Aproximácia politiky
(Strojové učenie II)

M. Mach

Katedra kybernetiky a umelej inteligencie, FEI, TUKE

február 2021 - marec 2023

Hodnotové funkcie vs politika

- Prístup založený na hodnotových funkciách
  - hodnotové funkcie učené
    - $v(s)$, $\hat{v}(s, \bar{w})$ - pre vyhodnotenie/porovnávanie politík
    - $q(s, a)$, $\hat{q}(s, a, \bar{w})$ - pre učenie politík
  - politika implicitná (ϵ-greedy, greedy)
- Prístup založený na politike
  - hodnotové funkcie neučené
  - politika učená: $\pi(a|s, \bar{\theta}) = P[A_t = a|S_t = s, \bar{\theta}_t = \bar{\theta}]$
  - *policy gradient algoritmy*
- Prístup založený na hodnotových funkciách a politike
  - hodnotové funkcie učené: $v(s)$, $\hat{v}(s, \bar{w})$
    - pre učenie politiky, nie pre výber akcií (preto iba v)
  - politika učená: $\pi(a|s, \bar{\theta}) = P[A_t = a|S_t = s, \bar{\theta}_t = \bar{\theta}]$
  - *actor-critic algoritmy* (actor - politika, critic - v)

Aproximácia politiky

- Parametrizovaná podoba stochastickej (diferencovateľnej) politiky $\pi(a|s,\bar{\theta}) \in (0,1)$
  - pre zabezpečenie explorácie hodnoty 0 a 1 sú vylúčené
- Ak priestor akcií diskrétny a nie veľmi veľký
  - preferencie akcií $h(s,a,\bar{\theta}) \in \mathcal{R}$
    - možná injektáž apriorných znalostí o forme politiky
  - soft-max transformácia na pravdepodobnosti
    - ľubovoľné približovanie deterministickej politike (žiadne ϵ)
    - najväčšej preferencii prislúcha najväčšia pravdepodobnosť

$$\pi(a|s,\bar{\theta}) = \frac{e^{h(s,a,\bar{\theta})}}{\sum_b e^{h(s,b,\bar{\theta})}}$$

  - ľubovoľná parametrizácia preferencií (aproximátor)
    - lineárna kombinácia príznakov

$$h(s,a,\bar{\theta}) = \bar{\theta}^T \bar{x}(s,a)$$

Gradientové učenie parametrov politiky $\bar{\theta}$

- Nech $J(\bar{\theta})$ je skalárna miera výkonu politiky s ohľadom na parametre politiky
- Cieľom je $J(\bar{\theta})$
  - hľadá sa také $\bar{\theta}$, ktoré maximalizuje $J(\bar{\theta})$
- Použitie gradientu
  - akualizačné pravidlo

$$\bar{\theta}_{t+1} = \bar{\theta}_t + \alpha \nabla J(\theta_t)$$

pohyb v smere gradientu

Učenie v epizodickom prostredí

- Meranie výkonu politiky $J(\bar{\theta}) \doteq v_{\pi\theta}(s_0)$ keď s_0 je prvý (štartovací) stav epizódy
  - výkon závisí na parametroch $\bar{\theta}$ pri
    - výbere akcií (závislosť známa)
    - distribúcii stavov, v ktorých sa akcie vyberajú (závislosť zvyčajne neznáma - funkcia prostredia)
- Policy gradient theorem

$$\nabla J(\bar{\theta}) \propto \sum_s \mu(s) \sum_a q_\pi(s,a) \nabla \pi(a|s,\bar{\theta})$$

  - kde $\propto$ je "proporcionálny k"
    - nevadí, vďaka α aj tak robíme iba malý posun v smere
  - nepotrebuje vyjadrovať deriváciu distribúcie stavov

