Table of Contents

In your report, mention what you see in the agent's behavior. Does it eventually make it to the target location?

The agent doesn't respect the traffic laws. It doesn't make it to the location neither. It just moves randomly and there is no improvement as time passes by.

Justify why you picked these set of states, and how they model the agent and its environment.

I think that are 4 fundamental states that the car can be in at every moment regarding the traffic rules: green_light_can_left, green_light_cant_left, red_light_can_right, red_light_cant_right. But these states are implicit in the inputs so we can use the text representation of the inputs as states. If we combine inputs with the direction that the planners want us to go (forward, left or right) we can represent all possible states in which the car can be at any intersection.

So a state would be formed in python like the following:

(next_waypoint, str(inputs))

Is the same as having a string like this as the state:

" forward {left: 'forward', right: None, left: None, light: green} "

At the end, the states will depend on 5 variables : next_waypoint, light, left, right, and forward. Every combination of these values is a different state. I think they model the world correctly since it is true that the state changes with time in a way that a decision taken at the same location might be good in time T but bad in time T+1, depending on the traffic and light states.

Another variable I could have taken into account is the time left before deadline. I didn't include this in the states because I don't think it will be a big contribution to the performance of the agent and it will create so much states that learning would be so much slower. For example, the default deadline is 100, so including it in the states would multiply the number of states by 100.

What changes do you notice in the agent's behavior?

After implementing Q learning and after a couple of steps, the agents begins to make better decisions. Also, it learns how avoid accidents if you give it enough time. It reaches the destination every time, without having to wait a lot of time (in less than 25 steps). I took the next metrics with a gamma value of 0.5 and a learning rate value of 0.5:

****

278 Steps completed for params gamma=0.5, learning_rate=0.5 :

Net Reward Positive Runs/ Runs with params : 1.0

Invalid Steps with params / Steps with params: 0.068345323741

Invalid Steps after policy learnt / Steps with params: 0.0666666666667

In time / All runs: 1.0

Steps before in time: 38

Runs before in time: 1

Percentage of runs not in time: 0.05

****

The reason for this enhancement is because we are now remembering the consequences of our acts in each state (taking into account immediate rewards and future rewards), instead of making random decisions.

Report what changes you made to your basic implementation of Q-Learning to achieve the final version of the agent. How well does it perform?

I changed the values of gamma and the learning and run the 20 trial per each combination of parameters. I compared the metrics of each pair of parameters and choose the one I consider the best.

The values that I used:

self.gamma_vals = [0.8, 0.5, 0.1, 0.01, 0.001, 0]

self.learning_rate_vals = [1, 0.8, 0.5, 0.3, 0.1, 0]

The combination of parameters I choose were : gamma=0.001, learning_rate=0.1. I choose these values because they produced the least value for the "Invalid Steps with params / Steps with params". This metric means the percentage of steps the algorithm took that had a negative reward in relation to all the steps it took. Also with these values, all the runs ended in positive net rewards and it was capable of getting to the destination on time 99.05% of the time.

Metrics for the chosen gamma and learning rate values:

****

279 Steps completed for params gamma=0.001, learning_rate=0.1 :

Net Reward Positive Runs/ Runs with params : 1.0

Invalid Steps with params / Steps with params: 0.0430107526882

Invalid Steps after policy learnt / Steps with params: 0.0244897959184

In time / All runs: 1.0

Steps before in time: 34

Runs before in time: 1

Percentage of runs not in time: 0.05

****

This is a list with all metrics (long scroll):

