Domain Adaptation under Missingness Shift

Helen Zhou, Sivaraman Balakrishnan, Zachary C. Lipton

Machine Learning Department, Carnegie Mellon University

Machine Learning Department

Carnegie Mellon University

School of Computer Science

The Problem

Motivation

Q: What % of adults had ≥1 dose of COVID-19 vaccine?

As of October 2021, in southwestern Pennsylvania,

40.5% — major regional healthcare provider

75.7–90.3% — Data.CDC.gov (in provider locations)

Suppose the provider adopts a new intake form…

Name: _____ Age: _____ DOB: _____ …	→	Name: _____ Age: _____ DOB: _____ …	Are you vaccinated for COVID-19? (required) ❑ Yes ❑ No

Absent any actual shift in patient status, the *distribution of observed data would shift*, due to clerical changes.

Setup

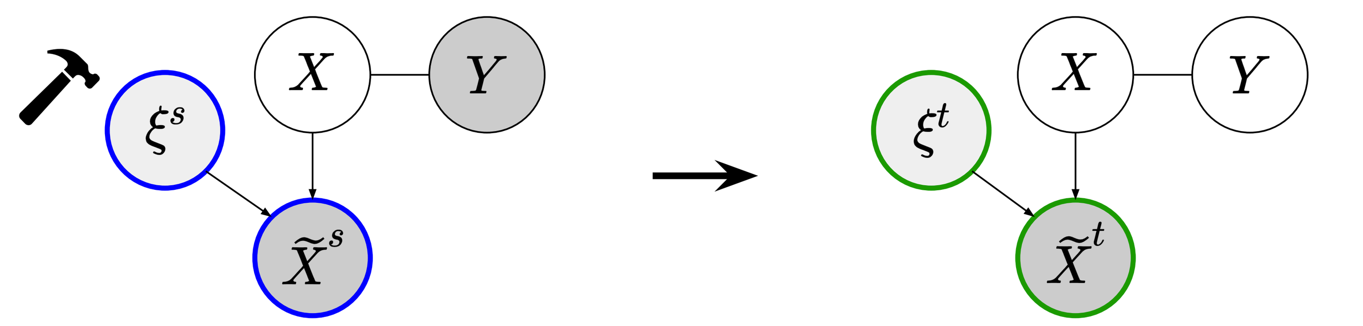
In any env. e with missing data, we do not observe *clean covariates* $X \in \mathbb{R}^d$ but instead observe *corrupted covariates*:

$$\widetilde{X} = X \odot \xi$$

where $\xi \in \{0,1\}^d$ and $(X, Y, \xi) \sim P^e$.

Assume $P(X,Y) = P^s(X,Y) = P^t(X,Y)$.

Missingness shift occurs when $P^s(\xi \mid \cdot) \neq P^t(\xi \mid \cdot)$.



source domain s — target domain t

UCAR: $\xi^s_j \sim \text{Bernoulli}(1 - m^s_j)$, $m^s \in [0,1]^d$ — $\xi^t_j \sim \text{Bernoulli}(1 - m^t_j)$, $m^t \in [0,1]^d$

Domain Adaptation under Missingness Shift (DAMS) goal: learn an optimal predictor on the corrupted target data.

Theoretical Results

The Cost of Non-Adaptivity

- The optimal source predictor can perform **arbitrarily worse** than simply guessing the label mean
- If missingness indicators are **observed** and depend on observed covariates, missingness shift can satisfy the covariate shift assumption:
  $$P^s\left(Y \mid \widetilde{X}' = \tilde{x}'\right) = P^t\left(Y \mid \widetilde{X}' = \tilde{x}'\right),$$
  where $\widetilde{X}' = \left(\widetilde{X}, \xi\right)$.
- For linear models, UCAR → form of L2 **regularization**.

