

Prediction-Powered Inference

Convex estimation

Consider estimands of the form

$$\theta^* = \underset{\theta \in \mathbb{R}^p}{\arg\min} \mathbb{E}\left[\ell_\theta\left(X_1, Y_1\right)\right],$$

where a loss function l_θ is convex.

From convexity, it holds that

$$\mathbb{E}\left[g_{\theta^*}\left(X_1, Y_1\right)\right] = 0$$

where $g_\theta : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^p$ is a subgradient of ℓ_θ with respect to θ.

Define the rectifier

$$\boldsymbol{\Delta}^f(\theta) = \mathbb{E}\left[g_\theta\left(X_1, Y_1\right) - g_\theta\left(X_1, f_1\right)\right].$$

Prediction-Powered Inference

Convex estimation

Create the confidence set for the rectifier, $\mathcal{R}_\delta(\theta)$,

$$P\left(\boldsymbol{\Delta}^f(\theta) \in \mathcal{R}_\delta(\theta)\right) \geq 1 - \delta.$$

Also, for every θ, we want a confidence set $\mathcal{T}_{\alpha-\delta}(\theta)$ for $\mathbb{E}\left[g_\theta\left(X_1, f_1\right)\right]$, satisfying

$$P\left(\mathbb{E}\left[g_\theta\left(X_1, f_1\right)\right] \in \mathcal{T}_{\alpha-\delta}(\theta)\right) \geq 1 - (\alpha - \delta).$$

Using the results, finally form the confidence set for θ^* as follows

$$P\left(\theta^* \in \mathcal{C}_\alpha^{\mathrm{PP}}\right) \geq 1 - \alpha$$

where $\mathcal{C}_\alpha^{\mathrm{PP}} = \{\theta : 0 \in \mathcal{R}_\delta(\theta) + \mathcal{T}_{\alpha-\delta}(\theta)\}$ where $+$ denotes the Minkowski sum.

Prediction-Powered Inference

Various algorithms

Algorithm 1 Prediction-powered mean estimation

Input: labeled data (X, Y), unlabeled features $\widetilde{X}$, predictor f, error level $\alpha \in (0,1)$

$\hat{\theta}^{\text{PP}} \leftarrow \tilde{\theta}^f - \hat{\mathbf{\Delta}}^f := \frac{1}{N}\sum_{i=1}^N \tilde{f}_i - \frac{1}{n}\sum_{i=1}^n (f_i - Y_i)$ ▷ prediction-powered estimator

$\hat{\sigma}^2_{\tilde{f}} \leftarrow \frac{1}{N}\sum_{i=1}^N (\tilde{f}_i - \tilde{\theta}^f)^2$ ▷ empirical variance of imputed estimate

$\hat{\sigma}^2_{f-Y} \leftarrow \frac{1}{n}\sum_{i=1}^n (f_i - Y_i - \hat{\mathbf{\Delta}}^f)^2$ ▷ empirical variance of empirical rectifier

$w_\alpha \leftarrow z_{1-\alpha/2}\sqrt{\frac{\hat{\sigma}^2_{f-Y}}{n} + \frac{\hat{\sigma}^2_{\tilde{f}}}{N}}$ ▷ normal approximation

Output: prediction-powered confidence set $\mathcal{C}^{\text{PP}}_\alpha = \left(\hat{\theta}^{\text{PP}} \pm w_\alpha\right)$

Algorithm 2 Prediction-powered quantile estimation

Input: labeled data (X, Y), unlabeled features $\widetilde{X}$, predictor f, quantile $q \in (0,1)$, error level $\alpha \in (0,1)$

Construct fine grid Θ_{grid} between $\min_{i\in[N]} \tilde{f}_i$ and $\max_{i\in[N]} \tilde{f}_i$

for $\theta \in \Theta_{\text{grid}}$ **do**

$\quad \hat{\mathbf{\Delta}}^f(\theta) \leftarrow \frac{1}{n}\sum_{i=1}^n (\mathbb{1}\{Y_i \le \theta\} - \mathbb{1}\{f_i \le \theta\})$ ▷ empirical rectifier

$\quad \hat{F}(\theta) \leftarrow \frac{1}{N}\sum_{i=1}^N \mathbb{1}\left\{\tilde{f}_i \le \theta\right\}$ ▷ imputed CDF