****

357 Steps completed for params gamma=0.8, learning_rate=1 :

Net Reward Positive Runs/ Runs with params : 0.952380952381

Invalid Steps with params / Steps with params: 0.0700280112045

Invalid Steps after policy learnt / Steps with params: 0.0535117056856

In time / All runs: 0.809523809524

Steps before in time: 58

Runs before in time: 1

Percentage of runs not in time: 0.047619047619

****

314 Steps completed for params gamma=0.8, learning_rate=0.8 :

Net Reward Positive Runs/ Runs with params : 1.0

Invalid Steps with params / Steps with params: 0.0541401273885

Invalid Steps after policy learnt / Steps with params: 0.0194552529183

In time / All runs: 1.0

Steps before in time: 57

Runs before in time: 1

Percentage of runs not in time: 0.05

****

262 Steps completed for params gamma=0.8, learning_rate=0.5 :

Net Reward Positive Runs/ Runs with params : 1.0

Invalid Steps with params / Steps with params: 0.0763358778626

Invalid Steps after policy learnt / Steps with params: 0.0704845814978

In time / All runs: 1.0

Steps before in time: 35

Runs before in time: 1

Percentage of runs not in time: 0.05

****

264 Steps completed for params gamma=0.8, learning_rate=0.3 :

Net Reward Positive Runs/ Runs with params : 1.0

Invalid Steps with params / Steps with params: 0.0681818181818

Invalid Steps after policy learnt / Steps with params: 0.0309278350515

In time / All runs: 0.95

Steps before in time: 70

Runs before in time: 1

Percentage of runs not in time: 0.05

****

317 Steps completed for params gamma=0.8, learning_rate=0.1 :

Net Reward Positive Runs/ Runs with params : 1.0

Invalid Steps with params / Steps with params: 0.0630914826498

Invalid Steps after policy learnt / Steps with params: 0.02

In time / All runs: 0.95

Steps before in time: 67

Runs before in time: 1

Percentage of runs not in time: 0.05

****

1786 Steps completed for params gamma=0.8, learning_rate=0 :

Net Reward Positive Runs/ Runs with params : 0.0

Invalid Steps with params / Steps with params: 0.591825307951

Invalid Steps after policy learnt / Steps with params: 0.778578784758

In time / All runs: 0.1

Steps before in time: 815

Runs before in time: 4

Percentage of runs not in time: 0.2

****

245 Steps completed for params gamma=0.5, learning_rate=1 :

Net Reward Positive Runs/ Runs with params : 1.0

Invalid Steps with params / Steps with params: 0.069387755102

Invalid Steps after policy learnt / Steps with params: 0.0348258706468

In time / All runs: 0.95

Steps before in time: 44

Runs before in time: 1

Percentage of runs not in time: 0.05

****

260 Steps completed for params gamma=0.5, learning_rate=0.8 :

Net Reward Positive Runs/ Runs with params : 1.0

Invalid Steps with params / Steps with params: 0.0653846153846

Invalid Steps after policy learnt / Steps with params: 0.0287081339713

In time / All runs: 1.0

Steps before in time: 51

Runs before in time: 1

Percentage of runs not in time: 0.05

****

278 Steps completed for params gamma=0.5, learning_rate=0.5 :

Net Reward Positive Runs/ Runs with params : 1.0

Invalid Steps with params / Steps with params: 0.068345323741

Invalid Steps after policy learnt / Steps with params: 0.0666666666667

In time / All runs: 1.0

Steps before in time: 38

Runs before in time: 1

Percentage of runs not in time: 0.05

****

284 Steps completed for params gamma=0.5, learning_rate=0.3 :

Net Reward Positive Runs/ Runs with params : 1.0

Invalid Steps with params / Steps with params: 0.0457746478873

Invalid Steps after policy learnt / Steps with params: 0.0243902439024

In time / All runs: 1.0

Steps before in time: 38

Runs before in time: 1

Percentage of runs not in time: 0.05

****

283 Steps completed for params gamma=0.5, learning_rate=0.1 :

Net Reward Positive Runs/ Runs with params : 1.0

Invalid Steps with params / Steps with params: 0.0742049469965

Invalid Steps after policy learnt / Steps with params: 0.0480349344978

In time / All runs: 0.95

Steps before in time: 54

Runs before in time: 1

Percentage of runs not in time: 0.05

****

1381 Steps completed for params gamma=0.5, learning_rate=0 :

Net Reward Positive Runs/ Runs with params : 0.5

Invalid Steps with params / Steps with params: 0.553946415641

Invalid Steps after policy learnt / Steps with params: 0.762865792129

In time / All runs: 0.3

Steps before in time: 390

Runs before in time: 1

Percentage of runs not in time: 0.05

****

240 Steps completed for params gamma=0.1, learning_rate=1 :

Net Reward Positive Runs/ Runs with params : 1.0

Invalid Steps with params / Steps with params: 0.0875

Invalid Steps after policy learnt / Steps with params: 0.0597014925373

In time / All runs: 1.0

Steps before in time: 39

