SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN

50.007 Machine Learning, Fall 2015
Lecture Notes for Week 13

Reinforcement Learning (II)

Last update: Sunday 6th December, 2015 21:10

We have learned value iteration in the last class, where we derived the update functions based on the following:

$$
\begin{aligned}
V^*(s) &= \max_a Q^*(s,a) = Q^*(s,\pi^*(s)) && (1)\\
Q^*(s,a) &= \sum_{s'} T(s,a,s')[R(s,a,s') + \gamma V^*(s')] && (2)\\
V^*(s) &= \max_a \sum_{s'} T(s,a,s')[R(s,a,s') + \gamma V^*(s')] && (3)\\
&= \sum_{s'} T(s,\pi^*(s),s')[R(s,\pi^*(s),s') + \gamma V^*(s')] && (4)
\end{aligned}
$$

Alternatively, we could replace $V^*(s')$ in Eq (2) by the right-hand side of Eq (1). We arrive at the following recurrence function that involves Q-values only:

$$
Q^*(s,a) = \sum_{s'} T(s,a,s')[R(s,a,s') + \gamma \max_{a'} Q^*(s',a')] \quad (5)
$$

The Q-Value Iteration Algorithm

- Start with $Q_0^*(s,a) = 0$ for all $s \in S, a \in A$.
- Given $Q_i^*(s,a)$, calculate the Q-values for all states (depth $i+1$) and for all actions a:

$$
Q_{i+1}^*(s,a) \leftarrow \sum_{s'} T(s,a,s')[R(s,a,s') + \gamma \max_{a'} Q_i^*(s',a')] \quad (6)
$$

This algorithm has the same convergence guarantees as its value iteration counterpart.
As before, the optimal policy can be easily recovered from the Q-values as:

$$
\pi(s) = \arg\max_a Q(s,a) \quad (7)
$$

Policy Iteration (optional)

Besides value iteration and Q-value iteration, there is also policy iteration algorithm that iteratively improves the policy directly. The algorithm is as follows.

- Randomly initialize the policy π.
- Find the values for the states under the policy π by solving the following system of linear equations:

$$V^{\pi}(s) = \sum_{s'} T(s, \pi(s), s') \Big(R(s, \pi(s), s') + \gamma V^{\pi}(s') \Big) \text{ for all } s. \tag{8}$$

  where $V^{\pi}(\cdot)$'s are unknowns, and all others variables have known values.

  This step is also called *policy evaluation*.

- Find the improved policy π'. For all s:

$$\pi'(s) \leftarrow \arg\max_{a} \sum_{s'} T(s, a, s') \Big(R(s, a, s') + \gamma V^{\pi}(s') \Big) \tag{9}$$

- If $\pi' = \pi$, return the final policy π (or π') as the optimal policy π^*. Otherwise, set $\pi = \pi'$ and repeat the above two steps until convergence ($\pi = \pi'$).

The algorithm involves finding the values for each state at each iteration, based on a particular policy π. This step is done via solving a system of linear equations, which can be performed analytically. The algorithm also comes with a guarantee: the policy always converges to *an* optimal policy eventually.

Reinforcement learning

Now, we will consider a set-up where neither reward no transitions are known a priori. Our robot can travel in the grid, moving from one state to another, collecting rewards along the way. The model of the world in unknown to the robot other than the overall Markov structure. The robot could do one of two things. First, it could try to learn the model, the reward and transition probabilities, and then solve the optimal policy using the algorithms for MDPs described above. Another option is to try to learn the Q-values directly.

Model-based learning We first assume that we can collect information about transitions involving any state s and action a. Under this assumption, we can learn T and R through experience, by collecting outcomes for each s and a.

$$T(s, a, s') = \frac{\text{Count}(s, a, s')}{\text{Count}(s, a)} \tag{10}$$

$$R(s, a, s') = \frac{\sum_t R_t(s, a, s')}{\text{Count}(s, a, s')} \tag{11}$$

where $R_t(s, a, s')$ is the reward we observed (for the t-th time) when starting in state s, taking action a, and transitioning to s'. If the reward is noisy, observed rewards R_t may vary from one instance to another. In reality, this naive approach is highly ineffective for any non-trivial state space. The best we can do is randomly explore, taking actions and moving from one state to another. Most likely, we will be unable to reach many parts of the state space in any complex environment. Moreover, the learned model would be quite large as we'd have to store all the states and possible transitions.

Model-free Learning Can we learn how to act without learning a full model? Remember:

$$\pi^*(s) = \arg\max_a Q^*(s, a) \tag{12}$$

We have shown that Q-values can be learned recursively, assuming we have access to T and R. Since this information is not provided to us, we will consider Q-learning algorithm, a sample based Q-value iteration procedure.