$\quad \hat{\sigma}^2_\Delta(\theta) \leftarrow \frac{1}{n}\sum_{i=1}^n \left(\mathbb{1}\{Y_i \le \theta\} - \mathbb{1}\{f_i \le \theta\} - \hat{\mathbf{\Delta}}^f(\theta)\right)^2$ ▷ empirical variance of empirical rectifier

$\quad \hat{\sigma}^2_{\tilde{f}}(\theta) \leftarrow \frac{1}{N}\sum_{i=1}^N \left(\mathbb{1}\left\{\tilde{f}_i \le \theta\right\} - \hat{F}(\theta)\right)^2$ ▷ empirical variance of imputed CDF

$\quad w_\alpha(\theta) \leftarrow z_{1-\alpha/2}\sqrt{\frac{\hat{\sigma}^2_\Delta(\theta)}{n} + \frac{\hat{\sigma}^2_{\tilde{f}}(\theta)}{N}}$ ▷ normal approximation

Output: prediction-powered confidence set $\mathcal{C}^{\text{PP}}_\alpha = \left\{\theta \in \Theta_{\text{grid}} : |\hat{F}(\theta) + \hat{\mathbf{\Delta}}^f(\theta) - q| \le w_\alpha(\theta)\right\}$

Prediction-Powered Inference

Various algorithms(cont.)

Algorithm 3 Prediction-powered logistic regression

Input: labeled data (X, Y), unlabeled features $\widetilde{X}$, predictor f, error level $\alpha \in (0,1)$

1: Construct fine grid $\Theta_{\text{grid}} \subset \mathbb{R}^d$ of possible coefficients

2: $\hat{\boldsymbol{\Delta}}_j^f \leftarrow \frac{1}{n}\sum_{i=1}^n X_{i,j}(f_i - Y_i), \quad j \in [d]$ ▷ empirical rectifier

3: $\hat{\sigma}_{\Delta,j}^2 \leftarrow \frac{1}{n}\sum_{i=1}^n \left(X_{i,j}(f_i - Y_i) - \hat{\boldsymbol{\Delta}}_j^f\right)^2, \quad j \in [d]$ ▷ empirical variance of empirical rectifier

4: **for** $\theta \in \Theta_{\text{grid}}$ **do**

5: $\quad \hat{g}_j^f(\theta) \leftarrow \frac{1}{N}\sum_{i=1}^N \widetilde{X}_{i,j}\left(\mu_\theta(\widetilde{X}_i) - \tilde{f}_i\right), \quad j \in [d], \quad$ where $\mu_\theta(x) = \frac{1}{1+\exp(-x^\top \theta)}$ ▷ imputed gradient

6: $\quad \hat{\sigma}_{g,j}^2(\theta) \leftarrow \frac{1}{N}\sum_{i=1}^N \left(\widetilde{X}_{i,j}(\mu_\theta(\widetilde{X}_i) - \tilde{f}_i) - \hat{g}_j^f(\theta)\right)^2, \quad j \in [d]$ ▷ empirical variance of imputed gradient

7: $\quad w_{\alpha,j}(\theta) \leftarrow z_{1-\alpha/(2d)}\sqrt{\frac{\hat{\sigma}_{\Delta,j}^2}{n} + \frac{\hat{\sigma}_{g,j}^2(\theta)}{N}}, \quad j \in [d]$ ▷ normal approximation

Output: prediction-powered confidence set $\mathcal{C}_\alpha^{\text{PP}} = \left\{\theta \in \Theta_{\text{grid}} : |\hat{g}_j^f(\theta) + \hat{\boldsymbol{\Delta}}_j^f| \leq w_{\alpha,j}(\theta), \forall j \in [d]\right\}$

Algorithm 4 Prediction-powered linear regression

Input: labeled data (X, Y), unlabeled features $\widetilde{X}$, predictor f, coefficient $j^* \in [d]$, error level $\alpha \in (0,1)$

1: $\hat{\theta}^{\text{PP}} \leftarrow \tilde{\theta}^f - \hat{\boldsymbol{\Delta}}^f := \widetilde{X}^\dagger \tilde{f} - X^\dagger(f - Y)$ ▷ prediction-powered estimator

2: $\tilde{\Sigma} \leftarrow \frac{1}{N}\widetilde{X}^\top \widetilde{X}$, $\tilde{M} \leftarrow \frac{1}{N}\sum_{i=1}^N (\tilde{f}_i - \widetilde{X}_i^\top \tilde{\theta}^f)^2 \widetilde{X}_i \widetilde{X}_i^\top$

