Bandits, MCTS & Games

Motivation

Monte Carlo Tree Search ist die Vereinigung von Tree Search und Wahrscheinlichkeiten.

Banditen

Banditen sind eine einfache Form von "Exploration-Exploitation-Problemen" (Erkundungs-Ausnutzungs-Probleme), bei denen das Ziel ist sowohl über seine Umgebung zu lernen als auch eine Belohnung zu bekommen. Die Zwickmühle ist, ob man eine Entscheidung trifft um zu lernen oder um einen möglichst hohen Reward zu bekommen.

Banditen sind ein klassisches Beispiel von

- sequenziellem Entscheidungstreffen
- Entscheidungen, die Wissen über die Umwelt und Rewards beeinflussen
- Exploration-Exploitation

Banditenprobleme stellen die Basis von Upper-Confidence-Bounds (UCB) dar. sie sind kommerziell sehr relevant.

Situation

- es gibt n einarmige Banditen
- jede Maschine i gibt einen Reward $y \sim P(y;\Theta_i$
- die Maschinenparameter Θ_i sind nicht bekannt
- das Ziel ist es, den Reward, beispielsweise nach T Versuchen zu maximieren.
- $a_t \in \{1,...,n\}$ stellt die Entscheidung zum Zeitpunkt t dar
- $y_t \in \realNumbers$ ist der Reward für den t-ten Schritt
- eine Policy oder Strategie bildet alle Entscheidungen und Rewards auf eine neue Entscheidung ab: $\pi: [(a_1,y_1),...,(a_{t-1},y_{t-1})] \rightarrow a_t$
- das Problem ist, eine Policy π zu finden, die
    - den Reward über alle T Schritte maximiert oder
    - den Reward im nächsten Schnitt maximiert

Exploration-Exploitation

Die Wahl einer Maschine hat zwei Folgen:

- man erlangt neues Wissen über die Umwelt
- man bekommt eine Belohnung

Exploration: Wähle die Maschine, die deine Umwissenheit über die Umwelt minimiert.

Exploitation: Wähle die Maschine, die deinen nächsten Reward maximiert.

Upper Confidence Bounds (UCB1)

```
play each machine once.

while(true)
  play the machine i that maximizes y_av_i + \beta * sqrt{2 ln(n) / n_i}
```

Wobei

- `y_av_i` der durchschnittliche Reward der Machine `i` ist
- `n_i` die Anzahl ist, wie oft `i` gewählt wurde
- `n` die Anzahl der Runden ist
- `\beta` ein Vorfaktor ist, der oft als 1 gewählt wird.

UCB Algorythmen

Ein Confidence-Intervall bei UCB ist

`y_av_i - \theta_i < y_i_mean < y_av_i + \theta_i`

wobei UCB die obere Grenze des Intervall wählt.

UCB Diskussion

- UCB unterschätzt - wie A* - die costs-to-go, aber in den probabilistischen Settings von Banditen
- UCB sind zu einer Kernmethode geworden, um herauszufinden, was erforscht werden soll: UCB wird verwendet um erauszufinden, welcher Branch im Entscheidungsbaum fortgeführt wird.

Monte-Carlo Tree Search (MCTS)

Monte Carlo Methoden

Generell wird bei Monte Carlo Simulationen eine große Zahl an zufälligen "Samples" generiert, anstatt jeden einzelnen Pfad zu betrachten oder etwas genau auszurechnen.

Beispiel: Wie groß sind die Chancen, bei einem Kartenspiel zu gewinnen? Statt alles auszurechnen, werden viele zufällige Spiele durchgerechnet und die Gewinne gezählt.

Prüfungsrelevant ist

- Upper Confidence Bounds
- Monte Carlo Tree Search