Treasure Hunt Game

To start a human being would solve this maze by using their past experience, as well as what they can see several squares ahead to solve the maze and find the treasure. The agent learns in a similar way, learning through actions made. Each state corresponds to a different location on the map, and at each state the agent maps a state to an action. After each action the agent is rewarded or punished, with the goal being the maximization of finding the treasure as quickly as possible. As the agent explores its environment, it updates an action value (determined by how efficient the action was) based on the negative or positive feedback the agent receives. Upon gaining more experience the agent attempts to get its action values to meet the optimal values, and as it learns more it becomes more accurate.

Exploration vs exploitation is sort of simple. After doing a bit of research, exploitation is a “purely greedy approach” (towardsdatascience.com”) Essentially the idea is that you force the agent to take the greediest route possible and hope it pays off and it makes this decision based on it’s current knowledge. Whereas exploration is used to essentially allow the agent to explore like Indiana Jones and hopefully it finds the correct route, this way however allows the agent to learn more about it’s surroundings, so it will pay off more later.

In terms of implementing deep Q-learning and using neural networks, Deep Q network was used to map action-state pairs. Using the following code, each action was punished or rewarded, to attempt to make the agent as accurate as possible:

def get\_reward(self):

pirate\_row, pirate\_col, mode = self.state

nrows, ncols = self.maze.shape

if pirate\_row == nrows-1 and pirate\_col == ncols-1:

return 1.0

if mode == 'blocked':

return self.min\_reward - 1

if (pirate\_row, pirate\_col) in self.visited:

return -0.25

if mode == 'invalid':

return -0.75

if mode == 'valid':

return -0.04  
  
The agent’s learning was based on a decaying learning pattern. I could have made these values a bit better as the agent did not being winning 100 percent of the time until epoch 624. The final path for the agent finding the treasure is shown in **figure 1a** below.

A black and white crossword puzzle

Description automatically generated with medium confidence  
**figure 1a**

Sources

Lindwurm, E. (2019, November 3). *Intuition: Exploration vs exploitation*. Medium. https://towardsdatascience.com/intuition-exploration-vs-exploitation-c645a1d37c7a