3: $\tilde{V} \leftarrow \tilde{\Sigma}^{-1}\tilde{M}\tilde{\Sigma}^{-1}$ ▷ "sandwich" variance estimator for imputed estimate

4: $\Sigma \leftarrow \frac{1}{n}X^\top X$, $M \leftarrow \frac{1}{n}\sum_{i=1}^n (f_i - Y_i - X_i^\top \hat{\boldsymbol{\Delta}}^f)^2 X_i X_i^\top$

5: $V \leftarrow \Sigma^{-1} M \Sigma^{-1}$ ▷ "sandwich" variance estimator for empirical rectifier

6: $w_\alpha \leftarrow z_{1-\alpha/2}\sqrt{\frac{V_{j^*j^*}}{n} + \frac{\tilde{V}_{j^*j^*}}{N}}$ ▷ normal approximation

Output: prediction-powered confidence set $\mathcal{C}_\alpha^{\text{PP}} = \left(\hat{\theta}_{j^*}^{\text{PP}} \pm w_\alpha\right)$

Prediction-Powered Inference

Various algorithms(cont.)

- mean: $\theta^* = \underset{\theta\in\mathbb{R}}{\arg\min}\mathbb{E}\left[\ell_\theta\left(Y_1\right)\right] = \underset{\theta\in\mathbb{R}}{\arg\min}\mathbb{E}\left[\frac{1}{2}\left(Y_1-\theta\right)^2\right]$
- quantile: $\theta^* = \underset{\theta\in\mathbb{R}}{\arg\min}\mathbb{E}\left[\ell_\theta\left(Y_1\right)\right] =$
  $\underset{\theta\in\mathbb{R}}{\arg\min}\mathbb{E}\left[q\left(Y_1-\theta\right)1\{Y_1>\theta\}+(1-q)\left(\theta-Y_1\right)1\{Y_1\leq\theta\}\right]$
- logistic regression:
  $\theta^* = \underset{\theta\in\mathbb{R}^d}{\arg\min}\mathbb{E}\left[\ell_\theta\left(X_1,Y_1\right)\right] = \underset{\theta\in\mathbb{R}^d}{\arg\min}\mathbb{E}\left[-Y_1\theta^\top X+\log\left(1+\exp\left(\theta^\top X_1\right)\right)\right]$
- linear regression: $\theta^* = \underset{\theta\in\mathbb{R}^d}{\arg\min}\mathbb{E}\left[\ell_\theta\left(X_1,Y_1\right)\right] = \underset{\theta\in\mathbb{R}^d}{\arg\min}\mathbb{E}\left[\left(Y_1-X_1^\top\theta\right)^2\right]$

Then powered confidence set in algorithms have valid coverage:

$$\liminf_{n,N\to\infty} P\left(\theta^*\in\mathcal{C}_\alpha^{\mathrm{PP}}\right)\geq 1-\alpha$$

.

Prediction-Powered Inference

Beyond convex estimation

1. estimate $\mathbb{E}\left[\ell_{\theta^*}\left(X_1, Y_1\right)\right]$ by approximating θ^* with an imputed estimate on the first $N/2$ unlabeled data points

$$\tilde{\theta}^f=\underset{\theta \in \Theta}{\arg \min } \frac{2}{N} \sum_{i=1}^{N/2} \ell_\theta\left(\tilde{X}_i, \tilde{f}_i\right), \quad \tilde{L}^f(\theta):=\frac{2}{N} \sum_{i=N/2+1}^{N} \ell_\theta\left(\tilde{X}_i, \tilde{f}_i\right).$$

2. construct $\left(\mathcal{R}_{\delta/2}^l(\theta), \mathcal{R}_{\delta/2}^u(\theta)\right)$ and $\mathcal{T}_{\alpha-\delta}(\theta)$ such that

$$P\left(\Delta^f(\theta) \leq \mathcal{R}_{\delta/2}^u(\theta)\right) \geq 1-\delta/2; \quad P\left(\Delta^f(\theta) \geq \mathcal{R}_{\delta/2}^l(\theta)\right) \geq 1-\delta/2;$$
$$P\left(\tilde{L}^f(\theta)-\mathbb{E}\left[\ell_\theta\left(X_1, f_1\right)\right] \leq \mathcal{T}_{\alpha-\delta}(\theta)\right) \geq 1-(\alpha-\delta).$$

