تمرین ۲-۷

سارا آذرنوش ۹۸۱۷۰۴۹۸

سوال ① [illegible] سوالات اول پیشگفته

Policy iteration ← ③ بخش ملی دیده شد

۱) مقایسه‌ی ساده‌ی محاسبات بین Q-value یا V-value

① در حالت value iteration با داشتن احتمال P(s',a,s)، R(s',a,s) reward، احتمال γ را داریم و هدف پیدا کردن reward با شروع از استیت s به صورت بهینه است که به صورت زیر نوشته می‌شود

$$V(s) = \max_{a\in A} \sum_{s'\in S} P(s'|a,s)\left(R(s',a,s) + \gamma V(s')\right)$$

برای جلوگیری از مشکلات ... سابق $V_t(s)$ که ... $V_0 = 0$ را اضافه می‌کنیم و به شکل زیر در می‌آید

$$V_{t+1}(s) = \max_{a\in A} \sum_{s'\in S} P(s'|a,s)\left(R(s',a,s) + \gamma V_t(s')\right)$$

② در حالت Q-iteration، $Q(s,a)$ را که برابر با مقدار reward مورد انتظار با شروع از استیت s و اکشن a است تعریف می‌کنیم

$$V(s) = \max_{a\in A} Q(s,a) \longrightarrow = \sum_{s'\in S} P(s'|a,s)\left(R(s',a,s) + \gamma \max_{a\in A} Q(s',a')\right)$$

مانند V-iteration به time component اضافه می‌کنیم

$$Q_0(s,a) = 0$$

$$Q_{t+1}(s,a) = \sum_{s'\in S} P(s'|a,s)\left(R(s',a,s) + \gamma \max_{a'\in A} Q_t(s',a')\right)$$

الگوریتم ساده‌تر با داشتن خط مشی متوالی در نظر می‌گیریم

$s_0, a_0, s_1, a_1, \dots$ را از سیاست $\pi: S \to A$ دنبال کنیم

$$Q(s_t,a_t) = (1-\alpha) Q_{old}(s_t,a_t) + \alpha\left(R(s_{t+1},a_t,s_t) + \gamma \max_{a\in A} Q_{old}(s_{t+1},a')\right)$$

شبیه به Q-iteration است

$$E_{s_{t+1}}\left(R(s_{t+1},a_t,s_t) + \gamma \max_{a'\in A} Q_{old}(s_{t+1},a')\right)$$

$$= \sum_{s'\in S} P(s'|a_t,s_t)\left(R(s',a_t,s_t) + \gamma \max_{a'\in A} Q(s',a')\right)$$

به دنبال Q-value بهینه است

Scanned by CamScanner

ب ۱۱ چرا ε-greedy ؟ مشکلات ؟

ε-greedy بهتر است زیرا به جست و جو ادامه می‌دهد تا شانس رسیدن به حالت بهینه را افزایش دهد

هر چه مقدار ε کوچک‌تر باشد نیز الگوریتم کندتر و بار پردازش بیشتری است

اگر واریانس reward زیاد باشد و به جست و جوی بیشتری نیاز دارد و این الگوریتم پاسخ بهتری می‌دهد

اما اگر ε باشد برای مثال الگوریتم greedy در دفعه اول مقدار واقعی هر اکشن را می‌فهمد و زودتر پاسخ می‌دهد

هر زمان که مقدارها زیاد احتمال ½ است و از مشکلات آن رفتن به همه جا است تا پیدا شود ممکن است

مسیرهای بیهوده نیز طی شود که ε کوچک‌تر و تابع exploration از راه حل‌های آن است.

در صورتی که از این الگوریتم استفاده کنیم ممکن است مقدارهایی که مناسب هستند را دور می‌کنیم و جواب

بهینه باشد

۱۲.

$$\forall s: E_{a\sim\pi'}[Q^{\pi}(s,a)] \geq E_{a\sim\pi}[Q^{\pi}(s,a)] \Rightarrow$$

$$\forall s: V^{\pi'}(s) \geq V^{\pi}(s)$$

$$V^{\pi} = Q^{\pi}(s,\pi(s)) = E_{a\sim\pi}[Q^{\pi}(s,a)] \leq Q^{\pi}(s,a')$$

$$\pi' = \underset{a\in A}{\operatorname{argmax}}\, Q^{\pi}(s,a)$$

$$V^{\pi'} - V^{\pi} = \sum_{i=0}^{\infty} (V^{\pi_{i+1}} - V^{\pi})$$

$$= \sum_{i=0}^{\infty} \gamma^{i} \sum_{s'\in S} P(s_{i+1} = s' | s_0 = s, \pi'] \left(Q^{\pi}(s',\pi'(s')) - Q^{\pi}(s',\pi(s'))\right)$$

$$= \sum_{i=0}^{\infty} \gamma^{i} \sum_{s'\in S} P[s_{i+1} = s' | s_0 = s, \pi'] \underbrace{A^{\pi}(s',\pi')}_{\geq 0 \rightarrow \text{طبق فرض}}$$

$$\rightarrow \frac{1}{1-\gamma} d^{\pi',s}(s')$$

$$V^{\pi'} - V^{\pi} \geq 0 \rightarrow V^{\pi'} \geq V^{\pi}$$