Off-policy learning with Recognizers

Lucas Berry

Comp 767

February 3rd, 2017

Off-policy Learning

- Uses two policies to estimate values, a behavior policy μ and a target policy π.
- Relies on importance sampling ratios:
$$\rho_t^T = \prod_{k=t}^{T-1} \frac{\pi(A_k|S_k)}{\mu(A_k|S_k)}.$$
- In ordinary importance sampling these weights can lead to high variance.

Recognizers

- A recognizer c is a function that takes a subset of the action space and maps it to 0 or 1. Let A$\subseteq \mathcal{A}$, where $\mathcal{A}$ is the action space. Then

$$c(a) = \begin{cases} 1 & \text{if } a \in A \\ 0 & \text{otherwise} \end{cases}$$

- Can be used to minimize the variance by helping to formulate a π.

Target Policy

A recognizer together with the behavior specifies the target policy with the minimum-variance one-step importance sampling corrections:

$$\pi(a) = \frac{c(a)\mu(a)}{\sum_x c(x)\mu(x)} = \frac{c(a)\mu(a)}{b}. \tag{1}$$

Note we have dropped the state for the theorem and proof, this we added back in later.

Theorem 1

Let A=$\{a_1, ..., a_k\} \subseteq \mathcal{A}$, A is a subset of all possible actions. Consider a fixed policy μ and let π_A be the class of policies that only choose actions from A, i.e. if $\pi_A > 0$ then $a \in A$. Then the policy induced by μ and c_A is the policy with minimum-variance one-step importance sampling corrections, among those in π_A:

$$\frac{c(a)\mu(a)}{b} = \arg\min_{\pi \in \pi_A} \mathbb{E}_\mu \left[\left(\frac{\pi(a_i)}{\mu(a_i)} \right)^2 \right].$$

Proof

Proof: The variance can be written as:

$$\mathbb{E}_\mu\left[\left(\frac{\pi(a_i)}{\mu(a_i)}\right)^2\right] - \mathbb{E}_\mu\left[\left(\frac{\pi(a_i)}{\mu(a_i)}\right)\right]^2 = \sum_i \frac{\pi(a_i)^2}{\mu(a_i)} - 1.$$

Note the summations over i are such that $a_i \in A$. Next since $\pi(a_i)$ is a probability distribution then $\sum_i \pi(a_i) = 1$. Making our problem a constrained optimization problem. Let λ be our Lagrange multiplier then,

$$L(\pi(a_i), \lambda) = \sum_i \frac{\pi(a_i)^2}{\mu(a_i)} - 1 + \lambda(\sum_i \pi(a_i) - 1). \tag{2}$$

Proof

Solving (2) yields,

$$\pi(a_i) = \frac{\mu(a_i)}{\sum_i \mu(a_i)}.$$

Which is exactly (1) as $c(a_i) = 1$ for $a_i \in A$.

MDP

Returning to our MDP setup we can write the target policy $\pi(a|s)$ induced by $\mu(a|s)$ and our recognizer $c(s,a)$ as:

$$\pi(a|s) = \frac{c(s,a)\mu(a|s)}{\sum_x c(s,x)\mu(x|s)} = \frac{c(s,a)\mu(a|s)}{b}.$$

Which implies,

$$\rho_t^T = \prod_{k=t}^{T-1} \frac{\pi(A_k|S_k)}{\mu(A_k|S_k)} = \prod_{k=t}^{T-1} \frac{c(S_k,A_k)}{b}.$$

Value Function

Applying our new ρ we can replace the old ρ in ordinary importance sampling to then find estimates of the value given state s at time t. The variance of the recognizer method, ordinary importance sampling and weighted importance sampling can then be compared.

Numerical Example