RL - H.W - 3 2017147

Divyam Anshumaan

Q3. Suppose G_t is return at time t for some episode following first visit MC.

We essentially want $E[\rho_{t:T-1} G_t \mid S_t, A_t]$

$= q_\pi(S_t, A_t)$, where

$$\rho_{t:T-1} = \prod_{k=t}^{T-1} \frac{\pi(A_k|S_k)}{b(A_k|S_k)}, \text{ where}$$

π is the target policy and b is the behaviour policy. Then,

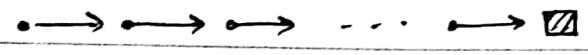
$$q_\pi(s,a) = \frac{\sum_{t \in \mathcal{T}(s,a)} \rho_{t:T(t)-1} G_t}{|\mathcal{T}(s,a)|} \text{ where}$$

$\mathcal{T}(s,a)$ = The set of all time steps where state s is visited and action a was taken (ie, $S_t, A_t = (s,a)$)

Neelgagan
Date
Page No.

Q2: Here 'o' represents a state, '.' represents a state action pair and ▨ is the terminal state

to estimate $q_\pi(s,a)$ we will have the follo-
-wing for an episode:



Q3. Suppose for a new episode our estimates of time to becomes very small. TD updates will give more weight to these new updates as comp-
-ared to MC (that is still marking against the ~~un~~upd unupdated G_t)

Neelgagan
Date:
Page No.

Q5. So it is clear that the first episode ends at the left terminal state (T). This is because following TD(0):

$$V(S_t) = V(S_t) + \alpha(R_{t+1} + \gamma V(S_{t+1}) - V(S_t))$$

for V(E) would have updated to

$$V(E) = 0.5 + 0.1(1 + 0 - 0.5)$$
$$= 0.55.$$

had the episode ended on the right.

Instead we have

$$V(X) = 0.5 + 0.1(0 + 1 \times 0.5 - 0.5)$$

for $X \in \{B, C, D, E\}$

and $V(A) = 0.5 + 0.1(0 + 1 \times 0 - 0.5)$
$= 0.45.$

Q6: Ex. 6.3: Yes. Suppose if our initial estimates were very off. We would have a significant RMSE. In such a case a larger α would quickly reduce RMSE before converging onto some value. At the same time, a smaller RMSE α would mean a longer time to converge, but the steady value (of RMSE) would also be smaller.

There is probably a value of α so that there is faster convergence and a lower RMSE value.

Ideally we would keep a small α and observe that RMSE becomes 0 after an infinite no. of episodes.

~~Yes~~ ~~NO~~ ~~No~~, this ~~would~~ is a would always ~~not~~ ~~to~~ initialization occur. Consider for any state x:

x.6.4.

$$V(X) = V(X) + \alpha \left(R_{t+1} + \gamma \cdot V(Y)\right) - V(X)$$

Then, $V(X) = V(X)(1-\alpha) + \alpha(R_{t+1} + \gamma V(Y))$

and in our situation: $V(X)(1-\alpha) + \alpha(R_{t+1} + V(Y))$

where Y is the ~~not~~ new state transition

Now suppose we had a reasonable estimate $V'(X) = V_\pi(X) + \delta$ and also some estimate $V'(Y)$

Here for most cases ($E \to$ Terminal) $R_{t+1} = 0$. Ignoring that,

$V(X) = V'(X)(1-\alpha) + \alpha \cdot V'(Y)$. For a high α, we would completely ~~lose~~ lose our estimate $V'(X)$ and instead obtain $V(X) \approx V'(Y)$.

Neelgagan
Date:
Page No.

which would give a high RMSE. Now suppose our initialization was perfect.

Now suppose our initialization was perfect. ie $V(x) = V_\pi(x)$. Our updates would very quickly make values approach $V(E)$.

$V(E) = V(E)_{init} + \alpha(1 - V(E)_{init})$.
$= V(E)(1 - \alpha)$, or

$V(D) = V(E)_{init} + \alpha(0 + \gamma(V(D)_{init}) - V(E)_{init})$
$= V(E)_{init}(1 - \alpha) + \alpha(V(D)_{init})$

for $\alpha \approx 1$ we give too much weight to our updates.

Supposing we had a bad initialization, (say, $V(x) = 0.9$ for $x \in \{A, B, C, D, E\}$)

RMSE after update for some $\alpha = 0.8$, would be $V(x)' = 0.9 \times 0.2$ or $0.9 \times 0.2 + 0.9 \times 0.8$
$= 0.18$ or 0.9

On average: $\frac{0.18 + 0.9}{2} = 0.54$

compared for same $\alpha = 0.2$, the same update gives : 0.9×0.8 or 0.9

$$= \frac{0.72 + 0.9}{2}$$

On average = 0.61

(When policy is arbitrary)

the decrease initially in RMSE is more but increases due to instability.

Ex. 6.5 In either case, we must know the ~~policy~~ model and policy π. By the codes of this it is likely the case that ~~Policy iteration~~ (policy eval.) was done and $v_\pi(\cdot)$ values were obtained. The model can be inferred ~~through~~ and we can calculate

$$v_\pi(s) = E_\pi(R_{t+1} + \gamma v_\pi(S_{t+1}) \mid S_t = s)$$

The other way to do this would be via Monte Carlo methods for an infinite no. of episodes.

Neelgagan
Date:
Page No.

Ex. 6.12: No they are still not the same. Even with greedy action selection, SARSA remains on-policy while Q-learning is off-policy.

Also, as per the pseudo-code SARSA updates Q after choosing S', A', while Q-learning is first updating Q, then choosing the next action.