Odvodenie náhrady $\nabla J(\bar{\theta})$ typu MC

$$
\begin{aligned}
\nabla J(\bar{\theta}) &\propto \sum_s \mu(s) \sum_a q_\pi(s,a) \nabla \pi(a|s,\bar{\theta}) \\
&= E_\pi \left[\sum_a q_\pi(S_t,a) \nabla \pi(a|S_t,\bar{\theta}) \right] \\
&= E_\pi \left[\frac{\pi(a|S_t,\bar{\theta})}{\pi(a|S_t,\bar{\theta})} \sum_a q_\pi(S_t,a) \nabla \pi(a|S_t,\bar{\theta}) \right] \\
&= E_\pi \left[\sum_a \pi(a|S_t,\bar{\theta}) q_\pi(S_t,a) \frac{\nabla \pi(a|S_t,\bar{\theta})}{\pi(a|S_t,\bar{\theta})} \right] \\
&= E_\pi \left[E_\pi \left(q_\pi(S_t,a) \frac{\nabla \pi(a|S_t,\bar{\theta})}{\pi(a|S_t,\bar{\theta})} \right) \right] = E_\pi \left[q_\pi(S_t,a) \frac{\nabla \pi(a|S_t,\bar{\theta})}{\pi(a|S_t,\bar{\theta})} \right] \\
&= E_\pi \left[E_\pi[G_t|S_t,A_t] \frac{\nabla \pi(A_t|S_t,\bar{\theta})}{\pi(A_t|S_t,\bar{\theta})} \right] = E_\pi \left[G_t \frac{\nabla \pi(A_t|S_t,\bar{\theta})}{\pi(A_t|S_t,\bar{\theta})} \right]
\end{aligned}
$$

Aktualizácia parametrov $\bar{\theta}$

- Aktualizačné pravidlo

$$\begin{aligned}\bar{\theta}_{t+1} &= \bar{\theta}_t + \alpha \nabla J(\theta_t) \\ &= \bar{\theta}_t + \alpha G_t \frac{\nabla \pi(A_t|S_t, \bar{\theta}_t)}{\pi(A_t|S_t, \bar{\theta}_t)} \\ &= \bar{\theta}_t + \alpha G_t \nabla \ln \pi(A_t|S_t, \bar{\theta}_t)\end{aligned}$$

- Intuitívna interpretácia aktualizačného pravidla
  - závislosť na $\nabla \pi$ - posun v smere najrýchlejšieho rastu pravdepodobnosti výberu A_t pri budúcich návštevách S_t
  - závislosť na G_t - ak úspešná akcia, tak väčšia kumulatívna odmena a posun o väčší usek
  - závislosť na $1/\pi$ - kompenzácia úspešnosti (úspešnejšie akcie sú častejšie vyberané a posun k nim je väčší ako k menej často vyberaným akciám)

Algoritmus REINFORCE

REINFORCE: Monte-Carlo Policy-Gradient Control (episodic) for π_*

Input: a differentiable policy parameterization $\pi(a|s, \boldsymbol{\theta})$
Algorithm parameter: step size $\alpha > 0$
Initialize policy parameter $\boldsymbol{\theta} \in \mathbb{R}^{d'}$ (e.g., to $\mathbf{0}$)

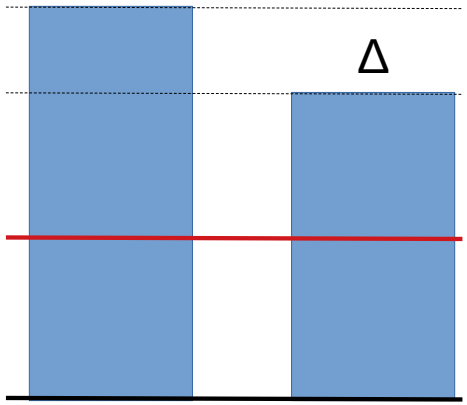
Loop forever (for each episode):
  Generate an episode $S_0, A_0, R_1, \ldots, S_{T-1}, A_{T-1}, R_T$, following $\pi(\cdot|\cdot, \boldsymbol{\theta})$
  Loop for each step of the episode $t = 0, 1, \ldots, T-1$:
    $G \leftarrow \sum_{k=t+1}^{T} \gamma^{k-t-1} R_k$ (G_t)
    $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha \gamma^t G \nabla \ln \pi(A_t|S_t, \boldsymbol{\theta})$

