Real Machine Learning

Объяснение целей через награду

Кумулятивная награда

$$G_t \triangleq R_t + R_{t+1} + R_{t+2} + \ldots + R_T$$

Дисконтирование награды

Решение позитивной обратной связи и расходимости

$$G_t \triangleq R_t + \gamma R_{t+1} + \gamma^2 R_{t+2} + \ldots = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}$$

Другая точка зрения: длительность эффекта награды

$$
\begin{aligned}
G_0 &= R_0 + \gamma R_1 + \gamma^2 R_2 + \ldots + \gamma^T R_T \\
&= (1-\gamma)R_0 \\
&\quad + (1-\gamma)\gamma(R_0 + R_1) \\
&\quad + (1-\gamma)\gamma^2(R_0 + R_1 + R_2) \\
&\quad \cdots \\
&\quad + \gamma^T \cdot \sum_{t=0}^{T} R_t
\end{aligned}
$$

Решаем вероятностность - мат. ожиданием

Наша политика и среда - вероятностные. Максимизируем мат. ожидание!

$$
\begin{aligned}
\mathbb{E}\left[G_0\right] &= \mathbb{E}\left[R_0 + \gamma R_1 + ... + \gamma^T R_T\right] \\
&= \mathbb{E}_{E,\pi_\theta}\left[\,G_0\,\right] \\
&= \mathbb{E}_{\pi_\theta}\left[\,G_0\,\right] \\
&= \mathbb{E}\left[G_0 \mid \pi_\theta\right] \\
&= \mathbb{E}_{\substack{s_{0:T} \\ a_{0:T}}}\left[\,G_0\,\right] \\
&= \mathbb{E}_{s_0}\left[\,\mathbb{E}_{a_0|s_0}\left[\,R_0 + \mathbb{E}_{s_1|s_0,a_0}\left[\,\mathbb{E}_{a_1|s_1}\left[\,\gamma R_1 + ...\,\right]\,\right]\,\right]\,\right] \\
&= \sum_{t=0}^{T} \mathbb{E}_{(s_t,a_t)\sim p_\theta}\left[\,\gamma^t R_t\,\right] \\
&= \mathbb{E}_{\tau\sim p_\theta(\tau)}\left[\,G(\tau)\,\right]
\end{aligned}
$$

$$
\begin{aligned}
\mathbb{E}\left[G_0\right] &= \mathbb{E}\left[R_0 + \gamma R_1 + \ldots + \gamma^T R_T\right] \\
&= \mathbb{E}_{E,\pi_\theta}\left[\, G_0 \,\right] \\
&= \mathbb{E}_{\pi_\theta}\left[\, G_0 \,\right] \\
&= \mathbb{E}\left[G_0 \mid \pi_\theta\right] \\
&= \mathbb{E}_{\substack{s_{0:T} \\ a_{0:T}}}\left[\, G_0 \,\right] \\
&= \mathbb{E}_{s_0}\left[\, \mathbb{E}_{a_0|s_0}\left[\, R_0 + \mathbb{E}_{s_1|s_0,a_0}\left[\, \mathbb{E}_{a_1|s_1}\left[\, \gamma R_1 + \ldots \,\right]\,\right]\,\right]\,\right] \\
&= \sum_{t=0}^{T} \mathbb{E}_{(s_t,a_t)\sim p_\theta}\left[\, \gamma^t R_t \,\right] \\
&= \mathbb{E}_{\tau\sim p_\theta(\tau)}\left[\, G(\tau) \,\right]
\end{aligned}
$$

$$
\tau \triangleq (s_0, a_0, s_1, \ldots, a_{T-1}, s_T)
$$

$$
p_\theta(\tau) = p(s_0) \prod_{t=0}^{T-1} \pi_\theta(a_t|s_t)\; p(s_{t+1}|s_t, a_t)
$$

Уравнения Беллмана

Для постановки задачи динамического программирования

V(s) - по политике

$$v_\pi(s) = \sum_a \pi(a \mid s) \sum_{r,\, s'} p(r, s' \mid s, a) \left[r + \gamma v_\pi(s') \right]$$

$$= \mathbb{E}_\pi \left[R_t + \gamma v_\pi(S_{t+1}) \mid S_t = s \right]$$

q(s, a) - по политике

$$
\begin{aligned}
q_\pi(s, a) &= \sum_{r,\, s'} p(r, s' \mid s, a) \left[r + \gamma v_\pi(s') \right] \\
&= \sum_{r,\, s'} p(r, s' \mid s, a) \Big[r + \gamma \sum_{a'} \pi(a' \mid s') q_\pi(s', a') \Big]
\end{aligned}
$$

