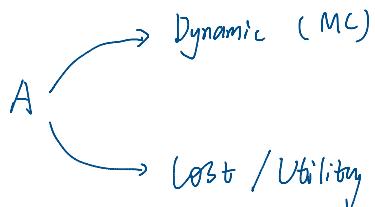


Markov Decision Process

2023年9月28日 星期四 下午9:33

建议的练习. 1.70. 1.74. 1.75

Markov Decision Process



定义：马尔可夫决策过程是一个4-tuple (S, A, P_a, R) . S 代表状态空间， A 是行动(actions)的集合

$P_a : A \times S \times S \rightarrow [0, 1]$ 代表转移概率，因此 $\sum_{y \in S} P_a(x, y) = 1$. $R : S \times A \rightarrow \mathbb{R}$ 为即时奖励函数 (immediate reward)

为了从中找到一个MC. 我们需要知道所有未来的行动.

$\mathcal{A} = \{\alpha : N \times S \rightarrow A\}$ 被称为 admissible control.
可采纳的. 可接受的.

对于任何给定的 $\alpha \in \mathcal{A}$ $P(X_{n+1} = y | X_n = x) = P_{\alpha(n, x)}(x, y)$.

我们把它记作 X_n^α .

objective: $U(x, \alpha) = \mathbb{E}_x \left[\sum_{n \geq 0} \lambda^n R(X_n^\alpha, \alpha(n, X_n^\alpha)) \right]$ 其中 $\lambda \in (0, 1)$ 是 discounting factor.

让MDP的值是 $V(x) = \sup_{\alpha} U(x, \alpha)$.

Bellman Equation.

$$\begin{aligned} \mathbb{E}_x \left[\sum_{n \geq 0} \lambda^n R(X_n^\alpha, \alpha(n, X_n^\alpha)) \right] &= R(x, \alpha(0, x)) + \mathbb{E}_x \left[\sum_{n \geq 1} \lambda^n R(X_n^\alpha, \alpha(n, X_n^\alpha)) \right] \\ &= R(x, \alpha(0, x)) + \lambda \mathbb{E}_x \left[\sum_{n \geq 1} \lambda^n R(X_{n+1}^\alpha, \alpha(n+1, X_n^\alpha)) \right] \quad \text{把从0到n的状价拿出来} \\ &= R(x, \alpha(0, x)) + \lambda \sum_{z \in S} P_{\alpha(0, x)}(z) \mathbb{E}_z \left[\sum_{n \geq 1} \lambda^n R(X_n^\alpha, \alpha(n+1, X_n^\alpha)) \right]. \end{aligned}$$

为了找到最优的 α .

$$V(x) = \sup_{\alpha \in \mathcal{A}} \left[R(x, \alpha) + \lambda \sum_{z \in S} P_{\alpha}(x, z) V(z) \right] \quad \text{first Bellman Equation}$$

Banach Fixed Point Theorem.

Let $T : S \rightarrow S$ be a contraction mapping. 即 $\exists 0 < \lambda < 1$. 使得 $d(T(x), T(y)) \leq \lambda d(x, y)$

则存在唯一的 T 的不动点 $T(x^*) = x^*$

证明：选择 $\forall x_0 \in S$. 让 $x_n = T(x_{n-1}) \quad \forall n \geq 1$.

$$d(x_{n+1}, x_n) = d(T(x_n), T(x_{n-1})) \leq \lambda d(x_n, x_{n-1}) \leq \lambda d(x_n, x_0)$$

$$d(x_m, x_n) \leq \sum_{k=n+1}^m d(x_k, x_{k-1}) \leq d(x_1, x_0) \sum_{k=n+1}^m \lambda^k = d(x_1, x_0) \lambda^n \frac{1-\lambda^{m-n}}{1-\lambda}$$

x_n 是一个 Cauchy Sequence 令它收敛至 x^*

$$\begin{aligned} d(T(x^*), x^*) &\leq d(T(x^*), x_n) + d(x_n, x^*) = d(T(x^*), T(x_{n-1})) + d(x_n, x^*) \\ &\leq \lambda d(x^*, x_{n-1}) + d(x_n, x^*) \rightarrow 0. \end{aligned}$$

对于唯一性 let x 是另一个不动点 $d(x, x^*) \leq d(T(x), T(x^*)) \leq \lambda d(x, x^*) \Rightarrow d(x, x^*) = 0$.

定理：假设奖励函数 R 是有界的. 则 V 定义为 (*) 是 Bellman 方程唯一有界的解

更进一步的如果存在 $I^*: S \rightarrow A$ 使得 $I^*(x) \in \arg\max_{a \in A} \left\{ R(x, a) + \sum_{z \in S} p_a(x, z)V(z) \right\}$

然后有 $x^*(n, x) = I^*(x)$ 是一个最优控制 (optimal control)

(由于这是在假设这个 Bellman 方程有且仅有唯一解).

让 $B(S, R)$ 是具备 supremum metric 的有界函数集合. 定义 $T: B(S, R) \rightarrow B(S, R)$

$$T(V)(x) := \sup_{a \in A} \left\{ R(x, a) + \lambda \sum_{z \in S} p_a(x, z)V(z) \right\}$$

$$\begin{aligned} \sup_{x \in S} |T(V)(x) - T(\tilde{V})(x)| &\leq \sup_x \sup_a \left\{ \lambda \sum_{z \in S} p_a(x, z) |V(z) - \tilde{V}(z)| \right\} \\ &\leq \lambda \left(\sup_{z \in S} |V(z) - \tilde{V}(z)| \right) \sup_x \sup_a \sum_{z \in S} p_a(x, z) = \lambda (\sup_z |V(z) - \tilde{V}(z)|) \end{aligned}$$