Maschinelles Lernen II

7. Übung

Prof. Tobias Scheffer
Dr. Niels Landwehr
Uwe Dick

Sommer 2014

Ausgabe am: 28.05.14
Besprechung am: 03.06.14

Aufgabe 1 *Approximate Policy Evaluation*

$$\phi_{s_1,a_1} = \begin{pmatrix}0.2\\0.4\end{pmatrix}, R_1 = 1, \phi_{s_2,a_2} = \begin{pmatrix}0.4\\0.4\end{pmatrix}, R_2 = 0.3, \phi_{s_3,a_3} = \begin{pmatrix}0.5\\0.2\end{pmatrix}, R_3 = 0.5, \phi_{s_4,a_4} = \begin{pmatrix}0.5\\0.1\end{pmatrix}$$

a) Stellen Sie das Optimierungskriterium auf, dass die Bewertungsfunktion $\hat{Q}$ für die Policy π schätzt, aus der obige Zustands-Aktions-Reward-Folge gezogen wurde. Dabei sei $\hat{Q}$ definiert als lineare Funktion $\hat{Q}(s,a;\theta) = \phi_{s,a}^\top \theta$. Verwenden Sie dazu den empirischen Schätzer $\hat{L}_{BRM}(\hat{Q};\pi,3)$ entsprechend dem Prinzip der Bellman Residuen Minimierung aus der Vorlesung. Der Discount-Faktor ist $\gamma = 0,5$.

b) Berechnen Sie eine Approximation des optimalen Parametervektors θ^*, indem Sie ausgehend von einem Startvektor $\theta_0 = \begin{pmatrix}1\\1\end{pmatrix}$ drei Berechnungsschritte mit Hilfe der approximativen (online) TD(0)-Methode durchführen. Als Schrittweitenfolge verwenden Sie $\alpha_t = 1/t$.

c) Berechnen Sie nun eine Approximation des optimalen Parametervektors θ^*, indem Sie ausgehend von einem Startvektor $\theta_0 = \begin{pmatrix}1\\1\end{pmatrix}$ drei Berechnungsschritte mit Hilfe der (online) Residual Gradient Methode durchführen. Als Schrittweitenfolge verwenden Sie $\alpha_t = 1/t$.

Aufgabe 2 *Policy Gradient*

Lernen Sie eine optimale stochastische Policy für einen unbekannten MDP mit kontinuierlichem Zustandsraum S und drei Aktionen $A = \{b_1, b_2, b_3\}$. Verwenden Sie dazu die Policy Gradient Methode. Nehmen Sie an, dass Sie mit einem Start-Parametervektor $\theta_0 = (1,1)$ gestartet sind, und dass mit Hilfe des online geupdateten Parametervektors die folgende Zustand-Aktion-Folge gezogen wurde.

$$(s_1, a_1 = b_1), (s_2, a_2 = b_2), (s_3, a_3 = b_1)$$

Die entsprechenden Featurevektoren für Zustand-Aktion-Paare können Sie der folgenden

Tabelle entnehmen.

$$\phi_{s_1,b_1} = \begin{pmatrix}0.2\\0.4\end{pmatrix}, \phi_{s_1,b_2} = \begin{pmatrix}0.6\\0.2\end{pmatrix}, \phi_{s_1,b_3} = \begin{pmatrix}0.5\\0.5\end{pmatrix}$$
$$\phi_{s_2,b_1} = \begin{pmatrix}0.1\\0.7\end{pmatrix}, \phi_{s_2,b_2} = \begin{pmatrix}0.6\\0.5\end{pmatrix}, \phi_{s_2,b_3} = \begin{pmatrix}0.4\\0.4\end{pmatrix}$$
$$\phi_{s_3,b_1} = \begin{pmatrix}0.4\\0.4\end{pmatrix}, \phi_{s_3,b_2} = \begin{pmatrix}0.1\\0.4\end{pmatrix}, \phi_{s_3,b_3} = \begin{pmatrix}0.6\\0.3\end{pmatrix}$$

In einer separaten Berechnung wurde die Bewertungsfunktion geschätzt und es wurde folgende Folge berechnet: $Q(s_1, b_1) = 1, Q(s_2, b_2) = 0.1, Q(s_3, b_1) = 0.4$. Berechnen Sie die zugehörige Parametervektorfolge $\theta_1, \theta_2, \theta_3$, wenn die Policy definiert ist als

$$\pi(s, b; \theta) = \frac{e^{\phi_{sb}^\top \theta}}{\sum_{i=1}^{3} e^{\phi_{sb_i}^\top \theta}}$$

für $s \in S$ und $b \in A$. Der log-Gradient der Policy ist definiert als

$$\nabla_\theta \log \pi(s, b; \theta) = \phi_{sb} - \sum_{i=1}^{3} \pi(s, b_i; \theta)\phi_{sb_i}$$

Als Schrittweitenfolge verwenden Sie $\alpha_t = 1/t$. Der Discountfaktor ist $\gamma = 0.9$.