Übungen zu "Deep Reinforcement Learning"

Wintersemester 2024 Zettel 1

Ausgabe: 14.10.2024, **Abgabe:** 20.10.2024, **Besprechung:** 23.10.2024

Wichtig - Die Übungen werden in kleinen Gruppen von 2 bis 3 Personen abgegeben. Bitte melden sie ihre Gruppe bei Abgabe (spätestens 20.10.2024) per E-Mail an Janosch Bajorath zurück (j.bajorath@uni-muenster.de).

Aufgabe 1.1 Der K-armige Bandit: (5 Punkte = 2 + 1 + 1 + 1)

Im Rahmen dieser Aufgabe ist ein stationärer k-armiger Bandit mit $k = 4$ Aktionen zu implementieren. Zur Lösung des Banditen-Problems verwenden Sie einen Algorithmus welcher zur Berechnung von $Q_s a$ die Methode des Stichprobenmittels nutzt und die ε-greedy Methode zur Auswahl der Aktionen verwendet. Der Algorithmus soll eine Aktion mit einer Wahrscheinlichkeit ε zufällig zur Erkundung auswählen, während er mit einer Wahrscheinlichkeit 1-ε diejenige Aktion wählt, die bislang den höchsten erwarteten Reward erbracht hat.

(a) **Auswerten des Banditen Problems** - Die gesammelte Belohnung (Reward), der Regret (Differenz zwischen dem maximal möglichen Rewards und dem tatsächlich erzielten Reward) sowie der prozentuale Anteil der Wahl des besten Arms über die Zeit hinweg sind zu plotten. Es sind mehrere Läufe mit durchzuführen, um die Resultate zu mitteln und somit verlässlichere Auswertungen und Grafiken zu erhalten. Erläutern Sie kurz Ihre Erkenntnisse.

(b) **Variation von** ϵ - Untersuchen Sie die Auswirkungen verschiedener ε-Werte auf das Verhalten des Algorithmus im zeitlichen Verlauf. Dazu sind drei verschiedene ε-Werte, beispielsweise 0.01, 0.1 und 0.2, zu testen. Analysieren Sie, wie die Wahl des ε-Parameters die Balance zwischen Erforschung und Ausbeutung sowie die Geschwindigkeit beeinflusst, mit der der Algorithmus die optimale Strategie erkennen und verfolgen kann.

(c) **Initialisierung der Schätzwerte** - Untersuche Sie welche Auswirkung die Initialisierung der Schätzwerte für die erwarteten Belohnungen der Arme hat. Vergleichen Sie das Verhalten des Algorithmus, wenn die Schätzwerte zu Beginn neutral (z.B. mit 0) oder optimistisch (mit einem hohen Wert) angesetzt werden. Diskutieren Sie, inwiefern optimistische Anfangswerte die anfängliche Erkundung anregen und den Algorithmus dazu befähigen, informierte Entscheidungen zu treffen.

(d) **Adaptive Erkundung** - Modifizieren Sie den Algorithmus dahingehend, dass der ε-Wert im Laufe der Zeit verkleinert wird. Implementieren Sie dafür eine geeignete Funktion um ε über die Zeit dynamisch anzupassen. Erstellen Sie vergleichende Abbildungen, die konstante und adaptive ε-Strategien gegenüberstellen, und interpretieren Sie die gewonnenen Daten hinsichtlich der Entscheidungsfindung des Algorithmus während der Lernphase.

Aufgabe 1.2 Das nicht-stationäre Bandit Problem: (5 Punkte = 1 + 1 + 2 + 1) Eine Voraussetzung für die erfolgreiche Ausnutzung von bisher gesammeltem Wissen ist, dass sich das untersuchte Bandit-Problem nicht verändert, das heißt, die Wahrscheinlichkeiten für das Vergeben von Rewards gleichbleibend sind. Wenn sich diese Wahrscheinlichkeiten im Laufe der Zeit ändern können, spricht man von einem nicht-stationären Problem. In dieser Aufgabe beschäftigen wir uns mit einem nicht-stationären K-armigen Banditen-Problemen, bei denen die Wahrscheinlichkeiten, eine Belohnung zu erhalten, im Laufe der Zeit variieren.

(a) **Formale Definition** - Verfassen Sie eine schriftliche Beschreibung der Schritte, die notwendig sind, um das Problem eines nicht-stationären K-armiken Banditen zu lösen. Das Ziel ist es, ein theoretisches Verständnis zu entwickeln, ohne eine konkrete Implementierung vorzunehmen. Zur Orientierung gehen Sie auf folgende Themen ein: Beschreibung des Bandit-Problems, Modellierung von Nicht-Stationarität, Algorithmus zur Lösung des Problems, Erkundung und Ausbeutung, sowie Evaluierungskriterien.

(b) **Anpassung des Banditen Problems** - Erstellen Sie eine modifizierte Version des ursprünglichen 4-armigen Bandit-Problems, bei dem alle Anfangswerte für die erwartete Belohnung $q_*(a)$ identisch sind. Im weiteren Verlauf sollen diese Werte unabhängig voneinander durch Hinzufügen von normalverteilten Zufallswerten (z.B. $\mu = 0.0$ und $\sigma = 0.01$) verändert werden.

(c) **Auswertung des Banditen Problems** - Erstellen Sie Diagramme, die den Verlauf der gesammelten Belohnungen sowie die Verteilung der gewählten Aktionen im Laufe der Zeit visualisieren, und beschreiben Sie deren Verhalten. Untersuchen Sie dabei den zuvor implementierten Algorithmus, der zur Berechnung von $Q_t(a)$ die Methode des Stichprobenmittels nutzt und die ε-greedy Methode zur Auswahl der Aktionen verwendet.

(d) **Mögliche Anpassungen** - Nicht-stationäre Banditen Probleme stellen eine Herausforderung für den bisher verwendeten Algorithmus dar. Überlegen Sie, welche Modifikationen möglich wären, um unter den Bedingungen eines nicht-stationären Bandit-Problems bessere Ergebnisse zu erzielen.

Literatur

Sutton, R. S. & Barto, A. G. (2018). *Reinforcement learning: An introduction* (Second Aufl.). The MIT Press.