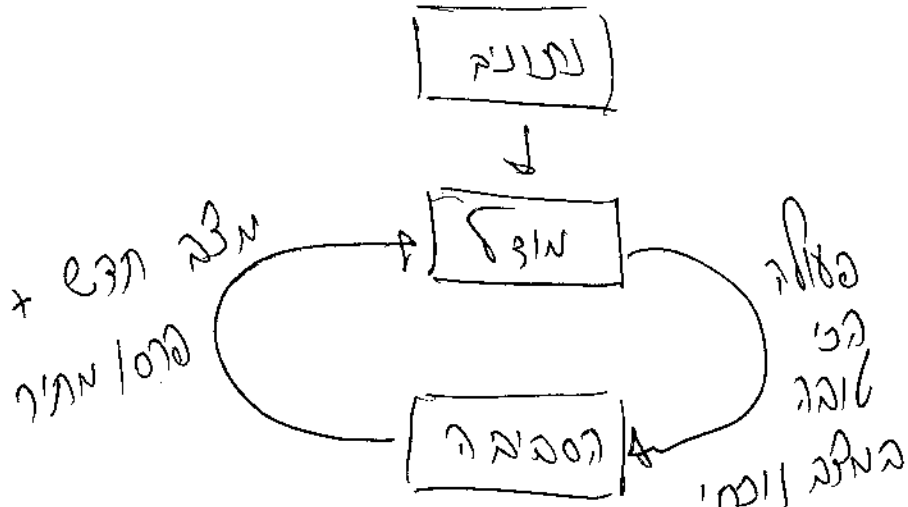
(1)

Reinforcement Learning

למידה מתחזקת:

הגדרה תאורטית:



מאמה כך שונה למידה של פונקציית מטרה שרגיל לזה?
מה מ"פ RL? למידה + תכנון.

Markov Decision Process

- אוסף מצבים S. S_t המצב בזמן t
- אוסף פעולות A. A_t פעולה שנבחרה בזמן t
- $P(s'|s,a)$ פונקציית מעברים
- R אוסף פרסים. R_t פרס שנקבל בזמן t שהיה שנבחרה בפעולה $R(s,a)$ לפעמים גם מסמנים $P(R,s'|s,a)$ לפונקציה

מה המטרה?

מקבלת מדיניות (policy) $\pi(s_t) \to A$ (יתן סטוכסטי)

כך שנמקסם את

$$R_t + R_{t+1} + R_{t+2} + \dots$$

מה נלמד? בהתלת כן, אם יתן כי נעבד שיקול בסיס מוקא טאורת כלומר המטרה היא

$$\pi^* = \arg\max_\pi E\left[\sum R_t \gamma^t\right]$$

Policy Gradient: נגדיר $\pi_\theta(a|s) = f(a,s;\theta)$

$$J(\theta) = E\left[\sum_{t=0}^{T} \gamma^t R(s_t, A_t)\right]$$

נתונים מהם טוב למדים.

ועכשיו ....

מיקי פשוטה יותר $f(\theta, s, a)$ כלי לחדות רשת מעקה שלנו עסוק בקורס. במקום זה, נתמקד בשיטות מבוססות value (מגדר מצב, פעולה ופעולה).

$$V_\pi(s) = E_\pi\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \mid S_t = s\right]$$

$$Q_\pi(s,a) = E_\pi\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \mid S_t = s, A_t = a\right]$$

(3)

משפט: $V^\pi(s) = r(s,\pi(s)) + \gamma \sum_{s'} p(s'|s,\pi(s)) V^\pi(s')$

הוכחה:

$$V^\pi(s) = E\left[\sum_t \gamma^t R_t \mid S_0 = s\right]$$

$$= E[R_0 \mid S_0 = s] + E\left[\sum_{t=1} \gamma^t R_t \mid S_0 = s\right]$$

$$= r(s,\pi(s)) + E\left[\sum_{t=1} \gamma^t R_t \mid S_0 = s\right]$$

$$= r(s,\pi(s)) + E\left[E\left[\sum_{t=1} \gamma^t R_t \mid S_0 = s, S_1\right]\right]$$

$$= r(s,\pi(s)) + \sum_{s'} P(S_1 = s' \mid S_0 = s) E\left[\sum_{t=1} \gamma^t R_t \mid S_0 = s, S_1 = s'\right]$$

$$= r(s,\pi(s)) + \sum_{s'} p(s'|s,\pi(s)) E\left[\sum_{t=1} \gamma^t R_t \mid S_1 = s'\right]$$

$$= r(s,\pi(s)) + \gamma \sum_{s'} p(s'|s,\pi(s)) V^\pi(s')$$

מהלך הוכחה זה מוכר?

בהגדרה דומה: של Q שלנו נעזרה שוב כדי להמשיך את Q^*.

