Reinforcement Learning April 10, 2019

Recitation 7

Lecturer:Yishay Mansour *TA: Lee Cohen*

Part 1: Algorithms

Algorithm 1. TD(0) - Temporal Difference(0)
Update value $V(S_t)$ toward estimated return $R_{t+1} + \gamma * V(S_{t+1})$:
$V(S_t) = V(S_t) + \alpha * (R_{t+1} + \gamma * V(S_{t+1}) - V(S_t))$
Where:

- $R_{t+1} + \gamma * V(S_{t+1})$ is called the TD target
- $\Delta_t = R_{t+1} + \gamma * V(S_{t+1}) - V(S_t)$ is called the TD error

Algorithm 2. n-Step Temporal Difference Learning
Update value $V(S_t)$ toward estimated return $G_t^{(n)}$:
$V(S_t) = V(S_t) + \alpha_t * (G_t^{(n)} - V(S_t))$
Where:

- The n-step return - $G_t^{(n)} = R_{t+1} + \gamma * R_{t+2} + ... + \gamma^{n-1} * R_{t+n} + \gamma^n * V(S_{t+n})$
- For $n = 1, 2, 3, ...$

Algorithm 3. λ Temporal Difference Learning - TD(λ) - Forward view
Update value $V(S_t)$ toward estimated return G_t^{λ}:
$V(S_t) = V(S_t) + \alpha_t * (G_t^{\lambda} - V(S_t))$
Where:

- $G_t^{\lambda} = (1 - \lambda) \sum_{n=1}^{\infty} \lambda^{n-1} G_t^{(n)}$
- For $\lambda \in (0, 1)$

Algorithm 4. λ Temporal Difference Learning - TD(λ) - Backward view
Eligibility traces:

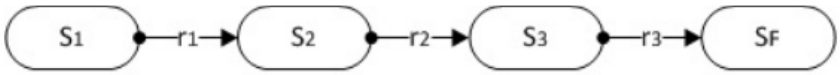
- $e_0(s) = 0$
- $e_t(s) = \gamma * \lambda * e_{t-1}(s) + \mathbf{1}(S_t = s)$

Keep an eligibility trace for every state s. Update value $V(S_t)$ with proportion to TD-error Δ_t and eligibility trace $e_t(s)$:

- $\Delta_t = R_{t+1} + \gamma * V(S_{t+1}) - V(S_t)$
- $V(S_t) = V(S_t) + \alpha_t * e_t(s) * \Delta_t$

Part 2: Exercises

Exercise 1. TD($\lambda = 1$) Given this episode, write down the update equations to the value function according to TD($\lambda = 1$) backward view algorithm:



Solution 1. TD(λ)

We start off at the beginning of an episode. We say that the eligibility for all the states is zero:

$e(s_1) = 0$
$e(s_2) = 0$
$e(s_3) = 0$

And the new value function is whatever the value function was at the end of the previous episode. So we just make a note of that. What we're going to do for this example is just keep track of the changes. So ultimately, this is all going to get added to whatever the previous value was: $V_{T-1}(s)$

Now we make a transition, and the first transition we make is from S_1 to S_2 with reward r_1.
Eligibility updates: $e(s_1) = 1$
$e(s_2) = 0$
$e(s_3) = 0$

States updates:
$\Delta V(s_1) = \alpha * (r_1 + \gamma * V_{T-1}(s_2) - V_{T-1}(s_1)) * 1$
$\Delta V(s_2) = 0$
$\Delta V(s_3) = 0$

Eligibility decay:

$e(s_1) = \gamma$
$e(s_2) = 0$
$e(s_3) = 0$

Now we take the next step, which is from S_2 to S_3 with reward r_2. Eligibility updates:
$e(s_1) = \gamma$
$e(s_2) = 1$
$e(s_3) = 0$

States updates:
$\Delta V(s_1) = \alpha * (r_1 + \gamma * V_{T-1}(s_2) - V_{T-1}(s_1)) + \alpha\,(r_2 + \gamma * V_{T-1}(s_3) - V_{T-1}(s_2)) =$
$= \alpha\left(r_1 + \gamma * r_2 + \gamma^2 * V_{T-1}(s_3) - V_{T-1}(s_1)\right) * \gamma$

$\Delta V(s_2) = \alpha * (r_2 + \gamma * V_{T-1}(s_3) - V_{T-1}(s_2)) * 1$
$\Delta V(s_3) = 0$

Eligibility decay:

$e(s_1) = \gamma^2$
$e(s_2) = \gamma$
$e(s_3) = 0$

The next step takes us from S_3 to S_F. We get reward r_3 when that happens.

