Monte Carlo Methods

Wednesday, June 17, 2020 7:03 PM

The underlying idea for Monte Carlo methods is that the value of a state is the average return observed after visits to that state

First-Visit Monte Carlo Method:

The value of a state is the average of all returns observed after the first visit to that state

Every-Visit Monte Carlo Method:

The value of a state is the average of all returns observed after every visit to that state

Both first-visit and every-visit Monte Carlo (MC) converge to $v_\pi(s)$ as the number of visits to s goes to infinity

Things to Note:

MC methods do not bootstrap (no estimates are based on other estimates)

The computational expense of estimating the value of a single state is independent of the number of states

MC methods can operate without complete knowledge of the dynamics of the environment

MC methods can learn from actual or simulated experience

MC methods may be adjusted to estimate $q_\pi(s, a)$ by keeping track of visits to state-action pairs (where you take a certain action from a certain state)

One problem with doing this is that if a policy is deterministic, some state-action pairs may never be visited

This problem is called the problem of maintaining exploration

One approach to solve this for episodic tasks is by starting episodes in a state-action pair and giving each state-action pair a nonzero probability of being selected as the start

Another approach is to only consider stochastic policies that have a nonzero probability of selecting all actions in each state

This is the standard approach

Determining a value function through MC methods allows general policy iteration as described in the previous section (alternating between policy evaluation and policy improvement) but with the policy evaluation step replaced by averaging returns from episodes (with exploring starts)

On-Policy Method:

A method to ensure all actions are selected infinitely many times as the number of episodes played goes to infinity by evaluating or improving the policy used to make decisions

Soft Policy:

A policy that gives nonzero probability to all actions in all states

ε-Greedy Policy:

A policy that gives a set nonzero probability $\frac{\varepsilon}{|A(s)|}$ to all nongreedy actions and the rest of the probability to the greedy action $1 - \varepsilon + \frac{\varepsilon}{|A(s)|}$

ε-Soft Policy:

A policy for which $\pi(a|s) \geq \frac{\varepsilon}{|A(s)|}$ for all states and actions for some $\varepsilon > 0$

Off-Policy Method:

A method to ensure all actions are selected infinitely many times as the number of episodes played goes to infinity by evaluating or improving a policy different from the one used to make decisions

The policy that is being learned about through exploration is called the target policy (this policy is trying to be optimal), while the policy used to make decisions (and therefore includes exploratory behavior) is called the behavior policy

The behavior policy is usually required to be soft to guarantee that all possibilities are explored

Note that this approach can also be applied to other sources of information about the problem such as a traditional controller or a human expert

Coverage:

The condition that the behavior policy must, with some probability, take all actions that the target policy would take

Note that this means that the behavior policy must be stochastic in all states where it isn't identical to the target policy

Importance Sampling:

A general technique for estimating properties of a particular distribution while having only samples generated from a different distribution

Almost all off-policy methods use importance sampling

This can be applied by weighting returns according to the relative probabilities of their trajectories occurring under the target and behavior policies (called the importance-sampling ratio)

Probability of a State-Action Trajectory:

$$\Pr\{A_t, S_{t+1}, A_{t+1}, \cdots S_T | S_t, A_{t:T-1} \sim \pi\} = \prod_{k=t}^{T-1} \pi(A_k|S_k) p(S_{k+1}|S_k, A_k)$$

Importance Sampling Ratio

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

Where π is the target policy and b is the behavior policy

Value Function for Target Policy:

$$v_\pi(s) = E[\rho_{t:T-1} G_t | S_t = s]$$

Where G_t is the return achieved by the behavior policy in state s

Ordinary Importance Sampling:

$$V(s) \equiv \frac{\sum_{t \in J(s)} \rho_{t:T(t)-1} G_t}{|J(s)|}$$

Where $V(s)$ is the state-value function being estimated, the time steps t disregard episodes, $T(t)$ denotes the first time of termination following time t, G_t denotes the return after time t through $T(t)$, and $J(s)$ denotes the set of all time steps state s is visited for every-visit MC and the set of all times steps where state s is visted for the first time in first-visit MC

Note that this is just an average of returns over all episodes weighted by their importance ratio

For first-visit MC this is unbiased (its expectation value is always $v_\pi(s)$) but its variance is unbounded (since the importance ratio is unbounded)

Note that this can be computed incrementally using the average increment formula from the multi-armed bandits section

Weighted Importance Sampling:

$$V(s) \equiv \frac{\sum_{t \in J(s)} \rho_{t:T(t)-1} G_t}{\sum_{t \in J(s)} \rho_{t:T(t)-1}}$$

Where the value is defined to be 0 if the denominator is 0

For first-visit MC this is biased (its expectation value won't be exactly $v_\pi(s)$, although it will converge to $v_\pi(s)$ in the limit of infinite samples) but its variance is bounded since each return can't be effectively weighted by more than 1

This is usually preferred in practice as its lower variance means it is usually faster to approach the true value

This is sometimes called normalized importance sampling

Incremental Update Formula:

$$V_{n+1} \equiv V_n + \frac{W_n}{C_n}[G_n - V_n]$$

Where W_n is the importance sampling ratio for the trajectory followed by the behavior policy in the episode being considered in the update, $C_{n+1} \equiv C_n + W_{n+1}$ (with $C_0 = 0$) is the cumulative sum of the weights given to the first n returns, and V_0

is arbitrary

Note that this could be used for the on-policy case by setting $\pi = b$ (in which case W is always 1)

For every-visit MC both ordinary and weighted importance sampling are biased, although this bias asymptotically decreases to 0 as the number of visits goes to infinity