Reinforcement Learning

tuyaux.winak.be/index.php/Reinforcement_Learning

Distributed AI / Reinforcement Learning

Richting	Informatica
Jaar	MINF
Studiepunten	6

Bespreking

Dit vak is een verplicht vak van de master.

Reinforcement Learning is in het jaar 2019-2020 helemaal veranderd tegenover de voorgaande jaren, met o.a. een nieuwe inhoud (Eerder heette dit van Distributed AI). Reinforcement Learning is een manier van computer gestuurd leren, waarbij de computer leert uit een 'environment' aan de hand van rewards. Misschien wel de bekendste toepassing hiervan is AlphaZero, een AI van Google die verschillende 'onmogelijke' taken succesvol heeft opgelost, zoals het spel Go tegen menselijke professionals winnen. Deze cursus begint opbouwend met basis technieken, maar tegen het einde van de cursus heb je een goede kennis van verschillende recente technieken in dit veld.

De evaluatie bestaat uit een theoretisch deel met schriftelijk gesloten boek examen en een praktisch deel in de vorm van wekelijkse taken waarin je de theoretische modellen moet in de praktijk brengen.

Puntenverdeling

Theorie : 10/20. Praktijk: 10/20. (Als 1 van de twee delen een buis is, dan wordt het totaal gedeeld door 2, dus onmogelijk om dan nog te slagen)

Examenvragen

Aangezien er nog geen examens van de nieuwe leerstof zijn geweest, heeft de prof wat voorbeeldvragen voorzien The exam will consist of a broad range of questions, where we aim at covering several chapters throughout the course: there will not be an emphasis on one particular chapter You can expect questions of the following types:

- Theoretical questions, where the answer can be found in the slides. Examples of these are

```
  - Explain algorithm X
  - What is the downside of approach A or B?
  - What type of techniques do you have available to boost e.g., exploration?
```

- Questions which compare different techniques:

```
    - why would you go for technique A vs B?
```

- Questions on applying an algorithm in a particular setting.

```
    - For example, given this MDP, how would value iteration/Q-learning/etc.
perform in these settings?
```

Juni 2022

1. UCB
    1. Explain how USB works and the impact of each of the terms used in the formula.
    2. What is the downside of UCB?
2. Discuss the following statement (that can be either right or wrong): "In a reinforcement learning task, an agent interacts with an environment defined by a Markov Decission Process. In each state, the agent performs an action, for which it receives a reward it aims to maximize."
3. Makov Decision Processes

There is a MDP with n states and a terminal state. From state i, you can go to state i+1 and get a reward of +1. You can also from state i to state 1 and get a reward of 0. From state n, you can also go to the terminal state and get a reward of +10.

1.
    1. What is the optimal policy?
    2. What is the optimal value of state n?
    3. What is the optimal value V∗V∗ for state 1, ... n-1?
    4. What is the value of the states for each nonzero value in the first and second iteration step with value iteration? (initial values are 0 for all stats)
    5. Assume the model is not known and we get the following samples:

- n-1, r, +1, n, r, +10, n
- n-2, r, +1, n-1, r, +1, n, r, +10, n

What is the value that Monte Carlo estimates based on those samples for all of the states?

1. Deadly Triad
    1. Explain the three elements of the deadly tried. What is the risk if you combine all three of them?
    2. Explain for each of the three elements why the combination of Q-Learning and Deep Neural Networks would lead to the deadly triad.
2. Yeet
    1. Explains how value-based, policy-based and actor-critic methods compare to each other in terms of policies and value functions.
    2. Give the pseudocode for Monte Carlo Policy Gradient (REINFORCE)

3. MCTS
    1. Give the 4 steps of MCTS, explain each of them and make a drawing for each of the steps.
    2. Is MCTS an example of background planning or decision time planning? Why?
4. Breakout

Assume the atari game of breakout. Your algorithm has choise from 3 actions: Move left, Move right or NoOp (No operation, don't move).

