動的治療計画と強化学習

江口　真透　　医療健康データ科学研究センター　教授

1. 研究の流れ

Logistic Regression, Cox (1958)

Perceptron, Rosenblatt (1958)

Dynamic programming, Bellman (1957)

Bellman equation (principle of optimality)

Q-learning, Watkins (1989)

Deep Learning, LeCun, Bengio, Hinton (2015)

Deep Q-learning, Silver, Hassabis,…

Dynamic treatment regime, Murphy (2003), Zhao, et al (2012)

2. ランダム臨床試験と動的治療計画

Random Clinical Trial

Treatment = Intervention

Random assignment for treatments

Dynamic treatment regimes

Treatment = Action

Adaptive biased coin, cf. Lavori-Dawson (2002)

Multiple Assignment Randomized Trial (**SMART**)

Classification-based approach, cf Zhao (2012, 2015)

Chakraborty & Moodie (2013). Statistical methods for dynamic treatment regimes, cf. Robins (1986); Murphy (2003).

3. 確率フレームワーク

- $$(\boldsymbol{X}, A, Y) \begin{cases} \text{state } \boldsymbol{X} \in \mathcal{X} \subseteq \mathbb{R}^p \\ \text{action } A \in \mathcal{A} = \{1, \cdots, M\} \\ \text{reward } Y \in \mathcal{Y} = \{y \in \mathbb{R} : y \geq 0\} \end{cases}$$

$$(\boldsymbol{X}, A, Y) \sim p(\boldsymbol{x}, a, y) = p(y|\boldsymbol{x}, a)p(a|\boldsymbol{x})p(\boldsymbol{x})$$

- Deterministic policy $d : \mathcal{X} \to \mathcal{A}$ has a value function

$$V_d = \mathbb{E}_d[Y] = \mathbb{E}\left[\frac{\mathbb{I}(d(\boldsymbol{X}) = A)}{p(A|\boldsymbol{X})} Y\right]$$

- Optimal policy $d^{\rm opt} = \underset{d \in \mathcal{D}}{\text{argmax}}\ V_d$ where $\mathcal{D}$ is the space of all policy functions

Cf. Supervised learning: A feature vector $\boldsymbol{X}$ predicts an outcome Y

$$p(\boldsymbol{x}, y) = p(y|\boldsymbol{x})p(\boldsymbol{x})$$

4. Q-関数

- Q-function $\quad Q(\boldsymbol{x}, a) = \mathbb{E}[Y|\boldsymbol{X} = \boldsymbol{x}, A = a]$

Optimal policy $\quad d^{\rm opt}(\boldsymbol{x}) = \underset{a \in \mathcal{A}}{\text{argmax}}\ Q(\boldsymbol{x}, a)$

(Sutton and Barto 1998)

- Decomposition of Q-function $\quad Q(\boldsymbol{x}, a) = \pi(a|\boldsymbol{x})\eta(\boldsymbol{x})$

where $\quad \pi(a|\boldsymbol{x}) = \dfrac{Q(\boldsymbol{x}, a)}{\sum_{b \in \mathcal{A}} Q(\boldsymbol{x}, b)}$ and $\eta(\boldsymbol{x}) = \sum_{a \in \mathcal{A}} Q(\boldsymbol{x}, a)$

Note: $\quad d^{\rm opt}(\boldsymbol{x}) = \underset{a \in \mathcal{A}}{\text{argmax}}\ \pi(a|\boldsymbol{x})$

We consider a parametric model of Q-function

$$Q(\boldsymbol{x}, a) = \pi(a|\boldsymbol{x}, \boldsymbol{\theta})\eta(\boldsymbol{x})$$

Here we call $\eta(\boldsymbol{x})$ nuisance function.

5. ガンマ・べきダイバージェンス

Cf. Fujisawa-Eguchi (2008)

$$D_\gamma(Q_1, Q_2) = \int_{\mathcal{X}} \frac{\sum_{a \in \mathcal{A}} Q_1(\boldsymbol{x}, a) Q_2(\boldsymbol{x}, a)^\gamma}{\{\sum_{b \in \mathcal{A}} Q_2(\boldsymbol{x}, b)^{\gamma+1}\}^{\frac{\gamma}{\gamma+1}}} p(\boldsymbol{x}) d\boldsymbol{x} - \int_{\mathcal{X}} \Big\{\sum_{a \in \mathcal{A}} Q_1(\boldsymbol{x}, a)^{\gamma+1}\Big\}^{\frac{1}{\gamma+1}} p(\boldsymbol{x}) d\boldsymbol{x}$$

We model a Q-function as $\quad Q_{\boldsymbol{\theta}}(\boldsymbol{x}, a) = \pi(a|\boldsymbol{x}, \boldsymbol{\theta})\eta(\boldsymbol{x})$. Then

$$D_\gamma(Q_0, Q_{\boldsymbol{\theta}}) = \sum_{a \in \mathcal{A}} \int_{\mathcal{X}} Q_0(\boldsymbol{x}, a) \frac{\pi(a|\boldsymbol{x}, \boldsymbol{\theta})^\gamma}{\{\sum_{b \in \mathcal{A}} \pi(b|\boldsymbol{x}, \boldsymbol{\theta})^{\gamma+1}\}^{\frac{\gamma}{\gamma+1}}} p(\boldsymbol{x}) d\boldsymbol{x} + \text{const.}$$

