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

When the action is set to pick from random the car moves around at random and would sometimes bump into the target location, but as soon as I changed the action to be the next waypoint, the agent would eventually make it to the target location.

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

I used stoplight and the next waypoint. This is a fairly basic representation of the agent's environment because it only takes these two variables into account, excluding what other drivers are doing. From my reward tracker output I can see the rewards for each time step for all trials, and I don't see negative rewards from either traffic violations or having accidents. For example if there is oncoming traffic turning left and I turn right on a green light, I get a reward of 2. From my trials I have not seen negative reward from potential accident-like situations. Therefore with the given traffic situation, only stoplight and waypoint is necessary to build a sufficient state.

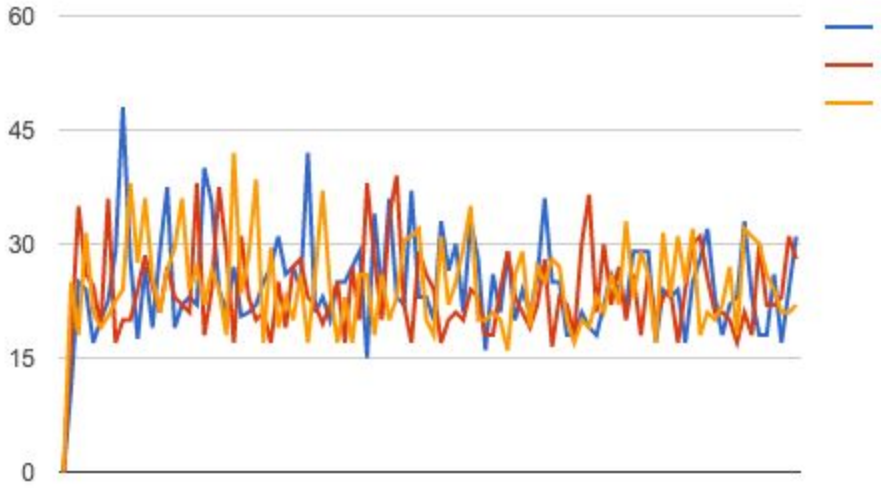
What changes do you notice in the agent's behavior?

Because I have initialized the Q table arbitrarily, the agent would sometimes miss the deadline early in the trials, but as the Q table builds up the agent would get to the target location without going over the deadline. Eventually the agent would decide between taking immediate negative reward but with a positive future reward, and sitting still at a stoplight. All of these contribute to getting the maximum total reward.

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 have added alpha the learning rate, gamma the discount rate, and epsilon the chance that the agent will explore to see if other actions would give more rewards. These improved the agent somewhat. For example the agent would choose immediate negative reward so that it can obtain a bigger long term reward.

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?



Yes. The result for my last 100 trials shows that the agent only missed the deadline early on and had always obtained a positive total trip reward. The graph shows rewards obtained in each trip for most recent 3 runs. It shows that as the agent gets more practice it performs more consistently in getting the rewards.