Tik-79.230 **Kevät 2002**
Agenttipohjaisen tietojenkäsittelyn perusteet
Laskuharjoitus 4
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:

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

   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 $\max_a \sum_j M_{ij}^a U(j)$ 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 $\rightarrow$ and in (2,1) it is $\leftarrow$. 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:

	$\uparrow$	+1
$\rightarrow$	$\leftarrow$	-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:

	$\rightarrow$	+1
$\rightarrow$	$\rightarrow$	-1

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

2. When we examine how the beliefs of an agent change as a function of time, we divide processing of each time step into two parts:

- *Estimation*, when we use our current set of beliefs ($Bel(\mathbf{X_t})$) to estimate what will the situation be in the next time step$t + 1$. We denote this probability distribution by $\widehat{Bel}(\mathbf{X_{t+1}})$.

$$\widehat{Bel}(\mathbf{X_{t+1}}) = \sum_{\mathbf{X_t}} \mathbf{P}(\mathbf{X_{t+1}} \mid \mathbf{X_t} = \mathbf{x_t}) Bel(\mathbf{X_t} = \mathbf{x_t})$$

- *Assessment*, when we get observations $\mathbf{E_{t+1}}$ of the time step $t+1$. They are added to the estimation using Bayesian updating:

  $$Bel(\mathbf{X_{t+1}}) = \alpha \mathbf{P}(\mathbf{E_{t+1}} \mid \mathbf{X_{t+1}}) \widehat{Bel}(\mathbf{X_{t+1}}).$$

  Here α is a normalization parameter that is used to ensure that the probabilities sum to unity.

Let us write the probabilities given in the exercise using a clearer notation:

	U_{t+1}	$\neg U_{t+1}$
U_t	0.6	0.4
$\neg U_t$	0.1	0.9

	V_{t+1}	$\neg V_{t+1}$
V_t	0.6	0.4
$\neg V_t$	0.3	0.7

We first estimate what the situation will be on t_1 when we know the probability distribution of t_0 $(Bel(U_0) = 0.1, Bel(V_0) = 0.3)$:

$$\begin{aligned}
\widehat{Bel}(U_1) &= P(U_1 \mid \neg U_0) Bel(\neg U_0) + P(U_1 \mid U_0) Bel(U_0) \\
&= 0.9 \cdot 0.1 + 0.1 \cdot 0.6 = 0.15 \\
\widehat{Bel}(\neg U_1) &= P(\neg U_1 \mid \neg U_0) Bel(\neg U_0) + P(\neg U_1 \mid U_0) Bel(U_0) \\
&= 0.85 \\
\widehat{Bel}(V_1) &= P(V_1 \mid \neg V_0) Bel(\neg V_0) + P(V_1 \mid V_0) Bel(V_0) \\
&= 0.39 \\
\widehat{Bel}(\neg V_1) &= P(\neg V_1 \mid \neg V_0) Bel(\neg V_0) + P(\neg V_1 \mid V_0) Bel(V_0) \\
&= 0.61
\end{aligned}$$

Then, on t_1 we observe that both doors are open, so our evidence is $A_1 \wedge B_1$. Next, we compute how probable it is that U is really open on t_1 when it is observed to be open:

$$\begin{aligned}
Bel(U_1) &= \alpha_{U_1} P(A_1 \mid U_1) \widehat{Bel}(U_1) \\
&= \alpha_{U_1} \cdot 0.9 \cdot 0.15 = 0.135 \alpha_{U_1} \\
Bel(\neg U_1) &= \alpha_{U_1} P(A_1 \mid \neg U_1) \widehat{Bel}(\neg U_1) \\
&= \alpha_{U_1} \cdot 0.1 \cdot 0.85 = 0.085 \alpha_{U_1}
\end{aligned}$$

The normalization constant is α_{U_1} is computed as follows:

$$\alpha_{U_1} = \frac{1}{0.135 + 0.085} = 4.54$$

This gives us:

$$\begin{aligned}
Bel(U_1) &= 0.61 \\
Bel(\neg U_1) &= 0.39
\end{aligned}$$

Now, we compute the same values for door V:

$$
\begin{aligned}
Bel(V_1) &= \alpha_{V_1} P(B_1 \mid V_1)\widehat{Bel}(V_1) \\
&= 0.27 \cdot \alpha_{V_1} \\
Bel(\neg V_1) &= \alpha_{V_1} P(B_1 \mid \neg V_1)\widehat{Bel}(\neg V_1) \\
&= 0.06 \cdot \alpha_{V_1} \\
\alpha_{V_1} &= \frac{1}{0.27 + 0.06} = 3.03 \\
Bel(V_1) &= 0.82 \\
Bel(\neg V_1) &= 0.18
\end{aligned}
$$

This concludes our assesment at t_1. Now we use exactly the same procedure to compute the beliefs at t_2 when we observe that both doors are closed $(\neg A_2 \wedge \neg B_2)$.

$$
\begin{aligned}
\widehat{Bel}(U_2) &= P(U_2 \mid U_1)Bel(U_1) + P(U_2 \mid \neg U_1)Bel(\neg U_1) \\
&= 0.6 \cdot 0.61 + 0.39 \cdot 0.1 = 0.41 \\
\widehat{Bel}(\neg U_2) &= 0.59 \\
Bel(U_2) &= \alpha_{U_2} P(\neg A_2 \mid U_2)\widehat{Bel}(U_2) \\
&= 0.04\alpha_{U_2} \\
Bel(\neg U_2) &= \alpha_{U_2} P(\neg A_2 \mid \neg U_2)\widehat{Bel}(\neg U_2) \\
&= 0.53\alpha_{U_2} \\
\alpha_{U_2} &= \frac{1}{0.04 + 0.53} \\
Bel(U_2) &= 0.07 \\
Bel(\neg U_2) &= 0.93 \\
\widehat{Bel}(V_2) &= P(V_2 \mid V_1)Bel(V_1) + P(V_2 \mid \neg V_1)Bel(\neg V_1) \\
&= 0.55 \\
\widehat{Bel}(\neg V_2) &= 0.45 \\
Bel(V_2) &= \alpha_{V_2} P(\neg A_2 \mid V_2)\widehat{Bel}(V_2) \\
&= 0.16\alpha_{V_2} \\
Bel(\neg V_2) &= \alpha_{V_2} P(\neg A_2 \mid \neg V_2)\widehat{Bel}(\neg V_2) \\
&= 0.41\alpha_{V_2} \\
\alpha_{V_2} &= \frac{1}{0.16 + 0.41} \\
Bel(V_2) &= 0.28 \\
Bel(\neg V_2) &= 0.72
\end{aligned}
$$

We were asked how certain we can be that on t_2 at least one of the doors is closed:

$$
\begin{aligned}
Bel(\neg(U_2 \wedge V_2)) &= 1 - Bel(V_2)Bel(U_2) \\
&= 1 - 0.01 = 0.99
\end{aligned}
$$