Project 2 - Constrained IRL

Kevin Gunn

April 26th 2019

1 Constrained Optimization Techniques

Examine the situation of linear decision rules, $I(\mathbf{x}^\mathsf{T}\boldsymbol{\eta} > 0)$. Let

$$f(\boldsymbol{\eta}, M, \lambda) \equiv \; V_Y(\boldsymbol{\eta}) - M\left[V_Z(\boldsymbol{\eta}) - \lambda\right]_+ . \tag{1}$$

We maximize $f(\boldsymbol{\eta}, M, \lambda)$ with respect to $\boldsymbol{\eta}$ to obtain the optimal constrained treatment decision, i.e. $\boldsymbol{\eta}^{opt} = \arg\max_{\boldsymbol{\eta}\in\mathbb{R}^p} f(\boldsymbol{\eta}, M, \lambda)$. If we apply Q-learning to estimate $V_Y(\boldsymbol{\eta})$ and $V_Z(\boldsymbol{\eta})$, we can use a linear model for the Q-function such as, $Q(Y, a, \mathbf{x}) = \theta_{0,Y}^\mathsf{T}\mathbf{x} + a\left(\theta_{1,Y}^\mathsf{T}\mathbf{x}\right)$ and $Q(Z, a, \mathbf{x}) = \theta_{0,Z}^\mathsf{T}\mathbf{x} - a\left(\theta_{1,Z}^\mathsf{T}\mathbf{x}\right)$. In this setting,

$$V_Y(\boldsymbol{\eta}) = E\left[\theta_{0,Y}^\mathsf{T}\mathbf{x} + I(\mathbf{x}^\mathsf{T}\boldsymbol{\eta} > 0)\left(\theta_{1,Y}^\mathsf{T}\mathbf{x}\right)\right] \text{ and } V_Z(\boldsymbol{\eta}) = E\left[\theta_{0,Z}^\mathsf{T}\mathbf{x} - I(\mathbf{x}^\mathsf{T}\boldsymbol{\eta} > 0)\left(\theta_{1,Z}^\mathsf{T}\mathbf{x}\right)\right].$$

Solving a constrained optimization problem of the form in Equation (1) can be done with penalty function methods. These methods approximate a solution to the constrained optimization problem with an unconstrained one. Afterwards, apply standard search techniques to obtain solutions.

1.1 OCTDs through AL-IRL

Inverse reinforcement learning has been utilized before in the optimal treatment regime literature before such as Lizotte et al. (2012) and Luckett et al. (2017). Lizotte et al. (2012) do not recover the "true" reward function, but rather attempt to learn the optimal treatment regime for a set of reward functions, where the data do not come from an expert. Luckett et al. (2017) adapt inverse reinforcement learning into the optimal treatment decision framework to estimate the optimal treatment decision to maximize an utility function comprised of composite outcomes of interest. Linn et al. (2015) and Wang et al. (2018) discuss how to approach optimal constrained treatment decisions, but techniques to recover the optimal constrained treatment decision when observing an "expert" behaviours has not been previously discussed.

In the optimal treatment decision (OTD) setting, let us assume the value function the clinician optimizes is $V(\delta) = E\left\{Y(\delta)\right\} - M\left[E\left\{Z(\delta)\right\} - \lambda\right]_+$. Assume larger values are preferred for both outcomes that the clinician's M and λ are approximately optimal. Denote $\left[E\left\{Z(\delta)\right\} - \lambda\right]_+$

as $V_{Z+}(\delta, \lambda)$. If we approach the problem from an apprenticeship learning (Abbeel and Ng, 2004; Syed and Schapire, 2010, 2008) point of view, we will recover a weight, w, as a tradeoff between $V_Y(\delta)$ and $V_{Z+}(\delta, \lambda)$. The expected constrained clinical value function under a treatment decision function, δ, becomes,

$$wV_Y(\delta) - (1-w)V_{Z+}(\delta, \lambda) \Longleftrightarrow V_Y(\delta) - \frac{1-w}{w} V_{Z+}(\delta, \lambda) \Longleftrightarrow V_Y(\delta) - MV_{Z+}(\delta, \lambda), \tag{2}$$

