Translation Sample, Nov 8, 2022

by Jo Havemann

Reinforcement learning

A sequential decision process (SDP) is any dynamic process that involves a time ordered sequence of decisions that influence some cost or reward that accrues over time. The idea is to determine the decision sequence that optimizes the defined cost or reward.

A third branch of machine learning, referred to as reinforcement learning, shares some of the essential characteristics of supervised and unsupervised learning, but is applied specifically to the problem of sequential decision making and therefore relies on a different form of data to that used by either supervised or unsupervised learning. Much of the theory of reinforcement learning is based on dynamic programming, which was formulated by Richard Bellman in the 1950s (Bellman, 1957). While this provides much of the mathematical foundation for reinforcement learning, as we will see below, dynamic programming has a limited computational capacity. Prac- tically, this means that dynamic programming, at least in its exact form, is not a viable solution for applications made to exploit the type of high throughput data that is often available today.

One especially compelling early application of reinforcement learning was the development of an SDP, known as TD-Gammon, which is able to play backgammon at a competitive level. More recently, reinforcement learning was used by DeepMind Technologies to create a chess-playing program known as AlphaZero. Reinforcement learning is particularly suited to develop game-playing applications, especially when coupled with deep neural network models (deep reinforcement learning). For example, it is currently used to create quite effective video game playing applications.

**Verstärkungslernen**

Ein sequentieller Entscheidungsprozess ist ein dynamischer Prozess, der eine zeitlich geordnete Abfolge von Entscheidungen beinhaltet, welche ihrerseits Einfluss auf anfallende Kosten oder die Anreize nehmen. Dadurch wird eine auf Kosten und Anreize optimierte Entscheidungssequenz bestimmt.

Ein dritter Zweig des maschinellen Lernens, das sogenannte Verstärkungslernen, beinhaltet einige der wesentlichen Merkmale des beaufsichtigten und unbeaufsichtigten Lernens, wird dabei jedoch speziell auf das Problem der sequenziellen Entscheidungsfindung angewandt und stützt sich entsprechend auf eine andere Form von Daten als das beaufsichtigte oder unbeaufsichtigte Lernen. Ein Großteil der Theorie des Verstärkungslernens basiert auf der dynamischen Programmierung, die von Richard Bellman in den 1950er Jahren formuliert wurde (Bellman, 1957). Auch wenn diese die mathematische Grundlage für das Verstärkungslernen ausmacht, wie wir weiter unten sehen werden, hat die dynamische Programmierung eine begrenzte Rechenkapazität. In der Praxis bedeutet dies, dass die dynamische Programmierung, zumindest in ihrer spezifischen Form, keine brauchbare Lösung für heutzutage häufig vorfindbare Anwendungen mit Hochdurchsatzdaten ist.

Eine besonders überzeugende frühe Anwendung des Verstärkungslernens war die Entwicklung eines sequentieller Entscheidungsprozesses, bekannt als TD-Gammon, der dazu fähig ist Backgammon auf einem konkurrenzfähigen Niveau zu spielen. Kürzlich wurde das Verstärkungslernen von DeepMind Technologies eingesetzt, um ein Schachspielprogramm namens AlphaZero zu entwickeln. Verstärkungslernen eignet sich besonders gut für die Entwicklung von Spielanwendungen, vor allem wenn es mit tiefen neuronalen Netzwerkmodellen gekoppelt wird (tiefes Verstärkungslernen). Derzeit wird es Beispielsweise zur Entwicklung sehr wirksame Anwendungen für Videospiele eingesetzt.