Intelligent Agents
2022
Quiz 1
20. October 2022

- place your student ID card (carte de legitimation) on the desk in front of you.
- this is a closed-book examination (no documents allowed).
- when choosing the right answer, consider that the given explanation also has to be correct.
- mark the number of your copy on the top of each page to make sure we identify all pages of your exam.
- for questions with a single answer, the correct answer gives you 2 points. For questions with multiple correct answers, each correct answer gives you one point.
- If you give an incorrect answer, the question gives 0 points (even if there one or more of the answers is correct).

Copy No:

1. Which statements are **not** true in a POMDP:

   a) states and actions are known and finite
   b) state transitions are deterministic
   c) the state is observed with certainty
   d) rewards are known with certainty

   Your answer:

2. Why is the discount factor important?

   a) because agents should not be influenced too much by future rewards.
   b) because it makes the sum of discounted rewards over an infinite time horizon computable by a recurrence.
   c) as a parameter to tune agent policies.
   d) to model the uncertainty of future rewards.

   Your answer:

3. How do we recognize that value iteration has converged?

   Between two successive iterations over all states:

   a) the value function does not change anymore
   b) changes in the value function are bounded by a certain value
   c) the average change in the value function is bounded by a certain value
   d) the optimal policy for this value function does not change.

   Your answer:

4. What is the definiton of cumulative regret?

   a) Average difference between the reward of the best possible action and the reward of the action actually taken.
   b) For a sequence of actions, difference between the sum of rewards of the best possible policy and the actual policy.
   c) For a sequence of actions, difference between the sum of rewards of the best possible fixed policy and the action actually taken.
   d) Average difference between the reward of the best possible policy and the reward of the actual policy.

   Your answer:

5. What are the assumptions underlying confidence bounds?

   a) rewards are distributed according to a Gaussian and bound holds with probability $(1 - \delta)$
   b) samples are statistically independent and bound holds with probablity $(1 - \delta)$.
   c) samples are statistically independent and rewards are distributed according to a Gaussian.
   d) optimism under uncertainty: reward of an action is equal to the upper confidence bound.

   Your answer:

6. Which of these learning algorithms require full observability ? (answer all that apply)?

   a) Q-learning
   b) regret matching
   c) multiplicative weight updates
   d) exponential weight updates

   Your answer:

7. When should we use a deliberative rather than a reactive agent (answer all that apply)?

   a) when rewards and state transitions are very uncertain.
   b) when only very few of the possible states are ever visited in the lifetime of the agent.
   c) when there are strict real-time constraints.
   d) when the rewards (goals) are changing frequently.

   Your answer:

8. When using A* to search for the fastest plan for moving 5 boxes from the same starting point to different destinations in a graph, which of the following heuristics is admissible? Assume that the agent can carry any number of packages and moves at a fixed speed of 1m/s.

   a) The sum of the lengths of the shortest paths from the starting point to the destinations, for the packages not yet delivered.
   b) The maximum of the lengths of the shortest paths from the starting point to the destinations, for the packages not yet delivered.
   c) The minimum of the lengths of the shortest paths from the starting point to the destinations of the packages.
   d) The maximum of the lengths of the shortest paths from the position of the agent to the destination of a package that has not yet been delivered.

   Your answer:

9. How are the actions chosen in counterfactual regret minimization?

   a) play each action a with probability proportional to the average of observed differences in reward for a and the actually played action a'.
   b) play the action a that maximizes the average of observed differences in reward for a and the actually played action a'
   c) play the action a that minmizes the average of observed differences in reward for a and the actually played action a'
   d) play the action a that in the past was most often the optimal one.

   Your answer: