DS 102 Discussion 8

Wednesday, March 30, 2022

1. **Propensity Scores**

   A propensity score $e(X)$ is defined as $\mathbb{P}[Z = 1|X]$. It turns out that propensity scores have many useful applications for causal inference in observational studies. In this question, we will explore some of these properties.

   (a) *Unconfoundedness and Propensity Scores*

   Show that if $Z \perp\!\!\!\perp \{Y(1), Y(0)\}|X$, then $Z \perp\!\!\!\perp \{Y(1), Y(0)\}|e(X)$. Why is this property useful?

(b) *Finding Propensity Scores*

Can we find the propensity score in a randomized experiment? How about the propensity score in an observational study? If it is unknown, is there a way to estimate the propensity score?

(c) *Propensity Score Matching*

Recall that for a particular subject i, we can only observe one of $Y_i(1)$ and $Y_i(0)$. To find the unknown potential outcome, we can use the method of matching. Specifically, we say that for two different subjects i and j, they can be matched if they have the same covariates (i.e. $X_i = X_j$). For example, if subject i and j match and i received the treatment, then we can estimate the treatment effect for subject i using

$$\hat{\tau}_i = Y_i(1) - Y_j(0)$$

Explain why this method of matching is very difficult to implement in practice. How can propensity scores be used to fix this issue?

Hint: Use the property we proved in Part (a).

(d) *The Overlap Assumption and Trimming*

Recall that the Inverse Propensity score Weighting estimator (IPW) is defined as

$$\hat{\tau}_{\text{IPW}} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{z_i y_i}{\hat{e}(x_i)} - \frac{(1-z_i)y_i}{1-\hat{e}(x_i)}\right)$$

where $\hat{e}(x_i)$ is the estimated propensity score. Explain what would happen to the IPW estimate if the covariates are very strong predictors of the observed treatment assignment? How can we adjust this estimator to avoid this problem?

2. **MDP Review**

In this problem, we'll review the concepts of the value function $V(s)$ and the Q-function $Q(s, a)$, as well as practice going through the computations needed to solve for them.

First, let's recap some terminology regarding Markov Decision Processes (MDPs):

- $s \in S$: the possible *states*
- $a \in A$: the possible *actions* we can take from states
- $T(s, a, s')$: the probability distribution over new states given that the agent starts at a particular state s taking action a
- $R(s, a, s')$: the reward the agent receives when moving to state s' using action a from state s
- $\gamma \in [0, 1]$: the discount factor for rewards received in the future
- $\pi : S \to A$: the policy, which describes a strategy of the action an agent should take from a given state

A *value function* gives the expected (discounted) reward received when starting from state s and using a particular strategy. Using the terminology above, we define the value function $V^\pi(s)$ of a policy π as:

$$V^\pi(s) = \sum_{a \in A} \pi(a \mid s) \sum_{s' \in S} T(s, a, s') \left[R(s, a, s') + \gamma V^\pi(s') \right]$$

This equation is also known as the **Bellman equation**. We are often interested in the value function of a particular policy: the one that is optimal from state s. This is the **optimal value function** $V^*(s)$, given by:

$$V^*(s) = \max_{a \in A} \sum_{s' \in S} T(s, a, s') \left[R(s, a, s') + \gamma V^*(s') \right]$$

Similarly, the **optimal Q-function** $Q^*(s, a)$ gives the expected (discounted) reward received when starting from state s, taking action a, then taking the optimal actions thereafter:

$$Q^*(s, a) = \sum_{s' \in S} T(s, a, s') \left[R(s, a, s') + \gamma V^*(s') \right].$$

A typical goal in reinforcement learning is to find a policy π^* that maximizes our expected discounted reward. Building up to that goal, we first need to understand how to evaluate the optimal value function and optimal Q-function.

In this problem, we will use the following gridworld environment:

			1
	×	`start`	-100

where `start` represents our initial state, × is a state we can't access, and the 1 and -100 states are terminal states with corresponding rewards. The reward received when moving to any other state is zero.

(a) *Computing $V^*(s)$*

Assume state transitions are deterministic, meaning that an action in a particular direction always moves us in that direction (unless it's toward the × state, in which case we stay in the same state). Compute the optimal value function at each state, where $\gamma = 0.9$.

(b) *Computing $Q^*(\texttt{start}, a)$*

Compute the optimal Q-function at our initial state for the actions of going up, down, left, and right.

(c) *Identifying an Optimal Action*

Based on the optimal Q-function you just computed, what would be the optimal move to make from `start`?

(d) *Identifying an Optimal Action with Stochastic Transitions*

Now suppose the state transitions are stochastic, such that there is a 0.8 probability of going in the direction you specified, and a 0.1 probability of going in either of the directions perpendicular to what specified. For example, if you decide to go up, you go up with 0.8 probability, go left with a 0.1 probability, and go right with a 0.1 probability. What is the best action to perform from `start`?