V(s) - оптимальное

$$
\begin{aligned}
v_*(s) &= \max_a \sum_{r,\, s'} p(r, s' \mid s, a) \left[r + \gamma v_*(s') \right] \\
&= \max_a \mathbb{E} \left[R_t + \gamma v_*(S_{t+1}) \mid S_t = s, A_t = a \right]
\end{aligned}
$$

q(s, a) - оптимальное

$$q_*(s,a) = \mathbb{E}\left[R_t + \gamma \max_{a'} q_*(S_{t+1}, a') \;\middle|\; S_t = s, A_t = a\right]$$

$$= \sum_{r,\, s'} p(r, s' \mid s, a)\left[r + \gamma \max_{a'} q_*(s', a')\right]$$

Операторный вид уравнений Беллмана

$$[\mathcal{T}^\pi V](s) = \mathbb{E}_{r,s'|s,a=\pi(s)}\left[\, r + \gamma V(s')\, \right]$$

$$[\mathcal{T}^\pi Q](s,a) = \mathbb{E}_{r,s'|s,a}\left[\, r + \gamma \mathbb{E}_{a'\sim\pi(s')}\left[\, Q(s',a')\, \right]\, \right]$$

$$[\mathcal{T} V](s) = \max_a \mathbb{E}_{r,s'|s,a}\left[\, r + \gamma V(s')\, \right]$$

$$[\mathcal{T} Q](s,a) = \mathbb{E}_{r,s'|s,a}\left[\, r + \gamma \max_{a'} Q(s',a') \right]$$

Все операторы порождают сжимающие отображения - значит имеют неподвижную точку

$$[\mathcal{T}^{\pi}V](s) = \mathbb{E}_{r,s'|s,a=\pi(s)}\left[r + \gamma V(s')\right]$$

$$[\mathcal{T}^{\pi}Q](s,a) = \mathbb{E}_{r,s'|s,a}\left[r + \gamma \mathbb{E}_{a'\sim\pi(s')}\left[Q(s',a')\right]\right]$$

$$[\mathcal{T}V](s) = \max_{a} \mathbb{E}_{r,s'|s,a}\left[r + \gamma V(s')\right]$$

$$[\mathcal{T}Q](s,a) = \mathbb{E}_{r,s'|s,a}\left[r + \gamma \max_{a'} Q(s',a')\right]$$

Не можете оценить – не можете улучшить

Оценка качества политики

$$v_\pi(s) = \sum_a \pi(a \mid s) \sum_{r,\, s'} p(r, s' \mid s, a) \left[r + \gamma v_\pi(s') \right]$$

$$= \mathbb{E}_\pi \left[R_t + \gamma v_\pi(S_{t+1}) \mid S_t = s \right]$$

Улучшение политики

$$\pi'(s) \leftarrow \underset{a}{\arg\max} \overbrace{\sum_{r,\, s'} p(r, s' \mid s, a)\,[r + \gamma v_\pi(s')]}^{q_\pi(s,\, a)}$$

$$v_{\pi'}(s) = \max_{a} \sum_{r,\, s'} p(r, s' \mid s, a)\,[r + \gamma v_\pi(s')]$$

1. Initialization
   $V(s) \in \mathbb{R}$ and $\pi(s) \in \mathcal{A}(s)$ arbitrarily for all $s \in \mathcal{S}$

2. Policy Evaluation
   Repeat
   $\quad \Delta \leftarrow 0$
   $\quad$ For each $s \in \mathcal{S}$:
   $\qquad v \leftarrow V(s)$
   $\qquad V(s) \leftarrow \sum_{s',r} p(s', r|s, \pi(s))\big[r + \gamma V(s')\big]$
   $\qquad \Delta \leftarrow \max(\Delta, |v - V(s)|)$
   until $\Delta < \theta$ (a small positive number)

3. Policy Improvement
   policy-stable $\leftarrow$ *true*
   For each $s \in \mathcal{S}$:
   $\quad$ *old-action* $\leftarrow \pi(s)$
   $\quad \pi(s) \leftarrow \arg\max_a \sum_{s',r} p(s', r|s, a)\big[r + \gamma V(s')\big]$
   $\quad$ If *old-action* $\neq \pi(s)$, then *policy-stable* $\leftarrow$ *false*
   If *policy-stable*, then stop and return $V \approx v_*$ and $\pi \approx \pi_*$; else go to 2

Initialize array V arbitrarily (e.g., $V(s) = 0$ for all $s \in \mathcal{S}^+$)

Repeat
    $\Delta \leftarrow 0$
    For each $s \in \mathcal{S}$:
        $v \leftarrow V(s)$
        $V(s) \leftarrow \max_a \sum_{s',r} p(s', r|s, a)\big[r + \gamma V(s')\big]$
        $\Delta \leftarrow \max(\Delta, |v - V(s)|)$
until $\Delta < \theta$ (a small positive number)

Output a deterministic policy, $\pi \approx \pi_*$, such that
    $\pi(s) = \arg\max_a \sum_{s',r} p(s', r|s, a)\big[r + \gamma V(s')\big]$