Thus, the expected/empirical loss is

$$L_\gamma(\boldsymbol{\theta}) = -\sum_{a \in \mathcal{A}} \int_{\mathcal{X}} Q_0(\boldsymbol{x}, a) \frac{\pi(a|\boldsymbol{x}, \boldsymbol{\theta})^\gamma}{\{\sum_{b \in \mathcal{A}} \pi(b|\boldsymbol{x}, \boldsymbol{\theta})^{\gamma+1}\}^{\frac{\gamma}{\gamma+1}}} p(\boldsymbol{x}) d\boldsymbol{x}$$

$$L_\gamma^{\rm emp}(\boldsymbol{\theta}, \mathcal{D}) = -\sum_{i=1}^n \frac{y_i}{p(a_i|\boldsymbol{x}_i)} \frac{\pi(a_i|\boldsymbol{x}_i, \boldsymbol{\theta})^\gamma}{\{\sum_{b \in \mathcal{A}} \pi(b|\boldsymbol{x}_i, \boldsymbol{\theta})^{\gamma+1}\}^{\frac{\gamma}{\gamma+1}}} \qquad \text{where } \mathcal{D} = \{(\boldsymbol{x}_i, a_i, y_i)_{i=1}^n\}$$

> Assume that the true Q-function is $Q_{\boldsymbol{\theta}_0}(\boldsymbol{x}, a)$ and let $\hat{\boldsymbol{\theta}}_\gamma = \text{argmin}_{\boldsymbol{\theta}} L_\gamma^{\rm emp}(\boldsymbol{\theta}, \mathcal{D})$.
> Then $\hat{\boldsymbol{\theta}}_\gamma$ is consistent with $\boldsymbol{\theta}_0$.

We note: Only the class of γ-estimators is free from η-dependence.

6. 決定関数

- Q-function $Q(x, a) = \mathbb{E}[Y|X = a, X = x]$

$$d^{\rm opt}(x) = \underset{a \in \mathcal{A}}{\text{argmax}}\ Q(x, a)$$

- Decision function $f(x, a)$ is defined to satisfy

$$\sum_{a \in \mathcal{A}} f(x, a) = 0 \quad (\forall x \in \mathcal{X})$$

> **Definition** Decision function $f(x, a)$ is said to be **D-consistent** if
> $$d^{\rm opt}(x) = \underset{a \in \mathcal{A}}{\text{argmax}}\ f(x, a)$$

NB1: Q-learning aims to estimate the optimal Q-function.

NB2: $Q(x, a)$ vs $f(x, a)$; regression vs prediction (Zhao, 2012,2015)

7. 決定一致性

- Let Ψ be a strictly decreasing and convex function on $\mathbb{R}$

Ψ-loss function $\quad L_\Psi(f) = \mathbb{E}\left[\dfrac{Y}{p(A|\boldsymbol{X})} \Psi(f(\boldsymbol{X}, A))\right]$

Let $\mathcal{D}$ be an emplical dataset, $\mathcal{D} = \{(\boldsymbol{x}_i, a_i, y_i)_{i=1}^n\}$

Ψ-empirical loss function $L_\Psi(f, \mathcal{D}) = \dfrac{1}{n} \sum_{i=1}^n \dfrac{y_i}{p(a_i|\boldsymbol{x}_i)} \Psi(f(\boldsymbol{x}_i, a_i))$

> **Proposition**
> Let $f^* = \underset{f \in \mathcal{F}}{\text{argmin}}\ L_\Psi(f)$. Then, $f^*(\boldsymbol{x}, a)$ is D-consistent.

NB3: Any $L_\Psi(f)$ leads to the optimal policy $d^{\rm opt}$

証明の概略

- It suffices to show

$$d^{\rm opt}(\boldsymbol{x}) = \underset{a \in \mathcal{A}}{\text{argmax}}\ f^*(\boldsymbol{x}, a) \text{ where } d^{\rm opt}(\boldsymbol{x}) = \underset{a \in \mathcal{A}}{\text{argmax}}\ Q(\boldsymbol{x}, a)$$

- By definition, $\quad L_\Psi(f) = \mathbb{E}\left[\dfrac{Y}{p(A|\boldsymbol{X})} \Psi(f(\boldsymbol{X}, A))\right]$

$$= \sum_{a \in \mathcal{A}} \int_{\mathcal{X}} \int_{\mathcal{Y}} \frac{y}{p(a|\boldsymbol{x})} \Psi(f(\boldsymbol{x}, a)) p(y|\boldsymbol{x}, a) p(a|\boldsymbol{x}) p(\boldsymbol{x}) dy d\boldsymbol{x}$$

$$= \sum_{a \in \mathcal{A}} \int_{\mathcal{X}} \Psi(f(\boldsymbol{x}, a)) Q(\boldsymbol{x}, a) d\boldsymbol{x}$$

Let $\quad \mathcal{L}(f) = L_\Psi(f) - \int_{\mathcal{X}} \lambda(\boldsymbol{x}) \sum_{a \in \mathcal{A}} f(\boldsymbol{x}, a) d\boldsymbol{x}$

- Then, we find the Euler-Lagrange equilibrium condition

$$\Psi'(f^*(\boldsymbol{x}, a)) Q(\boldsymbol{x}, a) = \lambda(\boldsymbol{x}) \quad (\forall (\boldsymbol{x}, a) \in \mathcal{X} \times \mathcal{A})$$

■