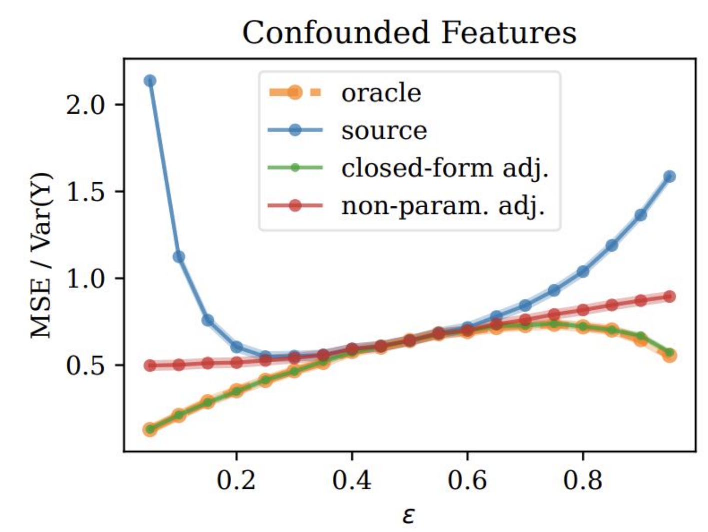
Identification

- Let $b \in \mathbb{R}^d$ be m-reachable from $a \in \mathbb{R}^d$ ($a \rightsquigarrow b$) if there exists some mask ξ such that $b = a \odot \xi$.
- Given missingness rates m, for any $x \in \mathbb{R}^d$ and $y \in \mathbb{R}$, we can identify the corrupted distribution from the clean one:
  $$\tilde{p}_{x,y} = \sum_{z: z \rightsquigarrow x,\, z \in \mathbb{R}^d} p_{z,y} \cdot \prod_{j=1}^{d} (1 - m_j)^{x_j} m_j^{z_j - x_j}$$
- Unfortunately, m is not in general identifiable.
- Instead, we can identify relative missingness rates:
  $$\frac{P^t(\tilde{x} \neq 0)}{P^s(\tilde{x} \neq 0)} = \frac{(1-m^t) \odot q}{(1-m^s) \odot q} = \frac{1-m^t}{1-m^s} \stackrel{\Delta}{=} 1 - r^{s \to t},$$
- ..and thus the labeled target distribution from labeled source.

Estimation

- To estimate relative missingness, compute $\hat{q}^s_j = \dfrac{\text{count}(\tilde{x}^s_j \neq 0)}{n_s}$, $\hat{q}^t_j = \dfrac{\text{count}(\tilde{x}^t_j \neq 0)}{n_t}$, $\hat{r}^{s \to t} = 1 - \dfrac{\hat{q}^t}{\hat{q}^s}$.
- w.p. $1 - \delta$, $\left|\hat{r}^{s \to t} - r^{s \to t}\right| \leq \dfrac{1}{\hat{q}^s}\left(\sqrt{\dfrac{\log(4/\delta)}{2n_t}} + (1 - r^{s \to t})\sqrt{\dfrac{\log(4/\delta)}{2n_s}}\right)$.
- Linear estimator: $\beta^*_t = \mathbb{E}[\widetilde{X}^{t\top}\widetilde{X}^t]^{-1}\left(r^{s \to t} \odot \mathbb{E}[\widetilde{X}^{s\top} Y^s]\right)$.
- Non-parametric adjustment: compute relative missingness $\tilde{r}^{s \to t} = \max\left(\hat{r}^{s \to t}, 0\right)$ and apply it to source.

Empirical Results



$$u_{x_2} \sim \mathcal{N}(0,1) \qquad m^s = [1 - \epsilon, \epsilon]$$
$$u_y \sim \mathcal{N}(0,1) \qquad m^t = [\epsilon, 1 - \epsilon]$$
$$X_1 \sim \text{Bernoulli}(0.5)$$
$$X_2 = \text{expit}(2X_1 + u_{x_2})$$
$$Y = X_1 - X_2 + u_y$$

Synthetic & semi-synthetic UCI MSE/Var(Y)

	Rednd.	Confnd.	Adult		Bank		Thyroid	
	m^s ? m^t	m^s ? m^t	$m^s \preceq m^t$	m^s ? m^t	$m^s \preceq m^t$	m^s ? m^t	$m^s \preceq m^t$	m^s ? m^t
Linear Regression Models								
Oracle	0.178	0.206	0.420	0.362	0.338	0.433	0.298	0.251
Source	1.259	1.103	0.437	0.380	0.371	0.480	0.350	0.320
Imputed	1.002	0.918	0.490	0.483	0.501	0.592	0.306	0.358
Closed-form	**0.186**	**0.209**	0.422	**0.363**	0.339	**0.442**	0.316	**0.291**
Non-param.	0.473	0.492	**0.420**	0.373	**0.338**	0.459	**0.293**	**0.291**
XGBoost Models								
Oracle	0.166	0.200	0.398	0.354	0.287	0.453	0.316	0.274
Source	**0.166**	0.475	**0.399**	**0.379**	0.305	**0.500**	**0.310**	**0.352**
Imputed	1.002	1.157	0.512	0.521	0.492	0.708	0.355	0.441
Non-param.	0.425	**0.473**	**0.399**	0.392	**0.287**	0.503	**0.310**	0.381
MLP Models								
Oracle	0.166	0.201	0.389	0.343	0.295	0.458	0.279	0.230
Source	**0.184**	**0.321**	0.399	0.357	0.322	0.499	0.320	0.303
Imputed	1.003	0.924	0.480	0.468	0.484	0.668	0.304	0.345
Non-param.	0.436	0.470	**0.389**	**0.355**	**0.294**	**0.487**	**0.278**	**0.272**

Conclusion

- Given missing data indicators, miss. shift can → covariate shift
- Provide identification & estimation results in DAMS with UCAR
- Next: missingness indicators to depend on covariates, or each other; real data expmt.