First Problem Set for Machine Learning B

708.068 KU Maschinelles Lernen B, WS 2006/07

Latest Submission Date: October 30, before the lecture at 13:15.

Guidelines:
- Write down clearly your name and matriculation number, as well as the names and matriculation numbers of all team members on the first page of your submission.
- Write down your solutions on separate sheets of paper for every exercise.
- If you use a theorem from the script in your proof, always cite its number or the number of the particular equation you are using.
- For proofs of theorems from the script you can use any results that appear before in the script. It is not allowed to use later results.
- Exercises marked with * are bonus exercises.

1. [10 Points] Prove Corollary 1.1. from the script.
2. [10 Points] Prove Corollary 1.3. from the script.
3. [10 Points] Assume that for a given continuing MDP with discount factor $\gamma < 1$ we modify the reward signal by either
   (a) adding a constant $d$ to all rewards
   (b) multiplying every reward with a constant $k > 0$
   (c) linearly transforming the reward signal to $k \cdot r + d$, $k > 0$

   Can this change the optimal policy of the MDP? Express for all three cases the new state values in terms of $V(s)$, $\gamma$ and the constants (where $V(s)$ is the optimal value of state $s$ under the original reward function).

Now consider the following modifications for deterministic MDPs:

(d) Let $(s_{max}, a_{max})$ be the state-action pair that leads to the highest possible immediate reward $r_{max} = \max_{s,a} r(s, a)$ in the MDP. Set $r(s_{max}, a_{max}) \leftarrow r_{max} + d$, $d > 0$

(e) Let $(s_{min}, a_{min})$ be the state-action pair that leads to the lowest possible immediate reward $r_{min} = \min_{s,a} r(s, a)$ in the MDP. Set $r(s_{min}, a_{min}) \leftarrow r_{min} - d$, $d > 0$

For simplicity, you can assume in both cases that the minimum/maximum is unique, i.e. it is taken on exactly at one state-action pair. Can you guarantee for arbitrary MDPs that the optimal policy stays the same? If not, show a counterexample.

4. [10 Points] Consider the following game: You have a random number generator that produces in every round an integer number from 1 to 3 with equal probability. You play 3 rounds and have to decide at which position of a 3 digit number you want to place the random digit.
   Your goal is to form the largest possible (decimal) number. Formulate this game as a Markov decision process and find an optimal policy. Also analyze the case where the numbers are drawn without replacement, i.e. if the digit 3 appears in the first round, it cannot appear anymore in the remaining two rounds.
5. *[10 *-Points] Suppose you have to solve the following real-world or simulated tasks with reinforcement learning. Define the components of a MDP (states, actions, rewards, transition, discount) that capture all aspects of the given task. For some tasks you may need to define additional sensors to generate the state signal. Does your state signal have the Markov property? Explain why your setup (especially the reward function) leads to a correct optimal policy. There are multiple solutions for every task.

a) An agent in a virtual car in a computer game is driving on a circuit without opponents. The goal is to finish one round on the track as fast as possible. The agent has a map of the track, a camera, and can sense the borders of the track, which are concrete walls (thus there is no shortcut). The car is damaged whenever the agent hits the wall.

b) A humanoid robot has to hit a golf ball with a club into a hole. The robot has a vision system to recognize the ball and the position of the hole. The goal is to hit the ball into the hole with a minimum number of strokes. Assume that the robot has a sub-routine available that executes a golf-swing, given the desired force and angle of the stroke (relative to some local coordinate system).

c) A single truck needs to deliver goods from a central depot to several customers. The demands of the customers are equal, and the truck may load maximally two units of goods. The driving times between the customers are known. The goal is to serve all customers in the shortest possible time. Consider also the case that a fleet of trucks is available, but the company wants to avoid using too many trucks in order to save costs.