Aktualizácia pre lineárny aproximátor

- Ak soft-max v spojení s lineárnou kombináciou príznakov, tak

$$
\begin{aligned}
\nabla \ln \pi(a|s,\bar{\theta}_t) &= \nabla \ln \frac{e^{\bar{\theta}^T \bar{x}(s,a)}}{\sum_b e^{\bar{\theta}^T \bar{x}(s,b)}} \\
&= \nabla \left(\ln e^{\bar{\theta}^T \bar{x}(s,a)} - \ln \sum_b e^{\bar{\theta}^T \bar{x}(s,b)} \right) \\
&= \bar{x}(s,a) - \frac{1}{\sum_b e^{\bar{\theta}^T \bar{x}(s,b)}} \nabla \sum_b e^{\bar{\theta}^T \bar{x}(s,b)} \\
&= \bar{x}(s,a) - \frac{1}{\sum_b e^{\bar{\theta}^T \bar{x}(s,b)}} \sum_b e^{\bar{\theta}^T \bar{x}(s,b)} \bar{x}(s,b) \\
&= \bar{x}(s,a) - \sum_b \frac{e^{\bar{\theta}^T \bar{x}(s,b)}}{\sum_c e^{\bar{\theta}^T \bar{x}(s,c)}} \bar{x}(s,b) \\
&= \bar{x}(s,a) - \sum_b \pi(b|s,\bar{\theta}) \bar{x}(s,b)
\end{aligned}
$$

Posun základne



- Rozdiel
  - absolútny
  - relatívny
- Posun nezávislý na akcii
- Očakáva sa zníženie variancie

$$
\begin{aligned}
\nabla J(\bar{\theta}) &\propto \sum_s \mu(s) \sum_a (q_\pi(s,a) - b(s)) \nabla \pi(a|s,\bar{\theta}) \\
&= \sum_s \mu(s) \sum_a q_\pi(s,a) \nabla \pi(a|s,\bar{\theta}) \\
&\quad - \sum_s \mu(s) \sum_a b(s) \nabla \pi(a|s,\bar{\theta}) \\
&= \ldots - \sum_s \mu(s) b(s) \sum_a \nabla \pi(a|s,\bar{\theta}) \\
&= \ldots - \sum_s \mu(s) b(s) \nabla \sum_a \pi(a|s,\bar{\theta}) \\
&= \ldots - \sum_s \mu(s) b(s) \nabla 1 \\
&= \sum_s \mu(s) \sum_a q_\pi(s,a) \nabla \pi(a|s,\bar{\theta}) - 0 \\
&= \sum_s \mu(s) \sum_a q_\pi(s,a) \nabla \pi(a|s,\bar{\theta})
\end{aligned}
$$

Algoritmus REINFORCE so základňou

REINFORCE with Baseline (episodic), for estimating $\pi_\theta \approx \pi_*$

Input: a differentiable policy parameterization $\pi(a|s,\boldsymbol{\theta})$
Input: a differentiable state-value function parameterization $\hat{v}(s,\mathbf{w})$
Algorithm parameters: step sizes $\alpha^{\boldsymbol{\theta}} > 0$, $\alpha^{\mathbf{w}} > 0$
Initialize policy parameter $\boldsymbol{\theta} \in \mathbb{R}^{d'}$ and state-value weights $\mathbf{w} \in \mathbb{R}^{d}$ (e.g., to $\mathbf{0}$)

Loop forever (for each episode):
  Generate an episode $S_0, A_0, R_1, \ldots, S_{T-1}, A_{T-1}, R_T$, following $\pi(\cdot|\cdot,\boldsymbol{\theta})$
  Loop for each step of the episode $t = 0, 1, \ldots, T-1$:
    $G \leftarrow \sum_{k=t+1}^{T} \gamma^{k-t-1} R_k$ $\qquad (G_t)$
    $\delta \leftarrow G - \hat{v}(S_t,\mathbf{w})$
    $\mathbf{w} \leftarrow \mathbf{w} + \alpha^{\mathbf{w}} \delta \nabla \hat{v}(S_t,\mathbf{w})$
    $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha^{\boldsymbol{\theta}} \gamma^t \delta \nabla \ln \pi(A_t|S_t,\boldsymbol{\theta})$

