A policy-based quiz

测验, 7 个问题

1
point

1。
In broad strokes, how do policy-based methods work?

● Parameterize the action-picking policy. Find such policy parameters that maximize expected returns.

○ Define a policy as an arg-max of Q-values learned by value-based methods.

○ Define exploration policy (e.g. epsilon-greedy). Then train Q-values in a way that accounts for current exploration policy.

○ Learn the optimal reward function given a fixed policy of a rational agent.

1
point

2。
Policy gradient -- it's a gradient of what function and with respect to what inputs?

● A gradient of expected reward w.r.t. action probabilities

○ A gradient of policy w.r.t. action probabilities

○ A gradient of policy w.r.t. actions

○ A gradient of policy w.r.t. states

1
point

3。
Which of those methods can learn from partial trajectories?

☑ SARSA

☑ Q-learning

□ Value Iteration

□ Crossentropy method

□ REINFORCE

☑ Advantage Actor-Critic

1
point

4。

What are valid reasons to use Q-learning and not REINFORCE

☑ Unlike reinforce, Q-learning can be trained much more efficiently with experience replay

☑ Unlike REINFORCE, Q-learning can be trained on partial experience (e.g. s,a,r,s')

□ Unlike REINFORCE, Q-learning can work with discounted rewards.

□ Unlike REINFORCE, Q-learning does not require exploration.

□ Unlike REINFORCE, Q-learning directly optimizes expected sum of rewards over session

1
point

5。

Which of the following is a valid expression for policy gradient J?

Legend:

- G(s,a) - discounted reward
- r(s,a) - immediate reward
- gamma - discount factor for discounted reward
- d(s) - a probability of being in this state at a random moment along random trajectory sampled with current policy
- pi(a|s) - agent's policy

● \nabla J = \underset{E}{s sim d(s), a sim \pi} \nabla log pi (a | s) * G(s, a)

○ \nabla J(s) = \underset{E}{s sim d(s), a sim \pi, s' sim P(s' | s,a)} r(s,a) + gamma * \nabla J(s')

○ \nabla J = \underset{E}{s sim d(s), a sim \pi, s' sim P(s' | s,a)} \nabla \pi (a | s) * G(s, a)

○ \nabla J = \underset{E}{s sim d(s), a sim \pi, s' sim P(s' | s,a)} \nabla \pi(a | s) * \nabla G(s,a)

1 point

6。
How does advantage actor critic works?

● It trains an agent (actor) with a help of human critic

○ It trains an ensemble of two models - Q-learning(critic) and REINFORCE(actor) - and picks actions by voting.

○ It trains a network to predict advantage A(s,a) = Q(s,a) - V(s) and picks action with highest predicted advantage

○ Actor is trained by the gradients propagated through the critic.

○ It uses learned state values(critic) as a baseline for policy gradient(actor)

1 point

7。

How do you train critic in Advantage Actor Critic?

○ In advantage actor-critic there's no need to train critic

○ With policy gradient \nabla J(s) = \underset{E}{s sim d(s), a sim \pi, s' sim P(s' | s,a)} r(s,a) + gamma * \nabla J(s')

○ \nabla J = \underset{E}{s sim d(s), a sim \pi, s' sim P(s' | s,a)} \nabla \pi (a | s) * G(s, a)

○ A critic predicts Q(s, a), we minimize [r + gamma * max(Q(s',a')) - Q(s,a)] ^2

● A critic predicts V(s), we minimize [r + gamma * const(V(s')) - V(s)] ^2

☑ 我（**伟臣 沈**）了解提交不是我自己完成的作业 将永远不会通过此课程或导致我的 Coursera 帐号被关闭。

了解荣誉准则的更多信息

Submit Quiz