Reinforcement Learning 2020: Sample Written Exam

Aske Plaat

December 25, 2019

This is a multiple choice exam. Each question has one correct answer. There are 12 questions, each can score 0 or 75 points. At most 900 points can be scored. Lowest possible grade is 1.0 (0 points), highest is 10.0 (900 points). Intermediate grades by linear interpolation.

Question 1

Describe the reinforcement learning model.

a. The agent has the current state, and the environment performs an action on the state, and updates the state to the agent. The environment reports a reward for this new state (which may be positive or negative).

b. The environment is the current state. The agent performs an action on the environment, resulting in a new state reported by the environment. Along with this new state comes a reward value (which may be positive or negative).

c. The agent knows the state and performs an action, resulting in a new state that it reports to the environment. Along with this new state comes a positive reward value.

d. The agent performs an action on the state, and reports it to the environment which returns a positive reward value for this state. Reinforcement learning is more powerful than supervised learning.

Question 2

What is the difference between on-policy learning and off-policy learning? Can you give one example algorithm for each?

a. On-policy learning uses two separate policy-arrays: one for exploratory behavior, and one to update as the current best policy. Example: SARSA. Off-policy learning updates actions directly on this single policy. Example: Q-learning

b. On-policy learning updates actions directly on its single policy. Example: SARSA Off-policy learning uses two separate policy-arrays: one for exploratory behavior, and one to update as the current best policy. Example: Q-learning.

c. On-policy learning uses two separate policy-arrays: one for exploratory behavior, and one to update as the current best policy. Example: Q-learning. Off-policy learning updates actions directly on this single policy. Example: SARSA

d. On-policy learning updates actions directly on its single policy. Example: Q-learning. Off-policy learning uses two separate policy-arrays: one for exploratory behavior, and one to update as the current best policy. Example: SARSA

Question 3

What causes the performance of alpha-beta to go from worst case to best case?

a. Move ordering determines the effectiveness of alpha-beta cut-offs. Iterative deepening, transposition tables, and heuristics all help move ordering.

b. Move ordering determines the effectiveness of alpha-beta cut-offs. Iterative deepening, piece-square tables, and null moves all help move ordering.

c. Null moves determine the effectiveness of alpha-beta cut-offs. Iterative deepening, transposition tables, and move ordering are frequently used heuristics.

d. Null moves determine the effectiveness of alpha-beta cut-offs. Iterative deepening, piece-square tables, and move ordering are frequently used heuristics.

Question 4

What is material balance?

a. The difference between the heuristic values of me and my opponent

b. The difference between the number of actions of my pieces and those of my opponent

c. The difference between the value function of me and of my opponent

d. The difference between the number and importance of my pieces and those of my opponent

Question 5

Describe two advantages of MCTS over rigid heuristic planning.

a. MCTS is inherently variable width variable depth, MCTS is a path-based algorithm

b. MCTS is a polynominal time algorithm, MCTS allows easy incorporation of heuristics

c. MCTS is an anytime algorithm, MCTS is a recursive algorithm

d. MCTS is an iterative algorithm, MCTS uses an evaluation function

Question 6

Give the UCT formula for node i, child j, visit count n and win count w

a.
$$\text{UCT}(j) = \frac{n_i}{n_j} + C_p\sqrt{\frac{\ln n}{w_j}}$$

b.
$$\text{UCT}(j) = \frac{w_j}{n_j} + C_p\sqrt{\frac{\ln w}{w_i}}$$

c.
$$\text{UCT}(j) = \frac{n_i}{n_j} + C_p\sqrt{\frac{\ln n}{n_j}}$$

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

Question 7

What is end-to-end learning? Do you know an alternative? What are the advantages of each?

a. The learning from output labels or actions based on the raw inputs, un-pre-processed, without intermediate hand crafted heuristics. End-to-end functions can overcome manual bias, but intermediate heuristics use less compute power.

b. The learning from output labels or actions based on heuristics, pre-processed, instead of un-processed pixel input. End-to-end functions are faster to compute, but may suffer from bias.

c. The learning from input labels or states based on pixel outputs, un-pre-processed, without intermediate hand crafted heuristics. End-to-end functions can overcome manual bias, but intermediate heuristics use less computee power.

d. The learning from input labels or states based on pixel outputs, un-pre-processed, with the help of intermediate hand crafted heuristics. End-to-end functions are faster to computer, but may suffer from bias.

Question 8

What is the role of the replay buffer?

a. It stops early during the training, in order to stabilize training by decorrelating the training examples in each batch used to update the neural network. It cannot overcome the deadly triad

b. It stores past experiences and is used to stabilize training by decorrelating the training examples in each batch used to update the neural network. It overcomes the deadly triad

c. It stores past experiences and is used to stabilize training by decorrelating the training examples in each batch used to update the neural network. It cannot overcome the deadly triad

d. It stops early during the training, in order to stabilize training by decorrelating the training examples in each batch used to update the neural network. It overcomes the deadly triad

Question 9

What is neural architecture search and how does it apply to reinforcement learning?

a. A technique for combining neural network optimization with search algorithms. Traditionally, search algorithms were used in reinforcement learning, NAS combines thee best of both worlds.

b. Combining Evolutionary Strategies with Artificial Neural Networks. Stochastic Gradient Descent is the most popular optimization strategy in neural networks, evolutionary strategies are winning in popularity.

c. A technique for automating the design of artificial neural networks. Designing a neural network for reinforcement learning is hard, and NAS automates part of the task.

d. Deep reinforcement learning uses neural networks for the evaluation function. Evaluation function architectures are thus optimized.

Question 10

Why are games with sparse rewards difficult for reinforcement learning?

a. The intersection of knowledge representation and reinforcement learning is an active area of research. This is driven by the desire for explainable AI. This field is related to learning Bayesian networks and belief networks.

b. Sparse rewards are often encountered in real life. Examples are highly prevalent in robotics, where an arm movement can be over a continuous angle.

c. Delayed credit assignment is one of the challenges in reinforcement learning. The problem is large periods in which little signal occurs to guide the search. Games that required more long range planning, such as Montezuma's Revenge, still are a problem.

d. You first start out with only easy examples of a task and then gradually increase the task difficulty.

Question 11

What is the main result of the AlphaZero paper?

a. It learns to play Go from scratch, tabula rasa, without any per-encoded heuristic knowledge.

b. It is able to learn to play from scratch in Go, Chess and Shogi, beating the best players, including conventional heuristic planning players.

c. It is the first program to beat human world champions.

d. It is the first to learn model-free combinatorial games.

Question 12

Explain the search-eval architecture briefly.

a. The search part is to determine the value of a state, recursively follow all actions to all successor states and then the eval part is to find their values using an evaluation function, that may be heuristic or learned. The architecture combines the strengths of exact algorithms, searching, with approximate algorithms, learning/heuristics.

b. Combine planning and learning in an open architecture.

c. Use approximation algorithms for small state spaces, and adaptive sampling for large state spaces. Their advantage is variable width, variable depth search.

d. Use minimax and heuristics in a combined architecture.

Correct answers: 1b, 2b, 3a, 4d, 5a, 6d, 7a, 8b, 9c, 10c, 11b, 12a