# 从强化学习到择时策略

朱大福

2024/07/09

# Outline

# 几类学习问题

机器学习：Data-driven

- 有监督学习
  - Data：$(x, y)$
  - 目标：$x \to y$

- 无监督学习
  - Data：$x$
  - 目标：$x_1$ 和 $x_2$ 是同类

- 强化学习
  - Data：$(s, a)$
  - 目标：回报最大化

# Motivation

有一家公司，怎样估值?

# Motivation

有一家公司，怎样估值?

DCF 模型（Discounted Cash Flow）

$$DCF_t = CF_t + \frac{CF_{t+1}}{1+r} + \frac{CF_{t+2}}{(1+r)^2} + \cdots$$

# Motivation

有一家公司，怎样估值?

DCF 模型（Discounted Cash Flow）

$$DCF_t = CF_t + \frac{CF_{t+1}}{1+r} + \frac{CF_{t+2}}{(1+r)^2} + \cdots$$

记作 $G_t = DCF_t$, $R_t = CF_t$, $\gamma = 1/(1+r)$

$$\begin{aligned} G_t &= R_t + \gamma R_{t+1} + \gamma^2 R_{t+2} + \cdots \\ &= R_t + \gamma(R_{t+1} + \gamma R_{t+2} + \cdots) \end{aligned}$$

$$\therefore G_t = R_t + \gamma G_{t+1}$$

# Motivation

有一家公司，怎样估值?

DCF 模型（Discounted Cash Flow）

$$DCF_t = CF_t + \frac{CF_{t+1}}{1+r} + \frac{CF_{t+2}}{(1+r)^2} + \cdots$$

记作 $G_t = DCF_t$, $R_t = CF_t$, $\gamma = 1/(1+r)$

$$\begin{aligned} G_t &= R_t + \gamma R_{t+1} + \gamma^2 R_{t+2} + \cdots \\ &= R_t + \gamma(R_{t+1} + \gamma R_{t+2} + \cdots) \end{aligned}$$

$$\therefore G_t = R_t + \gamma G_{t+1}$$

公司的价值为

$$V_t = \mathbb{E}[G_t]$$

Nothing special

# 状态（s），价值（V）

假设公司有两种状态

$$\mathcal{S} = \{\text{业绩好}, \text{业绩差}\}$$

当前的状态为 $S_t = s \in \mathcal{S}$。

显然业绩会影响公司的价值

$$V(S_t = s) = \mathbb{E}[G_t | S_t = s]$$

回忆 $G_t = R_t + \gamma G_{t+1}$

$$\therefore V(S_t = s) = \mathbb{E}[R_t | S_t = s] + \gamma V(S_{t+1} = s' | S_t = s)$$

简写为

$$V(s) = \mathbb{E}[R_t | s] + \gamma V(s' | s)$$

# Markov Decision Process

马尔可夫决策过程（MDP）：下一刻的状态只和当前时刻有关

$$P(S_{t+1}|S_t, S_{t-1}, \cdots) = P(S_{t+1}|S_t) = P(s'|s)$$

公司 $t+1$ 期的业绩情况只和 $t$ 期有关，状态转移用概率表示，比如 $P(S_{t+1} = \text{业绩好}|S_t = \text{业绩差}), \cdots$

$$\begin{aligned} V(s'|s) &= P(\text{好}|s) \cdot V(\text{好}) + P(\text{差}|s) \cdot V(\text{差}) \\ &= \sum_{s' \in \mathcal{S}} P(s'|s) \cdot V(s') \end{aligned}$$

# 动作（a）

作为公司老板，你的目标是什么

# 动作（a）

作为公司老板，你的目标是什么

你能做的事情是

$$\mathcal{A} = \{\text{研发}, \text{不研发}\}$$

策略是

$$\pi(\text{研发}|\text{好}) = p, \quad \pi(\text{不研发}|\text{好}) = 1 - p$$

$$\pi(\text{研发}|\text{差}) = q, \quad \pi(\text{不研发}|\text{差}) = 1 - q$$

我们很好奇 $p =?, q =?$

# 动作（a）

回顾

$$V(s) = \mathbb{E}[R_t|s] + \gamma V(s'|s)$$

$$V(s'|s) = \sum_{s' \in \mathcal{S}} P(s'|s) \cdot V(s')$$

# 动作（a）

回顾

$$V(s) = \mathbb{E}[R_t|s] + \gamma V(s'|s)$$

$$V(s'|s) = \sum_{s' \in \mathcal{S}} P(s'|s) \cdot V(s')$$

$\Rightarrow$

$$V(s) = \mathbb{E}[R_t|s] + \gamma \cdot \left[ \sum_{s' \in \mathcal{S}} P(s'|s) \cdot V(s') \right]$$

$a$ 会影响谁?

