(1)

למידה מנסיון

בידינו הפרקים (episodes) $s_0, a_0, r_0, s_1, a_1, r_1, \ldots$

מה ניתן ללמוד?

- $p(s'|s,a)$ } model based
- $p(r|s,a)$
- לנבא $V(s), Q(s,a)$ } model free
- לנבא פעולה $\pi(s,a)$
- לשלוט (control) π^*

Monte Carlo

נתונה π שמייצגת פוליסי (איך בוחרים בהתחלה?)

- מייצרים episode: $(s_0, a_0, r_0, \ldots, s_{T-1}, a_{T-1}, r_{T-1}, s_T)$
- לכל s_t ברצף מחשבים:

$$G_t = \sum_{k=t}^{T} \gamma^{k-t} r_k$$

- לכל s_t ברצף מוסיפים את הערך G_t: $U(s_t) = U(s_t) \cup \{G_t\}$

(first visit OR every visit)

עד התכנסות

- מחזירים $V^{\pi}(s) = \text{average}(U(s))$

כעת שבידינו שערוך ל-V^{π} נוכל שוב לנסות לשפר:

(1) Evaluation: $\pi_k \to Q^{\pi_k}(s,a)$ $\quad \forall s,a$

(2) Improvement: $\pi_{k+1}(s) = \arg\max_a Q^{\pi_k}(s,a)$ $\quad \forall s$

מה המגבלות? מה החסרונות?

(ב)

ε-Greedy Policy

מגדירים π_{k+1} בעזרת $Q^{\pi_k}(s,a)$ באופן הבא:
בהסתברות $1-\varepsilon$ בוחרים $\arg\max_{a'} Q^{\pi_k}(s,a')$
ובהסתברות ε בוחרים פעולה אקראית.

$$Q^{\pi_k}(s,\pi^{k+1}(s)) = \sum_a \pi_{k+1}(a|s) Q^{\pi_k}(s,a)$$

$$= \frac{\varepsilon}{|A|} \sum_a Q^{\pi_k}(s,a) + (1-\varepsilon) \max_a Q^{\pi_k}(s,a)$$

$$\geq \frac{\varepsilon}{|A|} \sum_a Q^{\pi_k}(s,a) + (1-\varepsilon) \cdot \sum_a \frac{\pi_k(a|s) - \varepsilon/|A|}{1-\varepsilon} \cdot Q^{\pi_k}(s,a)$$

$$= \sum_a \pi_k(a|s) \cdot Q^{\pi_k}(s,a) = V^{\pi_k}(s)$$

$$Q^{\pi_k}(s,\pi^{k+1}(s)) \geq V^{\pi_k}(s) \quad \Rightarrow$$

ומתקיים התנאים של משפט PI.

Temporal Difference

פורמולציה מעט שונה של MC היא כזאת בה נבצע עדכון

$$V(S_t) \leftarrow V(S_t) + \alpha\left[G_t - V(S_t)\right]$$

כאשר G_T הוא הרווח האמיתי שצפינו בו

α – "learning rate"; $\left[G_t - V(S_t)\right]$ – הטעות הנוכחית

אבל מדוע לחכות? במקום זאת נציג את אלגוריתם TD(0):

$$V(S_t) \leftarrow V(S_t) + \alpha\left[R_t + \gamma V(S_{t+1}) - V(S_t)\right]$$

$R_t + \gamma V(S_{t+1})$ – הערכה שלנו בזמן t, R_t זו המשלה.

$V(S_t)$ – הערכה ללא R_T

TD-error δ_t

תרגיל:

$$G_T - V(S_t) = \sum_{k=t} \gamma^{k-t}\delta_k$$

דוגמא: שני מצבים A ו-B. 8 episodes:

A, 0, B, 0	B, 1
B, 1	B, 1
B, 1	B, 1
B, 1	B, 0

מהו $V(A)$? $V(B)$?

Monte Carlo: $V(A)=0$ Minimize MSE on train

TD(0): $V(A)=V(B)=3/4$ Maximum Likelihood בדרך כלל מתכנס מהר יותר?

MC prediction

- מחכים עד הסוף
- עובד רק על episodes
- עדכון תלוי במספרים רבים עם שונות גבוהה
- שערוך לא מוטה
- לא תלוי באי-תלות מרקובית - יתרון!

TD prediction

- מעדכנים תוך כדי
- יכול גם להמשיך לנצח
- שונות קטנה המבוססת על מעבר בודד
- שערוך מוטה
- מסתמך על אי-תלות מרקובית

On policy SARSA

אלגוריתם control העוקב אחרי ה-policy (on policy)

$$Q_{t+1}(s,a) = Q_t(s,a) + \alpha_t\left(r(s,a) + \gamma Q_t(s',a') - Q_t(s,a)\right)$$

$a' = \pi(s')$

Off policy Q-Learning

$$Q_{t+1}(s,a) = Q_t(s,a) + \alpha_t\left[r(s,a) + \gamma \max_{a'} Q(s',a') - Q(s,a)\right]$$

behavior policy ≠ target policy

שווה למשוואת הרקורסיה של Q: $Q_{k+1}(s,a) = r(s,a) + \gamma \sum_{s'} p(s'|a,s) \max_{a'} Q_k(s',a')$

מה ניתן לומר על הקשר?

משפט: Q-Learning מתכנס (תחת תנאים מינימליים) ל-Q^*

__הוכחה למקרה הדטרמיניסטי__ $s'=g(s,a)$ (ונניח $\alpha=1$):

נסמן $\Delta_t = \|Q_t - Q^*\|_\infty = \max_{s,a} |Q_t(s,a) - Q^*(s,a)|$

$$|Q_{t+1}(s_t,a_t) - Q^*(s_t,a_t)| = |r(s_t,a_t) + \gamma \max_{a'} Q_t(s_{t+1},a')$$

$$- r(s_t,a_t) - \gamma \max_{a'} Q^*(g(s_t),a')|$$

(זה דטרמיניסטי אז אין מעבר על כל האפשרויות s')

$$= \gamma \left| \max_{a'} Q_t(s_{t+1},a') - \max_{a'} Q^*(s_{t+1},a') \right|$$

$$\leq \gamma \max_{a} |Q_t(s_{t+1},a) - Q^*(s_{t+1},a)|$$

$$\leq \gamma \max_{a} \max_{s} |Q_t(s,a) - Q^*(s,a)| \leq \gamma \Delta_t$$

כלומר, $Q(s,a)$ מתקרב עם הזמן ל- $Q^*(s,a)$ כל עוד מובטח שנבקר ב- (s,a) מספיק פעמים.