CSCE 475/875 Multiagent Systems

Examination - Solution

September 28, 2017

Name:

NUID:

☐ Undergraduate ☐ Graduate

You have 75 minutes to complete the examination.

1. (25 total points) Agency.

(a) (5 points) What is an agent?

Solution: An agent is an entity that senses its environment, makes autonomous decisions based on the sensory input, and actuates the decisions that in turn change the environment.

(b) (5 points) What is an intelligent agent?

Solution: An intelligent agent is an agent capable of flexible behavior: reactive (responsive in a timely manner), proactive (goal-directed) and social. Another definition calls for an agent to be capable of learning in order to be intelligent.

(c) (5 points) What is a multiagent system?

Solution: A multiagent system is an environment where there are multiple agents interacting directly or indirectly such that the system designer obtains an overall system result, usually one that cannot be achieved with each individual agent alone.

(d) (10 points) Under which ***five*** environment characteristics is it more appropriate to use an **agent-based** solution? Identify a problem and describe its environment with respect to these five characteristics.

Solution: Dynamic (environment changes while decisions are being made), Partially observable (or incomplete information), Uncertain (or stochastic such that the same action does not lead to the same outcome), Non-episodic (that state values depend on past history), and Continuous (where the number of states is infinite). When the environment is dynamic, or partially observable, or uncertain, or non-episodic, or continuous, it is more appropriate to use an agent-based solution so that the autonomy and decentralized reasoning can address these complex characteristics better. A smart grid simulation problem: dynamic, partially observable, uncertain,

probably episodic, and the states can be quite discrete (e.g., breaking a day's electricity usage to different phases: after getting up, away for work, back from work, sleep).

2. (25 points) Game Theory.

(a) (10 points) Consider the following Theorem:

Theorem 3.1.8 (von Neumann and Morgenstern, 1944) *If a preference relation* $\succcurlyeq$ *satisfies the axioms completeness, transitivity, substitutability, decomposability, monotonicity, and continuity, then there exists a function* $u: O \rightarrow [0, 1]$ *with the properties that*

1. $u(o_1) \geq u(o_2)$ *iff* $o_1 \succcurlyeq o_2$, *and*
2. $u([p_1: o_1, \ldots, p_k: o_k]) = \sum_{i=1}^{k} p_i u(o_i)$

What does $\sum_{i=1}^{k} p_i u(o_i)$ stand for? How can an agent make use of this for its decision making? (*Hint*: Use an example to illustrate.)

Solution: $\sum_{i=1}^{k} p_i u(o_i)$ stands for the expected utility of all outcomes. An agent can make use of this expected utility for its decision making by always choosing an action that will *maximize* this sum. For example, if an agent is given two choices: Option 1 and Option 2. Option 1 guarantees a reward of $10 with probability 1.0. Option 2 offers a reward of $100 with probability 0.05 and a reward of $0 with 0.95. Then, the expected reward from Option 1 is $10 while that from Option 2 is 0.05*$100 + 0.95*$0 = $5. An agent that maximizes the sum will thus choose Option 1 over Option 2.

(b) (10 points) Consider the normal form game ***Prisoner's Dilemma***:

	Player 2 No Betray	Player 2 Betray
Player 1 No Betray	1,1	-4,3
Player 1 Betray	3,-4	-3,-3

Why is it a ***difficult*** game to play? (*Hint*: What is each player's dominant strategy? Is there a Nash equilibrium? If yes, which one? What is the state with the best joint outcomes? What is the state with the worst joint outcomes?)

Solution: Each player's dominant strategy is to betray. Yes, there is a Nash equilibrium, the state where both players betray, resulting in -3,-3 to both players. The state with the worst joint outcomes is where both players betray, and the state with the best joint outcomes is where both players do not betray. This is a difficult game to play because each player's dominant strategy is to betray, as it yields better reward regardless of whether the other player betrays (-3 better than -4) or does not betray (3 better than 1). Yet, if both players do that, they will land in the state with the worst joint outcomes: (-3,-3), for a total of -6. And once in that state, each will NOT have incentive to deviate—if a player changes its action from Betray to No Betray, then it will receive even worse outcome while the other player benefits—from that outcome because it is also a Nash equilibrium. In a way, the players are motivated to do something better for themselves but

yet, they find themselves in a worse state, and to make things worse, once they are there, each is not motivated to move out of that state.

(c) (5 points) How is Nash Equilibrium related to agent design or reasoning. (*Hint*: Think about best response.)

Solution: When an agent decides on its best actions, it cannot do so without considering how other agents make their decisions. That is, an agent should consider the best responses of all agents to the best responses of all other agents. That is, an agent's decision is only viable if it is the best response to all other agents' decisions, such that every agent's decision is also its best response to all other agents' decisions. Only then will the joint decisions or policy be stable for agents to benefit from the rewards. A Nash equilibrium is where every agent's decision is its best response to the joint decision/action profiles of all other agents.

3. (25 points) Learning. Consider the Q-learning algorithm.

Initialize the Q-function and V values (arbitrarily, for example)
Repeat *until convergence*
  Observe the current state s_t.
  Select action a_t and take it.
  Observe the reward $r(s_t, a_t)$ and next state s_{t+1}
  Perform the following updates (and do *not* update any other Q-values):

$$Q_{t+1}(s_t, a_t) \leftarrow (1 - \alpha_t)Q_t(s_t, a_t) + \alpha_t(r(s_t, a_t) + \beta V_t(s_{t+1}))$$

$$V_{t+1}(s) \leftarrow \max_a Q_t(s, a)$$

End Repeat

(a) (7 points) Explain the purposes of the factors α_t and β.

