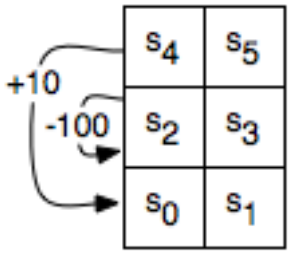
1 MDPs

13

Betrachten Sie das folgende Reinforcement-Learning-Problem:



Es gibt 6 Zustände $s_0, \ldots, s_5$ und 4 Aktionen UpC, Up, Left, Right. Die Aktionen funktionieren wie folgt:

- UpC: Der Agent geht hoch, außer in s_4 und s_5, wo nichts passiert. Reward: -1
- Right: Der Agent bewegt sich nach rechts, in s_0, s_2, s_4, mit Reward 0. In den anderen Zuständen passiert nichts, und der Reward ist -1.
- Left: Der Agent bewegt sich nach links, in s_1, s_3, s_5, mit Reward 0. In s_0 passiert nichts, und der Reward ist -1. In s_2, ist der Reward -100, und er bleibt in s_2. In s_4 ist der Reward 10, der Agent landet in s_0.
- Up: Mit Wahrscheinlichkeit 0.8, wie upC, mit Reward 0. Mit Wahrscheinlichkeit 0.1, gehe nach links, und mit Wahrscheinlichkeit 0.1 nach rechts (mit dem Reward, den die jeweilige Left oder Right Aktion hätte).

1. Geben Sie die Reward-Matrix R an. 3
2. Führen Sie Q-Learning durch, mit $\alpha = 1$, $\gamma = 0.5$. Der Startzustand sei s_0 und die Abfolge von Aktionen und Zuständen (upC,s_2),(left,s_2),(up,s_4),(left,s_0),(upC,s_2). Geben Sie für jeden Schritt die resultierende Q-Matrix an. 10

2 HMM Implementierung

50

Implementieren Sie das in Übung 6 beschriebene Hidden Markov Model in Python. Die Verteilung $p(x_t)$ soll durch ein Data Frame mit den Spalten i und $p(x_t = i)$ dargestellt werden ($i \in$ {Wach,Schlaeft} sind die Zustände). Alle weiteren Verteilungen sollen ebenfalls durch Data Frames dargestellt werden. Implementieren Sie folgende Funktionen:

3. trans(x): Berechne $p(x_{t+1}|x_t = x)$. 5
4. likelihood(y,x): Gibt $p(y|x)$, d.h. die Wahrscheinlichkeit einer bestimmten Beobachtung, gegeben ein Zustand x, aus. 5
5. predict($p(x_t)$): Gegeben eine Prior-Verteilung von Zuständen $p(x_t)$, berechne die Verteilung nach dem Predict-Schritt, d.h. $p(x_{t+1}|x_t)$. 15
6. update($p(x_t)$,y_t): Berechne $p(x_t|y_t)$. 10
7. filter($p(x_0)$,$y_{1:T}$): Gibt Liste von Zustands-Verteilungen $p(x_t)$ für $t = 1 \ldots T$ aus, gegeben eine Observation-Sequenz $y_{1:T}$ und eine initiale Verteilung $p(x_0)$. 10

Testen Sie Ihre Implementierung mit der Observations-Sequenz [Bewegung, Bewegung, Bewegung, Ruhe, Ruhe, Bewegung, Ruhe, Ruhe, Ruhe]. 5