Hochschule Bremerhaven

Exposé für eine Bachelorarbeit zum Thema:

Reinforcement Learning

Theoretische Grundlagen und praktische Umsetzung am Beispiel eines Ameisen-Agentenspiels oder eines lernenden Systems zur Bestimmung des optimalen Lenkverhaltens für einen autonom agierenden Einparkassistenten

Autor:	Jan Löwenstrom
Matrikelnr.:	34937
Erstprüfer:	Prof. Dr.-Ing. Henrik Lipskoch
Zweitprüfer:	Prof. Dr. Mathias Lindemann || Prof. Dr. Nadija Syrjakow

17. Januar 2020

Inhaltsverzeichnis

Abbildungsverzeichnis

1 Einleitung

Das ist eine Einleitung
(Fedjaev, 2017)

2 Grundlagen

Bei dem Bestärkten Lernen (*Reinforcement Learning*) interagiert ein Softwareagent (*Agent*) mit seiner Umwelt (*Environment*), die wiederum nach jeder Aktion (*Action*) Feedback an den Agenten zurückgibt. Dieses Feedback wird als Belohnung (*Reward*) bezeichnet, einem numerischen Wert, der sowohl positiv als auch negativ sein kann. Der Agent beobachtet zudem den Folgezustand (*State*) in dem sich die Umwelt nach der vorigen Aktion befindet, um so seine nächste Entscheidung treffen zu können. Ziel des Agenten ist es eine Strategie (*Policy*) zu entwickeln, so dass die Folge seiner Entscheidungen die Summe aller Belohnungen maximiert.

2.1 Markow Entscheidungsprozess

Die Umwelt wird in den allermeisten Fällen als Markow-Entscheidungsprozess (*Markov Decision Process, MDP*) definiert. //TODO Als *MDP* versteht sich die Formalisierung von sequentiellen Entscheidungsproblemen, bei denen eine Entscheidung nicht nur die sofortige Belohnung beeinflusst, sondern auch alle Folgezustände und somit auch alle zukünftigen Belohnungen (S. 47). Zudem bieten sie den mathematischen Rahmen für das *Reinforcement Learning* Problem, um z.B. Beweise über das Konvergenzverhalten eines Algorithmus hin zu einer optimalen Strategie führen oder andere theoretische Aussagen treffen zu können. Außerdem müssen Probleme die als *MDP* definiert werden zugleich die Markow-Eigenschaft erfüllen, die von essentieller Bedeutung ist und in Kapitel X näher erläutert wird.

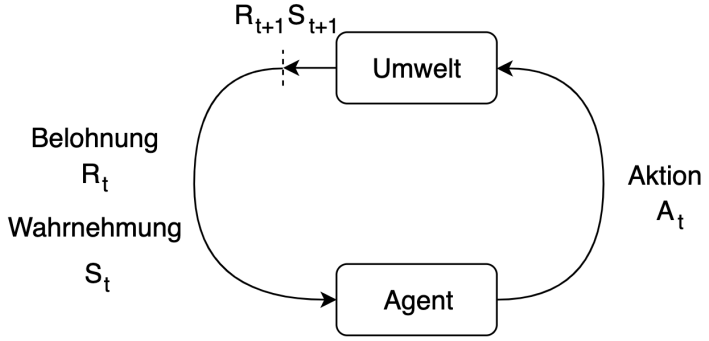

Abbildung 1: Agent-Umwelt Interface

Der Agent interagiert mit dem *MDP* jeweils zu diskreten Zeitpunkten $t = 0, 1, 2, 3, \ldots$. Zu jedem Zeitpunkt t beobachtet der Agent den Zustand seiner Umgebung $S_t \in \mathcal{S}$ und wählt aufgrund dessen eine Aktionen $A_t \in \mathcal{A}$. Als Konsequenz seiner Aktion erhält

er einen Zeitpunkt später eine Belohnung $R_{t+1} \in \mathcal{R} \subset \mathbb{R}$ und stellt den Folgezustand S_{t+1} fest. Das Zusammenspiel zwischen Agenten und MDP erzeugt also folgende Reihenfolge:

$$S_0, A_0, R_1, S_1, A_1, R_2, S_2, A_2, R_3, \ldots$$

Wird einfach nur von *MDPs* gesprochen, ist die endliche Variante (*finite MDP*) gemeint, bei dem die Mengen der Zustände, Aktionen und Belohnungen $(\mathcal{S}, \mathcal{A}, \mathcal{R})$ eine endliche Anzahl an Elementen besitzen. In diesem Fall haben die zufälligen Variablen R_t und S_t wohl definierte diskrete Wahrscheinlichkeitsverteilungen, die nur von dem vorigen Zustand und vorigen Aktion abhängig sind (S.48). Die Wahrscheinlichkeit, dass die bestimmten Werte für diese Variablen $s' \in \mathcal{S}$ und $r \in \mathcal{R}$ eintreten, für einen bestimmten Zeitpunkt t und dem vorigen Zustand s und Aktion a, kann somit durch folgende Funktion beschrieben werden:

$$p(s', r \mid s, a) \doteq Pr\{S_t = s', R_t = r | S_{t-1} = s, A_{t_1} = a\},$$