Solution: α_t is the learning rate. It is used to control the amount of "trust" in the knowledge that an agent has learned so far. If the value is low, then it means the agent will tend to adhere to what it has learned before. If the value is high, then the agent will be more willing to weigh current rewards more. β is the discount factor on the future "expected" value. A higher value of β means the agent is more willing to lookahead into the future, and vice versa. In other words, an agent with a higher learning rate will tend to explore more (more adaptive to the environment), and an agent while an agent with a lower discount factor will tend to be more myopic.

(b) (8 points) Explain the intuition behind the ***iterative update*** for the Q value, especially the ***roles*** of the three terms, $(1 - \alpha_t)Q_t(s_t, a_t)$, $\alpha_t(r(s_t, a_t))$, and $\alpha_t\beta V_t(s_{t+1})$, in the equations.

Solution: The intuition behind the update for the Q value is to combine what an agent has learned before, what it has just received as reward to its action, and then what it projects that it would receive in the future from being in the resulting state. The value of each state is also updated as the maximum of the Q value that the state is able to garner over all possible actions.

The first term is about the existing knowledge, allowing an agent to "reuse" what it has learned.

The second term is the immediate reward to an action that an agent has done while in a state, allowing an agent to focus and exploit the actual reward received. The third term is the value of the resulting state, allowing an agent to lookahead in the future when considering the Q-value of a state-action pair.

(c) (10 points) What is the central idea of the ***exploration vs. exploitation*** tradeoff? How is it related to Q-learning?

Solution: The central idea of the tradeoff lies in when to stop exploration and start exploiting knowledge gained up to now. Stop exploration too soon could lead to an agent executing or choosing to perform sub-optimal solutions. Start exploiting too late might not allow an agent to recover lost rewards and end up with less cumulative rewards in the end. It is related to Q-learning in several ways. First, how we select the actions to do in the Q-learning process allows us to recognize whether an agent is exploring or exploiting. Second, once the Q-table converges, an agent can start exploiting by choosing the action that gives the highest Q value for a given state. Third, during exploration, the learning rate should be set high; during exploitation, it should be set lower.

4. (25 points) Hodgepodge.

(a) (8 points) Consider the ***Matching Pennies*** game and the learning strategy ***Fictitious Play***. Consider also current status of the game:

		Player 2	
		Heads	Tails
Player 1	Heads	$1, -1$	$-1, 1$
	Tails	$-1, 1$	$1, -1$

Heads Tails

Round	1's action	2's action	1's beliefs	2's beliefs
0			(1.5,2)	(2,1.5)
1	T	T	(1.5,3)	(2,2.5)
2	T	H	(2.5,3)	(2,3.5)
3	T	H	(3.5,3)	(2,4.5)
4	H	H	(4.5,3)	(3,4.5)
5	H	H	(5.5,3)	(4,4.5)
6	H	H	(6.5,3)	(5,4.5)
7	H	T	(6.5,4)	(6,4.5)
⋮	⋮	⋮	⋮	⋮

For Round 8, what are Player 1's action, Player 2's action, Player 1's beliefs, and Player 2's beliefs? (Show all work.)

Solution: Player 1's action is H, because it believes that Player 2 will more likely play Heads than Tails (6.5 vs. 4). Player 1 gets more reward (1 vs. -1) from matching pennies. Player 2's action is T, because it believes that Player 1will more likely play Heads than Tails (6 vs. 4.5). Player 2 gets more reward (1 vs. -1) from *not* matching pennies. After both actions, Player 1 will update its beliefs to (6.5,5), and Player 2 will update its beliefs to (7,4.5).

(b) (7 points) ***Learning*** and ***Communication*** have a special relationship. Explain.

Solution: Learning can be used to improve communication by learning which agents to communicate, and how to communicate. Meanwhile, communication can be used to facilitate learning by sharing information or data between agents.

(c) (7 points) How can the ***Bayes theorem*** help agents execute ***Bayesian learning*** (a.k.a. rational learning)? ***Justify*** your response. (*Hint*: Think about "direct computability" of the "posterior" term; and use the *medical diagnosis* example.)

$$P(H|e) = \frac{P(e|H)P(H)}{P(e)}$$

Solution: The ***posterior*** term is *not* directly computable. This is because it is possible that for the same piece of evidence can be observed for multiple hypotheses. For example, in medical diagnosis, if a patient is observed to have the coughing symptom, it could be because of Cold, Flu, Allergy, Pneumonia, or other diseases. The Bayes theorem allows us to break down the posterior into three parts that are computable from historical data: P(e|H) is the ***likelihood***: How often did we observe coughing when a patient has a cold? How often did we observe coughing when a patient has flu? Etc. P(H) is the ***prior***: How many cases of coughing did we observe? How many cases of flu did we observe? P(e) is the ***marginal***: How often did we observe coughing as a symptom in all cases that we observed? An agent can observe its environment and collect data to directly compute for these three parts. And then using the Bayes theorem, it can then compute the posterior P(H|e), which in turn allows it to make a diagnosis or decision by choosing the posterior with the highest value.