Lecture D1. Markov Reward Process 1

Sim, Min Kyu, Ph.D., mksim@seoultech.ac.kr

서울과학기술대학교 데이터사이언스학과

I. Motivation

Recap

- In the first introduction of soda DTMC, the following question was posed.

  Given I drink coke today, what is likely my consumption for upcoming 10 days? (Pepsi is \$1 and Coke is \$1.5)

- In Lecture note C1, Section 4, we demonstrated Monte-Carlo method that generates 10,000 (`MC_N`) number of paths and found total expected cost to be approximately 13.36.
- This lecture builds more systematic approach rather than the previous time-consuming Monte-Carlo method.
- This lecture begins to introduce those daunting notations and mathematical treatment for reinforcement learning.

reward and return

- Let the spending on day-t is r_t. That is, r_t is *cost* or *reward* for time t.
- The reward r_t is determined by the state at time t with a function $R(\cdot)$ such as $r_t = R(s)$.

Definition 1 (reward function)

A real-valued function $R : S \to \mathbb{R}$ is called a *reward function* that determines the reward given the state. That is, $r_t = R(s)$, where $S_t = s$

- We were asked to find the expected value of $r_0 + r_1 + \cdots + r_9$.

Definition 2 (return)

The *return* G_t is the sum of remaining reward at time t.

- In our problem,
  - $G_0 = r_0 + r_1 + \cdots + r_9$
  - $G_1 = r_1 + \cdots + r_9$
  - $G_2 = r_2 + \cdots + r_9$
  - $\cdots$
  - $G_9 = r_9$

- In our problem, we were asked to find the expected value of G_0 starting from state c at time 0.
- At time 0, the value of r_0 is known, but $r_1, \ldots, r_9$ are random variables. So, G_0 is random variable as well.
- The random variable G_0 depends on the current state S_0 and the randomness along the stochastic path.
- In general, the random variable G_t depends on the last-known state S_t and some randomness along the remaining path.

- Since G_t is a random variable, we want to evaluate $\mathbb{E}[G_t]$. In particular, considering its dependence structure, we are interested in evaluating $\mathbb{E}[G_t|S_t = s]$.
- In this light, the current problem is $\mathbb{E}[G_0|S_0 = c]$, or $\mathbb{E}[r_0 + r_1 + \cdots + r_9|S_0 = c]$.
- This motivates the following definition.

state-value function

Definition 3 (state-value function)

A *state-value function* $V_t(s)$ is the expected return given state s at time t. That is, $V_t(s) = \mathbb{E}[G_t | S_t = s]$.

- Again, we are interested in finding $V_0(c) = \mathbb{E}[G_0 | S_0 = c] = \mathbb{E}[r_0 + \cdots + r_9 | S_0 = c]$.

II. Method 1 - Monte-Carlo simulation

Recap

- The MC simulation is a valid approach. We shall review our initial effort with newly introduced terminology.
- The algorithm includes$\cdots$
  1. Generate a single stochastic path starting from the *initial state*, $S_0 = c$.
  2. Collect a single value of *return*, $G_i, 1 \leq i \leq$ MC_N, by accumulating *rewards*, $\{r_0, r_1, ..., r_9\}$, along the path.
  3. Take an average of collected *returns* to evaluate state-value function, $V_0(c)$.

```r
MC_N <- 10000
spending_records <- rep(0, MC_N)
for (i in 1:MC_N) {
  path <- "c" # coke today (day-0)
  for (t in 1:9) {
    this_state <- str_sub(path, -1, -1)
    next_state <- soda_simul(this_state)
    path <- paste0(path, next_state)
  }
  spending_records[i] <- cost_eval(path)
}

cost_eval <- function(path) {
  cost_one_path <-
    str_count(path, pattern = "c")*1.5 +
    str_count(path, pattern = "p")*1
  return(cost_one_path)
}
```

MC simulation for estimating *state-value function*

- Formally, for a *finite-horizon MRP*, the following is MC simulation for estimating *state-value function*.

```
# MC evaluation for state-value function
# with state s, time 0, reward r, time-horizon H
episode_i <- 0
cum_sum_G_i <- 0
while episode_i < num_episode
  Generate an stochastic path starting from state s and time 0.
  Calculate return G_i <- sum of rewards from time 0 to time H-1.
  cum_sum_G_i <- cum_sum_G_i + G_i
  episode_i <- episode_i + 1
State-value-fn V_t(s) <- cum_sum_G_i/num_episode
return V_t(s)
```

