Tik-79.230 **Kevät 2005**

Agenttipohjaisen tietojenkäsittelyn perusteet

Laskuharjoitus 6

Ratkaisut

1. Let S be the state space of an gent and A the set of possible actions. A *policy* P is then a function $S \to A$. In other words, a policy attachs a unique action a to each state s, and the agent executes a every time it is in s. A policy P is *optimal policy*, if each chosen action maximises the agents expected utility.

   In the exercise $S = \{(1,1),(2,1),(3,1),(2,2),(3,2)\}$ and $A = \{\leftarrow, \uparrow, \rightarrow, \downarrow\}$.

■		+1
S		-1

   (a) In *value iteration* we compute on each step the expected utility values of each action in each state, and the action with maximal expected utility is chosen as a policy of that state. Then, the utility value of a state i is set as follows:

$$U(i) = R(i) + \max_a \sum_j M^a_{ij} U(j),$$

   where $R(i)$ is the *reward* of the state (here 1 in $(3,2)$, and -1 in $(3,1)$, and -0.2 in all other states), and $\max_a \sum_j M^a_{ij} U(j)$ is the expected utility of the chosen action.

   This process is continued until the utility values stabilize.

   Round 1.

State	a	$E(a)$	
$(2,2)$	$\leftarrow$	$1 \cdot (-0.2) = -0.2$	
	$\uparrow$	$0.9 \cdot (-0.2) + 0.1 \cdot 1 = -0.08$	
	$\rightarrow$	$0.8 \cdot 1 + 0.2 \cdot (-0.2) = 0.76$	$\times$
	$\downarrow$	$0.9 \cdot (-0.2) + 0.1 \cdot 1 = -0.08$	
$(2,1)$	$\leftarrow$	$1 \cdot (-0.2) = -0.2$	$\times$
	$\uparrow$	$0.9 \cdot (-0.2) + 0.1 \cdot (-1) = -0.28$	
	$\rightarrow$	$0.8 \cdot (-1) + 0.2 \cdot (-0.2) = -0.84$	
	$\downarrow$	$0.9 \cdot (-0.2) + 0.1 \cdot (-1) = -0.28$	
$(1,1)$	$\leftarrow$	$1 \cdot (-0.2) = -0.2$	
	$\uparrow$	$1 \cdot (-0.2) = -0.2$	
	$\rightarrow$	$1 \cdot (-0.2) = -0.2$	
	$\downarrow$	$1 \cdot (-0.2) = -0.2$	

   So, the policy in (2,2) is → and in (2,1) it is ←. Since all actions have the same expected utility in (1,1), its policy may be chosen arbitrarily.

■	→	+1
S	←	-1

The new expected utilities are:

$$
\begin{aligned}
U(2,2) &= -0.2 + 0.76 = 0.56 \\
U(2,1) &= -0.2 - 0.2 = -0.4 \\
U(1,1) &= -0.2 - 0.2 = -0.4
\end{aligned}
$$

Round 2.

State	a	$E(a)$	
$(2,2)$	$\leftarrow$	$0.9 \cdot 0.56 + 0.1 \cdot (-0.4) = 0.464$	
	$\uparrow$	$0.9 \cdot 0.56 + 0.1 \cdot 1 = 0.604$	
	$\rightarrow$	$0.8 \cdot 1 + 0.1 \cdot 0.56 + 0.1 \cdot (-0.4) = 0.816$	$\times$
	$\downarrow$	$0.8 \cdot (-0.4) + 0.1 \cdot 0.56 + 0.1 \cdot 1 = -0.164$	
$(2,1)$	$\leftarrow$	$0.9 \cdot (-0.4) + 0.1 \cdot 0.56 = -0.304$	
	$\uparrow$	$0.8 \cdot 0.56 + 0.1 \cdot (-1) + 0.1 \cdot (-0.4) = 0.308$	$\times$
	$\rightarrow$	$0.8 \cdot (-1) + 0.1 \cdot (-0.4) + 0.1 \cdot 0.56 = -0.784$	
	$\downarrow$	$0.9 \cdot (-0.4) + 0.1 \cdot (-1) = -0.46$	
$(1,1)$	$\leftarrow$	$1 \cdot (-0.4) = -0.4$	
	$\uparrow$	$1 \cdot (-0.4) = -0.4$	
	$\rightarrow$	$1 \cdot (-0.4) = -0.4$	
	$\downarrow$	$1 \cdot (-0.4) = -0.4$	

