Table of Contents

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

It doesn't make it. And if does, is just luck.

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. If we combine this 4 states with the direction that the planners want us to go (forward, left or right), we will end up with 3 new states for each one of 4 fundamental traffic states. So appending an underscore and the name of the direction given to us by the planner to the 4 fundamental traffic states, e.g, green_light_cant_left + _forward , I made a list of 12 possible states :

green_light_cant_left_forward,
green_light_cant_left_left,
green_light_cant_left_right,
green_light_can_left_forward,
green_light_can_left_left,
green_light_can_left_right,
red_light_cant_right_forward,
red_light_cant_right_left,
red_light_cant_right_right,
red_light_can_right_forward,
red_light_can_right_left,
red_light_can_right_right

They define the possible states where the car can be in any intersection. The current state depends on the values of the inputs, therefore the state can change with the time. 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.

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

It makes decisions that make sense after a couple of steps. 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).

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 alpha 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.alpha_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 : Alpha=0.5, learning_rate=0.3. 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 90% of the time.

This is a list with all the results (long scroll):

****

328 Steps completed for params Alpha=0.8, learning_rate=1 :

Net Reward Postive Runs/ Runs with params : 0.952380952381

Invalid Steps with params / Steps with params: 0.344512195122

Invalid Steps after policy learnt / Steps with params: 0.351170568562

In time / All runs: 0.952380952381

Steps before in time: 29

Runs before in time: 1

Percentage of runs not in time: 0.047619047619

****

280 Steps completed for params Alpha=0.8, learning_rate=0.8 :

Net Reward Postive Runs/ Runs with params : 1.0

Invalid Steps with params / Steps with params: 0.310714285714

Invalid Steps after policy learnt / Steps with params: 0.324324324324

In time / All runs: 0.95

Steps before in time: 21

Runs before in time: 1

Percentage of runs not in time: 0.05

****

273 Steps completed for params Alpha=0.8, learning_rate=0.5 :

Net Reward Postive Runs/ Runs with params : 1.0

Invalid Steps with params / Steps with params: 0.311355311355

Invalid Steps after policy learnt / Steps with params: 0.336032388664

In time / All runs: 1.0

Steps before in time: 26

Runs before in time: 1

Percentage of runs not in time: 0.05

****

235 Steps completed for params Alpha=0.8, learning_rate=0.3 :

Net Reward Postive Runs/ Runs with params : 1.0

Invalid Steps with params / Steps with params: 0.314893617021

Invalid Steps after policy learnt / Steps with params: 0.383783783784

In time / All runs: 0.95

Steps before in time: 50

Runs before in time: 1

Percentage of runs not in time: 0.05

****

291 Steps completed for params Alpha=0.8, learning_rate=0.1 :

Net Reward Postive Runs/ Runs with params : 1.0

Invalid Steps with params / Steps with params: 0.340206185567

Invalid Steps after policy learnt / Steps with params: 0.385245901639

In time / All runs: 0.95

Steps before in time: 47

Runs before in time: 1

Percentage of runs not in time: 0.05

****

306 Steps completed for params Alpha=0.8, learning_rate=0 :

Net Reward Postive Runs/ Runs with params : 1.0

Invalid Steps with params / Steps with params: 0.483660130719

Invalid Steps after policy learnt / Steps with params: 0.496402877698

In time / All runs: 0.95

Steps before in time: 28

Runs before in time: 1

Percentage of runs not in time: 0.05

****

362 Steps completed for params Alpha=0.5, learning_rate=1 :

Net Reward Postive Runs/ Runs with params : 0.95

Invalid Steps with params / Steps with params: 0.406077348066

Invalid Steps after policy learnt / Steps with params: 0.313725490196

In time / All runs: 0.95

Steps before in time: 158

Runs before in time: 1

Percentage of runs not in time: 0.05

****

291 Steps completed for params Alpha=0.5, learning_rate=0.8 :

Net Reward Postive Runs/ Runs with params : 1.0

Invalid Steps with params / Steps with params: 0.343642611684

Invalid Steps after policy learnt / Steps with params: 0.370212765957

In time / All runs: 1.0

Steps before in time: 56

Runs before in time: 1

Percentage of runs not in time: 0.05

****

316 Steps completed for params Alpha=0.5, learning_rate=0.5 :

Net Reward Postive Runs/ Runs with params : 1.0

Invalid Steps with params / Steps with params: 0.344936708861

Invalid Steps after policy learnt / Steps with params: 0.358156028369

In time / All runs: 0.95

Steps before in time: 34

Runs before in time: 1

Percentage of runs not in time: 0.05

****

367 Steps completed for params Alpha=0.5, learning_rate=0.3 :

Net Reward Postive Runs/ Runs with params : 1.0

