Fair Inference Through Semiparametric-Efficient Estimation Over Constraint-Specific Paths

for *New Developments in Nonparametric and Semiparametric Statistics*,
Joint Statistical Meetings; Vancouver, BC, Canada; 02 Aug. 2018

Nima Hejazi

Group in Biostatistics, and
Center for Computational Biology,
University of California, Berkeley

nimahejazi.org
Twitter: nshejazi
GitHub: nhejazi

slides: bit.ly/jsm_fairtmle_2018

Preview: Summary

- Recent work suggests that the widespread use of machine learning algorithms has had negative social and policy consequences.
- The widespread use of machine learning in policy issues violates human intuitions of bias.
- We propose a general algorithm for constructing “fair” optimal ensemble ML estimators via cross-validation.
- Constraints may be imposed as functionals defined over the target parameter of interest.
- Estimating constrained parameters may be seen as iteratively minimizing a loss function along a *constrained* path in the parameter space Ψ.

What’s fair if machines aren’t?

Bernard Parker, left, was rated high risk; Dylan Fugett was rated low risk. (Josh Ritchie for ProPublica)

Machine Bias

There’s software used across the country to predict future criminals. And it’s biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica

Fairness is machine learning?

```
Another potential result: a more diverse
workplace. The software relies on data
to surface candidates from a wide variety
of places...free of human biases. But
software is not free of human influence.
Algorithms are written and maintained by
people...As a result...algorithms can
reinforce human prejudices.

-Miller (2015)
```

Addressing bias in a technical manner

- The careless use of machine learning may induce *unjustified* bias.
- Problematic discrimination by ML approaches leads to solutions with *practical irrelevance*.
- Ill-considered discrimination by ML approaches leads to solutions that are *morally problematic*.
- Two doctrines of discrimination:
    1. Disparate treatment: formal or intentional
    2. Disparate impact: unjustified or avoidable

Background, data, notation

- An observational unit: $O = (W, X, Y)$, where W is baseline covariates, X a sensitive characteristic, Y an outcome of interest.
- Consider n i.i.d. copies $O_1, \ldots, O_n$ of $O \sim P_0 \in \mathcal{M}$.
- Here, $\mathcal{M}$ is an infinite-dimensional statistical model (i.e., indexed by an infinite-dimensional vector).
- We discuss the estimation of a target parameter $\psi : \mathcal{M} \to \mathbb{R}$, where

$$\Psi(P_0) = \arg\min_{\psi \in \Psi} \mathbb{E}_{P_0} L(\psi)$$

Just a few fairness criteria

- Let $C : (X, W) \to Y \in \{0, 1\}$ be a classifier; $X \in \{a, b\}$.
- Demographic parity: $\mathbb{P}_{(X=a)}(C = 1) = \mathbb{P}_{(X=b)}(C = 1)$
- Accuracy parity: $\mathbb{P}_{(X=a)}(C = Y) = \mathbb{P}_{(X=b)}(C = Y)$
- True positive parity:
  $\mathbb{P}_{(X=a)}(C = 1 \mid Y = 1) = \mathbb{P}_{(X=b)}(C = 1 \mid Y = 1)$
- False positive parity:
  $\mathbb{P}_{(X=a)}(C = 1 \mid Y = 0) = \mathbb{P}_{(X=b)}(C = 1 \mid Y = 0)$
- Positive predictive value parity:
  $\mathbb{P}_{(X=a)}(Y = 1 \mid C = 1) = \mathbb{P}_{(X=b)}(Y = 1 \mid C = 1)$
- Negative predictive value parity:
  $\mathbb{P}_{(X=a)}(Y = 1 \mid C = 0) = \mathbb{P}_{(X=b)}(Y = 1 \mid C = 0)$

Wait, where did the fairness go?

