Q-learning Assignment

Kyle du Plessis [DPLKYL002]

CSC3022H

9/26/19

The results had been written to a .csv file (results.csv) to be examined after running each test simulation.

"Results table: For tests 1, 2, and 3: maximum number of mines swept (from 50 runs) and average number of mines swept (over 50 runs)."

Mine-sweeper task performance	Test 1	Test 2	Test 3
Average number of mines swept	15.79	11.60	2.49
Maximum number of mines swept	23	21	5

"A brief justification (200 words maximum) of the Q-learning parameter values you selected and how these parameter values influenced learning and mine-sweeping task performance in each test simulation."

```
double CQLearningController::R(uint x, uint y, uint sweeper_no) {
	//TODO: roll your own here!
	double rewardValue = -10;
	int nearestObject = (m_vecSweepers[sweeper_no])->CheckForObject(
			m_vecObjects, CParams::dMineScale);
	if (nearestObject > -1) {
		switch (m_vecObjects[nearestObject]->getType()) {
		case CCollisionObject::Mine: {
			if (!m_vecObjects[nearestObject]->isDead()) {
				rewardValue = 100;
			}
			break;
		}
		case CCollisionObject::Rock: {
			rewardValue = -50;
			break;
		}
		case CCollisionObject::SuperMine: {
			rewardValue = -500;
			break;
		}
		}
	}
	return rewardValue;
}
```

Q-learning parameter values:

`double rewardValue = -10;`

- The initial reward value had been selected to be -10 to encourage exploration (i.e. finding more information about the environment). This small negative value initially penalises the agent for finding nothing and thereby encourages exploration of the rest of the unvisited grid cells in the testing environment.

`rewardValue = 100;`

- The reward value had been selected to be 100 if a sweeper collides with a mine. This large positive reward value encourages the collection of all the mines in the testing environment.

```
rewardValue = -50;
```

- The reward value had been selected to be -50 if a sweeper collides with a rock. This negative reward value encourages the agent to avoid rocks in the testing environment so that it does not incur this penalty.

```
rewardValue = -500;
```

- The reward value had been selected to be -500 if a sweeper collides with a super-mine. This large negative reward value encourages the agent to avoid super-mines in the testing environment so that it does not incur this penalty. This negative reward value is significantly more negative than the penalty for rocks, owing to the fact that if a sweeper collides with a super-mine both the sweeper and super-mine are destroyed.

```
const double discountFactor = 0.9;
const double learningRate = 0.5;
```

```
const double discountFactor = 0.9;
```

- The discount factor (gamma) had been selected to be 0.9 (as close as possible to 1) to enable the agent to care a lot about the distant future and to take the future effects of its actions into account as the goal is to collect all the mines in the test environment. This optimal value was found by varying gamma from 0 to 1 for each of the three test simulations.

```
const double learningRate = 0.5;
```

- The learning rate (eta) had been selected to be 0.5 to enable the agent to place equal value on new experiences, by adjusting the current state Q value for every move and thereby enabling learning. This also encourages the agent to start off exploring and end up exploiting the testing environment.

Test simulation 1 (1 mine-sweeper and 30 mines.):

- On average, 15.79 mines were swept with a maximum of 23. This is due to the fact that there were no super-mines in the testing environment and the agents learnt to collect the 30 mines which were in high density very rapidly.

Test simulation 2 (1 mine-sweeper, 25 mines and 5 super-mines):

- On average, 11.60 mines were swept with a maximum of 21. This is due to the fact that there were 5 super-mines in the testing environment and the agents learnt to collect the 25 mines with some difficulty due to the environment being more hostile.

Test simulation 3 (1 mine-sweeper, 5 mines and 25 super-mines):

- On average, 2.49 mines were swept with a maximum of 5. This is due to the fact that there were 25 super-mines in the testing environment and the agents learnt to collect the 5 mines with much difficulty due to the environment being very hostile.