Model-free

Не знаем

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

Монте карло

$$Q(s_t, a_t) \leftarrow \underset{r_t, s_{t+1}}{E} \; r_t + \gamma \cdot max_{a'} Q(s_{t+1}, a')$$

Q - можно приближать итеративно

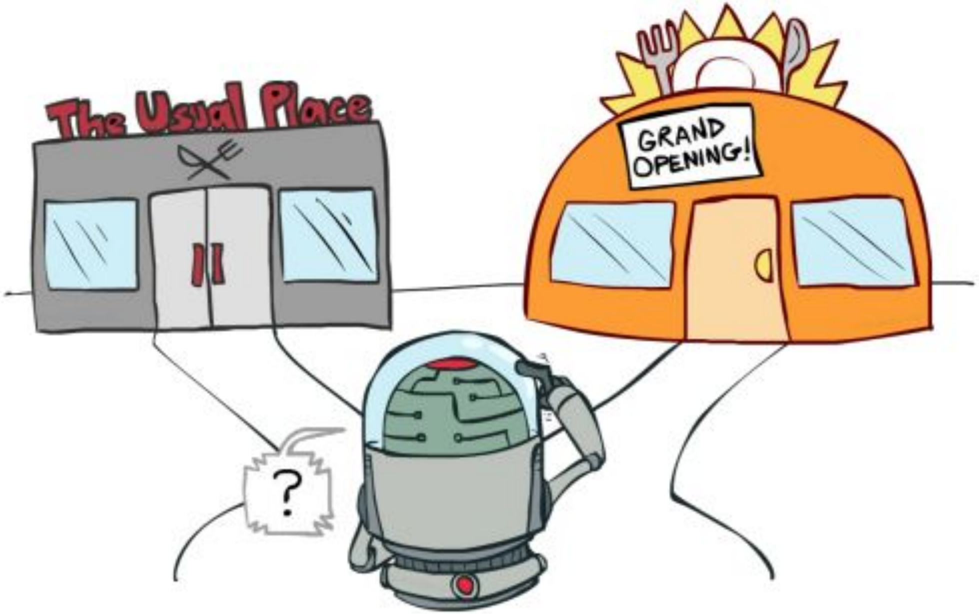
$$Q(s_t, a_t) \leftarrow \underset{r_t, s_{t+1}}{E}\, r_t + \gamma \cdot max_{a'} Q(s_{t+1}, a')$$

$$\underset{r_t, s_{t+1}}{E}\, r_t + \gamma \cdot max_{a'} Q(s_{t+1}, a') \approx \frac{1}{N} \sum_i r_i + \gamma \cdot max_{a'} Q(s_i^{next}, a')$$

$$Q(s_t, a_t) \leftarrow \alpha \cdot \hat{Q}(s_t, a_t) + (1 - \alpha) Q(s_t, a_t)$$

$$\pi(s): argmax_a Q(s, a)$$

Исследование и эксплуатация



Жадно и не очень

Жадно – эпсилон жадно!

$$\pi(a|s) = softmax\left(\frac{Q(s,a)}{\tau}\right)$$

on-policy & off-policy

REINFORCE

Expected reward: R(z) setting

$$J = \underset{\substack{s \sim p(s) \\ a \sim \pi_\theta(s|a) \\ \dots}}{E} R(s, a, s', a', \dots)$$

Expected discounted reward: R(s,a) = r + γ*R(s',a')

$$J = \underset{\substack{s \sim p(s) \\ a \sim \pi_\theta(s|a)}}{E} Q(s, a)$$

"true" Q-function

$$J = \mathop{E}_{\substack{s \sim p(s) \\ a \sim \pi_\theta(s|a)}} Q(s,a) = \int_s p(s) \int_a \pi_\theta(a|s) Q(s,a)\, da\, ds$$

$$J = \mathop{E}_{\substack{s \sim p(s) \\ a \sim \pi_\theta(s|a)}} Q(s,a) = \int_s p(s) \int_a \pi_\theta(a|s) Q(s,a) \, da \, ds$$

True action value

a.k.a. E[R(s,a)]

$$J \approx \frac{1}{N} \sum_{i=0}^{N} \sum_{s,a \in z_i} Q(s,a)$$

sample N sessions

$$\nabla \log \pi(z) = \frac{1}{\pi(z)} \cdot \nabla \pi(z)$$

$$\pi \cdot \nabla \log \pi(z) = \nabla \pi(z)$$

$$\nabla J = \int_s p(s) \int_a \nabla \pi_\theta(a|s) Q(s,a) \, da \, ds$$

$$\pi \cdot \nabla \log \pi(z) = \nabla \pi(z)$$

$$\nabla J = \underset{\substack{s \sim p(s) \\ a \sim \pi_\theta(s|a)}}{E} \nabla \log \pi_\theta(a|s) \cdot Q(s,a)$$