ANALYTICS CLUB

RL Playground

Project Selection Application

Adityan CG
EE19B003

Theory Part :

1Q.

Supervised Learning is about finding the formula R2.elation that could be used to relate the Input and Output by training the model through given examples. Supervised learning does classification or prediction by learning(training) on previous examples.

Reinforcement Learning is trained as an "Agent" and works based on action reward System. Where the agentinteracts with the environment and with each action it calculates reward, based on the change in states. In this way it performs actions and Learns by trial and error method to achieve towards the goal by maximizing the reward.

2Q.

Exploration and exploitation Dilemma :

This problem can be explained by an question of which action to take.
Ex : Do I continue my current job or look for a new one? Do I keep playing cricket or pursue badminton? Do I continue the current path or Look for a new path?

Exploration is important in RL. After each action the agent moves to next state in the environment with a reward. To attain maximum reward, the agent takes takes several actions through policies, and learns from previous actions.

The epsilon greedy policy can be associated with the exploration and exploitation dilemma. It takes a ‘k’ probability that the agent takes the best action provided by the knowledge of previous actions. And ‘1-k’ probability that it is a random action. In this way the agent can explore the possibilities along the way and helps better to maximise the reward.

Through a simulation we can see that the decaying epsilon greedy policy provides with more reward when compard to epsilon greedy. Clearly, here, As time passes, the exploration rate is more "k" decays.

3Q.

Experience Replay:(Replay Buffer)

The replay buffer contains a collection of experience tuples (S,A,R,S',D).

S - State ; A - Action ; S' = Next state ; R - Reward; D - Dones

The tuples are gradually added to the buffer as we are interacting with the Environment. The simplest implementation is a buffer of fixed size, with new data added to the end of the buffer so that it pushes the oldest experience out of it. And then a small batch size is randomly sampled from this buffer to learn from it, is called Experience replay.

We use deque to add the recent and remove the oldest and keep the constant buffer size. Instead of learning from only the previous experience we, learn from a samples experience over time.

Target Network:

Q(S,A) is calculated by Q(S',A') through bellman equation. This is very closer and can make our training unstable.
Hence we use Target network : We take a copy of our training network and use it to predict target Q values.And these target Q values are backpropagated to train the main Q network. The target netwok 's parameters are not trained, but they are periodically synchronised with the parameters of the main Q network.The idea is that using the target network's Q values to train the main Q-network will improve the stability of the training.

4Q.

The Question can mean many possibilities. As a part of a team can mean Each episode can be played by each individual and each reward can be maximised Independently. It has not been mentioned exactly as to how the game is played by the team. So I took this under MARL

The described situation is under MARL(Multi Agent RL) where if there are 2 independent agents. Let's take an example ;

In the first time(delta-t) step Agent1(A1) takes an action and in second time step(2*delta-t) Agent2(A2) takes an action and it alternatively goes on. Here the previous action is taken by A2. Hence we can't train A1 based on the previous action taken.(Markov Desicion Process : The next action depends on the previous action and state).Thus as a perspective of A1 we consider the step taken and presence of A2 as merged with the environment.

Since, Hint is MDP can be considered Static. We can Train each agent by considering 2 time steps altogether, And train them independently to maximize the score.

Code part is attached in the submission.