Reinforcement Learning 2022: Written Exam
SAMPLE 2022

Examiner: Aske Plaat

January 22, 2022

Reinforcement Learning Master Computer Science Leiden University 9 June 2022.

The exam text is in English. This is a multiple choice exam. Each question has one best answer. Indicate your answer clearly for each question using.

The exam starts at 09.00 hrs and ends at 12.00 hrs. Participation in the exam requires being present for at least 1 hrs.

There are 40 questions in the real exam, each best answer scores 1 point. The total score is converted to a grade afterwards. THERE ARE 20 QUESTIONS IN THIS SAMPLE EXAM.

No books, no electronic devices with communications functions, no smartphones, no smartwatches, no earphones, no cheating.

Write your name and student number clearly on your answer sheet.

Introduction

Question 1

Reinforcement Learning. Which of the following statements is true?

a. Reinforcement Learning learns a function from labeled examples in a pre-existing dataset.

b. Reinforcement Learning learns the inherent relations between items in a dataset.

c. Reinforcement Learning uses a number to score the quality of a state.

d. Reinforcement Learning environments are always programmed in Gym.

Tabular Value-Based

Question 2

What is the correct expression for the greedy policy?

a. $\pi(s) = \arg\max_a \sum_{s'} p(s'|s,a)Q(s,a)$

b. $\pi(s) = \arg\max_a V(s)$

c. $\pi(s) = \arg\max_a \sum_{s'} p(s'|s,a)r + \gamma \cdot Q(s,a)$

d. $\pi(s) = \arg\max_a Q(s,a)$

Question 3

You are teaching an algorithm to choose the correct action based on the state of an environment. Because you are lacking a dataset of states and the correct action to choose, you are instantly updating your values based on the state and reward pairs you are receiving from the environment. What kind of learning is this?

a. Online Reinforcement Learning

b. Supervised Learning

c. Offline Reinforcement Learning

d. Transfer Reinforcement Learning

Question 4

A function that, based on the current state of the environment and a taken action, determines the next state can be called:

a. Value-Function

b. Target-Function

c. Loss-Function

d. Transition-Function

Question 5

You have changed the gravity constant in the code of the Cartpole environment. What is affected by this?

a. Reward-Function

b. State Space

c. Transition-Function

d. Action Space

Question 6

Mike is facing a problem. Due to some data corruption his (S, a, r, S') tuple only retained S' and a. He would like to use this data to re-calculate S. What do you tell him?

a. That is impossible!

b. Use the environment formula to solve for the missing variables.

c. You're pretty sure you scrolled by some StackOverflow code recently that solved that problem.

d. None of the above

Question 7

Given the Q-table below, entry values are Q values. The agent just took action 2 from state s1 to state s2, and got a reward of 1(the episode is not terminated yet). According to the behavior policy, the agent will take action 2 in s2 for the next step. The discount factor is set to 1 and the learning rate to 0.5. What will the value for (s1, action2) be if you are using SARSA for updating?

state	action 1	action 2
s1	3	5
s2	5	3

a. 8

b. 5.5

c. 4.5

d. 7

Question 8

What is the difference between tabular Q-learning and SARSA?

a. Q-learning is on-policy and SARSA is off-policy.

b. The target policy in Q-learning is e-greedy, but in SARSA it is greedy.

c. The behavior policy in Q-learning is e-greedy, but in SARSA it is greedy.

d. The behavior policy and the target policy are different in Q-learning, but they are the same in SARSA.

Deep Value-Based

Question 9

Why is diversity important in learning?

a. Through de-correlation it improves stability in reinforcement learning

b. Through de-correlation it improves stability in supervised learning

c. Through correlation it prevents over-generalization in reinforcement learning

d. Through correlation it prevents over-generalization in supervised learning

Question 10

Which of the following DQN Extensions adresses overestimated action values?

a. Double DQN

b. Dueling DQN

c. Distributional DQN

d. Prioritized Action Replay

Question 11

Zhao is implementing a replay buffer for DQN and was wondering whether you had some tips regarding sampling methods. Your recommendation is:

a. Use uniform sampling

b. Use prioritized experience replay

c. Compare both to find out which one works best on his problem, as their performance varies per application

d. None of the above

Question 12

Which statement about the benefit of using DQN compared with tabular Q-learning is True?(pick the most convincing reason)

a. DQN can better deal with high-dimensional input.

b. DQN outperforms tabular Q-learning.

c. DQN is faster.

d. DQN is more data-efficient.

