Comparing the Value of Labeled and Unlabeled Data in Method-of-Moments Latent Variable Estimation

Mayee Chen*, Ben Cohen-Wang*, Steve Mussmann, Fred Sala, Chris Ré

Problem Setup

Training data:

Small amounts of **labeled** data $\in \mathcal{X} \times \mathcal{Y}$ (expensive, time-consuming)

vs

Large amounts of **unlabeled** data $\in \mathcal{X}$ (cheap, easily available)

?

+ weak sources (e.g. labeling functions) → Weakly labeled data (Ratner et. al., 2016)

Q: What are the tradeoffs of using labeled vs unlabeled data?

Our approach: theoretically analyze error of latent variable graphical model with labeled vs unlabeled input.

- Focus on the impact of **model misspecification** and how to reduce its effects in method-of-moments estimation.

Model

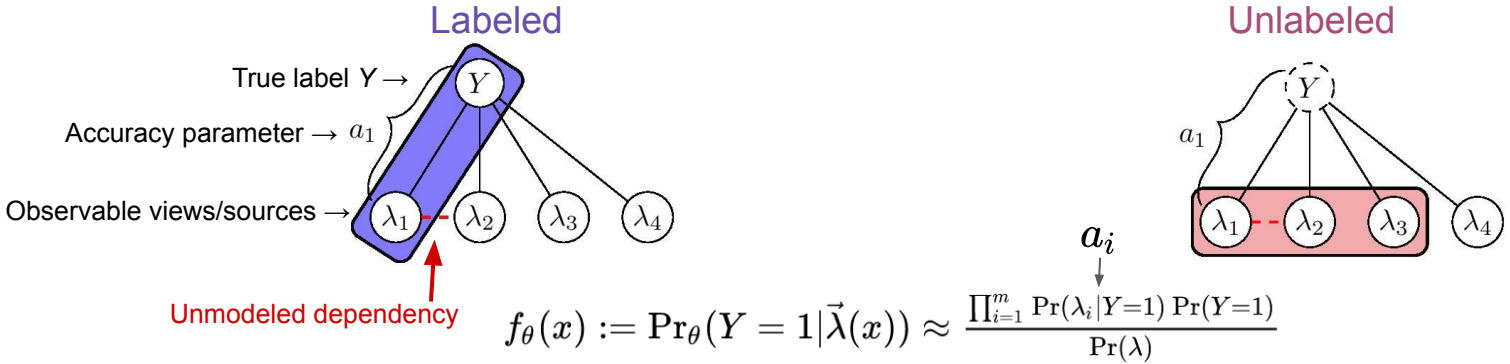
n_L labeled points $\in \mathcal{X} \times \mathcal{Y}$ and/or
n_U unlabeled points $\in \mathcal{X}$

+ *m* observable weak sources per point

$$\vec{\lambda} = \lambda_1, \ldots, \lambda_m$$

+ *dependency graph* ← misspecified!

$$G = \{(Y, \vec{\lambda}), E\}$$



$$f_\theta(x) := \Pr_\theta(Y = 1 | \vec{\lambda}(x)) \approx \frac{\prod_{i=1}^{m} \Pr(\lambda_i | Y=1) \Pr(Y=1)}{\Pr(\lambda)}$$

Labeled: directly estimate a_i

Unlabeled: use method-of-moments (Fu et. al., 2020) - relies on conditional independence of triples of sources

Model misspecification: *d* unmodeled dependencies among *m* sources

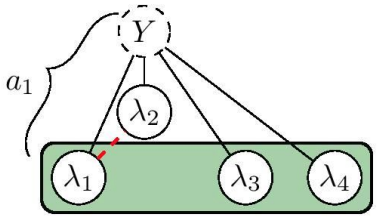
Results

1. Error Decomposition for $f_\theta(x)$

$\mathcal{L}_{CE}$ = Irreducible error + other sampling noise + inference bias + parameter estimation error

For labeled data: goes to 0
For unlabeled data: $\mathcal{O}(d/m)$ asymptotic bias!

2. Correcting misspecification for unlabeled data:



$\lambda_1, \lambda_2, \lambda_3 \to 0.73$ inconsistent
$\lambda_1, \lambda_3, \lambda_4 \to 0.78 = \hat{a}_1$
$\lambda_1, \lambda_2, \lambda_4 \to 0.81$ inconsistent

Select **median** accuracy parameter

- Median correction yields consistent estimates of a_i:

  Removes $\mathcal{O}(d/m)$ asymptotic bias and improves value of unlabeled data.

- True for other method of moments estimators (Chaganty and Liang, 2014; Anandkumar et. al., 2012)

Thank you!

Check out our paper for more details on:

- Theoretical framework for choosing between and combining labeled and unlabeled data
- Empirical results from application to weak supervision:
  - Verify our error decomposition and median correction approach
  - A little bit of labeled data (1%) combined with unlabeled data gives us performance close to a fully labeled dataset!

Paper: https://arxiv.org/abs/2103.02761

Contact: Mayee Chen, mfchen@stanford.edu

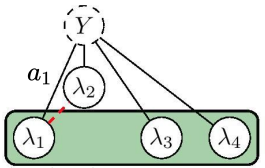
Results

Thm 1: Error Decomposition of Classifier $f_\theta(x)$

$\mathcal{L}_{CE}$ = Irreducible error + other sampling noise + inference bias + parameter estimation error

- Parameter estimation error goes to 0 (labeled) vs $\mathcal{O}(d/m)$ (unlabeled)
  - Labeled data generally better to use when model misspecification is unaddressed.

Correcting misspecification:



$\lambda_1, \lambda_2, \lambda_3 \to 0.73$ inconsistent
$\lambda_1, \lambda_3, \lambda_4 \to 0.78 = \hat{a}_1$
$\lambda_1, \lambda_2, \lambda_4 \to 0.81$ inconsistent

Select **median** accuracy parameter

- **Prop 1:** when enough unlabeled data, median correction yields consistent estimates of a_i and removes $\mathcal{O}(d/m)$ asymptotic bias → unlabeled data more valuable/useful now
- True for general method of moments estimators that exploit conditional independence (Chaganty and Liang, 2014; Anandkumar et. al., 2012)

⇒ *Theoretical framework* for choosing between and combining labeled and unlabeled data

- Application: weak supervision on binary sentiment classification (Maas et. al. 2011): median correction increases F1 score on unlabeled data by 3.31. Combining this with 1% labeled data comes within 0.75 points of F1 score if all samples were labeled.