Toolbox I.A. - ICM2A – ENSM-SE

Reinforcement Learning

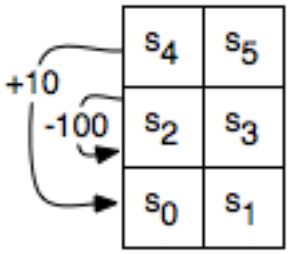
Practical Work

Friday 8th January 2016

Agenda :

- **half an hour : a small exercise below, paper copy, without IT help**
- around an hour and a half : online tutorial
- final hour : a second small exercise … IT help allowed :-)

A Tiny Game



Consider the tiny environment shown above. There are six states the agent could be in, labeled as $s_0,...,s_5$. The agent has four actions: *UpC*, *Up*, *Left*, *Right*. That is all the agent knows before it starts. It does not know how the states are configured, what the actions do, or how rewards are earned.

Suppose the actions work as follows:

- **UpC** (for "up carefully"): The agent goes up, except in states s_4 and s_5 where the agent stays still. In all the cases it has a reward of *-1*.
- **Right**: The agent moves to the right in states s_0,s_2,s_4 with a reward of 0, and stays still in the other states with a reward of *-1*.
- **Left**: The agent moves one state to the left in states s_1,s_3,s_5 with a reward of 0. In state s_0, it stays in state s_0 and has a reward of *-1*. In state s_2, it has a reward of *-100* and stays in state s_2. In state s_4, it gets a reward of *10* and moves to state s_0.
- **Up**: With a probability of *0.8* it acts like *upC*, except the reward is 0. With probability *0.1* it acts as a *left*, and with probability *0.1* it acts as *right*.

Assume the agent, starting in s_0, wants to act in order to gain the higher long-term discounted reward with a discount factor γ=0.9

According to you, what is the best (*or at least a good*) policy, and why ?

Around an hour and a half :

Online tutorial : http://www.aispace.org/exercises/exercise11-a-1.shtml

---

Final hour :

Q1 : From now on, according to you, what is the best policy?

Q2 : What is your strategy to make the agent learn this policy?