Eligibility updates:
$e(s_1) = \gamma^2$
$e(s_2) = \gamma$
$e(s_3) = 1$

States updates:
$\Delta V(s_1) = \alpha \left(r_1 + \gamma * r_2 + \gamma^2 * V_{T-1}(s_3) - V_{T-1}(s_1)\right) + \alpha \left(r_3 + \gamma * V_{T-1}(s_F) - V_{T-1}(s_3)\right) * \gamma^2 = \alpha \left(r_1 + \gamma * r_2 + \gamma^2 * r_3 + \gamma^3 * V_{T-1}(s_F) - V_{T-1}(s_1)\right)$
$\Delta V(s_2) = \alpha * (r_2 + \gamma * V_{T-1}(s_3) - V_{T-1}(s_2)) + \alpha \left(r_3 + \gamma * V_{T-1}(s_F) - V_{T-1}(s_3)\right) = = \alpha \left(r_2 + \gamma * r_3 + \gamma^2 * V_{T-1}(s_F) - V_{T-1}(s_2)\right)$
$\Delta V(s_3) = \alpha * (r_3 + \gamma * V_{T-1}(s_F) - V_{T-1}(s_3))$

Exercise 2. TD(λ) A rat is involved in an experiment. It experiences one episode. At the first step it hears a bell. At the second step it sees a light. At the third step it both hears a bell and sees a light. It then receives some food, worth +1 reward, and the episode terminates on the fourth step. All other rewards were zero. The experiment is undiscounted.

1. Represent the rat's state s by a vector of two binary features, bell(s) $\in \{0, 1\}$ and light(s) $\in \{0, 1\}$. Write down the sequence of feature vectors corresponding to this episode.
2. Approximate the state-value function by a linear combination of these features with two parameters: $b * bell(s) + l * light(s)$. If $b = 2$ and $l = -2$ then write down the sequence of approximate values corresponding to this episode.
3. Define the λ-return V_t^λ,
4. Write down the sequence of λ-returns V_t^λ corresponding to this episode, for $\lambda = 0.5$ and $b = 2, l = -2$.

Solution 2. TD(λ)

1. $\begin{bmatrix}1\\0\end{bmatrix}, \begin{bmatrix}0\\1\end{bmatrix}, \begin{bmatrix}1\\1\end{bmatrix}$
2. $2, -2, 0$ and also 0 for the terminal state

3. $v_t^{(n)} = r_{t+1} + \gamma * r_{t+2} + ... + \gamma^{n-1} * r_{t+n} + \gamma^n * v(s_{t+n})$
   $v_t^{\lambda} = (1-\lambda) \sum_{n=1}^{\infty} \lambda^{n-1} v_t^{(n)}$

4. $v_1^{(3)} = v_2^{(2)} = v_3^{(1)} = 1^{n-1} r_4 + 1^n v(s_4) = 1 + 0 = 1$ (where s_4 is the terminal state and r_4 is the "food" reward from the third step).
   $v_1^{(2)} = v_2^{(1)} = 0$
   $v_1^{(1)} = -2$
   $v_1^{\lambda} = (1-\lambda)(\lambda^0 v_1^{(1)} + \lambda^1 v_1^{(2)} + \lambda^2 v_1^{(3)}) = 0.5 * (1 * (-2) + 0.5 * 0 + 0.5^2 * 1) = -7/8$
   $v_2^{\lambda} = (1-\lambda)(\lambda^0 v_2^{(1)} + \lambda^1 v_2^{(2)}) = 0.5 * (1 * 0 + 0.5 * 1) = 1/4$
   $v_3^{\lambda} = (1-\lambda)(\lambda^0 v_3^{(1)}) = 0.5 * (1 * 1) = 1/2$

References:

1. RL course, UCL, David Silver, Lecture 4-5
2. Practical Reinforcement Learning by Dr. Engr. S.M. Farrukh Akhtar