Invalid Steps with params / Steps with params: 0.307901907357

Invalid Steps after policy learnt / Steps with params: 0.323699421965

In time / All runs: 0.9

Steps before in time: 21

Runs before in time: 1

Percentage of runs not in time: 0.05

****

303 Steps completed for params Alpha=0.5, learning_rate=0.1 :

Net Reward Postive Runs/ Runs with params : 1.0

Invalid Steps with params / Steps with params: 0.359735973597

Invalid Steps after policy learnt / Steps with params: 0.363957597173

In time / All runs: 0.95

Steps before in time: 20

Runs before in time: 1

Percentage of runs not in time: 0.05

****

239 Steps completed for params Alpha=0.5, learning_rate=0 :

Net Reward Postive Runs/ Runs with params : 1.0

Invalid Steps with params / Steps with params: 0.506276150628

Invalid Steps after policy learnt / Steps with params: 0.542452830189

In time / All runs: 1.0

Steps before in time: 27

Runs before in time: 1

Percentage of runs not in time: 0.05

****

241 Steps completed for params Alpha=0.1, learning_rate=1 :

Net Reward Postive Runs/ Runs with params : 1.0

Invalid Steps with params / Steps with params: 0.452282157676

Invalid Steps after policy learnt / Steps with params: 0.467889908257

In time / All runs: 1.0

Steps before in time: 23

Runs before in time: 1

Percentage of runs not in time: 0.05

****

255 Steps completed for params Alpha=0.1, learning_rate=0.8 :

Net Reward Postive Runs/ Runs with params : 1.0

Invalid Steps with params / Steps with params: 0.517647058824

Invalid Steps after policy learnt / Steps with params: 0.557939914163

In time / All runs: 1.0

Steps before in time: 22

Runs before in time: 1

Percentage of runs not in time: 0.05

****

266 Steps completed for params Alpha=0.1, learning_rate=0.5 :

Net Reward Postive Runs/ Runs with params : 1.0

Invalid Steps with params / Steps with params: 0.5

Invalid Steps after policy learnt / Steps with params: 0.512295081967

In time / All runs: 1.0

Steps before in time: 22

Runs before in time: 1

Percentage of runs not in time: 0.05

****

275 Steps completed for params Alpha=0.1, learning_rate=0.3 :

Net Reward Postive Runs/ Runs with params : 1.0

Invalid Steps with params / Steps with params: 0.538181818182

Invalid Steps after policy learnt / Steps with params: 0.562992125984

In time / All runs: 1.0

Steps before in time: 21

Runs before in time: 1

Percentage of runs not in time: 0.05

****

249 Steps completed for params Alpha=0.1, learning_rate=0.1 :

Net Reward Postive Runs/ Runs with params : 1.0

Invalid Steps with params / Steps with params: 0.566265060241

Invalid Steps after policy learnt / Steps with params: 0.581196581197

In time / All runs: 1.0

Steps before in time: 15

Runs before in time: 1

Percentage of runs not in time: 0.05

****

221 Steps completed for params Alpha=0.1, learning_rate=0 :

Net Reward Postive Runs/ Runs with params : 1.0

Invalid Steps with params / Steps with params: 0.461538461538

Invalid Steps after policy learnt / Steps with params: 0.502564102564

In time / All runs: 1.0

Steps before in time: 26

Runs before in time: 1

Percentage of runs not in time: 0.05

****

239 Steps completed for params Alpha=0.01, learning_rate=1 :

Net Reward Postive Runs/ Runs with params : 1.0

Invalid Steps with params / Steps with params: 0.531380753138

Invalid Steps after policy learnt / Steps with params: 0.550660792952

In time / All runs: 1.0

Steps before in time: 12

Runs before in time: 1

Percentage of runs not in time: 0.05

****

225 Steps completed for params Alpha=0.01, learning_rate=0.8 :

Net Reward Postive Runs/ Runs with params : 1.0

Invalid Steps with params / Steps with params: 0.493333333333

Invalid Steps after policy learnt / Steps with params: 0.522613065327

In time / All runs: 1.0

Steps before in time: 26

Runs before in time: 1

Percentage of runs not in time: 0.05

****

252 Steps completed for params Alpha=0.01, learning_rate=0.5 :

Net Reward Postive Runs/ Runs with params : 1.0

Invalid Steps with params / Steps with params: 0.535714285714

Invalid Steps after policy learnt / Steps with params: 0.543103448276

In time / All runs: 1.0

Steps before in time: 20

Runs before in time: 1

Percentage of runs not in time: 0.05

****

272 Steps completed for params Alpha=0.01, learning_rate=0.3 :

Net Reward Postive Runs/ Runs with params : 1.0

Invalid Steps with params / Steps with params: 0.525735294118

Invalid Steps after policy learnt / Steps with params: 0.54693877551

In time / All runs: 1.0

Steps before in time: 27

Runs before in time: 1

Percentage of runs not in time: 0.05

****

245 Steps completed for params Alpha=0.01, learning_rate=0.1 :

Net Reward Postive Runs/ Runs with params : 1.0

Invalid Steps with params / Steps with params: 0.497959183673

Invalid Steps after policy learnt / Steps with params: 0.54128440367

In time / All runs: 1.0

Steps before in time: 27

Runs before in time: 1

Percentage of runs not in time: 0.05

****

221 Steps completed for params Alpha=0.01, learning_rate=0 :

