Week 11 T10

Tuesday, 11 October 2022 10:20 AM

Reward Shaping

Adds a small reward for each transition based on domain knowledge to encourage positive actions (like a heuristic)

$$Q(s,a) = Q(s,a) + \alpha \cdot [r + \underbrace{F(s,s')}_{\text{shaped reward}} + \gamma Q(s',a') - Q(s,a)]$$

Potential Based Reward Shaping

Potential function $\Phi(s)$: Basically a heuristic for state s larger value means state is better

$$E[X] = \sum P(x) \cdot X$$

Can calculate shaped reward with Φ

$$F(s,s') = \gamma \Phi(s') - \Phi(s)$$

- Using potential functions guarantees alg will learn optimal policy

$\boxed{V}$

- However, no guarantee that it will converge faster

Policy Iteration

Evaluates the value of policies directly

Improves the policy over time

2 steps during each iteration:

1) policy evaluation
2) policy Improvement

- Repeat until policy converges

Policy evaluation

Find the value of each state $V^{\pi}(s)$ under the current policy

Algorithm – Policy evaluation

Input: π the policy for evaluation, V^π value function, and MDP $M = \langle S, s_0, A, P_a(s' \mid s), r(s,a,s') \rangle$
Output: Value function V^π

Repeat
  $\Delta \leftarrow 0$
  For each $s \in S$
    $\underbrace{V'^\pi(s) \leftarrow \sum_{s' \in S} P_{\pi(s)}(s' \mid s)\, [r(s,a,s') + \gamma\, V^\pi(s')]}_{\text{Policy evaluation equation}}$ V(s) based only on policy action
    $\Delta \leftarrow \max(\Delta, |V'^\pi(s) - V^\pi(s)|)$
  $V^\pi \leftarrow V'^\pi$
Until $\Delta \leq \theta$ Until convergence

If done by hand: Use a table Or solve simultaneous equations

eg)

Store Policy

$\pi(s_0)$	$\pi(s_1)$	$\pi(s_2)$
a_0	a_1	a_2

	$V(s_0)$	$V(s_1)$	$V(s_2)$
Init	0	0	0
1	2	-1	3
2	1	-2	4
3	0.5	-2.7	4.3
4	0.5	-2.7	4.3

Values on each line depend on values in prev line

Until 2 lines are the same (or close enough)

Once $V^\pi(s)$ found for all states, move to policy Improvement

Policy Improvement

1) Calculate $Q(s,a)$ for all actions a in state s
2) Set policy for s $\pi(s)$ = action with highest $Q(s,a)$
3) Repeat for all states

$$Q^\pi(s,a) = \sum_{s' \in S} P_a(s' \mid s)\, [r(s,a,s') + \gamma\, \underline{V^\pi(s')}]$$

$V^\pi(s)$ from the table

$$Q^\pi(s,a) = \sum_{s' \in S} P_a(s' \mid s)\,[r(s,a,s') + \gamma\, V^\pi(s')]$$ $V^\pi(s)$ from the table

Put Together

Algorithm – Policy Iteration

Input: MDP $M = \langle S, s_0, A, P_a(s' \mid s), r(s,a,s')\rangle$
Output: Policy π

Set V^π to arbitrary value function; e.g., $V^\pi(s) = 0$ for all s.

Set π to arbitrary policy; e.g. $\pi(s) = a$ for all s, where $a \in A$ is an arbitrary action.

Repeat
  Compute $V^\pi(s)$ for all s using policy evaluation
  For each $s \in S$
    $\pi(s) \leftarrow \text{argmax}_{a \in A(s)} Q^\pi(s,a)$
Until π does not change