Introduction to Artificial Intelligence (INFO8006)

Exercises 7 – Reinforcement learning

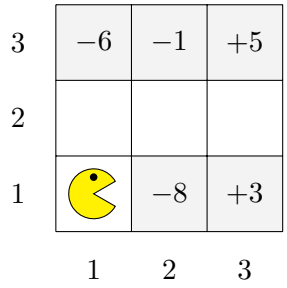
December 8, 2021

Learning outcomes

At the end of this session you should be able to

- differentiate passive and active RL;
- define and apply direct utility estimation and temporal-difference learning;
- define and apply Q-learning.

Exercise 1 Passive RL



Consider the grid-world given above and an agent who is trying to learn the optimal policy. The agent starts from the bottom-left corner and can take the actions `north` (N), `south` (S), `west` (W) and `east` (E). Rewards are only awarded for reaching the terminal (shaded) states. You observe the following trials, whose trajectories are sequences of tuples $(s_t^i, r_t^i, a_t^i, s_{t+1}^i)$.

t	Trial 1	Trial 2	Trial 3	Trial 4	Trial 5
0	$(1,1), 0, N, (1,2)$	$(1,1), 0, N, (1,2)$	$(1,1), 0, N, (1,2)$	$(1,1), 0, N, (1,2)$	$(1,1), 0, N, (1,2)$
1	$(1,2), 0, E, (2,2)$	$(1,2), 0, E, (2,2)$	$(1,2), 0, E, (2,2)$	$(1,2), 0, E, (2,2)$	$(1,2), 0, E, (2,2)$
2	$(2,2), 0, N, (2,3)$	$(2,2), 0, E, (3,2)$	$(2,2), 0, S, (2,1)$	$(2,2), 0, E, (3,2)$	$(2,2), 0, E, (3,2)$
3	$(2,3), -1, \varnothing, \varnothing$	$(3,2), 0, N, (3,3)$	$(2,1), -8, \varnothing, \varnothing$	$(3,2), 0, W, (2,2)$	$(3,2), 0, S, (3,1)$
4		$(3,3), +5, \varnothing, \varnothing$		$(2,2), 0, N, (2,3)$	$(3,1), +3, \varnothing, \varnothing$
5				$(2,3), -1, \varnothing, \varnothing$	

Assuming a discount factor $\gamma = 1$,

1. Perform direct utility estimation of the expected utilities $V^\pi(s)$, given the four first trials.
2. Update the estimated expected utilities with respect to the fifth trial using temporal-difference learning. Assume a learning rate $\alpha = 0.5$.

Exercise 2 Q-learning

An agent is in an unknown environment where there are three states $\{A, B, C\}$ and two actions $\{0, 1\}$. We are given the following tuples (s, a, r, s'), generated by taking actions in the environment.

s	a	r	s'
A	0	$+2$	A
C	1	-2	A
B	1	$+1$	B
A	0	-1	B
B	1	-2	C
C	0	$+4$	B
B	0	$+1$	A

Assuming a discount factor $\gamma = 0.5$ and a learning rate $\alpha = 0.75$,

1. Apply the Q-learning algorithm to obtain state-action-value $Q(s, a)$ estimates. Estimates are initialized to 0.
2. We now switch to a feature-based estimator $\hat{Q}(s, a) = w_0 + w_1 f_1(s, a)$, with $f_1(s, a) = 2a - 1$. Starting from weights $w_0 = w_1 = 0$, update the weights according to the approximate Q-learning algorithm.

Exercise 3 Football

ULiège's football team is playing against UCL's team next week. With a lot of losses this season, Liège needs to improve their attack strategy to win the game and increase their popularity. Luckily, the team captain follows INFO8006 and knows how to model the attack as a Markov Decision Process. The captain considers four states `close`, `away`, `fail`, and `goal` and two actions `pass` and `shoot`. Although the transition probabilities are unsure, the possible transitions (s, a, s') are known. To each transition is associated an increase/decrease of the team's popularity.

s	a	s'	$R(s, a, s')$
close	pass	close	$+1$
close	pass	away	-1
close	pass	fail	-2
close	shoot	close	$+3$
close	shoot	fail	-5
close	shoot	goal	$+10$
away	pass	close	$+2$
away	pass	away	0
away	pass	fail	-3
away	shoot	close	$+3$
away	shoot	fail	-10
away	shoot	goal	$+20$

The current strategy of the team is to always shoot. Last match, they had several attack opportunities, resulting in the following actions.

s	a	s'
close	shoot	goal
close	shoot	close
close	shoot	goal
close	shoot	fail
away	shoot	fail
away	shoot	close
away	shoot	fail
away	shoot	fail

Assuming a discount factor $\gamma = 0.75$ and a learning rate $\alpha = 0.25$,

1. Build an estimator $\hat{P}(s'|s, a)$ of the transition model and, from it, determine the expected utility V^π of the team's current policy π.
2. Perform direct utility estimation of the expected utility V^π. Do you observe a difference with the previous estimation? Why?

The captain found the tapes of the previous season where they had much more success. Together with the team, the captain selects the following instructive actions.

s	a	s'
close	pass	fail
close	pass	close
close	pass	away
close	pass	fail
away	pass	away
away	pass	close
away	pass	close
away	pass	fail

3. Apply the Q-learning algorithm to obtain state-action-value $Q(s, a)$ estimates. Estimates are initialized to 0.
4. Determine the optimal policy according to the state-action-value estimates.

Supplementary materials

- Playing Atari with Deep Reinforcement Learning

- Chapter 21 of the reference textbook.