Runs before in time: 1

Percentage of runs not in time: 0.05

****

270 Steps completed for params gamma=0.1, learning_rate=0.8 :

Net Reward Positive Runs/ Runs with params : 1.0

Invalid Steps with params / Steps with params: 0.0851851851852

Invalid Steps after policy learnt / Steps with params: 0.0646766169154

In time / All runs: 0.95

Steps before in time: 69

Runs before in time: 1

Percentage of runs not in time: 0.05

****

287 Steps completed for params gamma=0.1, learning_rate=0.5 :

Net Reward Positive Runs/ Runs with params : 1.0

Invalid Steps with params / Steps with params: 0.0487804878049

Invalid Steps after policy learnt / Steps with params: 0.00892857142857

In time / All runs: 0.95

Steps before in time: 63

Runs before in time: 1

Percentage of runs not in time: 0.05

****

274 Steps completed for params gamma=0.1, learning_rate=0.3 :

Net Reward Positive Runs/ Runs with params : 1.0

Invalid Steps with params / Steps with params: 0.0693430656934

Invalid Steps after policy learnt / Steps with params: 0.0480349344978

In time / All runs: 1.0

Steps before in time: 45

Runs before in time: 1

Percentage of runs not in time: 0.05

****

306 Steps completed for params gamma=0.1, learning_rate=0.1 :

Net Reward Positive Runs/ Runs with params : 1.0

Invalid Steps with params / Steps with params: 0.0653594771242

Invalid Steps after policy learnt / Steps with params: 0.0413533834586

In time / All runs: 1.0

Steps before in time: 40

Runs before in time: 1

Percentage of runs not in time: 0.05

****

1783 Steps completed for params gamma=0.1, learning_rate=0 :

Net Reward Positive Runs/ Runs with params : 0.1

Invalid Steps with params / Steps with params: 0.59338194055

Invalid Steps after policy learnt / Steps with params: 0.690224570674

In time / All runs: 0.25

Steps before in time: 269

Runs before in time: 1

Percentage of runs not in time: 0.05

****

360 Steps completed for params gamma=0.01, learning_rate=1 :

Net Reward Positive Runs/ Runs with params : 1.0

Invalid Steps with params / Steps with params: 0.0694444444444

Invalid Steps after policy learnt / Steps with params: 0.0434782608696

In time / All runs: 0.95

Steps before in time: 107

Runs before in time: 1

Percentage of runs not in time: 0.05

****

285 Steps completed for params gamma=0.01, learning_rate=0.8 :

Net Reward Positive Runs/ Runs with params : 1.0

Invalid Steps with params / Steps with params: 0.0631578947368

Invalid Steps after policy learnt / Steps with params: 0.0217391304348

In time / All runs: 0.95

Steps before in time: 55

Runs before in time: 1

Percentage of runs not in time: 0.05

****

286 Steps completed for params gamma=0.01, learning_rate=0.5 :

Net Reward Positive Runs/ Runs with params : 1.0

Invalid Steps with params / Steps with params: 0.0594405594406

Invalid Steps after policy learnt / Steps with params: 0.0316742081448

In time / All runs: 0.95

Steps before in time: 65

Runs before in time: 1

Percentage of runs not in time: 0.05

****

215 Steps completed for params gamma=0.01, learning_rate=0.3 :

Net Reward Positive Runs/ Runs with params : 1.0

Invalid Steps with params / Steps with params: 0.0697674418605

Invalid Steps after policy learnt / Steps with params: 0.0306748466258

In time / All runs: 1.0

Steps before in time: 52

Runs before in time: 1

Percentage of runs not in time: 0.05

****

321 Steps completed for params gamma=0.01, learning_rate=0.1 :

Net Reward Positive Runs/ Runs with params : 1.0

Invalid Steps with params / Steps with params: 0.0591900311526

Invalid Steps after policy learnt / Steps with params: 0.0272373540856

In time / All runs: 0.95

Steps before in time: 64

Runs before in time: 1

Percentage of runs not in time: 0.05

****

1548 Steps completed for params gamma=0.01, learning_rate=0 :

Net Reward Positive Runs/ Runs with params : 0.3

Invalid Steps with params / Steps with params: 0.581395348837

Invalid Steps after policy learnt / Steps with params: 0.775065387969

In time / All runs: 0.2

Steps before in time: 401

Runs before in time: 1

Percentage of runs not in time: 0.05

****

237 Steps completed for params gamma=0.001, learning_rate=1 :

Net Reward Positive Runs/ Runs with params : 1.0

Invalid Steps with params / Steps with params: 0.0759493670886

Invalid Steps after policy learnt / Steps with params: 0.0576923076923

In time / All runs: 1.0

Steps before in time: 29

Runs before in time: 1

Percentage of runs not in time: 0.05

****

316 Steps completed for params gamma=0.001, learning_rate=0.8 :

Net Reward Positive Runs/ Runs with params : 1.0

Invalid Steps with params / Steps with params: 0.0632911392405