Net Reward Postive Runs/ Runs with params : 1.0

Invalid Steps with params / Steps with params: 0.524886877828

Invalid Steps after policy learnt / Steps with params: 0.544554455446

In time / All runs: 1.0

Steps before in time: 19

Runs before in time: 1

Percentage of runs not in time: 0.05

****

257 Steps completed for params Alpha=0.001, learning_rate=1 :

Net Reward Postive Runs/ Runs with params : 1.0

Invalid Steps with params / Steps with params: 0.548638132296

Invalid Steps after policy learnt / Steps with params: 0.572033898305

In time / All runs: 1.0

Steps before in time: 21

Runs before in time: 1

Percentage of runs not in time: 0.05

****

264 Steps completed for params Alpha=0.001, learning_rate=0.8 :

Net Reward Postive Runs/ Runs with params : 1.0

Invalid Steps with params / Steps with params: 0.534090909091

Invalid Steps after policy learnt / Steps with params: 0.558333333333

In time / All runs: 1.0

Steps before in time: 24

Runs before in time: 1

Percentage of runs not in time: 0.05

****

230 Steps completed for params Alpha=0.001, learning_rate=0.5 :

Net Reward Postive Runs/ Runs with params : 1.0

Invalid Steps with params / Steps with params: 0.486956521739

Invalid Steps after policy learnt / Steps with params: 0.513888888889

In time / All runs: 1.0

Steps before in time: 14

Runs before in time: 1

Percentage of runs not in time: 0.05

****

258 Steps completed for params Alpha=0.001, learning_rate=0.3 :

Net Reward Postive Runs/ Runs with params : 1.0

Invalid Steps with params / Steps with params: 0.484496124031

Invalid Steps after policy learnt / Steps with params: 0.510548523207

In time / All runs: 1.0

Steps before in time: 21

Runs before in time: 1

Percentage of runs not in time: 0.05

****

240 Steps completed for params Alpha=0.001, learning_rate=0.1 :

Net Reward Postive Runs/ Runs with params : 1.0

Invalid Steps with params / Steps with params: 0.483333333333

Invalid Steps after policy learnt / Steps with params: 0.50462962963

In time / All runs: 1.0

Steps before in time: 24

Runs before in time: 1

Percentage of runs not in time: 0.05

****

251 Steps completed for params Alpha=0.001, learning_rate=0 :

Net Reward Postive Runs/ Runs with params : 1.0

Invalid Steps with params / Steps with params: 0.513944223108

Invalid Steps after policy learnt / Steps with params: 0.556074766355

In time / All runs: 1.0

Steps before in time: 37

Runs before in time: 1

Percentage of runs not in time: 0.05

****

239 Steps completed for params Alpha=0, learning_rate=1 :

Net Reward Postive Runs/ Runs with params : 1.0

Invalid Steps with params / Steps with params: 0.510460251046

Invalid Steps after policy learnt / Steps with params: 0.536697247706

In time / All runs: 1.0

Steps before in time: 21

Runs before in time: 1

Percentage of runs not in time: 0.05

****

231 Steps completed for params Alpha=0, learning_rate=0.8 :

Net Reward Postive Runs/ Runs with params : 1.0

Invalid Steps with params / Steps with params: 0.510822510823

Invalid Steps after policy learnt / Steps with params: 0.543269230769

In time / All runs: 1.0

Steps before in time: 23

Runs before in time: 1

Percentage of runs not in time: 0.05

****

261 Steps completed for params Alpha=0, learning_rate=0.5 :

Net Reward Postive Runs/ Runs with params : 1.0

Invalid Steps with params / Steps with params: 0.486590038314

Invalid Steps after policy learnt / Steps with params: 0.497816593886

In time / All runs: 1.0

Steps before in time: 32

Runs before in time: 1

Percentage of runs not in time: 0.05

****

212 Steps completed for params Alpha=0, learning_rate=0.3 :

Net Reward Postive Runs/ Runs with params : 1.0

Invalid Steps with params / Steps with params: 0.452830188679

Invalid Steps after policy learnt / Steps with params: 0.477386934673

In time / All runs: 1.0

Steps before in time: 13

Runs before in time: 1

Percentage of runs not in time: 0.05

****

237 Steps completed for params Alpha=0, learning_rate=0.1 :

Net Reward Postive Runs/ Runs with params : 1.0

Invalid Steps with params / Steps with params: 0.523206751055

Invalid Steps after policy learnt / Steps with params: 0.559241706161

In time / All runs: 1.0

Steps before in time: 26

Runs before in time: 1

Percentage of runs not in time: 0.05

****

243 Steps completed for params Alpha=0, learning_rate=0 :

Net Reward Postive Runs/ Runs with params : 1.0

Invalid Steps with params / Steps with params: 0.481481481481

Invalid Steps after policy learnt / Steps with params: 0.5

In time / All runs: 1.0

Steps before in time: 13

Runs before in time: 1

Percentage of runs not in time: 0.05

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. After my agent learned how to make it in time to the destination (after the frist trial on time), it was able to get in time to the destination 100% of the time and the percentage steps with negative rewards was 32.37%. In conclution, the policy learnt by my agend was not the optimal policy, but it does the job of getting you to the destination in time.