1.
    1. Is this a discrete or continious action space?
    2. Should we normally use a guassian or softmax distribution for this?
    3. Which of those models can you use to solve this (select all that apply): DQN, A2C/A3C, PPO

2020 Zit 2

Each question was about a different chapter, the questions were in order of chapters

1. Use UCB on the following three one-armed bandits in a sufficient number of iterations to show that they converge to the expected value. You can assume the sample average. The c value is 2.

```
Bandit 1 : 100% r = 5.
Bandit 2 : 100% r = 2.
Bandit 3 : 100% r = 1.
```

2. a) Use policy iteration on the following grid world, the positions with values are terminal states with the given reward. Assume the following parameters : noise = 0.3, living reward = 0, discounting factor = 1.

```
--------------------------------
|       | -100| -100| -100| 1000|
|       |     |     |     |     |
| start | -100| -100| -100| -100|
--------------------------------
```

b) If we set the discounting factor to 1.5, what should be the expected result and explain why

3. Explain Temporal Difference learning given a gridworld example.

4. Explain Q-learning given a similar gridworld example as question 3.

5. a) Explain Dueling DQN

b) Give the formules for Dueling DQN

6. a) Explain policy and the value function.

b) Explain how policy based, value based and actor critic compares to policy, and value function.

7. a) Give 2 uses for Monte Carlo Methods

b) Give the pseudocode from Every-visit Monte Carlo prediction

8. a) Give the difference between Intrinsic and extrinsic reward

b) Explain how intrinsic reward is used

2020 Zit 1

Each question was about a different chapter, the questions were in order of chapters

1. Use optimistic epsilon Greedy on the following three one-armed bandits in a sufficient number of iterations to show that they converge to the expected value. You can assume the sample average. The starting optimistic value is 4 for all machines. The epsilon value is 1/3.

```
Bandit 1 : 50% r = 2, 50% r = 1.
Bandit 2 : 90% r = 2, 10% r = -10.
Bandit 3 : 100% r = 1
```

2. a) Use value iteration on the following grid world, the positions with values are terminal states with the given reward. Assume the following parameters : noise = 0.3, living reward = 0, discounting factor = 1.

```
--------------------------------
|       | 10  | 10  | 10  | 1000|
|       |     |     |     |     |
| start | -100| -100| -100| -100|
--------------------------------
```

b) If we set the living reward to -5, what should be the expected result and explain why

3. The game donkey kong is a game where mario needs to rescue the princess from Donkey Kong, while avoiding barrels and other obstacles. The image is a static background. It is not feasible to use Q learning, nor do we want to use deep learning algorithms. How would you solve this problem?

4. Explain the use of Inverse Probability weighting in contextual bandits

5. a) What is a problem associated with DQN? Give two things we can do to mitigate this

b) Give the loss function of DQN and indicate those two solutions in the loss function.

6. a) Give two advantages of using policy networks instead of value networks

b) Give the pseudocode for the REINFORCE algorithm

7. a) Give the 4 steps of MCTS and explain

b) give an illustration for each step

8. An 8x8 gridworld, with a treasure, bombs. The input is presented as a 8x8x4 matrix, OBS[x,y,0] = true if there is a wall at position (x,y). OBS[x,y,1] = true if the position at (x,y) is free. OBS[x,y,2] = true if there is a bomb in position (x,y). OBS[x,y,3] = true if the treasure is at (x,y). We need to find the shortest path from the start location to the treasure without touching a bomb. Assume an episodic setting.

a) For an MDP, how would you construct the reward function and explain?

b) Which type Deep Neural Network would you use to solve this problem, explain?

c) Which deep algorithm would you use + explain?

9. Multiple choice : Add an X in the cells that are correct

	Discrete	Continue
Soft Max		
Gaussian policy		

	Values	Policy	Discrete	Continuous
DQN				
A3C				
PPO				

10. True / False? If true, motivate, if false explain why?

a) Monte Carlo methods update their values each timestep

b) Value iteration is faster than policy iteration, as the values would be calculated before the policy would be finished.

c) Exponential weighted moving average is is useful in a situation with non stationary distributions.

Categorieën:

- Informatica
- MINF