Tutorial 9

CS 337 Artificial Intelligence & Machine Learning, Autumn 2019

October, 2019

1 Example: Optimal Action Plan

Lets define a state machine with two states s_1, s_2. Formally, in the example depicted in Figure 1 we have:

time: $T = \{1...N\}$

states: $S = \{s_1, s_2\}$

possible actions in each state: $A_{s_1} = \{a_{11}, a_{12}\}, A_{s_2} = \{a_{21}\}$

reward: $r(s_1, a_{11}) = 5, r(s_1, a_{12}) = 10, r(s_2, a_{21}) = -1$

transition function: $\Pr(s_1|s_1, a_{11}) = 0.5$, $\Pr(s_2|s_1, a_{11}) = 0.5$, $\Pr(s_2|s_1, a_{12}) = 1$, $\Pr(s_2|s_2, a_{22}) = 1$

Now we want to maximize the return. First we compare two deterministic policies:

π_1- always chooses a_{11} when in state s_1.

π_2 - always chooses a_{12} when in state s_1.

Determine the best policy, that is the policy that maximizes the value function. Explain the procedure.

Solution:

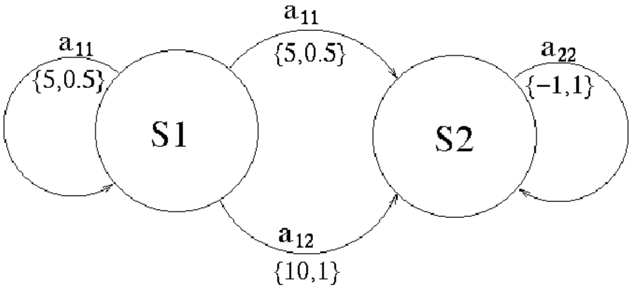

Figure 1: State Diagram

As an aside, recall that $V_\pi(s,N)$ is the expected return of policy Π in N time steps.

$$\begin{aligned}
V_{\pi_1}(s_1,3) &= \frac{1}{2}(5-1+0)+\frac{1}{2}(5+5+0)=7 \\
V_{\pi_2}(s_1,3) &= 10-1=9 \\
V_{\pi_1}(s_1,5) &= \frac{1}{2}(5-3)+\frac{1}{4}(10-2)+\frac{1}{8}(15-1)+\frac{1}{8}(20)=7.25 \\
V_{\pi_2}(s_1,5) &= 10-3=7
\end{aligned}$$

Notice how in shorter time frames π_2 is the preferable policy, while in longer ones, π_1 brings on the higher returns.

Finding a π that maximizes $V_\pi(s,N)$ is called a finite horizon problem. When $N\to\infty$, we get the infinite horizon problem.

We solve by looking first at the reward gained in the last step, for N=2:

$$\begin{aligned}
Q_2(s_1,a_{11}) &= 5 \\
Q_2(s_1,a_{12}) &= 10
\end{aligned}$$

In this case:

$$\begin{aligned}
V_2(s_1) &= 10 \\
V_2(s_2) &= -1
\end{aligned}$$

When N=2, The optimal action from s_1 is a_{12}.

For N=3, we use the optimal policy for N=2, after we do one action.

$$\begin{aligned}
Q_3(s_1,a_{11}) &= 5+\frac{1}{2}V_2(s_1)+\frac{1}{2}V_2(s_2)=5+\frac{1}{2}*10+\frac{1}{2}*(-1)=9.5 \\
Q_3(s_1,a_{12}) &= 10+V_2(s_2)=10-1=9
\end{aligned}$$

Hence, when N=3 the optimal action from s_1 is a_{11}.

2 MDP and Optimal Policy for the Recruitment Problem

A manager has to recruit a new employee and he can serially interview a finite group of candidates. There is a total order defined on the candidates' fitness to the opening, and there are no two employees with the same skill level. The manager is able to sort the candidates' fitness level after a short interview. After each interview the manager has two alternatives:

- recruit the last interviewed candidate
- continue to the next interview (and give up the chance of recruiting the previous candidate)

The goal is to maximize the probability of recruiting the best candidate. Formulate this as a Markov Decision Process (MDP) and develop its optimal policy. Analyze what happens when N (size of candidate set being interviewed) is 5. What happens when $N \to \infty$?

Solution: We will first construct a corresponsing MDP for the problem, and then develop the optimal policy for it.
Let $A = \{Continue, QuitAndHire\}$ the possible actions, where $Continue$ stands for 'continue' and $QuitAndHire$ for 'quit and recruit'. Let $S = \{MaxSoFar, Other, 1, 0\}$ be the group of states of the MDP, standing for:

- MaxSoFar - the last interviewed candidate was the best so far
- Other - the last interviewed candidate was NOT the best so far
- 1 - the last interviewed candidate was THE best candidate in the entire group and was hired
- 0 - the last interviewed candidate was NOT the best candidate in the entire group and was hired

The resultant MDP, uses the following transition probabilities:

