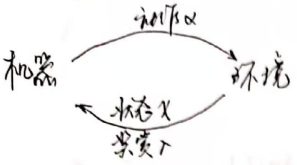
第16章 强化学习



马尔可夫决策过程：

环境 E，状态空间 X，$x \in X$

动作空间 A，$a \in A$

a 作用在状态 x 上，潜在的转移函数 P 将使环境状态发生变化（一定概率）

同时会以潜在的"奖赏"函数 R 反馈给机器一个奖赏

四元组 $E = \langle X, A, P, R \rangle$

$P: X \times A \times X \to \mathbb{R}$ 指定状态转移概率

$R: X \times A \times X \to \mathbb{R}$ 或 $X \times A \to \mathbb{R}$ } 奖赏

注意机器与环境的界限

- 状态转移，奖赏返回不受机器控制
- 机器只能通过实执行动作影响环境
- 通过感知状态和奖赏来感知环境

策略 $\pi: X \to A$ 确定性

$: X \times A \to \mathbb{R}$ 概率性

目标：最优策略

长期累积奖赏最大化

16.2 探索与利用

① 探索：获取每个动作的期望奖赏估计

利用：根据目前最优的动作（但实际可能错）

常要折中

② ϵ-贪心

以 ϵ 概率探索，$1-\epsilon$ 概率利用

均匀概率随机选取一个摇臂 / 选取当前平均奖赏最高

$Q(k) = \frac{1}{n}\sum_{i=1}^{n} v_i$ 摇臂 k 的平均奖赏

$$Q_{n+1}(k) = \frac{1}{n+1}\left((n+1) \times Q_{n+1}(k) + v_n\right)$$
$$= Q_{n+1}(k) + \frac{1}{n+1}\left(v_n - Q_{n+1}(k)\right)$$

③ Softmax

基于当前已知的摇臂奖赏来对探索和利用进行折中。

若各摇臂的平均奖赏相当，选取概率也相当

若某些更高，则选取概率也更高

$$P(k) = \frac{e^{\frac{Q(k)}{\tau}}}{\sum_{i=1}^{K} e^{\frac{Q(k)}{\tau}}} \qquad \tau > 0 \text{ (温度)}$$

τ 趋小则平均奖赏越高的摇臂选取概率也越高，$\begin{cases} \tau \to 0 \Rightarrow \text{仅利用} \\ \tau \to \infty \Rightarrow \text{仅探索} \end{cases}$

将 K-摇臂赌博问题算法用于每个状态上动作的选择，用强化学习的累积奖赏来代替 K-摇臂奖赏函数

16.3 有模型学习

$E = \langle X, A, P, R \rangle$ 已知

$V^{\pi}(x)$ 状态值函数

$Q^{\pi}(x,a)$ 状态-动作值函数

$$\begin{cases} V_T^{\pi}(x) = E_{\pi}\left[\frac{1}{T}\sum_{t=1}^{T} r_t \mid x_0 = x\right] & T\text{步} \\ V_{\gamma}^{\pi}(x) = E_{\pi}\left[\sum_{t=0}^{\infty} \gamma^t r_{t+1} \mid x_0 = x\right] & \gamma\text{折扣} \end{cases}$$

$$\begin{cases} Q_T^{\pi}(x,a) = E_{\pi}\left[\frac{1}{T}\sum_{t=1}^{T} r_t \mid x_0 = x, a_0 = a\right] \\ Q_T^{\pi}(x,a) = E_{\pi}\left[\frac{1}{T}\sum_{t=0}^{T} r_{t+1} \mid x_0 = x, a_0 = a\right] \end{cases}$$

MDP 具有马尔可夫性

系统下一时刻状态仅由当前时刻状态决定，不依赖于以往其他状态

$$V_T^{\pi}(x)=E_{\pi}\left[\frac{1}{T}\sum_{t=1}^{T}r_t \mid x_0=x\right]$$

$$=E_{\pi}\left[\frac{1}{T}r_1+\frac{T-1}{T}\frac{1}{T-1}\sum_{t=2}^{T}r_t \mid x_0=x\right]$$

动态规划-全概率展开

$$=\sum_{a\in A}\pi(x,a)\sum_{x'\in X}P_{x\to x'}^{a}\left(\frac{1}{T}R_{x\to x'}^{a}+\frac{T-1}{T}E_{\pi}\left[\frac{1}{T-1}\sum_{t=1}^{T-1}r_t \mid x_0=x'\right]\right)$$

$$=\sum_{a\in A}\pi(x,a)\sum_{x'\in X}P_{x\to x'}^{a}\left(\frac{1}{T}R_{x\to x'}^{a}+\frac{T-1}{T}V_{T-1}^{\pi}(x')\right)$$

$$V_{\gamma}^{\pi}(x)=\sum_{a\in A}\pi(x,a)\sum_{x'\in X}P_{x\to x'}^{a}\left(R_{x\to x'}^{a}+\gamma V_{\gamma}^{\pi}(x')\right)$$