III. Method 2 - Iterative solution

Motivation

- Same as previous section, our goal is still to estimate $V_0(c) = \mathbb{E}[G_0|S_t = c]$.
- Since $G_t = \sum_{i=t}^{9} r_i$ has less number of terms when t is high number, we shall start from $t = 9$ and work backward, i.e. from $V_9(s)$, then $V_8(s)$, then $V_7(s)$,$\cdots$
- For $t = 9$,
  - From the general formula $V_t(s) = \mathbb{E}[G_t|S_t = s]$, it is easy to see that $V_9(s) = \mathbb{E}[G_9|S_9 = s] = \mathbb{E}[\sum_{i=9}^{9} r_i|S_9 = s] = \mathbb{E}[r_9|S_9 = s] = R(s)$.
  - In other words,
    - $V_9(c) = \mathbb{E}[r_9|S_9 = c] = R(c) = 1.0$ and
    - $V_9(p) = \mathbb{E}[r_9|S_9 = p] = R(p) = 1.5$.
  - In general,

$$V_9(s) = R(s) + V_{10}(s) \tag{1}$$

,where $V_{10}(s) = 0, \ \forall s$

- For $t = 8$,
  - From the general formula $V_t(s) = \mathbb{E}[G_t | S_t = s]$, *(watch below carefully)*

$$\begin{aligned} V_8(s) &= \mathbb{E}[G_8 | S_8 = s] \\ &= \mathbb{E}\left[\sum_{i=8}^{9} r_i \mid S_8 = s\right] \\ &= \mathbb{E}[r_8 + r_9 | S_8 = s] + \\ &= \mathbb{E}[r_8 | S_8 = s] + \mathbb{E}[r_9 | S_8 = s] \\ &= R(s) + \mathbb{E}[r_9 | S_8 = s] \end{aligned} \tag{2}$$

  - Here, let's consider $\mathbb{E}[r_9 | S_8 = c]$ first.
    - This is expected spending on day-9 given that I drink coke on day-8. This value is conditioned on what I drink on day-9. If coke on day-9 with probability 0.7, $r_9 = 1.5$. If pepsi w/ prob. 0.3, $r_9 = 1.0$. This expectation is 1.35 (= 0.7 · 1.5 + 0.3 · 1.0).
    - Formally,

$$\begin{aligned} \mathbb{E}[r_9 | S_8 = c] &= \mathbf{P}_{cc}\mathbb{E}[r_9 | S_8 = c, S_9 = c] + \mathbf{P}_{cp}\mathbb{E}[r_9 | S_8 = c, S_9 = p] \\ &= \mathbf{P}_{cc}\mathbb{E}[r_9 | S_9 = c] + \mathbf{P}_{cp}\mathbb{E}[r_9 | S_9 = p] \ (\because \textit{Markov property}) \\ &= \mathbf{P}_{cc}\mathbb{E}[G_9 | S_9 = c] + \mathbf{P}_{cp}\mathbb{E}[G_9 | S_9 = p] \\ &= \mathbf{P}_{cc}V_9(c) + \mathbf{P}_{cp}V_9(p) \end{aligned}$$

- (Cont'd for $t = 8$)
  - Now, let's consider $\mathbb{E}[r_9|S_8 = s]$ for generalized state s. With the notation assuming a transition from this state s to the next state s',

$$\begin{aligned}
\mathbb{E}[r_9|S_8 = s] &= \sum_{s' \in S} \mathbf{P}_{ss'} \mathbb{E}[r_9|S_8 = s, S_9 = s'] \\
&= \sum_{s' \in S} \mathbf{P}_{ss'} \mathbb{E}[r_9|S_9 = s'] \ (\because \textit{Markov property}) \\
&= \sum_{s' \in S} \mathbf{P}_{ss'} \mathbb{E}[G_9|S_9 = s'] \ (\because \textit{Markov property}) \\
&= \sum_{s' \in S} \mathbf{P}_{ss'} V_9(s') \qquad (3)
\end{aligned}$$

  - We shall now summarize for $t = 8$,

$$\begin{aligned}
V_8(s) &= \mathbb{E}[G_8|S_8 = s] = \mathbb{E}[r_8 + G_9|S_8 = s] \\
&= R(s) + \mathbb{E}[G_9|S_8 = s] \\
&= R(s) + \sum_{s' \in S} \mathbf{P}_{ss'} V_9(s') \qquad (4) \\
\textit{(expected return at time 8)} &= \textit{(reward at time 9)} + \textit{(expected return at time 9)}
\end{aligned}$$