- $q_t = Prob[\max\{x_1, ..., x_t\} = \max\{x_1, ..., x_N\}] = \frac{t}{N}$
- $r_t = Prob[x_{t+1} \geq \max\{x_1, ..., x_t\}] = \frac{1}{t+1}$

where N is the finite time horizon (the size of the candidates group) and t is the current time (the number of candidates interviewed so far).
Writing the optimality equations for this problem we get:

$$U_t^*(1) = 1, \;\; U_t^*(0) = 0, \;\; U_N^*(MaxSoFar) = U_N^*(Other) = 0$$

$$\begin{aligned}
U_t^*(Other) &= \max\{0, r_t U_{t+1}^*(MaxSoFar) + (1 - r_t)U_{t+1}^*(Other)\} \\
&= r_t U_{t+1}^*(MaxSoFar) + (1 - r_t)U_{t+1}^*(Other) \\
U_t^*(MaxSoFar) &= \max\{q_t U_{t+1}^*(1), r_t U_{t+1}^*(MaxSoFar) + (1 - r_t)U_{t+1}^*(Other)\} \\
&= \max\{q_t \cdot 1, r_t U_{t+1}^*(MaxSoFar) + (1 - r_t)U_{t+1}^*(Other)\} \\
&= \max\{q_t, U_t^*(Other)\}
\end{aligned}$$

Assigning the terms for q_t and r_t we get:

$$U_t^*(Other) = \frac{1}{t+1}U_{t+1}^*(MaxSoFar) + \frac{t}{t+1}U_{t+1}^*(Other) \tag{1}$$

$$U_t^*(MaxSoFar) = \max\{\frac{t}{N}, U_t^*(Other)\} \tag{2}$$

An optimal decision rule has the following properties:

- In $s = Other$, always choose $a_s = Continue$ (otherwise the return is promised to be zero)
- In $s = MaxSoFar$, choose $a_s = Continue$ if $\frac{t}{N} < U^*_t(Other)$ and $a_s = QuitAndHire$ otherwise (in order to maximize $U^*_t(MaxSoFar)$)

The optimal policy is of the form:

- Interview τ candidates, performing $a_t = Continue,\ t \leq \tau$
- Quit and hire the first candidate after time τ, which is the best so far

2.1 When $N = 5$

Let $N = 5$, we get
$\frac{1}{3} + \frac{1}{4} < 1$ and $\frac{1}{2} + \frac{1}{3} + \frac{1}{4} > 1$
and therefore we choose $\tau = 2$.

2.2 When $N \to \infty$

Let $N \to \infty$, we get

$$\sum_{j=\tau}^{N-1} \frac{1}{j} \sim \ln \frac{N}{\tau}.$$

We therefor shearch for τ such that
$\ln \frac{N}{\tau+1} < 1$ and $\ln \frac{N}{\tau} > 1$,
which leads to:

$$\tau \sim \frac{N}{e}$$

OPTIONAL Optimality Proof We will now show that the described policy agrees with the optimality equations.
We start by showing that if there is a time τ such that $\frac{\tau}{N} < U^*_\tau(MaxSoFar)$ then $\forall t, t < \tau$ we get $\frac{t}{N} < U^*_t(MaxSoFar)$. That is, if there exists a time $t = \tau$ in which it is preferred to *Continue*, then for each earlier time it is also preferred to continue. Later on, we will show that such a time, τ, does exist.
Let $\frac{t}{N} < U^*_t(MaxSoFar)$, then according to equation 2 $U^*_\tau(MaxSoFar) = U^*_\tau(Other)$. Performing backword induction steps, and using equations 1 and 2 we show that for $t < \tau$ the inequality remains true:
$U^*_{\tau-1}(Other) = \frac{1}{\tau}U^*_\tau(MaxSoFar) + \frac{\tau-1}{\tau}U^*_\tau(Other) = U^*_\tau(Other) > \frac{\tau}{N}$
$U^*_{\tau-1}(MaxSoFar) = \max\{\frac{\tau-1}{N}, U^*_{\tau-1}(Other)\} > \frac{\tau}{N} > \frac{\tau-1}{N}$
We now show that for $t > \tau$ and $s = MaxSoFar$, it is preferred to *QuitAndHire*.
Let $t > \tau$, then according to the first half of the proof,
$U^*_t(MaxSoFar) = \frac{t}{N}$

and,

$$\begin{aligned} U_t^*(Other) &= \frac{1}{t+1}U_{t+1}^*(MaxSoFar) + \frac{t}{t+1}U_{t+1}^*(Other) \\ &= \frac{1}{N} + \frac{t}{t+1}U_{t+1}^*(Other) \\ &= \frac{1}{N} + \frac{t}{t+1}\frac{1}{N} + \frac{t}{t+2}\frac{1}{N} + ... \\ &= \frac{1}{N}\sum_{i=0}^{N-t}\frac{t}{t+i} \sim \ln(\frac{N}{t}) \end{aligned}$$

