Reinforcement Learning Exercise 0

Team: Nico Ott 4214197, Lior Fuks 4251285, Hendrik Vloet 4324249

Date: November 8, 2017

Due Date: November 9, 2017

1 Introduction to RL

- **Model**: A model is an agent's internal representation of the environment that predicts its behavior, i.e. forecasts the environment's evolution, regarding an action of the agent.
  A model may help a policy to find the best next action and it consists of 2 components:
  - State-Transition model $\mathcal{P}$. It predicts the next state of the agent after executing an action. Formalized:

    $$\mathcal{P}^a_{SS'} = \mathbb{P}\left\{S_{t+1} = s' | S_t = s, A_t = a\right\}$$

  - Reward model $\mathcal{R}^a_s$. It predicts the next expected (immediate) reward for the agent after performing some action, formalized:

    $$\mathcal{R}^a_s = \mathbb{E}\{R_{t+1} | S_t = s, A_t = a\}$$

- **Policy**: the policy is the behavior of the agent, i.e. it is a mapping function from state to an action. Or in other words: it is a strategy how the agent will choose its next action, taking into account in which state it is. Nevertheless, a policy does not guarantee optimal performance, since it just represents one possible behavior of the agent out of infinitely many. Briefly, it can be good or bad and thus increase or decrease performance.
  We can divide them into two overall classes:
  - deterministic policy, where the outcome of an evaluation of the state is exactly known:

    $$a = \pi(s)$$

  - stochastic policy, where the outcome of an evaluation of the state is not exactly known but can be described with the help of probabilistic calculus:

    $$\pi(a|s) = \mathbb{P}(A_t = a | S_t = s)$$

- **Value-function**: predicts the future reward of the agent and is used to evaluate the goodness/badness of states, i.e. the value function can be used to pick actions, depending of the outcome. The value functions "unrolls" future possible rewards under some policy π. Usually, the value function cannot predict the future with absolute certainty. Therefore, expected rewards of the future will have less influence on the value of the actual state("Myopic Evaluation"):

  $$v_\pi^{(s)} = \mathbb{E}_\pi\{R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + ... | S_t = s\}$$

2 GYM

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.