Maschinelles Lernen II

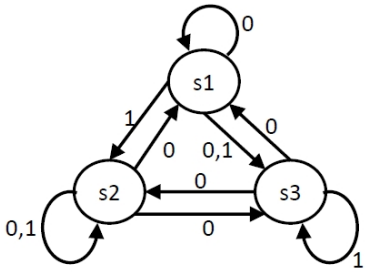
6. Übung

Prof. Tobias Scheffer
Dr. Niels Landwehr
Uwe Dick

Sommer 2014

Ausgabe am: 20.05.14
Besprechung am: 27.05.14

Aufgabe 1 *Policy Iteration*



Es sei der folgende MDP (S, A, P, R, γ) gegeben. Eine Zustandsmenge $S = \{s_1, s_2, s_3\}$ und Aktionsmenge $A = \{a_1, a_2, a_3\}$ mit deterministischen Zustandsübergängen

$$P(s_i|s, a_j) = 1, \text{ falls } i = j \text{ und } 0 \text{ sonst.}$$

und direktem Gewinn R wie aus der Graphik ersichtlich. Der Discount-Faktor ist $\gamma = 0,5$. Das Ziel ist es, die Bewertungsfunktion für eine deterministische Policy π zu lernen, die folgerndermaßen definiert ist:

$$\pi(s_1) = a_2, \pi(s_2) = a_1, \pi(s_3) = a_2$$

a) Berechnen sie eine Approximation der Bewertungsfunktion $Q^\pi(s, a), \forall s, a$, ausgehend von einem initialen $\hat{Q}_0^\pi(s, a) = 0, \forall s, a$ mit Hilfe von Value Iteration für Policy Evaluation unter der Annahme eines vollständig bekannten Modells. Stoppen sie die Berechnung nach 2 vollständigen Iterationen.

b) Berechnen sie eine Approximation von Q^π, falls die Beispiele von einer Verhaltenspolicy π_b gezogen wurden, mit der folgende Zustands-Aktionsfolge gezogen wurde:

$$s_1, a_2, s_2, a_1, s_1, a_3, s_3, a_3, s_3, a_2, s_2, a_3, s_3, a_1, s_1, a_1, s_1, a_3, s_3$$

c) Berechnen sie eine Approximation von Q^π nach 10 Schritten, falls die Zustands-Aktions-Sequenz On-Policy gezogen wurde, ausgehend von s_3.

Aufgabe 2 *Value Iteration*

Verwenden sie nun den obigen MDP um eine Approximation der optimalen Policy zu finden.

a) Berechnen sie eine Approximation der optimalen Bewertungsfunktion $Q^*(s,a), \forall s,a$, ausgehend von einem initialen $\hat{Q}_0(s,a) = 0, \forall s,a$ mit Hilfe von Value Iteration unter der Annahme eines vollständig bekannten Modells. Stoppen sie die Berechnung nach 2 vollständigen Iterationen.

b) Berechnen sie eine Approximation von Q^*, falls die Beispiele von einer Verhaltenspolicy π_b gezogen wurden, mit der folgende Zustands-Aktionsfolge gezogen wurde:

$$s_1, a_2, s_2, a_1, s_1, a_3, s_3, a_3, s_3, a_2, s_2, a_3, s_3, a_1, s_1, a_1, s_1, a_3, s_3$$

Aufgabe 3 *TD(λ)*

Berechnen sie nun eine Approximation on Q^{π_2} mit Hilfe von TD(λ) und $\lambda = 0,5$, wobei die folgende on-policy Samplefolge von π_2 verwendet werden soll.

$$s_1, a_2, s_2, a_2, s_2, a_1, s_1, a_3, s_3, a_3, s_3, a_2, s_2, a_1, s_1, a_3, s_3, a_3, s_3$$

Verwenden sie dazu die beiden folgenden Möglichkeiten der Updateregel für den e-Speicher der Eligibility Traces.

$$e(s) \leftarrow e(s) + 1 \quad \text{Accumulating Traces} \tag{1}$$

$$e(s) \leftarrow 1 \quad \text{Replacing Traces} \tag{2}$$

Was für Probleme könnten sich ergeben, wenn man den off-policy Lernalgorithmus Q-Learning um Eligibility Traces erweitern würde?