where $w = \frac{1}{1+M}$ and $M \geq 0$. The empirical value function is,

$$V(\delta, \lambda) = \left\{ \frac{1}{n} \sum_{i=1}^{n} I(A_i = \delta(\mathbf{x}_i))\ Q(Y_i, A_i) \right\} - \frac{1-w}{w} \left\{ \frac{1}{n} \sum_{i=1}^{n} I(A_i = \delta(\mathbf{x}_i))\ Q(Z_i, A_i) - \lambda \right\}_+ . \tag{3}$$

The optimal treatment decision is determined through a constrained regression or constrained classification algorithm (Linn et al., 2015; Wang et al., 2018). Estimating the treatment decision rule with this algorithm recovers the decision rule the clinician used with respect to maximizing $V(\delta)$.

The weight vector in SVMs have support in $\mathbb{R}^p$, but we need to constrain the solution space to $w \in (0, 1)$. Therefore we need to modify the QP problem to satisfy these constraints. Let $\boldsymbol{\mu}_V(\delta, \lambda) = (V_Y(\delta), -V_{Z+}(\delta, \lambda))^\mathsf{T}$. The new QP problem is,

$$\begin{aligned} &\min_{\mathbf{w}, q} \ \ q \\ &\text{subject to} \ \ \mathbf{w}^\mathsf{T} \boldsymbol{\mu}_V(\delta^C, \lambda) - \mathbf{w}^\mathsf{T} \boldsymbol{\mu}_V(\widehat{\delta}, \lambda) \geq q; \\ &\text{where} \ \ \sum_{i=1}^{2} \mathbf{w}_i = 1;\ \mathbf{w}_i \geq 0\ \forall i \in \{1, 2\}\,. \end{aligned} \tag{4}$$

Algorithm 1: Estimation of δ^{opt} with risk constraint

Set a grid, $\lambda_1 < \lambda_2 < \cdots < \lambda_T$;
Denote the clinician's treatment decision rule as $\delta^{(0)}$;
for $t = 1, ..., T$ **do**
    Calculate empirical estimate for $\boldsymbol{\mu}_V(\delta^{(0)}, \lambda_t)$;
    Provide initial treatment decision policy, $\delta^{(1)}$;
    Set k = 1;
    while $q_k > \epsilon$ **do**
        Calculate empirical estimate for $\boldsymbol{\mu}_V(\delta^{(k)}, \lambda_t)$;
        Obtain $\mathbf{w}_k$ and q_k by solving Equation 4 ;
        If $|q_k| \leq \epsilon$, then **break**;
        Set $M_k = \frac{1-w_k}{w_k}$;
        Set k = k+1;
        Obtain $\delta^{(k)} = I(\mathbf{x}^\mathsf{T} \boldsymbol{\eta}_k > 0)$ by solving: $\max_\eta\ V_Y(\boldsymbol{\eta}) - M_{k-1}\left[V_Z(\boldsymbol{\eta}) - \lambda_t\right]_+$;
    end
    Set $\delta_t = \delta^{(k)}$;
end
Set $\delta^{opt} = \arg\max_{t \in \{1, ..., T\}} V(\delta_t, \lambda_t)$;

Note 1: Algorithm is adapted from Abbeel and Ng (2004).

Note 2: Classification step is the inverse reinforcement learning (IRL) step, as we are trying to learn the "weights" being optimized by the expert.

Note 3: The tolerance level ϵ allows us to decide on a decision rule with performance comparable to the clinician's decision rule with difference ϵ to allow for deviation from the clinicians' decision rule, $\delta^{(0)}$.

1.2 Simulation Settings

The data will be generated in the following way:

- $\mathbf{X} \sim \mathcal{N}(0, I_p)$
- $\epsilon \sim \mathcal{N}(0, 0.5)$
- $\mathbf{A} \equiv I(\mathbf{X}^\mathsf{T}\boldsymbol{\eta}^{opt} > 0)$

The following models will be tested with $p = 5$ and $n = 2000$ for 500 replications. The different models examined will be as follows.

