Belohnung

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 Belohnung ist eine beobachtete Größe, die es einer machine learning (ML)-Methode ermöglicht, den Verlust abzuschätzen, der durch die Vorhersage einer Hypothese entsteht. Ein Beispiel für eine solche Größe ist das Label eines Datenpunkt selbst. Viele ML-Anwendungen erlauben es jedoch nicht, das Label zu beobachten, sondern nur eine verrauschte Schätzung davon; einige erlauben nur die Beobachtung des Vorhersage-Fehlers einer spezifischen Vorhersage. Bestärkendes Lernen (RL)-Methoden lernen eine Folge von Hypothese-Abbildungen aus den Belohnungen, die sie durch ihre Vorhersagen erhalten. Zum Beispiel prognostiziert ein selbstfahrendes Fahrzeug zu jedem Zeitpunkt eine Lenkrichtung und beobachtet als Belohnung, ob diese Richtung das Fahrzeug auf seiner Spur hält, wobei eine niedrige Belohnung für eine Richtung vergeben wird, die es von der Straße abführt.

Definition

Ein selbstfahrendes Auto muss zu jedem Zeitpunkt eine Lenkrichtung wählen, und die optimale Richtung wird nie beobachtet; was beobachtet wird, ist, ob die gewählte Richtung das Auto auf seiner Spur hält (Abb. 1). Das ultimative Ziel von machine learning (ML) ist es, eine Hypothese zu lernen, die das Label eines Datenpunkt gut aus seinen Merkmale vorhersagt, und die Wahl zwischen verschiedenen Hypothese-Abbildungen erfordert ein Maß für den Vorhersage-Fehler, den eine Hypothese verursacht. Ein solches Maß ist der Verlust, der beispielsweise durch den Vergleich der Vorhersage $h(\mathbf{x})$ mit dem tatsächlichen Label-Wert y definiert werden kann. Für das Auto ist ein Datenpunkt der aktuelle Zeitpunkt und sein Label ist die optimale Lenkrichtung, die überhaupt nicht zugänglich ist.

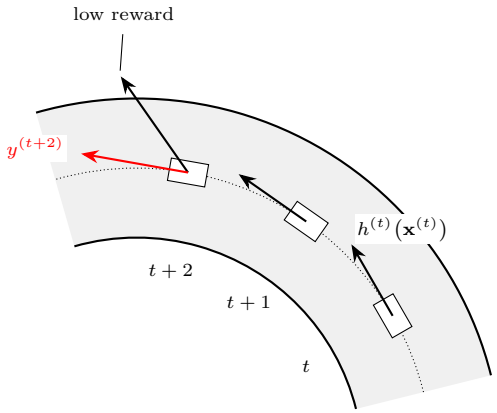


Abbildung 1: Ein selbstfahrendes Auto zu drei Zeitpunkten, zwischen den beiden Fahrbahnkonturen. Zu jedem Zeitpunkt prognostiziert die Hypothese eine Lenkrichtung (schwarze Pfeile); zum Zeitpunkt t ist diese Vorhersage $h^{(t)}(\mathbf{x}^{(t)})$. Der rote Pfeil ist das Label $y^{(t+2)}$, die optimale Lenkrichtung, die der Spur folgt. Dieses Label ist selten zugänglich: Um es zu erhalten, müsste ein Fahrlehrer jedes mögliche Fahrszenario analysieren. Stattdessen wird eine Belohnung beobachtet: zu t und $t+1$ hält die vorhergesagte Richtung das Auto auf seiner Spur, während sie zu $t+2$ das Auto über die äußere Kontur hinausführt und eine niedrige Belohnung erhält

Eine Belohnung ist eine beobachtete Größe, aus der eine ML-Methode den Verlust abschätzt, der durch die Vorhersage einer Hypothese entsteht. Ein Beispiel für eine solche Größe ist das Label eines Datenpunkt selbst. Viele ML-Anwendungen erlauben es jedoch nicht, das Label zu beobachten, sondern nur eine verrauschte Schätzung davon. Einige erlauben nur die Beobachtung des Vorhersage-Fehlers einer spezifischen Vorhersage. Das selbstfahrende Auto von Abb. 1 gehört zur letzten Art: Die Belohnung ist, ob die zu einem Zeitpunkt vorhergesagte Lenkrichtung $h(\mathbf{x})$ das Auto auf seiner Spur hält, und eine niedrige Belohnung wird einer Richtung zugewiesen, die das Auto über eine Fahrbahnkontur hinausführt.

Eine Belohnung ersetzt das Label des überwachtes Lernen: Sie offenbart nicht die korrekte Vorhersage, sondern nur, wie gut die gewählte Vorhersage, die Aktion, war, und sie kann verzögert, mehrere Zeitschritte nach der Aktion, die sie verursacht hat, oder durch Rauschen verfälscht eintreffen (Sutton and Barto, 2018, Sect. 1.1). Bestärkendes Lernen (RL)-Methoden lernen eine Folge von Hypothese-Abbildungen aus den Belohnungen, die sie durch ihre Vorhersagen erhalten: Die negative Belohnung ist der Verlust der ausgeführten Aktion, und die gewählten Aktionen sind diejenigen, die die kumulative Belohnung, den Return, maximieren (Sutton and Barto, 2018, Sect. 3.2). Im selbstfahrenden Auto von Abb. 1 ist der Zustand s_t die Position und Ausrichtung des Autos auf der Straße zum Zeitpunkt t, und die Aktion a_t ist die dort vorhergesagte Lenkrichtung. Abb. 2 zeigt, wie Zustände, Aktionen und Belohnungen über zwei

Zeitschritte aufeinanderfolgen.

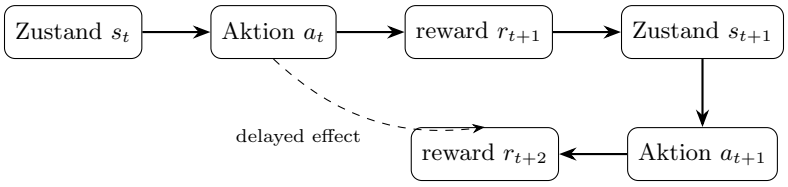


Abbildung 2: Das Funktionsprinzip des RL. Bei jedem Zeitschritt t wird der Zustand s_t beobachtet, eine Aktion a_t gewählt und eine Belohnung r_{t+1} empfangen (durchgezogene Pfeile). Die Belohnung bewertet die gewählte Aktion, ohne zu offenbaren, welche Aktion die beste gewesen wäre, und eine Aktion kann auch eine spätere Belohnung beeinflussen (gestrichelter Pfeil), weshalb der Return, die kumulative Belohnung, und nicht nur die nächste Belohnung maximiert wird

Siehe auch: Verlust, multiarmed bandit (MAB), RL, Return, Aktion, Label.

Querverweise

Literatur

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