CS 188 – Artificial Intelligence Notes (for Reinforcement Learning)

Pranay Pasula

July 19, 2019

Lecture 8: Markov Decision Processes I

1. How is an MDP defined?
2. How does an MDP relate to a general search problem?

3. What is Markovian about MDPs?
4. What does this imply about how one should design an MDP?

5. What is usually the main goal in a deterministic search problem?
6. What is usually the main goal in an MDP?

7. What is a policy?
8. What does an explicit policy define? Explain what this is.
9. How does expectimax differ from finding an optimal policy?
10. How does expectimax relate to finding an optimal policy?

11. What is an MDP search tree?
12. How does an MDP search tree relate to an expectimax search tree?
13. How does an MDP search tree differ from an expectimax search tree?

14. What is discounting?
15. What issue may arise without discounting? What other condition must hold for this to be possible?

16. What are stationary preferences?
17. If we assume stationary preferences, what is implied about the ways to define utilities?

18. What is a potential issue if an MDP can last forever (and accrue infinite rewards)? What are three ways to address this? What are the two most common ways to address this? How does the third way relate to the probability of the MDP terminating?

19. What are the three optimal values of interest when solving MDPs? Explain the meaning of each one.

20. How many Q-values are associated with each state?
21. How do Q-values relate to values?

22. How can we find the value of a state?
23. What is the formula for the value of a state?
24. What is the formula for the Q-value of a pair (s, a)?
25. What are two potential issues with using just expectimax to solve an MDP (hint: consider an infinite-length MDP)? How can we address these?

26. What are time-limited values? How can we quickly compute $V_i(s)$ for any $i \in [0, 1, ..., k]$? In general, for what value of i is $V_i(s) = V^*(s)$?

27. Explain the value iteration algorithm.
28. What is its computational complexity?
29. Is the solution unique?
30. Is the solution optimal?
31. What is the main problem with value iteration?
32. Does value iteration always converge?
33. How does the convergence of the values relate to the convergence of the policy?
34. Does value iteration converge to the optimal policy?
35. How does adjusting γ affect the convergence rate?

Lecture 9 - Markov Decision Processes II

36. What is policy evaluation?
37. What is a fixed policy?
38. Explain two ways to find the values for a fixed policy.
39. How does this compare to value iteration?
40. What is the complexity of the iterative approach to policy evaluation?
41. For the non-iterative approach, what kind of matrix is usually involved? (Hint: usually enables a computer to find the solution quickly)

42. Given the optimal value for each state, can we determine the optimal policy without any calculations?
43. Explain how to determine the optimal policy given optimal value for each state.
44. Explain how to determine the optimal policy given the Q-values for each state.

45. State three problems with value iteration.
46. How does the rate of changes of "max" actions differ for early, middle, and late stages of value iteration?
47. What are the two main steps of policy iteration? Explain each.
48. Does policy iteration converge to the optimal policy? Does it converge to the optimal values? Explain why.
49. Why use policy iteration instead of value iteration? For what type of MDP is policy iteration especially preferred over value iteration?
50. Does policy iteration result in a better solution than value iteration?
51. Explain how to find the optimal policy using value iteration.
52. Explain how to find the optimal values using policy iteration.
53. When does policy iteration terminate? How does $\gamma \in (0, 1)$ affect this?
54. What is a multi-armed bandit?

55. What kind of rewards do choices in the multi-armed bandit problem usually have? How does this affect the corresponding MDP?
56. Explain how to find the optimal policy of a multi-armed bandit MDP given all of the details of the MDP. What is this called? State why this is advantageous to having only partial information of the MDP.
57. Explain how to find the optimal policy of a multi-armed bandit MDP given only partial details of the MDP. What is this called?