Advanced Data Analysis II

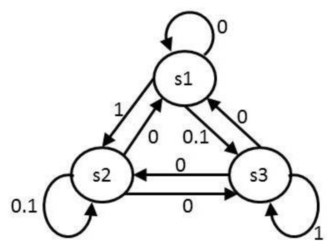
Exercise Sheet 13

Prof. Tobias Scheffer
Dr. Niels Landwehr
Uwe Dick

Winter term 2015/2016

Handed out: 26.01.16
Tutorial: 02.02.16

Exercise 1 *Policy Iteration*



Let a Markov decision process (S, A, P, R, γ) be defined as follows. Let state space $S = \{s_1, s_2, s_3\}$, action space $A = \{a_1, a_2, a_3\}$, and deterministic transition probabilities

$$P(s_i|s, a_j) = 1, \text{ if } i = j \text{ and } 0 \text{ otherwise.}$$

The immediate reward R is defined in the graph and the discount factor is $\gamma = 0.5$.
The goal is to learn the value function of a deterministic policy π_1, which is defined as:

$$\pi_1(s_1) = a_2, \pi_1(s_2) = a_1, \pi_1(s_3) = a_2$$

a) Compute an approximation of the value function $Q^{\pi_1}(s, a), \forall s, a$, starting from an initial $\hat{Q}_0(s, a) = 0, \forall s, a$ using value iteration for policy evaluation. Assume the model is fully defined. Stop the computation after 2 full iterations.

b) Assume that the following state action sequence is observed while using a behavior policy π_b.

$$s_1, a_2, s_2, a_1, s_1, a_3, s_3, a_3, s_3, a_2, s_2, a_3, s_3, a_1, s_1, a_1, s_1, a_3, s_3$$

Compute an approximation of Q^{π_1}.

c) Compute an approximation of Q^{π_1} after 10 steps if the state-action sequence is drawn on-policy, starting from s_3.

d) Compute the next policy π_2 using the greedy policy improvement step based on each of the three approximations of Q^{π_1}.

Exercise 2 *Value Iteration*

Use the above MDP to approximate the optimal value function.

a) Compute an approximation of the optimal value function $Q^*(s,a), \forall s,a$, starting from an initial $\hat{Q}_0(s,a) = 0, \forall s,a$ using Value Iteration. Assume the model is fully defined. Stop the computations after 2 full iterations.

b) Compute an approximation of Q^* when the following sequence is drawn off-policy by a behavior policy π_b:

$$s_1, a_2, s_2, a_1, s_1, a_3, s_3, a_3, s_3, a_2, s_2, a_3, s_3, a_1, s_1, a_1, s_1, a_3, s_3$$

Exercise 3 *TD(λ)*

Compute an approximation of Q^π using TD(λ) and $\lambda = 0.5$, where the following on-policy samples are drawn using π.

$$s_1, a_2, s_2, a_2, s_2, a_1, s_1, a_3, s_3, a_3, s_3, a_2, s_2, a_1, s_1, a_3, s_3, a_3, s_3$$

Try both update rules for the additional table e of the eligibility traces and show the differences.

$$e(s) \leftarrow e(s) + 1 \quad \text{Accumulating Traces} \tag{1}$$
$$e(s) \leftarrow 1 \quad \text{Replacing Traces} \tag{2}$$

What problems would appear if one would augment the off-policy method Q-Learning with eligibility traces?