אלגוריתם 1: חישוב V^π ו-Q^π במדיניות π (Policy Evaluation):

$\forall s$ $V^*_{t+1}(s) = r(s,\pi(s)) + \gamma \sum_{s'} p(s'|s,\pi(s)) V^*_t(s')$

משפט: התהליך מתכנס ל-$V^\pi(s)$

(4)

תזכורת: נשים $T(V_t)$ ונסמן V_{t+1} ונמשיך הלאה. מה שחשוב לנו הוא ש-T הוא אופרטור מצומצם (contracting).

$$0 \le c \le 1 \qquad d(T(x), T(x')) \le c \cdot d(x, x')$$

Banach: כל סדרה x_t שמקיימת את התנאי שמתחת מתכנסת לנקודת שבת.

$$d(x_{t+1}, x^*) \le c \cdot d(x_t, x^*)$$

$$\Downarrow$$

$$d(x_{t+1}, x^*) = c^t d(x_0, x^*)$$

$$\Downarrow$$

$$t = \log \varepsilon / d(x_0, x^*)$$

$$d(T[V], T[\tilde{V}]) = \| T(V) - T(\tilde{V}) \|_\infty$$

$$= \max_s \left| r(s, \pi(s)) + \gamma \sum_{s'} p(s'|s, \pi(s)) V(s') - \left(r(s, \pi(s)) + \gamma \sum_{s'} p(s'|s, \pi(s)) \tilde{V}(s') \right) \right|$$

$$\le \gamma \max_s \left| \sum_{s'} p(s'|s, \pi(s)) [V(s') - \tilde{V}(s')] \right|$$

$$\le \gamma \max_s \sum_{s'} p(s'|s, \pi(s)) \left| V(s') - \tilde{V}(s') \right|$$

$$\le \gamma \max_s \max_{s'} \left| V(s') - \tilde{V}(s') \right| = \gamma \| V - \tilde{V} \|_\infty$$

הערה: אם $\gamma = 1$ אז אנחנו לא מובטחים שנגיע לנקודת שבת.

ובמקרה כזה, אפשר למשל לחשב את $Q^\pi(s,a)$

Optimal Policy

שנו מדיניות π לפי $V_\pi(s) \geq V_{\pi^*}(s)$ מתקיים π^* מדיניות אופטימלית
אחת לפחות s. סמנו π^* נסמן את הפונקציות

$$V_*(s) = \max_\pi V_\pi(s) \qquad Q^*(s,a) = \max_\pi Q_\pi(s,a)$$

$$V_*(s) = \max_a Q^*(s,a)$$

מייצג פעולה a כלשהי ב-s ואח"כ
מאמינים ב-π^*

אם נתון $Q^*(s,a)$ נוכל למצוא את π ע"י:

(A) $$\boxed{\pi^*(s) = \arg\max_a Q^*(s,a)}$$

משוואת בלמן: $Q^*(s,a) = \max_\pi E\left[\sum_{t=0}^{\infty} \gamma^t R_t \mid s_0 = s, A_0 = a\right]$

$$= \max_\pi r(s,a) + \sum_{s'} p(s'|s,a) E\left[\sum_{t=1}^{\infty} \gamma^t R_t \mid s_1 = s'\right]$$

$$= r(s,a) + \gamma \max \sum_{s'} p(s'|s,a) E\left[\sum_{t=1}^{\infty} \gamma^{t-1} R_t \mid s_1 = s'\right]$$

סמלנו ב-s שהגענו ל- $V^*(s')$

$$= r(s,a) + \gamma \sum_{s'} p(s'|s,a) V^*(s')$$

$$= r(s,a) + \gamma \sum_{s'} p(s'|s,a) \max_a Q^*(a,s)$$

(B) $$\boxed{V^*(s) = \max_a r(s,a) + \gamma \sum_{s'} p(s'|s,a) V^*(s')}$$ נקודת בונוס: $V^*(s) = \max_a Q^*(s,a)$

Valve Iteration

$$V_{t+1}(s) = \max_a r(s,a) + \gamma \sum_{s'} p(s'|s,a) V_t(s')$$

משפט: VI מתכנס ל-V^*. מספיק שיעשה שם מתכנס לשם
ומתי ב' שלבים יהיה מספיק כדי שיהיה שלב מתכנס
מלמן הפולינומיות.

הסבר: מחושב ע"י policy evaluation

Policy Iteration

② שיפור מדיניות policy improvement:

$$\pi_{t+1} = \arg\max_a Q^{\pi_t}(s,a)$$

טענה: לכל s מתקיים $V^{\pi_{t+1}}(s) \geq V^{\pi_t}(s)$

① חישוב הערך של π_{t+1}: $V^{\pi_{t+1}}, Q^{\pi_{t+1}}$

שאלה: כמה מחזורים מספיק עד שמתכנס למדיניות π^*
הטוב ביותר? מהיר?
מתי מובטח שהגענו למדיניות הטובה ביותר?