Chapter 5 - Monte Carlo Methods

Atul Balaji - EE16B002

February 25, 2019

- In DP, the complete transition probability distributions are required for the model. However, in most cases, explicit distributions are not available. To solve this issue, Monte Carlo methods are introduced, where learning is done using only experience (interaction with the environment).
- Before understanding the Monte Carlo methods, we must define the concept of **episodes**. An experience is divided into episodes which start and terminate at regular intervals. On the completion of an episode the value estimates and policies are changed.
- In Monte Carlo methods, the estimates for each state are independent. The estimate for one state does not build upon the estimate of other states, as is the case in DP. This means that there is **no bootstrapping** in MC methods.
- This implies that the computational expense of estimating the value of a single state is independent of the number of states. So, Monte Carlo methods are preferred over DP methods when we require the value of only a few states.

1 Monte Carlo Prediction

We want to learn the state-value function $v_\pi(s)$ for the given policy π and a set of episodes obtained by following π and visiting s. But, s may be visited multiple times in an episode. There are two ways to estimate $v_\pi(s)$:

- **First-Visit** - $v_\pi(s)$ is estimated as the average of the returns following the first visits to state s.
- **Every-Visit** - $v_\pi(s)$ is estimated as the average of the returns following all the visits to state s.

2 Monte Carlo Estimation of Action Values

- When a model is not available, state values(π) alone are not sufficient to determine a policy. Action values are also required.
- A state-action pair (s,a) is said to be visited in an episode if the state s is visited and action a is taken in it. The action-value function is estimated using returns obtained when a particular (state, action) pair is visited. Both first-visit and every-visit variants can be used for action value estimation.
- **Exploring starts** - In MC methods, many state-action pairs may never be visited. If a deterministic policy π is followed, then we will observe returns only for one of the actions from each state. Therefore, the Monte Carlo estimates of the other actions will not improve with experience. However, continued exploration is required for policy evaluation to work. We can overcome this issue by specifying that the episodes start in some (s,a) pair, and that every pair has a non-zero probability of being selected as the start. This guarantees that all state-action pairs will be visited an infinite number of times in the limit of infinite no.of episodes.
- The assumption of exploring starts is useful but is not true generally and hence is not reliable.

3 On-Policy Monte Carlo Control

- To approximate optimal policies, we use GPI, but with action-values (q) rather than value function (v). Let us assume that we observe an infinite no.of episodes and these are generated with exploring starts.
- Action-values are estimated for π. π is computed greedily with respect to current estimate of action-values. This is repeated until convergence. We have:

$$\pi(s) = \underset{a}{argmax}\ q(s,a)\ and \tag{1}$$

$$q_{\pi_k}(s, \pi_{k+1}(s)) = q_{\pi_k}(s, \underset{a}{argmax}\ q_{\pi_k}(s,a)) \tag{2}$$

- However, when a fixed policy π is followed, some (state, action) pairs will never occur. This problem can be addressed using:

  - Exploring starts (ES)
  - Without ES - ES is an unlikely assumption since it cannot be simulated practically. The alternate approach is to use ϵ-soft policies. The disadvantage with this method is that even though the optimal policy obtained is the best among all the possible ϵ-soft policies, it may not be the overall best optimal policy.

- We can implement policy evaluation in an incremental manner using cumulative sum of weights for the visited states.

4 Off-Policy Prediction via Importance Sampling

- In Monte Carlo methods, we need to obtain samples following policy π. But, for the estimate to be reliable, exploration is required.
- This can be solved using off-policy methods where we use two policies - behaviour policy (used to generate behaviour) and target policy (which is being learned). We generate the episodes following a behaviour policy b and estimate values for target policy π.
- Coverage - to use episodes from b to estimate values for π, we require that every action taken under π can also be taken under b. Therefore, b should be stochastic and hence have a non-zero probability of selecting actions which might be selected when we follow π.
- Ordinary importance sampling is done as a simple average. An alternative to this is the weighted importance sampling, where weighted average is taken. Estimates from weighted sampling are more reliable even though the estimates don't correspond to an expectation over policy π. But given a large number of episodes both these estimates are close enough and hence converge to the same true optimal value. Ordinary sampling is unbiased but has large variance. Weighted sampling has lower variance, but is generally biased.

5 Off-Policy Monte Carlo Control

- Off policy MC control methods use GPI. Importance sampling is used for prediction, where a behaviour policy is followed while learning and improving the target policy.
- Policy improvement is done in a greedy manner. π is estimated greedily using the current action-value estimates.
- It is preferred that b is an ϵ-soft policy, so as to ensure sufficient exploration. But, there is a potential problem here, that this method learns only from the tails of episodes, when all the remaining actions in the episode are greedy. If there are lot of exploratory actions, learning is slow and hence it takes a long time converge. This problem can be addressed using TD Learning.