Biological Learning: Exercises RL3

RL3.1: Bellman equation

Policy

- actions: right, up, left, down
- for each action a, policy: $\pi(a|s) = p(A_t = a | S_t = s) = 0.25$

→ every action has same probability

Probability $p(s', r | s, a)$

$s = (1, 1)$:

$$p(s', r | s = (1,1), a = \text{'right'}) = \begin{cases} 1 & \text{for } s' = (1,2) \text{ and } r = 0 \\ 0 & \text{for all other values of } s' \text{ and } r \end{cases}$$

$$p(s', r | s = (1,1), a = \text{'up'}) = \begin{cases} 1 & \text{for } s' = (0,1) \text{ and } r = 0 \\ 0 & \text{for} \dots \end{cases}$$

$$p(s', r | s = (1,1), a = \text{'left'}) = \begin{cases} 1 & \text{for } s' = (1,0) \text{ and } r = 0 \\ 0 & \text{for} \dots \end{cases}$$

$$p(s', r | s = (1,1), a = \text{'down'}) = \begin{cases} 1 & \text{for } s' = (2,1) \text{ and } r = 0 \\ 0 & \text{for} \dots \end{cases}$$

$s = (0, 2)$:

$$p(s', r | s = (0,2), a = \text{'right'}) = \begin{cases} 1 & \text{for } s' = (0,3) \text{ and } r = 0 \\ 0 & \text{for} \dots \end{cases}$$

$$p(s', r | s = (0,2), a = \text{'up'}) = \begin{cases} 1 & \text{for } s' = (0,2) \text{ and } r = -1 \\ 0 & \text{for} \dots \end{cases}$$

$$p(s', r | s = (0,2), a = \text{'left'}) = \begin{cases} 1 & \text{for } s' = (0,1) \text{ and } r = 0 \\ 0 & \text{for} \dots \end{cases}$$

$$p(s', r | s = (0,2), a = \text{'down'}) = \begin{cases} 1 & \text{for } s' = (1,2) \text{ and } r = 0 \\ 0 & \text{for} \dots \end{cases}$$

$s = A$:

$$p\left(s', r \,\middle|\, s = A, a = \begin{matrix}\text{'right' or} \\ \text{'up' or} \\ \text{'left' or} \\ \text{'down'}\end{matrix}\right) = \begin{cases} 1 & \text{for } s' = (4,1) \text{ and } r = 10 \\ 0 & \text{for} \dots \end{cases}$$

Bellman equation

$$v_\pi(s) = \sum_a \pi(a|s) \sum_{s',r} p(s', r|s, a)\left[r + \gamma\, v_\pi(s')\right],$$

for all $s \in S$

$s = (1, 1)$:

$$\begin{aligned}
v_\pi(s=(1,1)) &= 0.25 \cdot (0 + \gamma\, v_\pi(s'=(1,2)) \\
&\quad + 0.25 \cdot (0 + \gamma\, v_\pi(s'=(0,1)) \\
&\quad + 0.25 \cdot (0 + \gamma\, v_\pi(s'=(1,0)) \\
&\quad + 0.25 \cdot (0 + \gamma\, v_\pi(s'=(2,1))
\end{aligned}$$

with $\gamma = 0.9$ | $= 0.25 \cdot 0.9\,(2.3 + 0.7 + 1.5 + 8.8)$

$= 2.9925$ //

$s = (0, 2)$:

$$\begin{aligned}
v_\pi(s=(0,2)) &= 0.25 \cdot (0 + \gamma\, v_\pi(s'=(0,3)) \\
&\quad + 0.25 \cdot (-1 + \gamma\, v_\pi(s'=(0,2)) \\
&\quad + 0.25 \cdot (0 + \gamma\, v_\pi(s'=(0,1)) \\
&\quad + 0.25 \cdot (0 + \gamma\, v_\pi(s'=(1,2)) \\
&= 0.25 \cdot 0.9\,(5.3 + 4.4 + 8.8 + 2.3) - 0.25 \\
&= 4.43
\end{aligned}$$

$s = A$:

$$\begin{aligned}
v_\pi(s=A) &= 4 \cdot 0.25 \cdot (10 + \gamma\, v_\pi(s'=(4,1)) \\
&= 10 + 0.9 \cdot (-1.3) \\
&= 8.83
\end{aligned}$$

RL3.2: Compute the values of states using pen & paper

Policy 1

6	7	8	9
7	8	9	10

Policy 2

6	7	8	9
5	4	6	10

$s = (1,2)$

$$v(s) = 0.5 \cdot (-1) + 0.5 \cdot 4 + 0.5 \cdot (-1) + 0.5 \cdot 10$$

$$= -0.5 + 2 - 0.5 + 5 = 6$$

→ Policy 1 is better: each state in Policy 1 has a higher value than or similar value as the corresponding state in Policy 2

Policy 3

$a = 5$	$d = 6$	8	9
$b = 5$	$c = 6$	9	10

$c = 0.5 \cdot (-1+9) + 0.5 \cdot (-1+b)$

$= 0.5 \cdot (b+7)$

$b = c - 1$

$d = 0.5(-1+8) + 0.5(-1 + 0.5(b+7))$

$= 0.5(6 + 0.5(b+7))$

$= 0.5(6 + 0.5b + 3.5)$

$= 0.5(9.5 + 0.5b)$

$a = d - 1$

c in b:

$b = 0.5(b+7) - 1 = 0.5b + 3.5 - 1 = 0.5b + 2.5 \Leftrightarrow$

$0.5b = 2.5 \Leftrightarrow$

$b = 5$

b in c:

$b = c - 1 \Leftrightarrow$

$c = b + 1 \Leftrightarrow$

$c = 5 + 1 \Leftrightarrow$

$c = 6 \Leftrightarrow$

b in d:

$d = 0.5(9.5 + 0.5b)$

$= 4.75 + 0.25b$

$= 4.75 + 0.25 \cdot 5$

$= 4.75 + 1.25$

$= 6$

d in a:

$a = d - 1$

$= 6 - 1$

$= 5$

<u>Optimal policy</u>

- Policy is best out of the three
- It is also optimal: optimizes number of steps to reach the goal but other equivalent policies are there, e.g.

→	↓	↓	↓
→	→	→	10

Bellmann optimality equation

v_*

6	7	8	9
7	8	9	10

$$\arg\max_a \sum_{s',r} p(s',r|s,a)[r+\gamma v_*(s')]$$

→ maximal for policies with movements in row 0 either right or down and movement in row 1 right

→ fits with policy 1 but not policy 3 (s = (1,1) has 0.5 probability of going left)