- Goal: estimate $\Psi(P_0) = \mathbb{E}_{P_0}(Y \mid X, W)$.
- Let $Y \in \{0, 1\}$ and use negative log-likelihood loss:

  $$L(\psi) = -(Y \log(\mathbb{P}(Y \mid X, W)) + (1-Y) \log(1-\mathbb{P}(Y \mid X, W)))$$

- Fairness criterion — *equalized odds*:

  $$\Theta_\psi(P_0) = \sum_y \{\mathbb{E}_{P_0}(L(\psi)(O) \mid X = 1, Y = y) \\ - \mathbb{E}_{P_0}(L(\psi)(O) \mid X = 0, Y = y)\}^2$$

- Let $\Theta_\psi(P_0) : \mathcal{M} \to \mathbb{R}$ be a pathwise differentiable *functional* for each $\psi \in \Psi$.

Constrained functional parameters

- Estimate target parameter under a constraint:

$$\Psi(P_0) = \underset{\psi \in \Psi, \Theta_\psi(P_0)=0}{\arg\min} \mathbb{E}_{P_0} L(\psi)$$

- Goal: estimate $\Psi^*(P_0)$, the projection of $\Psi(P_0)$ onto the subspace $\Psi^*(P_0) = \{\psi \in \Psi : \Theta_\psi(P_0) = 0\}$:

$$(\Psi^*, \lambda) = (\Psi^*(P_0), \Lambda(P_0)) \equiv \underset{\psi \in \Psi, \lambda}{\arg\min} \mathbb{E}_{P_0} L(\psi) + \lambda \Theta_\psi(P_0).$$

- *Lemma*: If $\widetilde{\Psi}(P_0) = (\Psi^*(P_0), \Lambda(P_0))$ is the minimizer of the Lagrange multiplier penalized loss, then

$$\Psi^*(P_0) = \underset{\psi \in \Psi, \Theta_\psi(P_0)=0}{\arg\min} \mathbb{E}_{P_0} L(\psi).$$

Learning with constrained parameters

- Risk function: $R(\widetilde{\psi} \mid P) \equiv P_n L(\psi^*) + \lambda \Theta(\psi^* \mid P)$, where $\widetilde{\psi} = (\psi^*, \lambda)$
- For $\widetilde{\psi}(P_n) = (\hat{\Psi}^*(P_n), \hat{\lambda}(P_n))$ of $\widetilde{\Psi}(P_0)$, and sample splitting scheme $B_n \in \{0,1\}^n$:

$$R_0(\widetilde{\psi}, P_n) = \mathbb{E}_{B_n} P_0 L(\hat{\Psi}^*(P^0_{n,B_n})) + \hat{\Lambda}(P_n) \mathbb{E}_{B_n} \Theta(\hat{\Psi}^*(P^0_{n,B_n}) \mid P_0)$$

Learning with constrained parameters

- Cross-validated risk:

$$R_{n,CV}(\tilde{\psi}, P_n) = E_{B_n} P^1_{n,B_n} L(\hat{\Psi}^*(P^0_{n,B_n})) \tag{1}$$

$$+ \hat{\Lambda}(P_n) E_{B_n} \Theta(\hat{\Psi}^*(P^0_{n,B_n}) \mid P^*_{n,B_n}) \tag{2}$$

- Given candidate estimators $\widetilde{\psi}_j(P_n) = (\hat{\Psi}^*_j(P_n), \hat{\Lambda}_j(P_n))$, $j = 1, \ldots, J$, the CV selector is given by: $J_n = \arg\min_j R_{n,CV}(\widetilde{\psi}_j, P_n)$.

- We may define an optimal estimate of $\widetilde{\Psi}$ by $\widetilde{\psi}_n \equiv \widetilde{\psi}_{J_n}(P_n) = (\hat{\Psi}_{J_n}(P_n), \hat{\lambda}_{J_n}(P_n))$

Mappings with constrained learners

A straightforward approach to generating estimators of the constrained parameter would be to simply generate a mapping according to the following simple process:

1. Generate an unconstrained estimate ψ_n of the unconstrained parameter ψ_0,
2. Map an estimator $\Theta_{\psi_n,n}$ of the constraint $\Theta_{\psi_n}(P_0)$ into the path $\psi_{n,\lambda}$. The corresponding solution $\psi_n^* = \psi_{n,\lambda_n}$ of $\Theta_{\psi_{n,\lambda_n},n} = 0$ generates an estimator of the constrained parameter.