To better understand the difference between model-based and model-free estimation, consider the task of computing the expected value of a function $f(x) : E[f(x)] = \sum_x p(x)f(x)$

- **Model-based computation**: First estimate $p(x)$ from samples and then compute expectation:

$$x_i \approx p(x), i = 1, \ldots, k \tag{13}$$

$$\hat{p}(x) = \frac{\text{Count}(x)}{k} \tag{14}$$

$$E[f(x)] \approx \sum_x \hat{p}(x)f(x) \tag{15}$$

- **Model-free estimation**: estimate expectation directly from samples

$$x_i \approx p(x), i = 1, \ldots, k \tag{16}$$

$$E[f(x)] \approx \frac{1}{k}\sum_{i=1}^{k} f(x_i) \tag{17}$$

Now we will apply the model-free learning approach to the estimation of Q-values. Recall,

$$Q^*_{i+1}(s, a) \leftarrow \sum_{s'} T(s, a, s')[R(s, a, s') + \gamma \max_{a'} Q^*_i(s', a')] \tag{18}$$

We will repeatedly take one action at a time, and observe the reward and the next state. We can compute:

$$\text{Sample 1}: \quad R(s,a,s'_1) + \gamma \max_{a'} Q_i(s'_1, a') \tag{19}$$

$$\text{Sample 2}: \quad R(s,a,s'_2) + \gamma \max_{a'} Q_i(s'_2, a') \tag{20}$$

$$\dots \tag{21}$$

$$\text{Sample } k: \quad R(s,a,s'_k) + \gamma \max_{a'} Q_i(s'_k, a') \tag{22}$$

Now we can average all the samples, to obtain the Q-value estimate:

$$Q_{i+1}(s,a) \leftarrow \frac{1}{k} \sum_{l=1}^{k} \left[R(s,a,s'_l) + \gamma \max_{a'} Q_i(s'_l, a') \right] \tag{23}$$

which, for large k, would be very close to the Q-value iteration step. We are almost there. In practice, we only observe the states when we actually move. Therefore, we cannot really collect all these sample at once. Instead, we will update the Q-values after every experience (s, a, s', r), where r is the reward.

Assume we have observed $k-1$ samples related to (s,a) so far, and now we just observed the k-th sample and we would like to update the Q-value.

The update function is:

$$Q_{new}(s,a) \leftarrow \frac{(k-1)Q_{old}(s,a) + R(s,a,s'_k) + \gamma \max_{a'} Q_{old}(s'_k, a')}{k} \tag{24}$$

Simplifying the formula, we arrive at the following update function:

$$Q(s,a) \leftarrow Q(s,a) + \frac{1}{k} \left[R(s,a,s'_k) + \gamma \max_{a'} Q(s'_k, a') - Q(s,a) \right] \tag{25}$$

This fact leads to the following Q-learning algorithm.

Q-learning Algorithm

- Collect a sample: s, a, s' and $R(s,a,s')$.
- Update Q-values, by incorporating the new sample into a running average over samples:

$$Q(s,a) \leftarrow (1-\alpha)Q(s,a) + \alpha \left[R(s,a,s') + \gamma \max_{a'} Q(s', a') \right] \tag{26}$$

$$= Q(s,a) + \alpha \left[R(s,a,s') + \gamma \max_{a'} Q(s', a') - Q(s,a) \right] \tag{27}$$

where the learning rate α takes the role of $1/k$ in the earlier sample average example. During the iterations of the algorithm, likely s' will contribute more often to the Q-values estimate. As the algorithm progresses, old estimates fade, making the Q-value more consistent with more recent samples.

You may have noticed that the form of the update closely resembles stochastic gradient decent. In fact, it has the same convergence conditions as the gradient ascent algorithm. Each sample corresponds to (s, a), *i.e.*, being in state s and taking action a. We can assign a separate learning rate for each such case, *i.e.*, $\alpha = \alpha_k(s, a)$, where k is the number of times that we saw (s, a). Then, in order to ensure convergence, we should have

$$\sum_k \alpha_k(s, a) \to \infty \tag{28}$$

$$\sum_k \alpha_k^2(s, a) < \infty \tag{29}$$

Obviously $\alpha_k(s, a) = 1/k$ satisfies the above two conditions.

Exploration/Exploitation Trade-Off In the Q-learning algorithm, we haven't specified how to select an action for a new sample. One option is to do it fully randomly. While this exploration strategy has a potential to cover a wide spectrum of possible actions, most likely it will select plenty of suboptimal actions, and leads to a poor exploration of the relevant (high reward) part of the state space. Another option is to exploit the knowledge we have already obtained during previous iterations. Remember that once we have Q estimates, we can compute a policy. Since our estimates are noisy in the beginning, and the corresponding policy is weak, we wouldn't want to follow this policy completely. To allow for additional exploration, we select a random action every once in a while. Specifically, with probability ϵ, the action is selected at random and with probability $1-\epsilon$, we follow the policy induced by the current Q-values. Over time, we can decrease ϵ, to rely more heavily on the learned policy as it improves based on the Q-learning updates.

Learning Objectives

You need to know:

1. What is Q-value iteration and how to do Q-value iteration for a simple MDP problem
2. What is Q-learning and how to perform Q-learning for a simple reinforcement learning problem