Policy (bestärkendes Lernen)

Autor und Herausgeber

Alexander Jung*, Department of Computer Science, Aalto University, Espoo, Finland (ORCID 0000-0001-7538-0990)
Editors: Konstantina Olioumtsevits (ORCID 0000-0003-2697-8659), Ekkehard Schnoor (ORCID 0000-0001-8252-1499), Aalto University, Espoo, Finland
*Corresponding author: alex.jung@aalto.fi

Zusammenfassung

Eine Policy (auch Strategie genannt) beim bestärkendes Lernen (RL) ist eine Funktion, die den aktuellen Zustand auf eine Aktion abbildet. Manchmal ist eine Policy stochastisch und wird durch eine bedingte Wahrscheinlichkeitsverteilung über die Aktionen bei gegebenem Zustand definiert. Die Policy eines Verfahrens des RLs kann als die Hypothese eines maschinelles Lernen (ML)-Verfahrens aufgefasst werden. Sie nimmt die Merkmale, die den aktuellen Zustand des Agenten charakterisieren, als Eingabe entgegen und gibt eine Vorhersage der besten auszuführenden Aktion aus. Anstelle eines durchschnittlichen Verlusts auf einem Trainingsdatensatz nutzen Verfahren des RLs die nach der Ausführung von Aktionen beobachteten Belohnungen, um das Erlernen einer Policy zu steuern.

Definition

Ein selbstfahrendes Auto beobachtet seine Position und Fahrtrichtung, wählt eine Lenkrichtung und erfährt, wie gut es der Straße folgt; Abb. 1 zeigt drei Zeitpunkte davon. Beim bestärkendes Lernen (RL) interagiert ein solcher Agent mit einer Umgebung: Zu jedem Zeitpunkt $t = 1, 2, \ldots$ beobachtet der Agent den Zustand s_t der Umgebung, führt eine Aktion a_t aus und erhält den nächsten Zustand s_{t+1} zusammen mit einer Belohnung r_{t+1}. Diese Interaktion wird als ein Markow-Entscheidungsprozess (MDP) modelliert (Bellman, 1957). Für das Auto ist der Zustand die Position und Fahrtrichtung des Autos, die Aktion ist die Lenkrichtung, und die Belohnung gibt an, wie gut das Auto der Straße folgt.

In ihrer einfachsten Form ist eine Policy eine Funktion, die angibt, welche Aktion a_t in diesem MDP auszuführen ist, wenn der aktuelle Zustand s_t vorliegt (Sutton and Barto, 2018, Sect. 3.5). Eine Policy unterscheidet sich von einem Plan, einer festen Folge von Aktionen: Eine Policy reagiert auf den beobachteten Zustand. Eine Policy dieser Art ist deterministisch. Sie ist ein Spezialfall einer stochastischen Policy, die eine bedingte Wahrscheinlichkeitsverteilung $\mathbb{P}^{(a|s)}$ über die Aktionen bei gegebenem Zustand definiert, eine Wahrscheinlichkeitsverteilung, aus der die Aktion gezogen wird.

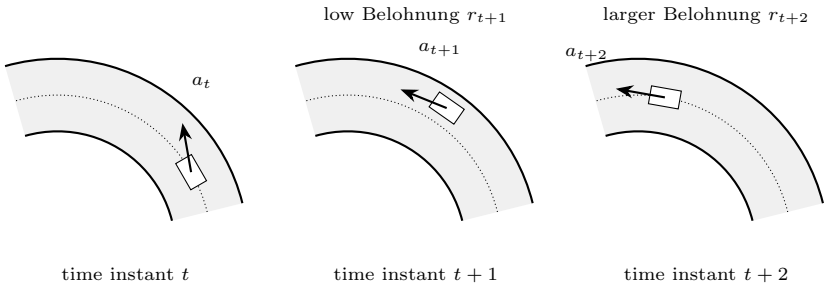


Abbildung 1: Ein selbstfahrendes Auto zu drei Zeitpunkten zwischen den beiden Begrenzungen einer Straße. Der Zustand s_t ist die Position und Fahrtrichtung des Autos; die von der Policy ausgewählte Aktion a_t ist die Lenkrichtung (Pfeile). Zum Zeitpunkt t weicht die Richtung leicht von der Straßenmitte ab, sodass das Auto in Richtung der äußeren Begrenzung abdriftet und die Belohnung r_{t+1} gering ausfällt. Die Richtung zum Zeitpunkt $t+1$ ist besser auf die Mitte ausgerichtet, sodass das Auto dorthin zurückkehrt und die Belohnung r_{t+2} größer ist

Eine Policy kann als eine Hypothesenabbildung $h(\mathbf{x})$ aufgefasst werden, die die beste auszuführende Aktion vorhersagt. Der in die Hypothese eingespeiste Merkmalsvektor $\mathbf{x}$ ist entweder der Zustand selbst, wenn der Zustand vollständig beobachtet wird, oder es sind indirekte Messungen davon, wenn der Zustand nur teilweise beobachtet wird. Was das überwachtes Lernen vom RL unterscheidet, ist die Art und Weise, wie die Nützlichkeit einer Hypothese bewertet wird. Anstatt den durchschnittlichen Verlust auf einem Datensatz aus gelabelte Datenpunkte zu berechnen, verwenden Verfahren des RLs den erwarteten Return, der erzielt wird, wenn die Policy auf einen MDP angewendet wird.

Der Return zum Zeitpunkt t ist die diskontierte Summe der danach erhaltenen Belohnungen,

$$G_t := \sum_{k=0}^{\infty} \gamma^k r_{t+k+1},$$

mit einem Diskontierungsfaktor $\gamma \in [0, 1]$. Die Summe hat einen endlichen Wert, wenn $\gamma < 1$ gilt und die Belohnungen beschränkt sind (Sutton and Barto, 2018, Sect. 3.3). Der Fall $\gamma = 1$ wird nur für Episoden verwendet, die nach endlich vielen Zeitpunkten enden, sodass die Summe endlich viele von null verschiedene Terme hat (Sutton and Barto, 2018, Sect. 3.4). Der erwartete Return, als eine Funktion des Ausgangs-Zustands, ist die Wertfunktion der Policy (Sutton and Barto, 2018, Sect. 3.5). Sie kann durch Ausführen der Policy geschätzt werden: Der Agent startet mehrmals vom Zustand aus, folgt der Policy für eine Reihe von Zeitpunkten, zeichnet die Belohnungen auf und mittelt die beobachteten Returns (Sutton and Barto, 2018, Ch. 5).

Siehe auch: Aktion, MDP, Zustand, RL, Belohnung, Wertfunktion, Return, Policy-Gradient, Agent.

Querverweise

Literatur

1. Bellman (1957). Dynamic Programming. Princeton Univ. Press, Princeton, NJ, USA.

2. Sutton and Barto (2018). Reinforcement Learning: An Introduction. 2nd ed., MIT Press, Cambridge, MA, USA.