We conclude this proof by showing that such a τ exists, that is, for $N \geq 2$ there exists $\tau \geq 1$ that meets the above requiremints. Let us assume $\tau = 0$ we get:
$U_1^*(Other) = \frac{1}{2} + \frac{1}{3} + ... + \frac{1}{N} > \frac{1}{2} \geq \frac{1}{N} = U_1^*(MaxSoFar) \geq U_1^*(Other)$
which is a circular inequality, and thus leads to cnotradiction.
Note that for $N \geq 2$ we always get $\tau \geq 1$:
$U_1^*(Other) = ... = U_\tau^*(Other)$
$||$
$U_1^*(MaxSoFar) = ... = U_\tau^*(MaxSoFar)$
and for $t > \tau$
$U_t^*(MaxSoFar) = \frac{t}{N}$
$U_t^*(Other) = \frac{t}{N}(\frac{1}{t} + \frac{1}{t+1} + ... + \frac{1}{N-1})$.
We therefore choose $Continue$ if $\frac{1}{t} + ... + \frac{1}{N-1} > 1$ and $QuitAndHire$ otherwise.

3 Bound concerning Value Iteration

Consider an MDP $(S, A, \Pr, R, \gamma)$, with notations being as usual and discount factor $\gamma \in [0, 1)$. Recall that the Value Iteration algorithm produces a sequence V_0,V_1, V_2, ... each element being a mapping from S to R, which converges to the optimal value function V_*

1. For $t = 0, 1, ...$, write down how V_{t+1} is obtained from V_t
2. Assume that there is a scalar $R_{max} > 0$ such that each individual reward obtained from R lies in $[0, R_{max}]$. Also assume that Value Iteration is initialised with the zero vector: that is, $V_0 = 0$. Show that for $t = 0, 1, ...$ and for $s \in S$:

$$V_*(s) - V_t(s) \leq \frac{\gamma^t R_{max}}{1-\gamma}$$

Solution:

1. For $t = 0, 1, ...$ and $s \in S$:

$$V_{t+1}(s) \Leftarrow \max_{a \in A}\sum_{s' \in S}\Pr(s'|a, s)R(s, a, s') + \gamma V_t(s')$$

2. We prove the result by induction on t. For $t = 0$, we have to show that for $s \in S$,

$$V^*(s) \leq \frac{R^{max}}{1-\gamma}$$

which is evident since $V^*(s)$ is the expected infinite discounted reward obtained by a policy, in this case an optimal policy. Even if each reward is maximum, $V^*(s)$ can at most be $R^{max} + \gamma R^{max} + \gamma^2 R_{max} + ... = \dfrac{R_{max}}{1-\gamma}$

Assume the result is true for some $t \geq 0$. Take π^* to be any optimal policy. We have, for $s \in S$:

$$\begin{aligned} V^{t+1}(s) &= \max_{a \in A} \sum_{s' \in S} \Pr(s'|a,s)\{R(s,a,s') + \gamma V^t(s')\} \\ &\geq \sum_{s' \in S} \Pr(s'|\pi^*(s),s)\left\{R(s,\pi^*(s),s') + \gamma V^t(s')\right\} \\ &\geq \sum_{s' \in S} \Pr(s'|\pi^*(s),s)\left\{R(s,\pi^*(s),s') + \gamma\left(V^*(s') - \frac{\gamma^t R_{max}}{1-\gamma}\right)\right\} \qquad (3) \end{aligned}$$

wherein the last step applies the induction hypothesis and also the fact that the transition probabilities are non-negative. By expanding out, we get

$$\begin{aligned} V^{t+1}(s) &\geq \sum_{s' \in S} \Pr(s'|\pi^*(s),s)\{R(s,\pi^*(s),s') + \gamma V^*(s')\} - \sum_{s' \in S} \Pr(s'|\pi^*(s),s)\left\{\frac{\gamma^{t+1} R_{max}}{1-\gamma}\right\} \qquad (4) \\ &= V^*(s) - \frac{\gamma^{t+1} R_{max}}{1-\gamma} \qquad (5) \end{aligned}$$

obtained by invoking Bellman's Equations for π^* and equating the sum of transition probabilities from $(s, \pi^*(s))$ to 1.

A second, direct approach would be to split

$$V^*(s) = \mathbb{E}_{\pi^*}\left\{r^0 + \gamma r^1 + \gamma^2 r^2 + ... \middle| s^0 = s\right\}$$

into a sum of

$$T_1 = \mathbb{E}_{\pi^*}\left\{r^0 + \gamma r^1 + \gamma^2 r^2 + ... \gamma^{t-1} r^{t-1} \middle| s^0 = s\right\}$$

and

$$T_2 = \mathbb{E}_{\pi^*}\left\{\gamma^t r^t + \gamma^{t+1} r^{t+1} + ... \middle| s^0 = s\right\}$$

It so happens that $V^t(s)$ is the maximum expected discounted t-step reward that can be possibly obtained; in general one would have to follow a non-stationary (time-dependent) policy in order to achieve it. T_1 is the expected discounted t-step reward obtained by following π^* and so cannot exceed $V^t(s)$. Since each individual reward is upper-bounded by R_{max}, we see that T_2 is at most $\dfrac{\gamma^t R_{max}}{1-\gamma}$. Hence, $V^*(s) \leq V^t(s) + \dfrac{\gamma^t R_{max}}{1-\gamma}$