Policy-Based

Question 13

A3C. Which is true?

a. A3C is an efficient, distributed, implementation of Actor Critic

b. A3C is an asynchronous algorithm, is calculates multiple results in parallel

c. A3C can be used for Atari Learning Environment

d. All of the above is true

Question 14

Pick the correct statement.

a. Value-based RL is primarily applicable to discrete action space, policy-based RL is applicable to both discrete and continuous action spaces.

b. Value-based RL is primarily applicable to continuous action space, policy-based RL is applicable to both discrete and continuous action spaces.

c. Policy-based RL is primarily applicable to discrete action space, value-based RL is applicable to both discrete and continuous action spaces.

d. Policy-based RL is primarily applicable to continuous action space, value-based RL is applicable to both discrete and continuous action spaces.

Model-Based

Question 15

Latent model. Which is true?

a. Latent variables are confounding variables; latent models in reinforcement learning are based on these variables

b. Latent models train the model on value prediction

c. Latent models forego the actual models, and can therefore miss the policy

d. Latent models are like heuristics, they are based on rules of thumb

Two-Agent Self-Play

Question 16

Tabula Rasa vs Supervised. Which is true?

a. AlphaGo uses grandmaster games to learn by supervised learning. It also uses supervised learning for MCTS rollouts. AlphaGo Zero is based on this approach, and uses self-play.

b. AlphaGo uses grandmaster games to learn by supervised learning. It also uses supervised learning for MCTS rollouts, and reinforcement learning based on self-play games. AlphaGo Zero is based on this approach, and uses self-play.

c. AlphaGo uses grandmaster games to learn by supervised learning. It also uses supervised learning for MCTS rollouts. AlphaGo Zero is not based on this approach, and only uses self-play.

d. AlphaGo Zero is a clean sheet software engineering design, which caused the term: Tabula Rasa.

Question 17

How does UCT achieve trading off exploration and exploitation, which inputs does it use?

$$\mathrm{UCT}(j) = \frac{w_j}{n_j} + C_p\sqrt{\frac{\ln n}{n_j}}$$

a. The UCT formula balances *winrate* and *"newness"* in the selection of nodes to expand. A low C_p is more exploitation, a high C_p is more exploration.

b. The UCT formula balances *visits* and *"newness"* in the selection of nodes to expand. A high C_p is more exploitation, a low C_p is more exploration.

c. The UCT formula balances *visits* and *"newness"* in the selection of nodes to expand. A low C_p is more exploitation, a high C_p is more exploration.

d. The UCT formula balances *winrate* and *"newness"* in the selection of nodes to expand. A high C_p is more exploitation, a low C_p is more exploration.

Multi-Agent

Question 18

Counterfactual Regret Minimization. Which is true?

a. CFR achieves the minimax point in a two-agent competitive situation

b. The strongest two-agent Poker program, Libratus, is based on CFR. CFR is a probabilistic algorithm.

c. The strongest multi-agent Poker program, Pluribus, is based on CFR

d. All of the above are true

Hierarchical

Question 19

What is intrinsic motivation?

a. An inner drive to explore

b. Named so to contrast it with classic extrinsic motivation (the conventional RL reward signal)

c. Often related to model curiosity

d. All of the above

Meta-Learning

Question 20

What is pretraining?

a. Pretraining is what comes before posttraining: it initializes the network of weights

b. Pretraining is using a part of the knowledge of a network for the target network. It is followed by finetuning the target network

c. Pretraining is the opposite of posttraining. Posttraining finished the domain adaptation of the network

d. Pretraining is a difficult topic in deep learning, that can be solved by transfer learning

Future

- No questions in the sample exam -

Answers

1c
2d
3a
4d
5c
6a
7c
8d
9a
10a
11c
12a
13d
14a
15b
16c
17a
18d
19d
20b