Samuel Bailey

CS-370

Cartpole Assignment:

**Explain how the cartpole problem can be solved using the REINFORCE algorithm**

**The CartPole question exist of a post join by an un-activated joint to a truck, that then moves along a streaming way. The framework happens restrictively by solicitation of power +1 or - 1 to the truck. The something used to quantify substantialness fires straight-up, and the reason for an activity search out keep it from dropping over. A prize of +1 happen help each timestep that the bar remaining part noteworthy.**

def \_\_init\_\_(self, actions, epsilon, alpha, gamma):

self.q = {}

self.epsilon = epsilon # exploration constant

self.alpha = alpha # discount constant

self.gamma = gamma # discount factor

self.actions = actions

def getQ(self, state, action):

return self.q.get((state, action), 0.0)

def learnQ(self, state, action, reward, value):

'''

Q-learning:

Q(s, a) += alpha \* (reward(s,a) + max(Q(s') - Q(s,a))

'''

oldv = self.q.get((state, action), None)

if oldv is None:

self.q[(state, action)] = reward

else:

self.q[(state, action)] = oldv + self.alpha \* (value - oldv)

def chooseAction(self, state, return\_q=False):

q = [self.getQ(state, a) for a in self.actions]

maxQ = max(q)

if random.random() < self.epsilon:

minQ = min(q); mag = max(abs(minQ), abs(maxQ))

# add random values to all the actions, recalculate maxQ

q = [q[i] + random.random() \* mag - .5 \* mag for i in range(len(self.actions))]

maxQ = max(q)

count = q.count(maxQ)

# In case there're several state-action max values

# we select a random one among them

if count > 1:

best = [i for i in range(len(self.actions)) if q[i] == maxQ]

i = random.choice(best)

else:

i = q.index(maxQ)

action = self.actions[i]

if return\_q: # if they want it, give it!

return action, q

return action

def learn(self, state1, action1, reward, state2):

maxqnew = max([self.getQ(state2, a) for a in self.actions])

self.learnQ(state1, action1, reward, reward + self.gamma\*maxqnew)

**Explain how the cartpole problem can be solved using the A2C algorithm**.

This instructional exercise utilizes model subclassing to delimit the model. During the football pass toward adversary's objective, the model will comprehend US of America as the proposal and will yield two together individual deed probabilities and critic benefit V, that models the state-controlled by worth capacity. The point search out trains a model that select direct dependent on a system π that augments expected return. For Cartpole-v0, talented exist four standards addressing US of America: truck position, truck speed, post point, and bar speed independently. The specialist can take two direct to push the truck deserted (0) and right (1) independently.

**Explain how policy gradient approaches differ from value-based approaches, such as Q-learning**.

The two plans are in principle constrained by the Markov Choice Cycle collect, and appropriate way utilize comparative composed comments and thought. Moreover, in regular logical conditions you surrender plausibility wish the two methodology to impact something very similar - or not totally same - ideal strategies. Nonetheless, they exist different inside. The most major contrasts center from two focuses the methodologies occur in how or way they approach something done choice, two together if learning, and as the something delivered (the all around educated arrangement). In Q-information, the objective pursuit out finds a solitary deterministic something done from a singular arrangement of activities by decision the greatest benefit. With strategy slopes, and added direct strategies look, the objective pursuit out decides a guide from state to something done, which perhaps accepted, and works stylish ceaseless activity scope. Accordingly, strategies slope means can answer issues that advantage-found techniques can't: Enormous and consistent activity space. In any case, going with esteem found course of action, this can in any case be approximated going with discretization - and this happen not a terrible decision, in light of the fact that the arrangement work in strategy incline has expected not many sort of approximator basically.

Stochastic strategies are a worth found strategy can't answer an environment where the ideal technique is speculated requiring explicit probabilities, to a degree Scissor/Paper/Stone. That is on the grounds that there happen no educable boundaries stylish Q-schooling that control probabilities of activity, the inquiry definition trendy TD instruction expects that a deterministic force possibly ideal. Notwithstanding, monetary worth-found strategies like Q-schooling bear some advantage too: Straightforwardness. You can execute Q capacities as straightforward individual tables, and this gives a few certifications of association. There happen no even record of an event of strategies angle, cause you need an arrangement work p(a∣s,θ)p(a∣s,θ) which notwithstanding should bear a smooth inclination concerning θθ.Speed. TD instruction strategies that begin working framework happen frequently a lot quicker to get data an approach than plan that must simply test from the environmental factors so that assess progress.

**Explain how actor-critic approaches differ from value- and policy-based approaches**.

In the method of information private frameworks, addressing in parts works with in familiar practice procedures and the best approach to see as a legitimate entertainer. Training exists underneath power. Displaying limits the wide assortment of techniques a craftsman has, and the methodologies are demonstrated explicitly. Assembling the 2, there might be a phase of accentuation and expense in arrangement. The strategy at the improvement level is taken from gauges from tables that aren't provided autonomously, thinking about the exact techniques of whether they're subjective. America's parametric skills are requested the utilization of vital rakish making arrangements rather than a grouping of expense decisions, but the job betting technique obligations can't substitute delight other than while it is achievable to choose the slant. It is the person that executes the nation and the system, recognizing that almost certainly, there might be a seeking with the limit that controls the association with the nation, and that there are attainable external cutoff points to advantage the accentuation of the records increase plan is normally withinside the gift valuation. This makes a fine technique, which may be presently as of now not a genuine, very much depicted portrayal of a recruit r deciding through the quality, there aren't any external cutoff points to which the expert exists, and adequate records is to be had for the ideal development plan.

Resources:

Blog, Guest. “Reinforce Algorithm: Taking Baby Steps in Reinforcement Learning.” *Analytics Vidhya*, 24 Nov. 2020, www.analyticsvidhya.com/blog/2020/11/reinforce-algorithm-taking-baby-steps-in-reinforcement-learning/.

Yoon, Chris. “Understanding Actor Critic Methods.” *Medium*, Towards Data Science, 17 July 2019, https://towardsdatascience.com/understanding-actor-critic-methods-931b97b6df3f.

Yoon, Chris. “Deriving Policy Gradients and Implementing Reinforce.” *Medium*, Medium, 23 May 2019, https://medium.com/@thechrisyoon/deriving-policy-gradients-and-implementing-reinforce-f887949bd63.