- For $t = 7$,
  - From the general formula $V_t(s) = \mathbb{E}[G_t | S_t = s]$,

$$
\begin{aligned}
V_7(s) &= \mathbb{E}[G_7 | S_7 = s] \\
&= \mathbb{E}\left[\sum_{i=7}^{9} r_i \mid S_7 = s\right] \\
&= \mathbb{E}[r_7 + r_8 + r_9 | S_7 = s] \\
&= \mathbb{E}[r_7 | S_7 = s] + \mathbb{E}[r_8 + r_9 | S_7 = s] \\
&= R(s) + \mathbb{E}[G_8 | S_7 = s] \qquad (5)
\end{aligned}
$$

  - You get the hint? From here, we want to use $V_8(s) = \mathbb{E}[G_8 | S_8 = s]$ to express this as a recursive formula for state-value function just like Eq. (4).

$$
\begin{aligned}
V_7(s) &= R(s) + \sum_{s' \in S} \mathbf{P}_{ss'} \mathbb{E}[G_8 | S_7 = s, S_8 = s'] \\
&= R(s) + \sum_{s' \in S} \mathbf{P}_{ss'} V_8(s') \qquad (6)
\end{aligned}
$$

- For general t, *(exercise)*

- So far,

$$
\begin{aligned}
V_{10}(s) &= 0 \\
V_9(s) &= R(s) + \sum_{s' \in S} \mathbf{P}_{ss'} V_{10}(s') \textit{ from Eq. (1)} \\
V_8(s) &= R(s) + \sum_{s' \in S} \mathbf{P}_{ss'} V_9(s') \textit{ from Eq. (4)} \\
V_7(s) &= R(s) + \sum_{s' \in S} \mathbf{P}_{ss'} V_8(s') \textit{ from Eq. (6)} \\
\cdots &= \cdots \\
V_t(s) &= R(s) + \sum_{s' \in S} \mathbf{P}_{ss'} V_{t+1}(s') \\
\cdots &= \cdots \\
V_0(s) &= R(s) + \sum_{s' \in S} \mathbf{P}_{ss'} V_1(s')
\end{aligned}
$$

- Note that the array of equations can be solve from the top to the bottom.
- This iterative method is called as *backward induction* that works well with finite horizon problem.
- This iterative method (and its painful derivaion) is the most important mathematical essence of Markov decision process.

Implementation strategy

- Summary so far

$$
\begin{aligned}
V_{10}(s) &= 0 \\
V_t(s) &= R(s) + \sum_{s' \in S} \mathbf{P}_{ss'} V_{t+1}(s') \;\; (\textit{for} \;\; t \in \{0, 1, ..., 9\})
\end{aligned}
$$

- Strategy
  - Column vector v_t for $V_t(s)$
  - Column vector R for $R(s)$
  - The term $\sum_{s' \in S} \mathbf{P}_{ss'} V_{t+1}(s')$ can be written as $\mathbf{P} v_{t+1}$.
  - It follows $v_t = R + \mathbf{P} v_{t+1}$.

```r
P <- array(c(0.7,0.5,0.3,0.5), dim=c(2,2))
R <- array(c(1.5,1.0), dim=c(2,1))
H <- 10 # time-horizon
v_t1 <- array(c(0,0), dim=c(2,1)) # v_{t+1}

t <- H-1
while (t >= 0) {
  v_t  <- R + P %*% v_t1
  t    <- t-1
  v_t1 <- v_t
}
v_t
```

```
##          [,1]
## [1,] 13.35937
## [2,] 12.73438
```

- Thus, we have the following *state-value function*.
  - $V_0(c) = 13.359375$
  - $V_0(p) = 12.734375$

Backward induction for estimating *state-value function*

- Formally, for a *finite-horizon MRP*, the following is *backward induction* for estimating *state-value function*.

```
# Backward induction for state-value function
# with transition prob mat P, reward vector R, time-horizon H, state-value vector v_{}
1: v_H <- zero-column vector
2: t <- H-1
3: while t >= 0
4:   v_t <- R + P*v_{t+1}
5:   t <- t-1
9: return v_t # this is v_0(s) for all s, because t=0 at this point
```

```r
cat(str)
```

```
## If I only had an hour to chop down a tree, I would spend the first 45 minutes sharpening my axe. -
A. Lincoln
```