COMP4418 17s2 • Week 12 • Exercises

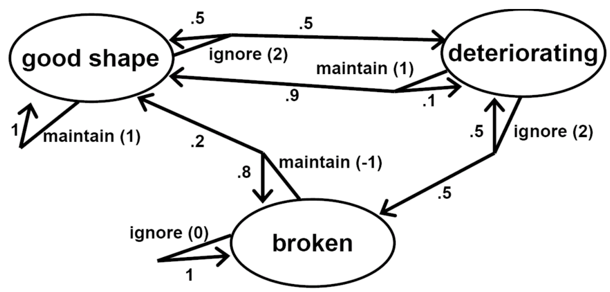
Decision Making

---

1. (Markov Decision Process)
   Let $g \mathrel{\hat{=}}$ "good shape", $d \mathrel{\hat{=}}$ "deteriorating" and $b \mathrel{\hat{=}}$ "broken."

   Consider a discount factor of $\delta = 0.9$.

   Starting with $v_0(g) = 0$, $v_0(d) = 0$ and $v_0(b) = 0$, apply three steps of the value iteration algorithm towards computing the optimal policy for the MDP below.



   What does the algorithm suggest as the optimal policy at this point?

COMP4418 17s2 • Week 12 • Sample Solutions

Decision Making

---

1. (Markov Decision Process)

$$
\begin{aligned}
v_{i+1}(g) = \max\{&u(g,\text{ignore}) + 0.9 \cdot \textstyle\sum_{s'} P(g,\text{ignore},s') \cdot v_i(s'),\\
&u(g,\text{maintain}) + 0.9 \cdot \textstyle\sum_{s'} P(g,\text{maintain},s') \cdot v_i(s')\}\\
v_{i+1}(d) = \max\{&u(d,\text{ignore}) + 0.9 \cdot \textstyle\sum_{s'} P(d,\text{ignore},s') \cdot v_i(s'),\\
&u(d,\text{maintain}) + 0.9 \cdot \textstyle\sum_{s'} P(d,\text{maintain},s') \cdot v_i(s')\}\\
v_{i+1}(b) = \max\{&u(b,\text{ignore}) + 0.9 \cdot \textstyle\sum_{s'} P(b,\text{ignore},s') \cdot v_i(s'),\\
&u(b,\text{maintain}) + 0.9 \cdot \textstyle\sum_{s'} P(b,\text{maintain},s') \cdot v_i(s')\}
\end{aligned}
$$

Hence, ;

$$
\begin{aligned}
v_1(g) &= \max\{2;\ 1\} = 2\\
v_1(d) &= \max\{2;\ 1\} = 2\\
v_1(b) &= \max\{0;\ -1\} = 0
\end{aligned}
$$

$$
\begin{aligned}
v_2(g) &= \max\{2 + 0.9 \cdot (0.5 \cdot 2 + 0.5 \cdot 2);\ 1 + 0.9 \cdot (1 \cdot 2)\} = 3.8\\
v_2(d) &= \max\{2 + 0.9 \cdot (0.5 \cdot 2 + 0.5 \cdot 0);\ 1 + 0.9 \cdot (0.9 \cdot 2 + 0.1 \cdot 2)\} = 2.9\\
v_2(b) &= \max\{0 + 0.9 \cdot (1 \cdot 0);\ -1 + 0.9 \cdot (0.8 \cdot 0 + 0.2 \cdot 2)\} = 0
\end{aligned}
$$

$$
\begin{aligned}
v_3(g) &= \max\{2 + 0.9 \cdot (0.5 \cdot 3.8 + 0.5 \cdot 2.9);\ 1 + 0.9 \cdot (1 \cdot 3.8)\} = 5.015\\
v_3(d) &= \max\{2 + 0.9 \cdot (0.5 \cdot 2.9 + 0.5 \cdot 0);\ 1 + 0.9 \cdot (0.9 \cdot 3.8 + 0.1 \cdot 2.9)\} = 4.339\\
v_3(b) &= \max\{0 + 0.9 \cdot (1 \cdot 0);\ -1 + 0.9 \cdot (0.8 \cdot 0 + 0.2 \cdot 3.8)\} = 0
\end{aligned}
$$

Optimal policy = best action taken in each state in the last step:

$$
\begin{aligned}
\pi(g) &= \text{ignore}\\
\pi(d) &= \text{maintain}\\
\pi(b) &= \text{ignore}
\end{aligned}
$$

Decision Making

---

1. (Monty Hall Game as Markov Decision Process, POMDP)

   Only show one of three actions in S_0 – the other two are symmetric

S_0	a_0	$P(S_0,a_0,S_1)$	S_1	a_1	$P(S_1,a_1,S_2)$	S_2	a_2	$P(S_2,a_2,S_3)$	S_3	$u(S_2,a_2)$
()	choose(2)	1/3	$(2,1)$	noop	1	$(2,1,3)$	noop	1	$(2,1,3)$	0
							switch	1	$(2,2,3)$	100
()	choose(2)	1/3	$(2,2)$	noop	1/2	$(2,2,1)$	noop	1	$(2,2,1)$	100
							switch	1	$(3,2,1)$	0
					1/2	$(2,2,3)$	noop	1	$(2,2,3)$	100
							switch	1	$(1,2,3)$	0
()	choose(2)	1/3	$(2,3)$	noop	1	$(2,3,1)$	noop	1	$(2,3,1)$	0
							switch	1	$(3,3,1)$	100

Like colours in a column indicate states with identical observations. The agent cannot distinguish these states from each other. Some probabilities from the belief states over S_0, S_1 and S_2:

$P(S_0 = ()) \;=\; 1.0$

$P(S_1 = (2,1) \,|\, a_0 = \text{choose}(2)) \;=\; P(S_1 = (2,2) \,|\, a_0 = \text{choose}(2)) \;=\; P(S_1 = (2,3) \,|\, a_0 = \text{choose}(2)) \;=\; \frac{1}{3}$

$$
\begin{aligned}
P(S_2 = (2,1,3) \,|\, a_0 = \text{choose}(2), a_1 = \text{noop}, o_1 = \text{door 3 opened}) &= (1/3)/(1/3 + 1/3 * 1/2) = 2/3 \\
P(S_2 = (2,2,1) \,|\, a_0 = \text{choose}(2), a_1 = \text{noop}, o_1 = \text{door 1 opened}) &= (1/3 * 1/2)/(1/3 * 1/2 + 1/3) = 1/3 \\
P(S_2 = (2,2,3) \,|\, a_0 = \text{choose}(2), a_1 = \text{noop}, o_1 = \text{door 3 opened}) &= (1/3 * 1/2)/(1/3 + 1/3 * 1/2) = 1/3 \\
P(S_2 = (2,3,1) \,|\, a_0 = \text{choose}(2), a_1 = \text{noop}, o_1 = \text{door 1 opened}) &= (1/3)/(1/3 * 1/2 + 1/3) = 2/3
\end{aligned}
$$

It follows that the optimal policy is: any action in $b(S_0)$, noop in $b(S_1)$, switch in $b(S_2)$. The expected value is (1/3)*0+(2/3)*100 = 66.667