Prediction-Powered Inference

Beyond convex estimation

3. combining 1 2, obtain

$$\mathcal{C}_\alpha^{\mathrm{PP}} = \left\{ \theta \in \Theta : \tilde{L}^f(\theta) \leq \tilde{L}^f\left(\tilde{\theta}^f\right) - \mathcal{R}^l_{\delta/2}(\theta) + \mathcal{R}^u_{\delta/2}\left(\tilde{\theta}^f\right) + \mathcal{T}_{\alpha-\delta}(\theta) \right\}$$

such that

$$P\left(\theta^* \in \mathcal{C}_\alpha^{\mathrm{PP}}\right) \geq 1 - \alpha.$$

Prediction-Powered Inference

Experiments

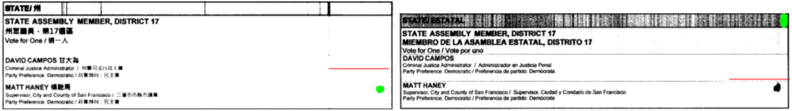


Figure 3: **Examples of ballots** correctly and incorrectly classified. The raw ballot is black and white, the voter's marking is automatically identified by a computer vision algorithm with a green annotation, and markings below the red line annotation will be considered votes for Matt Haney (and vice versa). The instructional portion of the ballots was cropped out.

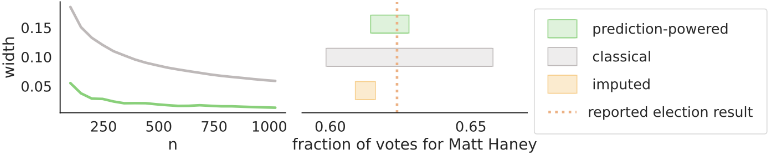


Figure 4: **Election results** produced by prediction-powered inference and the classical and imputed baselines at level 95%. Left: width of intervals as a function of n. Right: confidence intervals with $n = 1024$.

Prediction-Powered Inference

Experiments

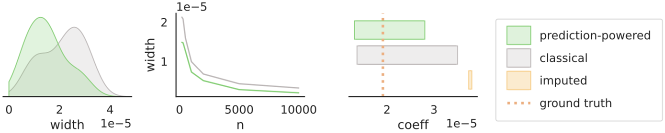


Figure 9: **Confidence intervals for the logistic regression coefficient** relating income and private health insurance coverage at the 95% level. Left: distribution of interval widths with $n = 200$. Middle: mean width as a function of n. Right: intervals with $n = 200$.

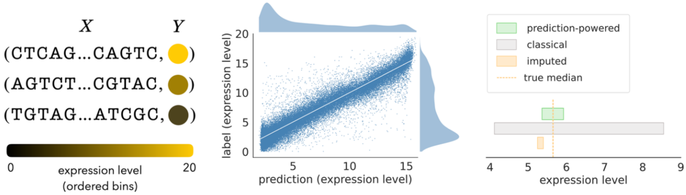


Figure 10: **Predicting gene expression levels from a promoter sequence** [34]. Left: each data point consists of a promoter sequence, X_i, and an expression level, Y_i. Middle: predictive performance of the transformer model on the native yeast promoters used in our experiments (RMSE 2.18, Pearson 0.963, Spearman 0.946). Right: confidence intervals for the median native yeast promoter expression level with $n = 75$ and $\alpha = 0.1$.

Further Studies

- High-dimension setting: what if $p \to \infty$?
- Using sample splitting?
- General framework for other estimators such as kernel mean embedding, MMD and other U-statistics
- Applying such for estimating mean difference or other test statistics
- Computing statistical validity for other semi-/self-supervised learning models

$\vdots$

Reference

- Kim, I. (2023). Selective Topics in Mathematical Statistics. 33-35.
- Zhang, A., Brown, L. D., Cai, T. T. (2019). Semi-supervised inference: General theory and estimation of means.
- Zhang, Y., Bradic, J. (2022). High-dimensional semi-supervised learning: in search of optimal inference of the mean. Biometrika, 109(2), 387-403.
- Tony Cai, T., Guo, Z. (2020). Semisupervised inference for explained variance in high dimensional linear regression and its applications. Journal of the Royal Statistical Society Series B: Statistical Methodology, 82(2), 391-419.
- Angelopoulos, A. N., Bates, S., Fannjiang, C., Jordan, M. I., Zrnic, T. (2023). Prediction-powered inference. arXiv preprint arXiv:2301.09633.
- Zhu, B., Ding, M., Jacobson, P., Wu, M., Zhan, W., Jordan, M., Jiao, J. (2023). Doubly Robust Self-Training. arXiv preprint arXiv:2306.00265.