für alle $s', s \in \mathcal{S}, r \in \mathcal{R}$ und $a \in \mathcal{A}(s)$. Diese Funktion p definiert die sog. Dynamiken (*Dynamics*) eines *MDP*. Sie ist eine gewöhnliche deterministische Funktion mit vier Parametern $p : \mathcal{S} \times \mathcal{R} \times \mathcal{A} \to [0, 1]$. Das „|“ Zeichen kommt ursprünglich aus der Notation für bedingte Wahrscheinlichkeiten, soll hier aber nur andeuten, dass es sich um eine Wahrscheinlichkeitsverteilung handelt für jeweils alle Kombinationen von s und a:

$$\sum_{s' \in \mathcal{S}} \sum_{r \in \mathcal{R}} p(s', r \mid s, a) = 1 \; \forall s \in \mathcal{S}, a \in \mathcal{A}(s)$$

Ist die Zustandsüberführungsfunktion nicht stochastisch, so ist p immer nur für ein bestimmtes Triplet (s, a, r) für jedes $s' \in \mathcal{S}$ gleich 1, für alle andere jeweils 0. Mit anderen Worten, wird im Zustand s die Aktion a gewählt, führt dies immer zu einem bestimmten Folgezustand s'.

Das MDP Framework gilt als extrem flexibel und kann auf die unterschiedlichsten Probleme angewendet werden. Es bietet die nötige Abstraktion für Probleme, bei denen unter Vorgabe eines Ziels mittels Interaktionen gelernt wird. Dabei sind die Einzelheiten über eigentliche Ziel, die Zustände oder die Form des Agenten unerheblich, denn jedes zielgerichtete Lernen kann auf drei Signale reduziert werden, die zwischen dem Agenten und der Umwelt ausgetauscht werden. Ein Signal repräsentiert die Entscheidung, die der Agent getroffen hat (die Aktion), ein Signal repräsentiert die Basis, auf der er diese Entscheidung getroffen hat (der Zustand) und ein Signal definiert das zu erreichende Ziel (die Belohnung).

3 Markow-Eigenschaft und Zustandsmodellierung

Die Markow-Eigenschaft, obwohl relativ simpel, erhält ein eigenes Kapitel, da sie von fundamentaler Wichtigkeit ist und bei der Modellierung eines Reinforcement Learning Problems eine besondere Rolle spielt. Verbinden lässt sich dies sehr gut mit einem Einblick über die generelle Modellierung von Zuständen bei einem RL Problem.

> The future is independent of the past given the present

Dieser Satz erscheint oft in Büchern und Papern, wenn es um die Markow-Eigenschaft geht, denn er versucht zusammenzufassen, was diese aussagt. Im Zusammenhang von MDPs lässt sich dieser Satz so übersetzen, dass ein Folgezustand nicht abhängig von Aktionen bzw. Zuständen in der Vergangenheit ist, sondern ausschließlich von dem aktuellen Zustand und der aktuell gewählten Aktion. In der Literatur gibt es unterschiedliche Auffassungen darüber, ob die Markow-Eigenschaft an den MDP direkt geknüpft ist oder an den Zustand, den der Agent zur Abwägung der Entscheidung zur Verfügung hat. Bei der ersten Annahme wird davon ausgegangen, dass der Zustand, der von der Umwelt ausgeliefert wird direkt die Markow-Eigenschaft besitzen muss. (Sutton S.49) hingegen bindet die Eigenschaft an den Zustand und nicht an den Entscheidungsprozess als solches. Ein Zustand ist somit die Menge aller notwendigen Informationen der Vergangenheit, die für die Zukunft relevant sind. Statt den gegebenen Zustand der Umwelt direkt zu übernehmen, werden hier Beobachtungen der Umwelt zu einer internen Repräsentation von Markow-Zuständen verarbeitet.

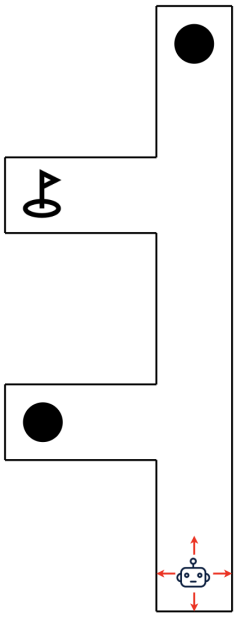

Abbildung 2: Zwei-Wege Beispiel zu der Markow-Eigenschaft

Literatur

Fedjaev, J. (2017). *Decoding eeg brain signals using recurrent neural networks.* Zugriff auf https://www.spektrum.de/kolumne/eine-waffe-gegen-malaria/1525481

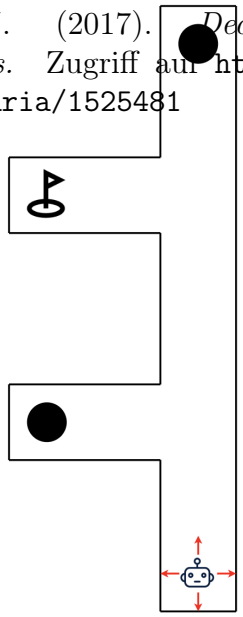

Abbildung 3: Birds