注意：P，R 均已知

16.3.2 策略改进

$$\pi^{*}=\arg\max_{\pi}\sum_{x\in X}V^{\pi}(x)$$

$$V_T^{*}(x)=\max_{a\in A}\sum_{x'\in X}P_{x\to x'}^{a}\left(\frac{1}{T}R_{x\to x'}^{a}+\frac{T-1}{T}V_{T-1}^{*}(x')\right)$$

$$V_{\gamma}^{*}(x)=\max_{a\in A}\sum_{x'\in X}P_{x\to x'}^{a}\left(R_{x\to x'}^{a}+\gamma V_{\gamma}^{*}(x')\right)$$

$$V^{*}(x)=\max_{a\in A}Q^{\pi^{*}}(x,a)$$

$$Q_T^{*}(x,a)=\sum_{x'\in X}P_{x\to x'}^{a}\left(\frac{1}{T}R_{x\to x'}^{a}+\frac{T-1}{T}\max_{a'\in A}Q_{T-1}^{*}(x',a')\right)$$

$$Q_{\gamma}^{*}(x,a)=\sum_{x'\in X}P_{x\to x'}^{a}\left(R_{x\to x'}^{a}+\gamma\max_{a'\in A}Q_{\gamma}^{*}(x',a')\right)$$

$$V^{\pi}(x)\le Q^{\pi}(x,\pi'(x))$$

$$=\sum_{x'\in X}P_{x\to x'}^{\pi(x)}\left(R_{x\to x'}^{\pi(x)}+\gamma V^{\pi}(x')\right)$$

$$\le\sum_{x'\in X}P_{x\to x'}^{\pi(x)}\left(R_{x\to x'}^{\pi'(x)}+\gamma Q^{\pi}(x',\pi'(x'))\right)$$

$$=V^{\pi'}(x)$$

$$\pi'(x)=\arg\max_{a\in A}Q^{\pi}(x,a)$$

16.3.3 策略迭代 值迭代

16.4 免模型学习

环境：状态，状态转移概率，奖赏函数往往很难直接得知

同策略蒙特卡罗强化学习

策略评估和改进是一个策略

异策略：

于在 p 分布下的期望：

$$E[f]=\int_x p(x)f(x)\,dx$$

从 p 上采样 $x_1,x_2,\cdots,x_m$

$$\hat{E}[f]=\frac{1}{m}\sum_{i=1}^{m}f(x_i)$$

若从另一分布 q，则于在 p 下的期望

$$E[f]=\int_x q(x)\frac{p(x)}{q(x)}f(x)\,dx$$

看作 $\frac{p(x)}{q(x)}f(x)$ 在分布 q 下的期望

$$\hat{E}[f]=\frac{1}{m}\sum_{i=1}^{m}\frac{p(x_i)}{q(x_i)}f(x_i)$$

策略不同采样轨迹来评估 π，（累积奖赏估计期望）

$$Q(x,a)=\frac{1}{m}\sum_{i=1}^{m}r_i$$

若用 π' 的采样来评估 π，则仅需对累积奖赏加权

$$Q(x,a)=\frac{1}{m}\sum_{i=1}^{m}\frac{P_i^{\pi}}{P_i^{\pi'}}r_i$$

$$P^{\pi}=\prod_{i=0}^{T-1}\pi(x_i,a_i)P_{x_i\to x_{i+1}}^{a_i}$$

$$\frac{P^{\pi}}{P^{\pi'}}=\prod_{i=0}^{T-1}\frac{\pi(x_i,a_i)}{\pi'(x_i,a_i)}$$

10.4.2

时序差分学习

蒙特卡罗强化学习每次学试后求平均来作为期望积累奖赏的近似

"批处理式"：在一完整的采样轨迹完成后再对所有的状态-动作对更新

增量式进行：

$$Q_t^{\pi}(x,a)=\frac{1}{t}\sum_{i=1}^{t} r_i$$

$$Q_{t+1}^{\pi}(x,a)=Q_t^{\pi}(x,a)+\underbrace{\frac{1}{t+1}\left(r_{t+1}-Q_t^{\pi}(x,a)\right)}_{\text{增量}}$$

$$\Downarrow$$

$$\underbrace{\frac{\alpha_{t+1}}{\alpha}}_{\text{更新步长}}\left(r_{t+1}-Q_t^{\pi}(x,a)\right)$$

α越大，越靠后的累积奖赏越重要