Einführung in die
Künstliche Intelligenz
WS14/15 - Prof. Dr. J. Fürnkranz

TECHNISCHE UNIVERSITÄT DARMSTADT

7. Übungsblatt (03.02.2015)

Aufgabe 1 Reinforcement Learning

Ein Agent bewegt sich in einer einfachen deterministischen Welt, die wie folgt angeordnet ist:

a	b	c
d	e	f

Der Agent kann sich jeweils ein Feld nach unten, oben, links, oder rechts bewegen, falls dort ein Feld ist. Ein Schritt verursacht keine Kosten. Wenn der Agent im Feld f landet, erhält er einen Reward von 1 Punkt und kann sich von dort nicht mehr wegbewegen. Auf allen anderen Feldern erhält er einen Reward von 0 Punkten.

Benutzen sie im Folgenden als Discountfaktor $\gamma = 0.8$.

a) Formulieren Sie zunächst die Reward-Funktion, d.h. geben Sie für alle möglichen Zustands-Aktions Paare (s, a) die unmittelbare Belohnung $r(s, a)$ an.

b) Berechnen Sie die Bewertungsfunktion $V^\pi(s)$ für alle Zustände s, wobei Policy π wie folgt definiert ist:

- wenn dies möglich ist, gehe nach oben; ansonsten:
- wenn dies möglich ist, gehe nach rechts; ansonsten:
- wenn dies möglich ist, gehe nach unten; ansonsten:
- gehe nach links

→	→	↓
↑	↑	

c) Welche Änderung würde PolicyImprovement an der Strategie π aus b) für das Feld e vornehmen ? (Benutzen Sie Teilaufgabe a und b)

d) Überlegen Sie sich für jedes Feld s, welches ein optimaler Weg zum Ziel wäre. Berechnen Sie damit die optimale Bewertung $V^*(s)$ für dieses Feld. Bestimmen Sie zusätzlich die optimale $Q(s, a)$-Funktion für alle möglichen Zustands-Aktion Paare (s, a).

e) Gegeben sei nur die optimale Q-Funktion aus d). Bestimmen bzw. rekonstruieren Sie damit eine optimale Policy.

f) Versuchen Sie, mittels Q-Learning die Q-Funktion direkt zu lernen, indem Sie den Agenten auf ein zufällig gewähltes Anfangsfeld stellen und die jeweils beste Aktion nach der momentanen Q-Funktion ausführen (bei Gleichheit zufällige Auswahl), die Update Regel anwenden, bis der Agent am Ziel angekommen ist und das ganze bis zur Konvergenz wiederholen. Benutzen Sie als Lernrate $\alpha = 1$.
Alternativ können Sie sich überlegen, welche Simulationssequenzen hier mit minimaler Anzahl an Updates konvergieren würde.

g) Gehen Sie nun davon aus, dass der Agent einen Defekt hat mit einer Wahrscheinlichkeit von 10% in jedem Schritt explodiert und somit die Sequenz beendet. Wie ändert sich nun die Bewertungsfunktion $V^\pi(s)$ und die Werte aus Aufgabe b? Gehen Sie davon aus, dass die Reward Funktion aus Aufgabe a weiterhin gültig ist. Das heißt, der Reward wird auch vergeben wenn der Zielzustand nicht erreicht wird, aber die korrekte Aktion gewählt wurde.

h) Welches Problem tritt auf wenn man den Reward nur vergeben will wenn der Zielzustand auch wirklich erreicht wurde ? Wie kann man die Bewertungsfunktion $V^\pi(s)$ verändern um dieses Verhalten abzubilden ?