- *Model 1 (Linear Model for Y)*: $\mathbf{Y} = \boldsymbol{\theta}_{0,Y}^\mathsf{T}\mathbf{X} + \mathbf{A}\left(\boldsymbol{\theta}_{1,Y}^\mathsf{T}\mathbf{X}\right) + \epsilon$
- *Model 2 (Linear Model for Z)*: $\mathbf{Z} = \boldsymbol{\theta}_{0,Z}^\mathsf{T}\mathbf{X} - \mathbf{A}\left(\boldsymbol{\theta}_{1,Z}^\mathsf{T}\mathbf{X}\right) + \epsilon$

The models will be tested with following true regression coefficients:

- $\boldsymbol{\theta}_{0,Y} = (1, 0, 1, 0, 1)^\mathsf{T}$
- $\boldsymbol{\theta}_{0,Z} = (1, 0.5, 0.5, 1, 1)^\mathsf{T}$
- $\boldsymbol{\theta}_{1,Y} = (0, 0, 1, 1, 0)^\mathsf{T}$
- $\boldsymbol{\theta}_{1,Z} = (0.25, 1, 1, 0, 1)^\mathsf{T}$

Assume clinician acts optimally, i.e. $\delta^{clinician} \equiv \delta^{opt}$, where $\delta^{opt} = I(\mathbf{x}^\mathsf{T}\boldsymbol{\eta}^{opt} > 0)$. We will use the Nelder-Mead algorithm for derivative-free optimization algorithm to obtain an estimator for $\boldsymbol{\eta}^{opt}$. The unconstrained optimal treatment decision for Y is $\delta^Y(\mathbf{x}) = I(\theta_{1,Y}^\mathsf{T}\mathbf{x} > 0)$, and the unconstrained optimal treatment decision for Z is $\delta^Z(\mathbf{x}) = I(\theta_{1,Z}^\mathsf{T}\mathbf{x} > 0)$.

Performance Measures:

- MSE of $\boldsymbol{\eta}^{opt}$ in $\delta^{opt}(\mathbf{x}) = I(\mathbf{x}^\mathsf{T}\boldsymbol{\eta}^{opt} > 0) \implies \frac{1}{500}\sum_{i=1}^{500}\|\widehat{\boldsymbol{\eta}}^{opt} - \boldsymbol{\eta}^{opt}\|^2$
- $\frac{1}{500}\sum_{i=1}^{500}(\widehat{M} - M^{opt})^2$
- PCD (SD) and Value function (SD).

1.3 Results

True values for $\boldsymbol{\eta}$ were generated from a Monte Carlo data set of size 100,000. The simulations were repeated for 500 repetitions. The two treatment decisions, δ^Y and δ^Z overlap for approximately 74% of patients in the Monte Carlo data set. A value for λ was chosen such that $E[Z(\delta^Z)] < \lambda < E[Z(\delta^Y)]$. We assume λ_{opt} is known and set at -0.75. Three values of M were chosen to examine, 0.1, 0.5 and 1. We set $\boldsymbol{\eta}_1 = (1, 0, 0, 1, 1)^\mathsf{T}$ for both settings.

	Clinician		Apprentice	
	True Value	-	Bias	SD
$\boldsymbol{\eta}_1$	1.33	-	0.03	0.14
$\boldsymbol{\eta}_2$	0.78	-	0.04	0.09
$\boldsymbol{\eta}_3$	0.78	-	0.04	0.09
$\boldsymbol{\eta}_4$	1.29	-	0.03	0.14
$\boldsymbol{\eta}_5$	1.42	-	0.06	0.16
	True Value	-	Estimate	SD
M	0.10	-	0.12	0.02
PCD	1.00	-	0.993	0.002
V_Y	0.75	0.06	0.75	0.06
V_{Z+}	-0.24	0.03	-0.23	0.03

Table 1: Parameters of Interest with $M_{opt} = 0.1$ and $\lambda_{opt} = -0.75$.

