Universität Hamburg
Department Informatik, AB TAMS
Norman Hendrich, Jianwei Zhang

Sekretariat: T. Tetsis
Raum F-311
Tel 42883-2430, Fax 42883-2397

Übungen zum Modul WPM6: Algorithmisches Lernen

SS 2011 Blatt 4

Ausgabe: 06.07.2011, **Besprechung:** 15.07.2011

Aufgabe 4.1 Labyrinth:
Stellen Sie sich vor, Sie entwickeln einen Roboter, der auf dem kürzesten Weg einem Labyrinth entfliehen soll. Sie entscheiden sich dafür, Aktionen mit einem *Reward* von $+1$ zu belohnen, wenn der Roboter dem Labyrinth dadurch entflieht. In allen anderen Fällen ist der *Reward* 0. Diese Aufgabe lässt sich in Episoden trennen - die einzelnen Läufe durch das Labyrinth - mit dem Ziel den gesamt zu erwartenden *Reward* zu maximieren. Nachdem der lernende Roboter einige Durchläufe durch das Labyrinth gemacht hat, stellen Sie fest, dass dieser sich bei seiner Suche nicht verbessert und nicht unbedingt den kürzesten Weg gefunden hat.

a) Warum findet der Roboter nicht den kürzesten Weg?

b) Was müsste man ändern, damit der Roboter im Laufe der Lernphase den optimalen Weg findet?

Aufgabe 4.2 Bellman-Gleichung:
Die Bellman-Gleichung (Folie 56) muss für jeden Zustand der Wertefunktion V^π aus der Abbildung von Folie 58 stimmen. Zeigen Sie als ein Beispiel numerisch, dass die Gleichung für den mittleren Zustand mit dem Wert $+0.7$ unter Berücksichtigung der vier Nachbarzustände stimmt (die in der Abbildung angegebenen Werte sind Rundungen auf die erste Nachkommastelle).

Aufgabe 4.3 Bellman-Gleichung:
Die Abbildung von Folie 64 gibt den optimalen Wert des besten Zustandes des Gridworld Beispiels mit 24.4 an (auf eine Nachkommastelle gerundet). Nutzen Sie Ihr Wissen über die optimale *Policy* sowie die Gleichung von Folie 55 um diesen Wert zuerst verbal zu beschreiben und danach auf drei Nachkommastellen genau zu berechnen.

Aufgabe 4.4 Konstante in kontinuierlicher Task:
In dem Gridworld Beispiel waren die *Rewards* positiv für ein Ziel, negativ für einen Schritt aus dem Feld, sowie 0 in allen übrigen Situationen. Sind die Vorzeichen dieser *Rewards* wichtig oder nur die Intervalle zwischen ihnen? Beweisen Sie mittels der Gleichung von Folie 51, dass eine Addition einer Konstanten C zu allen *Rewards* lediglich eine Konstante K zu den Werten aller Zustände addiert und somit die relativen Werte der Zustände unter einer beliebigen *Policy* nicht verändert. Wie sieht K in Abhängigkeit von C und γ aus?

Aufgabe 4.5 Konstante in episodischer Task:
Stellen Sie sich nun vor, dass Sie eine Konstante C zu allen *Rewards* einer episodischen Aufgabe, wie z.B. der Labyrinthsuche, addieren. Hätte dies irgend einen Effekt oder würde es die Aufgabe wie in dem kontinuierlichen Fall aus Aufgabe 8.1 unberührt lassen? Warum oder warum nicht? Geben Sie ein Beispiel an.

Norman Hendrich
(hendrich@informatik...)
Jianwei Zhang
(zhang@informatik...)

Vorlesung:
Mi 8–12, F-334
Übungen:
Mi 12:30-14:00, F-334

Dienstgebäude
Universität Hamburg
Vogt-Kölln-Str. 30
22527 Hamburg

Öffentliche Verkehrsmittel
Buslinien 181, 281
Haltestelle: Informatikum