# 动作（a）

回顾

$$V(s) = \mathbb{E}[R_t|s] + \gamma V(s'|s)$$

$$V(s'|s) = \sum_{s' \in \mathcal{S}} P(s'|s) \cdot V(s')$$

$\Rightarrow$

$$V(s) = \mathbb{E}[R_t|s] + \gamma \cdot \left[ \sum_{s' \in \mathcal{S}} P(s'|s) \cdot V(s') \right]$$

$a$ 会影响谁?

$$\mathbb{E}[R_t|s] \rightarrow \mathbb{E}[R_t|s, a]$$

$$P(s'|s) \rightarrow P(s'|s, a)$$

$$V(s) \rightarrow V(s|a)$$

# 动作价值函数（Q）

$$V(s) = \mathbb{E}[R_t|s] + \gamma \cdot \left[\sum_{s' \in \mathcal{S}} P(s'|s) \cdot V(s')\right]$$

$$\Downarrow$$

$$V(s|a) = \mathbb{E}[R_t|s, a] + \gamma \cdot \left[\sum_{s' \in \mathcal{S}} P(s'|s, a) \cdot V(s')\right]$$

# 动作价值函数（Q）

$$V(s) = \mathbb{E}[R_t|s] + \gamma \cdot \left[\sum_{s' \in \mathcal{S}} P(s'|s) \cdot V(s')\right]$$

$$\Downarrow$$

$$V(s|a) = \mathbb{E}[R_t|s, a] + \gamma \cdot \left[\sum_{s' \in \mathcal{S}} P(s'|s, a) \cdot V(s')\right]$$

定义

$$Q(s, a) = V(s|a)$$

# 动作价值函数（Q）

$$V(\text{好}) = \pi(\text{研发}|\text{好})Q(\text{好}, \text{研发}) + \pi(\text{不研发}|\text{好})Q(\text{好}, \text{不研发})$$

# 动作价值函数（Q）

$$V(\text{好}) = \pi(\text{研发}|\text{好})\,Q(\text{好}, \text{研发}) + \pi(\text{不研发}|\text{好})\,Q(\text{好}, \text{不研发})$$

$$V(s) = \sum_{a \in \mathcal{A}} \pi(a|s)\,Q(s, a)$$

# 动作价值函数（Q）

$$V(\text{好}) = \pi(\text{研发}|\text{好})\,Q(\text{好}, \text{研发}) + \pi(\text{不研发}|\text{好})\,Q(\text{好}, \text{不研发})$$

$$V(s) = \sum_{a \in \mathcal{A}} \pi(a|s)\,Q(s, a)$$

$$Q(s, a) = \mathbb{E}[R_t|s, a] + \gamma \cdot \left[ \sum_{s' \in \mathcal{S}} P(s'|s, a) \cdot V(s') \right]$$

# 动作价值函数（Q）

$$V(\text{好}) = \pi(\text{研发}|\text{好})\,Q(\text{好}, \text{研发}) + \pi(\text{不研发}|\text{好})\,Q(\text{好}, \text{不研发})$$

$$V(s) = \sum_{a \in \mathcal{A}} \pi(a|s)\,Q(s, a)$$

$$Q(s, a) = \mathbb{E}[R_t|s, a] + \gamma \cdot \left[ \sum_{s' \in \mathcal{S}} P(s'|s, a) \cdot V(s') \right]$$

$$Q(s, a) = \mathbb{E}[R_t|s, a] + \gamma \cdot \left[ \sum_{s' \in \mathcal{S}} P(s'|s, a) \cdot \sum_{a' \in \mathcal{A}} \pi(a'|s')\,Q(s', a') \right]$$

# 随机性的两个来源

$$Q(s, a) = \mathbb{E}[R_t|s, a] + \gamma \cdot \left[ \sum_{s' \in \mathcal{S}} P(s'|s, a) \cdot \sum_{a' \in \mathcal{A}} \pi(a'|s') Q(s', a') \right]$$

1. 状态 s 下，做什么动作 a?

$$\pi(a|s)$$

2. 状态 s 下且做出动作 a，会到达什么状态 s'?

$$P(s'|s, a)$$

# Outline

# 最优化问题

$$Q(\text{业绩好}, \text{研发}) = 130, \quad Q(\text{业绩好}, \text{不研发}) = -50$$