Computational Learning Theory

Sommersemester 2017

Mario Holldack, M. Sc.
Prof. Dr. Georg Schnitger
Hannes Seiwert, M. Sc.

GOETHE
UNIVERSITÄT
FRANKFURT AM MAIN

Institut für Informatik
AG Theoretische Informatik

Übungsblatt 8

Ausgabe: 12.06.2017
Abgabe: 19.06.2017

Dieses Blatt ist ein Bonusblatt.

Aufgabe 8.1 *Mehrarmige Banditen* (8 + 12 + 12 Punkte)

Im *Mehrarmige-Banditen-Problem* sind n „Arme" $1, \dots, n$ eines Spielautomaten gegeben. Jeder Arm j besitzt eine eigene (uns unbekannte) Verteilung f_j auf $[0, 1]$. In jeder Runde $t = 1, 2, \dots, T$ müssen wir einen der Arme J_t auswählen und erhalten einen zufälligen Gewinn g_t gemäß der Verteilung f_{J_t}.

Ziel ist es, den erwarteten Gesamtgewinn $\sum_{t=1}^{T} \mathbb{E}[g_t]$ zu maximieren.

Um das zu erreichen, muss zwischen *Exploration* und *Exploitation* abgewogen werden: Ziehe ich einen Arm, über den ich wenig weiß, um zu lernen; oder ziehe ich den nach bisherigem Wissen vielversprechendsten Arm?

Wir definieren: Es sei J_t der im Schritt t gezogene Arm. Es sei μ_j sei der erwartete Gewinn des j-ten Arms, $\mu^* := \max_j \mu_j$ der erwartete Gewinn des optimalen Arms und $\Delta_j := \mu^* - \mu_j$.
Ferner gebe $T_{j,t}$ an, wie oft der j-te Arm bis zum Zeitpunkt t gezogen wurde, und es sei $\hat{\mu}_{j,t} = \frac{1}{T_{j,t}} \sum_{i=1}^{t} g_i \cdot \mathbf{1}_{[J_i=j]}$ der beobachtete durchschnittliche Gewinn des j-ten Arms zum Zeitpunkt t basierend auf allen bisherigen Aktionen.

Als Qualitätsmaß eines Algorithmus A betrachten wir hier den *(Pseudo[1]-)Regret*

$$\mathsf{Regret}_T(A) := \mu^* T - \sum_{t=1}^{T} \mathbb{E}[g_t] \,.$$

Beachte, dass gilt: $\mathsf{Regret}_T(A) = \sum_{j:\mu_j<\mu^*} \Delta_j \mathbb{E}[T_{j,T}]$. Um den Regret eines Algorithmus zu bestimmen, genügt es also, die Häufigkeiten $T_{j,T}$ der suboptimalen Arme zu bestimmen.

Wir betrachten zunächst zwei einfache Algorithmen:

- Die ε-*Greedy*-Strategie: Mit Wahrscheinlichkeit ε wähle gleichverteilt einen zufälligen Arm und mit Wahrscheinlichkeit $(1–\varepsilon)$ den Arm j, der bisher den höchsten durchschnittlichen Gewinn $\hat{\mu}_{j,t}$ erbracht hat.
- Die *Exploration-first*-Strategie: Ziehe die ersten εT Runden gleichverteilt einen zufälligen Arm und die restlichen $(1-\varepsilon)T$ Runden jeweils den Arm j, der bisher den höchsten durchschnittlichen Gewinn $\hat{\mu}_{j,t}$ erbracht hat.

a) Zeigen Sie, dass beide Strategien einen linearen Regret $\Omega(T)$ haben, falls $\varepsilon = \Theta(1)$.

Hinweis: Eine Rechnung ist nicht nötig. Eine qualitative Argumentation genügt.

Bitte wenden!

[1] Im Gegensatz zur Definition des Regrets bei der Auswahl der Experten vergleichen wir den Algorithmus hier mit dem besten Arm, *nachdem* wir den Erwartungswert über die Gewinnverteilung der Arme bilden.

Der folgende Algorithmus UCB[2] trifft eine sorgfältige Abwägung zwischen Exploration und Exploitation:

1. Initialisierung: Ziehe jeden Arm einmal.
2. In allen folgenden Schritten t:
   2.1. Für jeden Arm j berechne $w_{j,t} := \sqrt{2\ln(t)/T_{j,t}}$
   2.2. Wähle J_t als den Arm j, der $\hat{\mu}_{j,t} + w_{j,t}$ maximiert.

Seien die Gewinnverteilungen $f_1, \ldots, f_n$ fest. Dann erreicht UCB einen Regret von $\mathcal{O}(\log(T))$ für $T \to \infty$.

b) Zeigen Sie dazu die folgenden Schritte:

   i) Für einen beliebigen Arm j gilt $\text{prob}\,[\,|\hat{\mu}_{j,t} - \mu_j| > w_{j,t}\,] = \mathcal{O}(1/t^4)$.
   Hinweis: Hoeffdings Ungleichung

   ii) Für einen suboptimalen Arm j gilt: $\text{prob}[J_t = j] \;\leq\; \mathcal{O}(1/t^4) + \text{prob}\Big[\, w_{j,t} > \frac{\Delta_j}{2} \,\Big]$

   iii) Für einen suboptimalen Arm j gilt: $\mathbb{E}[T_{j,T}] = \mathcal{O}\Big(\frac{\log(T)}{\Delta_j^2}\Big)$

   iv) $\mathsf{Regret}_T(\text{UCB}) = \mathcal{O}(\log(T))$

Wir wollen nun zeigen, dass ein logarithmischer Regret für jeden Algorithmus nicht zu vermeiden ist. Dazu nehmen wir $n = 2$ an. Der erste Arm sei $(1/2)$-bernoulliverteilt (fairer Münzwurf), der zweite Arm $(1/2{+}\xi\Delta)$-bernoulliverteilt, wobei $\xi \in \{-1, +1\}$ unbekannt und $\Delta > 0$ fest sei.

c) i) Angenommen, der zweite Arm wurde τ mal gezogen. Wie groß ist die Wahrscheinlichkeit, dass die Fälle $\xi = -1$ und $\xi = +1$ nicht unterschieden werden können bzw. verwechselt werden? Geben Sie eine untere Schranke an.
   Hinweis: Aufgabe 2.2 b)

   ii) Sei A ein Algorithmus, der unter Annahme des Falles $\xi = -1$ den zweiten Arm bis zum Zeitpunkt t genau $\tau(t)$ mal zieht. Zeigen Sie

$$\mathsf{Regret}_T(A \mid \xi{=}{-1}) + \mathsf{Regret}_T(A \mid \xi{=}{+1}) = \Omega(\log(T))\,.$$

[2] für *upper confidence bound*