Advanced AI Techniques

Prof. Dr. Burgard, Prof. Dr. Nebel, Dr. Kersting — University of Freiburg
M. Ragni, A. Rottmann — Department of Computer Science
WS 2006/2007

Exercise Sheet 4

Due: Tuesday, 28. November 2006

Exercise 4.1 (MiniMax)

Consider the game of kicking a penalty. Assume that you are player A and have to kick the ball. You can choose between three different actions. You can kick the ball in the left corner (action $= L$), in the middle (M), or in the right corner (R) of the goal. Correspondingly the goalkeeper (player B) can do the following actions: jump to the left (action $= l$), stay in the middle(m) or jump to the right (r). Lets assume if the goalkeeper choose the right action, he will always catch the ball. Furthermore we assume that kicking the ball in a corner is more difficult and therefore you will receive a higher reward (at least in your mind). Finally you will get the following game-table:

	l	m	r
L	-2	1	2
M	2	-1	2
R	2	1	-2

Table 1: Reward for kicking a penalty

Determine the result of the minimax algorithm for this game and compare it with a reinforcement learning approach. What would change if you recognize after some kicks that the goalkeeper is very lazy and remains most of the time in the middle of the goal (he choose action m).

This is a typical situation in ordinary life. Often it is not clear which strategy your fellows follow. In our case, if you do not know which strategy the keeper uses, you can learn a rewarding behavior strategy by kicking repeatedly and observing what the keeper is going to do. What are the advantages and disadvantages of reinforcement learning in such real life situations?

Exercise 4.2 (3-armed Bandit Problem)

Consider the three actions a, b, and c. The reward distributions are normal Gaussian distributions with a standard deviation of 1 and the following mean values:

action	$E(r\|action)$
a	5
b	4
c	2

Table 2: Mean expected reward

In order to find out which is the best action, implement the action evaluation method "sample average". Start with several initial estimated qualities and apply in different runs of your program the following three action selection methods:

- greedy,
- ϵ-greedy with $\epsilon = 0.01$,
- ϵ-greedy with $\epsilon = 0.1$

Plot for each run the averages $Q_t(a)$ for all actions. Compare your final quality estimation for the three actions with the real values. Which of the three runs yielded the best action quality estimation and why?

Exercise 4.3 (Bellman Equation)

Derive the Bellman equation for Q^π without the expectation operator from the form

$$Q^\pi(s, a) = E_\pi\{\Sigma_{k=0}^{\infty}\gamma^k r_{t+k+1}|s_t = s, a_t = a\}.$$

This derivation can be done analogously to the one for $V^\pi(s)$ presented in the lecture.