Markov Decision Process

Chao Lan
Introduction
<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>-0.04</td>
<td>-0.04</td>
<td>-0.04</td>
</tr>
<tr>
<td>2</td>
<td>-0.04</td>
<td>-0.04</td>
<td>-0.04</td>
</tr>
<tr>
<td>1</td>
<td>START</td>
<td>-0.04</td>
<td>-0.04</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
</tbody>
</table>
### Example

<table>
<thead>
<tr>
<th>Step</th>
<th>State(S)</th>
<th>Reward(R)</th>
<th>Action(A)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>(1,1)</td>
<td>-0.04</td>
<td>right</td>
</tr>
<tr>
<td>1</td>
<td>(2,1)</td>
<td>-0.04</td>
<td>right</td>
</tr>
<tr>
<td>2</td>
<td>(3,1)</td>
<td>-0.04</td>
<td>up</td>
</tr>
<tr>
<td>3</td>
<td>(3,2)</td>
<td>-0.04</td>
<td>right</td>
</tr>
<tr>
<td>4</td>
<td>(4,2)</td>
<td>-1</td>
<td>stop</td>
</tr>
<tr>
<td><strong>Total Reward</strong></td>
<td></td>
<td><strong>-1.16</strong></td>
<td></td>
</tr>
</tbody>
</table>
Two Questions

1. What is the optimal policy to take actions for maximizing total reward?

2. Given an arbitrary policy, what is the total reward we can receive by starting from any state?
Stochastic Move
Q1: Optimal policy to move round?
Q2: Expected total reward from each state?
Model
1. Transition Probability

We will characterize results of stochastic moves using transition probability.

\[ \Pr\{S_t = s', R_t = r \mid S_{t-1} = s, A_{t-1} = a\} \]

\( t \) is step index.
Example

Pr\{ (2,3), -0.04 \mid (1,3), \text{right} \} = \begin{array}{cccc}
\text{START} & -0.04 & -0.04 & -0.04 \\
1 & 1 & 2 & 3 & 4 \\
2 & -0.04 & & -0.04 & -1 \\
3 & & -0.04 & -0.04 & +1 \\
\end{array}
Example

\[ \text{Pr}\{ (2,3), -0.04 \mid (1,3), \text{right} \} = \]

\[ \text{Pr}\{ (1,2), -0.04 \mid (1,3), \text{right} \} = \]
Example

\[ \text{Pr}\{(2,3), -0.04 \mid (1,3), \text{right}\} = \]

\[ \text{Pr}\{(1,2), -0.04 \mid (1,3), \text{right}\} = \]

\[ \text{Pr}\{(1,3), -0.04 \mid (1,3), \text{right}\} = \]
Example

Pr\{(2,3), -0.04 \mid (1,3), \text{right}\} =

Pr\{(1,2), -0.04 \mid (1,3), \text{right}\} =

Pr\{(1,3), -0.04 \mid (1,3), \text{right}\} =

Pr\{(2,3), -0.04 \mid (2,3), \text{right}\} =
Example

Pr\{ (2,3), -0.04 \mid (1,3), \text{right} \} = 0.8

Pr\{ (1,2), -0.04 \mid (1,3), \text{right} \} = 0.1

Pr\{ (1,3), -0.04 \mid (1,3), \text{right} \} = 0.1

Pr\{ (2,3), -0.04 \mid (2,3), \text{right} \} = 0.2
[Q] True or False?

This probability accurately models state-transition in real-world problems.

\[ \Pr\{S_t = s', R_t = r \mid S_{t-1} = s, A_{t-1} = a\} \]
Example
Markov Property

If the next state depends only on the current state (and no earlier), transition probability fully characterizes the dynamics of the environment.

\[ \Pr\{S_t = s', R_t = r \mid S_{t-1} = s, A_{t-1} = a\} \]
2. **Policy**

We will characterize each solution using a policy $\pi$, which is a function mapping from a state $s$ to its action $\pi(s)$.

$$\pi(s) = \text{action to take at state } s$$

A solution is a set of actions for all states in the decision process.
Example

\[ \pi(1,1) = \]

\[ \pi(3,3) = \]

\[ \pi(4,2) = \]
Example

\[ \pi(1,1) = \text{UP} \]

\[ \pi(3,3) = \text{RIGHT} \]

\[ \pi(4,2) = \text{LEFT} \]
[Q] True or False?

The following are two different policies.
3. (Expected) Utility

The expected utility of policy $\pi$ is the cumulative discounted rewards which can be collected by following $\pi$ from any state $s$ until termination.

$$U^\pi(s) = E \left[ \sum_{t=0}^{\infty} \gamma^t R(S_t) \right]$$

$\gamma$ is the discount factor, chosen in $[0,1]$. 
Example
Count all possible paths to (t).

1. 0.8 x 0.8 = 0.64

2. 0.1 x 0.8 = 0.08

3. 1 x 0.1 x 0.8 = 0.08

\[ E[\text{Reward}] = 0.76 \times 0.8 + 0.568 \times 0.08 + 0.4144 \times 0.008 + \ldots \]

\[ = 0.918 \]
[Q] True or False?

Changing action at a state will not change its utility.
4. Optimal Policy

An optimal policy $\pi^*$ is one that maximizes the expected utility at every state.

$$\pi^*_s = \arg\max_{\pi} U^\pi(s)$$
[Q] True or False?

This policy is not optimal. Action at (1,3) should be UP for reaching +1 faster.

Assume reward of every non-terminal state is -0.04.
[Q] What factors will affect optimal policy?
Impact of Reward on Optimal Policy

1. If we change reward from -0.04 to -2...
Impact of Reward on Optimal Policy

1. If we change reward from -0.04 to -2...
2. If we change reward from -0.04 to +10...
Impact of Reward on Optimal Policy

2. If we change reward from -0.04 to +10...

A star means an arbitrary action.
[Q] Will the change remain?

If we change reward from -0.04 to +0.04 (or, any positive value)...

![Diagram showing movement changes](image-url)