The new policy is:

■	→	+1
S	↑	-1

and the new utility values are:

$$
\begin{aligned}
U(2,2) &= -0.2 + 0.816 = 0.616 \\
U(2,1) &= -0.2 + 0.308 = 0.108 \\
U(1,1) &= -0.2 - 0.4 = -0.6
\end{aligned}
$$

Continuing the iteration, the policies of (2,2) and (2,1) stay unchanged. The state (1,1) finally gets a policy since now (2,1) has higher utility than (1,1). Thus, the policy is:

■	→	+1
→	↑	-1

This is actually the optimal policy, but it takes still several iterations until the utility values finally stabilize.

(b) In *policy iteration* we start by creating a random policy. Then, we compute the utility values of states given that policy, update the policy by choosing the actions with highest expected utilities, and compute new utility values. This continues until the policy stabilizes. Suppose that the random policy is as follows:

■	↑	+1
→	←	-1

The utilities given the policy can be computed analytically by solving the following group of equations: (Here u_{ij} denotes the utility of the state (i, j)).

$$\begin{aligned}
u_{11} &= 0.2u_{11} + 0.8u_{21} - 0.2 \\
u_{21} &= 0.8u_{11} + 0.1u_{21} + 0.1u_{22} - 0.2 \\
u_{22} &= 0.9u_{22} + 0.1 \cdot 1 - 0.2
\end{aligned}$$

The solution is:

$$\begin{aligned}
u_{11} &= -5.25 \\
u_{21} &= -5 \\
u_{22} &= -1
\end{aligned}$$

Now we compute the expected utilities for different actions:

State	a	$E(a)$	
$(2, 2)$	$\leftarrow$	$0.9 \cdot (-1) + 0.1 \cdot (-5) = -1.4$	
	$\uparrow$	$0.9 \cdot (-1) + 0.1 \cdot 1 = -0.8$	
	$\rightarrow$	$0.8 \cdot 1 + 0.1 \cdot (-1) + 0.1 \cdot (-5) = 0.2$	$\times$
	$\downarrow$	$0.8 \cdot (-5) + 0.1 \cdot (-1) + 0.1 \cdot 1 = -4$	
$(2, 1)$	$\leftarrow$	$0.9 \cdot (-5.25) + 0.1 \cdot (-1) = -4.825$	
	$\uparrow$	$0.8 \cdot (-1) + 0.1 \cdot (-1) + 0.1 \cdot (-5.25) = -1.425$	
	$\rightarrow$	$0.8 \cdot (-1) + 0.1 \cdot (-5) + 0.1 \cdot (-1) = -1.4$	$\times$
	$\downarrow$	$0.8 \cdot (-5) + 0.1 \cdot (-1) + 0.1 \cdot (-5.25) = -4.625$	
$(1, 1)$	$\leftarrow$	$1 \cdot (-5.25) = -5.25$	
	$\uparrow$	$0.9 \cdot (-5.25) + 0.1 \cdot (-5) = -5.225$	
	$\rightarrow$	$0.8 \cdot (-5) + 0.2 \cdot (-5.25) = -5.05$	$\times$
	$\downarrow$	$0.9 \cdot (-5.25) + 0.1 \cdot (-5) = -5.225$	

The new policy is:

■	→	+1
→	→	-1

In the next step the policy of (2,1) changes to the optimal action ↑.