	Clinician		Apprentice	
	True Value	-	Bias	SD
$\boldsymbol{\eta}_1$	1.07	-	0.07	0.34
$\boldsymbol{\eta}_2$	1.15	-	0.09	0.36
$\boldsymbol{\eta}_3$	1.15	-	0.09	0.37
$\boldsymbol{\eta}_4$	0.88	-	0.06	0.28
$\boldsymbol{\eta}_5$	1.60	-	0.12	0.50
	True Value	-	Estimate	SD
M	0.80	-	0.82	0.04
PCD	1.00	-	0.995	0.003
V_Y	0.70	0.06	0.70	0.06
V_{Z+}	-0.13	0.03	-0.13	0.03

Table 2: Parameters of Interest with $M_{opt} = 0.8$ and $\lambda_{opt} = -0.75$.

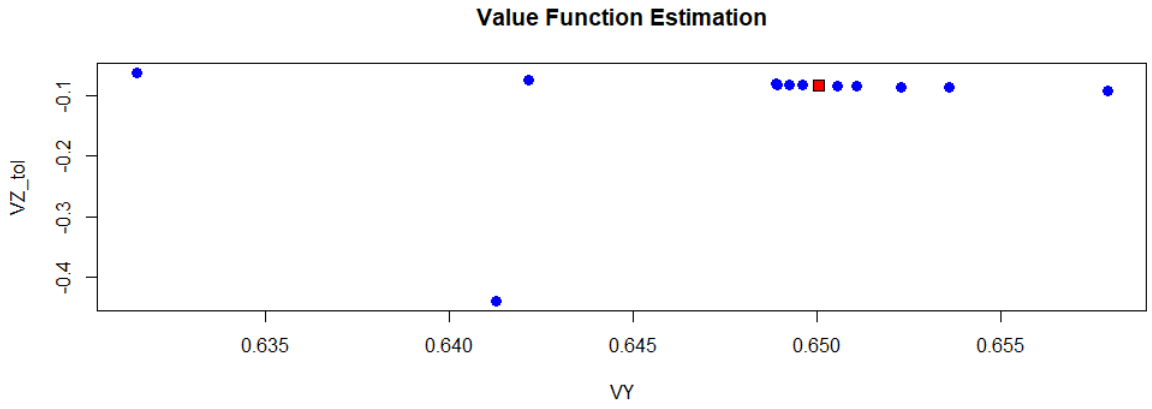

Figure 1: Example of Iterative IRL Algorithm for Estimating Value Functions with $M = 0.8$.

References

Abbeel, P. and Ng, A. Y. (2004). Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the twenty-first international conference on Machine learning*, page 1. ACM.

Linn, K. A., Laber, E. B., and Stefanski, L. A. (2015). Chapter 15: Estimation of dynamic treatment regimes for complex outcomes: Balancing benefits and risks. In *Adaptive Treatment Strategies in Practice: Planning Trials and Analyzing Data for Personalized Medicine*, pages 249–262. SIAM.

Lizotte, D. J., Bowling, M., and Murphy, S. A. (2012). Linear fitted-q iteration with multiple reward functions. *Journal of Machine Learning Research*, 13(Nov):3253–3295.

Luckett, D. J., Laber, E. B., and Kosorok, M. R. (2017). Estimation and optimization of composite outcomes. *arXiv preprint arXiv:1711.10581*.

Syed, U. and Schapire, R. E. (2008). A game-theoretic approach to apprenticeship learning. In *Advances in neural information processing systems*, pages 1449–1456.

Syed, U. and Schapire, R. E. (2010). A reduction from apprenticeship learning to classification. In *Advances in Neural Information Processing Systems*, pages 2253–2261.

Wang, Y., Fu, H., and Zeng, D. (2018). Learning optimal personalized treatment rules in consideration of benefit and risk: with an application to treating type 2 diabetes patients with insulin therapies. *Journal of the American Statistical Association*, 113(521):1–13.