Dynamisches Programmieren

Motivation

Dynamisches Programmieren beschäftigt sich mit zeitlich ändernden Umwelten. Sie sind auch wichtig, weil Reinforcement Learning direkt darauf aufbaut.

Markov Entscheidungsprozesse

Generell ist die Welt so aufgebaut:

- sie hat einen Zustand s_o
- ich kann eine Aktion a wählen
- dafür bekomme ich einen Reward r

Allerdings "macht die Welt nicht immer, was ich will", wie wenn ich beim Dart spielen z.B. in 70% der Fälle das gewünschte Feld der Scheibe treffe. Daher hat die Welt...

- eine Initial State Verteilung $P(s_o)$
- eine Transition Probability $P(s_{t+1}|s_t, a_t)$
- eine Reward Probability Tabelle $P(r_t|s_t, a_t)$
- einen Agenten mir der Policy $\pi(a_t, s_t) = P(a_0 | s_0, \pi)$

Stationary MDP: Reward und Transition sind zeitunabhängig. Ein stationary MDP ist durch die vier Punkte oben definiert.

Dynamic Programming

- die Value V der Policy π ist der erwartete discountete Reward, wenn er im Zustand s beginnt: $V^\pi(s) = E{r_o + \gamma*r_1 + \gamma^2*r_2 .... | s_0 = s}$, wobei $\gamma \in [0,1]$ der Discounting-Faktor ist.
- Eine Policy ist Optimal wenn sie für alle States V maximiert.

Value Function

$$V^\pi(s) = R(s|\pi(s)) + \gamma \sum{s'} P(s'|s,\pi(s)) * V^\pi(s')$$

Bedeutet: Die Value eines States mit einer Policy ist der Reward des States plus die Value der States wo die Policy hin will oder ausversehen landet.

Value Iteration

Kurz: Wiederhole die Value Funktion so lange bis sich nichts mehr wirklich ändert.

$$V_k+1(s) = max{a} [R(s,a) + \gamma \sum{s'} P(s'|s,a) V_k(s's)$$

Bedeutet: Für den neuen Wert rechnet man nicht die Value der anderen States in diesem Zeitschritt, sondern "eins davor". Das konvergiert irgendwann.

Q-Funktion (**State-Action** value function)

$$Q^\pi(s,a) = R(s|\pi(s)) + \gamma \sum{s'} P(s'|s,a) * Q^\pi(s', \pi(s'))$$

Also: $V^\pi(s) = Q^\pi(s,\pi(s))$.

Q-Iteration

Analog zu Value Iteration

Dynamic Programming in Belief Space

Wird vom Autor als nicht relevant erachtet.

Prüfungsrelevant ist vor allem

- Markov Entscheidungsprozesse auf jeden Fall, mini kleines Problem wo man Value Iteration 3 mal durchführen muss
- Value und Q-Iteration muss man wissen