Constraint-specific paths

- Consider $\psi_{0,\lambda} = \arg\max_{\psi \in \Psi} \mathbb{E}_{P_0} L(\psi) + \lambda \Theta_0(\psi)$.
- $\{\psi_{0,\lambda} : \lambda\}$ represents a path in the parameter space Ψ through ψ_0 at $\lambda = 0$.
- This is a *constraint-specific path*, as it produces an estimate under the desired functional constraint.
- Leverage this construction to map an initial estimator of the unconstrained parameter ψ_0 into its corresponding constrained version ψ_0^*.

Future work

- Further generalization of constraint-specific paths: the solution path $\{\psi_{0,\lambda} : \lambda\}$ in the parameter space Ψ through ψ_0 at $\lambda = 0$.
- Further develop relation between constraint-specific paths and universal least favorable submodels.
- Integration of the approach of constraint-specific paths with classical classical targeted maximum likelihood estimation — in particular, what, if any, are the implications for inference?

Review: Summary

- Recent work suggests that the widespread use of machine learning algorithms has had negative social and policy consequences.
- The widespread use of machine learning in policy issues violates human intuitions of bias.
- We propose a general algorithm for constructing "fair" optimal ensemble ML estimators via cross-validation.
- Constraints may be imposed as functionals defined over the target parameter of interest.
- Estimating constrained parameters may be seen as iteratively minimizing a loss function along a *constrained* path in the parameter space Ψ.

References I

Bang, H. and Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973.

Breiman, L. (1996). Stacked regressions. *Machine Learning*, 24(1):49–64.

DiCiccio, T. J. and Romano, J. P. (1990). Nonparametric confidence limits by resampling methods and least favorable families. *International Statistical Review/Revue Internationale de Statistique*, pages 59–76.

Duchi, J., Glynn, P., and Namkoong, H. (2016). Statistics of robust optimization: A generalized empirical likelihood approach. *ArXiv e-prints*.

References II

Dudoit, S. and van der Laan, M. J. (2005). Asymptotics of cross-validated risk estimation in estimator selection and performance assessment. *Statistical Methodology*, 2(2):131–154.

Hardt, M., Price, E., Srebro, N., et al. (2016). Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, pages 3315–3323.

Severini, T. A. and Wong, W. H. (1992). Profile likelihood and conditionally parametric models. *The Annals of Statistics*, pages 1768–1802.

References III

Stein, C. (1956). Efficient nonparametric testing and estimation. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*. The Regents of the University of California.

Tsiatis, A. (2007). *Semiparametric Theory and Missing Data*. Springer Science & Business Media.

van der Laan, M. J. and Dudoit, S. (2003). Unified cross-validation methodology for selection among estimators and a general cross-validated adaptive epsilon-net estimator: Finite sample oracle inequalities and examples.

References IV

van der Laan, M. J., Dudoit, S., and Keles, S. (2004). Asymptotic optimality of likelihood-based cross-validation. *Statistical Applications in Genetics and Molecular Biology*, 3(1):1–23.

van der Laan, M. J., Polley, E. C., and Hubbard, A. E. (2007). Super Learner. *Statistical Applications in Genetics and Molecular Biology*, 6(1).

van der Laan, M. J. and Rubin, D. (2006). Targeted maximum likelihood learning. *The International Journal of Biostatistics*, 2(1).

Acknowledgments

Mark J. van der Laan — University of California, Berkeley
David C. Benkeser — Emory University

Funding source:
National Institutes of Health: T32-LM012417–02

Thank you.

Slides: bit.ly/jsm_fairtmle_2018

`nimahejazi.org`

`Twitter: nshejazi`

`GitHub: nhejazi`