SmartCab Reinforcement Learning Report

QUESTION: Observe what you see with the agent’s behavior as it takes random actions. Does the smart cab eventually make it to the destination? Are there any other interesting observations to note?

When taking turns randomly, the car does not reliably reach its destination. Sometimes it gets lucky, but it usually runs out of turns before successfully reaching the destination. After watching the visual simulation in Pygame, it appears rare for the car to encounter traffic from the left or right. These inputs are unlikely to be useful when implementing the learning algorithm.

After running 100 iterations (enforce_deadline=True) of the trial with random actions, only 18 (18%) reached the destination.

QUESTION: What states have you identified that are appropriate for modeling the smart cab and environment? Why do you believe each of these states to be appropriate for this problem?

I set the state to a tuple with the with the indices representing Light, Waypoint, and Oncoming traffic. In total, this provides 96 possible combinations for the state. I chose this pattern because it maximizes information while minimizing complexity. It also accounts for all reasonable traffic laws and collision avoidance in the environment.

I could have added right traffic to the tuple to bring it to 384 possible combinations. However, in the United States oncoming traffic from the right would not cause an issue for driver, so this has been left out. I should also note that Deadline was left out because it would increase the number of states beyond a reasonable number. Adding a deadline or right traffic would cause the model to suffer from the Curse of Dimensionality.

```
example_state = ( light, waypoint, oncoming, left )
>>> ('red', 'left', None, None)
```

- Light: Necessary for training the cab to obey traffic rules.
- Waypoint: Necessary for providing a general sense of direction. I imagine this input like a passenger in the back seat giving left/right/forward directions to the driver. Without this data, the driver would still be randomly searching for the endpoint.
- Oncoming Traffic: Necessary for avoiding traffic accidents. Of the three traffic inputs, oncoming appears to contain the most information.
- Left Traffic: Necessary for obeying right of way rules. Oncoming traffic to the left would have the right of way in this scenario.

OPTIONAL: How many states in total exist for the smart cab in this environment? Does this number seem reasonable given that the goal of Q-Learning is to learn and make informed decisions about each state? Why or why not?

Total features can be determined by multiplying the total count of possible values from each input.

inputs = light(2), waypoint(3), oncoming(4), left(4), right(4)

With the inputs above, possible state combinations are 2x3x4x4x4 = 384.

If you add another state for each number of turns left in deadline(50), possible states go up to 19,200.

With 100 turns to learn, 19,200 is not reasonable. The driver would not enter many of the possible states during the course of the trials. Most of this information should be consolidated into a simplified structure.

QUESTION: What changes do you notice in the agent's behavior when compared to the basic driving agent when random actions were always taken? Why is this behavior occurring?

Updating the Q table with the current reward value creates a much more effective driver. The algorithm does not consider the impact of future moves, but it still has a much higher success rate compared to complete randomness. Visually, the smartcab appears to be more patient and deliberate with its actions. It drives in straight lines and is not afraid to wait at traffic lights. After 100 trials, the driver failed to reach the destination only 10 times for a 90% success rate.

QUESTION: Report the different values for the parameters tuned in your basic implementation of Q-Learning. For which set of parameters does the agent perform best? How well does the final driving agent perform?

- α (alpha) is the learning rate. A higher alpha value will make the algorithm more sensitive to changes in rewards.
- γ (gamma) is the discount rate on future rewards. A higher gamma will make the algorithm more sensitive to rewards from future actions.
- ε (epsilon) is the exploration rate. A higher epsilon will cause the algorithm to make more random choices.
- λ (lambda/decay) is the decay constant to reduce epsilon for (K) number of trials. A higher lambda will cause epsilon to decay faster, reducing randomness earlier in the trial run.

I ran experiments with the tuning parameters listed below. Each experiment ran with 100 trials.

α=0.3, γ=0.3, ε=1.0, λ=0.01 (Random/Explorer Learner)

The parameters above are designed to create a random explorer. At the beginning, the model will be almost completely random, gradually decaying by 1% each turn. This policy resulted in a 46 failed runs, or a success rate of 54%. This is an improvement, but I suspect the slow decay rate of epsilon is causing the algorithm too behave with too much randomness.

α=0.5, γ=0.5, ε=1.0, λ=0.1 (Balanced Learner)

For the second trial run, the epsilon decay rate was increased to 10% per trial. This combination eliminates much of the randomness in the previous example. For instance, the model from 100% random at turn 0, to 31% random by turn 10, and 0.4% by turn 50, and so on. In terms of learning, it puts an equal weight on the current action reward and the future action reward.

The failed runs were reduced to only 16, for a total success rate of 84%. The balanced learner puts the same amount of weight on the reward in the current move as it does for the future move.

α=0.8, γ=0.2, ε=0.8, λ=0.1 (Short-Term/Greedy Learner) WINNER

The lowered epsilon value makes this a greedy combination that does make exploratory moves often. The alpha was increased to put a heavier weight on new information, while the gamma was decreased to make future rewards less import to the driver.

Short term learning worked well, with only 4 failed runs or a 96% success rate. Comparatively, this combination leads to a policy with more negative Q value rewards, or discouragement from performing bad moves. Overall, the Short-Term/Greedy Learner appears to the optimal policy for the smart cab problem.

α=0.2, γ=0.8, ε=0.8, λ=0.1 (Long-Term/Greedy Learner)

To test the relationship between alpha and gamma, I left the epsilon values unchanged for this trial run. The learning rate was decreased to make it harder to change established values, while gamma was increased to put a higher weight on future rewards. This model should be more farsighted and not as

This policy leads to larger positive Q values, or encouragement for successful moves. It failed 13 runs, for a 87% success rate. By the end of the 100 trials, this policy was still receiving frequent negative rewards because it failed to recognize issues or opportunities in the current state.

QUESTION: Does your agent get close to finding an optimal policy, i.e. reach the destination in the minimum possible time, and not incur any penalties? How would you describe an optimal policy for this problem?

The optimal policy turned out to **Short-Term/Greedy** strategy.

Towards the end of the 100 trials, the agent is far less likely fail and the total reward received was always greater than 0. Most failed runs are found within the first 20 attempts. In fact, it is common for the driver to avoid all negative rewards during many of the later trials, which is optimal behavior.

The agent learns very quickly with the input states provided, which leads me to believe that a short term policy is optimal. It only takes about 5 to 10 turns for the Q table to be filled with effective knowledge about the most common turns. There does not appear to be any intricate patters between the states that justify a more complex model. In other words, the learning rate (α) should be high, while the gamma (γ) and epsilon (ε) should be low. This combination will force the algorithm to weigh its decisions mostly on the current move. Only during times when the next move has a significant reward will the resulting action be influenced.

The sample of results below show the difference between the rewards received on trials 1, 10, 50, and 100. The mean reward received gradually increases with the number of trials and negative rewards become less common.

Trial 1 Rewards

Mean: –0.125

Min: –1.0

Max: 2.0

Trial 10 Rewards

Mean: 0.71

Min: –1.0

Max: 2.0

Trial 50 Rewards

Mean: 1.43

Min: 0.0

Max: 2.0

Trial 100 Rewards

Mean: 1.64

Min: 0.0

Max: 2.0