Invalid Steps after policy learnt / Steps with params: 0.0592592592593

In time / All runs: 0.95

Steps before in time: 46

Runs before in time: 1

Percentage of runs not in time: 0.05

****

273 Steps completed for params gamma=0.001, learning_rate=0.5 :

Net Reward Positive Runs/ Runs with params : 1.0

Invalid Steps with params / Steps with params: 0.0586080586081

Invalid Steps after policy learnt / Steps with params: 0.0398406374502

In time / All runs: 1.0

Steps before in time: 22

Runs before in time: 1

Percentage of runs not in time: 0.05

****

307 Steps completed for params gamma=0.001, learning_rate=0.3 :

Net Reward Positive Runs/ Runs with params : 1.0

Invalid Steps with params / Steps with params: 0.0586319218241

Invalid Steps after policy learnt / Steps with params: 0.0541516245487

In time / All runs: 1.0

Steps before in time: 30

Runs before in time: 1

Percentage of runs not in time: 0.05

****

279 Steps completed for params gamma=0.001, learning_rate=0.1 :

Net Reward Positive Runs/ Runs with params : 1.0

Invalid Steps with params / Steps with params: 0.0430107526882

Invalid Steps after policy learnt / Steps with params: 0.0244897959184

In time / All runs: 1.0

Steps before in time: 34

Runs before in time: 1

Percentage of runs not in time: 0.05

****

1867 Steps completed for params gamma=0.001, learning_rate=0 :

Net Reward Positive Runs/ Runs with params : 0.7

Invalid Steps with params / Steps with params: 0.560257096947

Invalid Steps after policy learnt / Steps with params: 0.743491577335

In time / All runs: 0.15

Steps before in time: 561

Runs before in time: 1

Percentage of runs not in time: 0.05

****

314 Steps completed for params gamma=0, learning_rate=1 :

Net Reward Positive Runs/ Runs with params : 1.0

Invalid Steps with params / Steps with params: 0.0541401273885

Invalid Steps after policy learnt / Steps with params: 0.0236220472441

In time / All runs: 0.9

Steps before in time: 60

Runs before in time: 1

Percentage of runs not in time: 0.05

****

297 Steps completed for params gamma=0, learning_rate=0.8 :

Net Reward Positive Runs/ Runs with params : 1.0

Invalid Steps with params / Steps with params: 0.0740740740741

Invalid Steps after policy learnt / Steps with params: 0.0441767068273

In time / All runs: 0.95

Steps before in time: 48

Runs before in time: 1

Percentage of runs not in time: 0.05

****

293 Steps completed for params gamma=0, learning_rate=0.5 :

Net Reward Positive Runs/ Runs with params : 1.0

Invalid Steps with params / Steps with params: 0.0648464163823

Invalid Steps after policy learnt / Steps with params: 0.0423728813559

In time / All runs: 0.95

Steps before in time: 57

Runs before in time: 1

Percentage of runs not in time: 0.05

****

322 Steps completed for params gamma=0, learning_rate=0.3 :

Net Reward Positive Runs/ Runs with params : 1.0

Invalid Steps with params / Steps with params: 0.0527950310559

Invalid Steps after policy learnt / Steps with params: 0.0176211453744

In time / All runs: 0.9

Steps before in time: 95

Runs before in time: 2

Percentage of runs not in time: 0.1

****

267 Steps completed for params gamma=0, learning_rate=0.1 :

Net Reward Positive Runs/ Runs with params : 1.0

Invalid Steps with params / Steps with params: 0.0524344569288

Invalid Steps after policy learnt / Steps with params: 0.0379746835443

In time / All runs: 1.0

Steps before in time: 30

Runs before in time: 1

Percentage of runs not in time: 0.05

****

1528 Steps completed for params gamma=0, learning_rate=0 :

Net Reward Positive Runs/ Runs with params : 0.05

Invalid Steps with params / Steps with params: 0.577879581152

Invalid Steps after policy learnt / Steps with params: 0.672116257947

In time / All runs: 0.25

Steps before in time: 427

Runs before in time: 3

Percentage of runs not in time: 0.15

****

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?

The optimal policy would be the one that gets in time to the destination 100% of the time without making any step with negative reward. Also the optimal policy shouldn't violate any traffic law ever (as in this scenario we want to avoid it at all cost), and it might choose an alternative route if the final net reward is going to be better. In contrast, my policy never takes alternative routes on purpose since I am not taking into account the deadline. Nevertheless, if you get my agent enough time to learn, it will learn a policy that gets you in time to the destination 100% of the time, it will avoid accidents and the percentage of steps with negative rewards will be 2.45% or lower (see metrics in the previous question). In conclusion, the policy learnt was not the optimal policy, but it performs really well.