©*Sutton-Barto: Reinforcement Learning, 2nd ed., 2018*

Pridanie kritika

- REINFORCE je v podstate MC metóda
- MC $\rightarrow$ TD pre zrýchlenie konvergencie

$$
\begin{aligned}
\bar{\theta}_{t+1} &= \bar{\theta}_t + \alpha \nabla J(\theta_t) \\
&= \bar{\theta}_t + \alpha (G_t - \hat{v}(S_t, \bar{w})) \frac{\nabla \pi(A_t|S_t, \bar{\theta}_t)}{\pi(A_t|S_t, \bar{\theta}_t)} \\
&= \bar{\theta}_t + \alpha (R_{t+1} + \gamma \hat{v}(S_{t+1}, \bar{w}) - \hat{v}(S_t, \bar{w})) \frac{\nabla \pi(A_t|S_t, \bar{\theta}_t)}{\pi(A_t|S_t, \bar{\theta}_t)} \\
&= \bar{\theta}_t + \alpha \delta_t \frac{\nabla \pi(A_t|S_t, \bar{\theta}_t)}{\pi(A_t|S_t, \bar{\theta}_t)}
\end{aligned}
$$

Algoritmus AC

One-step Actor–Critic (episodic), for estimating $\pi_{\boldsymbol{\theta}} \approx \pi_*$

Input: a differentiable policy parameterization $\pi(a|s,\boldsymbol{\theta})$
Input: a differentiable state-value function parameterization $\hat{v}(s,\mathbf{w})$
Parameters: step sizes $\alpha^{\boldsymbol{\theta}} > 0$, $\alpha^{\mathbf{w}} > 0$
Initialize policy parameter $\boldsymbol{\theta} \in \mathbb{R}^{d'}$ and state-value weights $\mathbf{w} \in \mathbb{R}^{d}$ (e.g., to $\mathbf{0}$)
Loop forever (for each episode):
  Initialize S (first state of episode)
  $I \leftarrow 1$
  Loop while S is not terminal (for each time step):
    $A \sim \pi(\cdot|S,\boldsymbol{\theta})$
    Take action A, observe S', R
    $\delta \leftarrow R + \gamma\,\hat{v}(S',\mathbf{w}) - \hat{v}(S,\mathbf{w})$ (if S' is terminal, then $\hat{v}(S',\mathbf{w}) \doteq 0$)
    $\mathbf{w} \leftarrow \mathbf{w} + \alpha^{\mathbf{w}}\,\delta\,\nabla\hat{v}(S,\mathbf{w})$
    $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha^{\boldsymbol{\theta}}\,I\,\delta\,\nabla \ln \pi(A|S,\boldsymbol{\theta})$
    $I \leftarrow \gamma I$
    $S \leftarrow S'$

Učenie v kontinuálnom prostredí

- Meranie výkonu politiky $J(\bar{\theta})$ priemernou odmenou

$$\begin{aligned} J(\bar{\theta}) = r(\pi) &= \lim_{h\to\infty} \frac{1}{h} \sum_{t=1}^{h} E[R_t | S_0, A_{0:t-1} \sim \pi] \\ &= \sum_s \mu_\pi(s) \sum_a \pi(a|s) \sum_r r \sum_{s'} p(s', r|s, a) \end{aligned}$$

kde $\mu_\pi(s)$ je ustálená distribúcia (ergodický MDP)

$$\begin{aligned} \mu_\pi(s) &= \lim_{t\to\infty} P[S_s = s | A_{0:t-1} \sim \pi] \\ \mu_\pi(s') &= \sum_s \mu_\pi(s) \sum_a \pi(a|s, \bar{\theta}) p(s'|s, a) \end{aligned}$$

Kritik pre kontinuálne úlohy

- Diferenčná odmena

$$G_t = R_{t+1} - r(\pi) + R_{t+2} - r(\pi) + \dots$$

- Diferenčná hodnotová funkcia

$$v_\pi(s) = E_\pi[G_t | S_t = s]$$

- Diferenčná forma TD chyby

$$\delta_t = \left(R_{t+1} - \overline{R_t} + \hat{v}(S_{t+1}, \bar{w}_t)\right) - \hat{v}(S_t, \bar{w}_t)$$

kde $\overline{R_t}$ je odhad $r(\pi)$ v čase t

- Policy gradient theorem ostáva naďalej platný aj pre kontinuálny prípad

Algoritmus AC (kontinuálny)

- *Zmeny voči epizodickému algoritmu AC:*

Parameters: step sizes $\alpha^{\theta} > 0$, $\alpha^{\mathbf{w}} > 0$, $\alpha^{\bar{R}} > 0$

Initialize $\bar{R} \in \mathcal{R}$ (napr. na 0)

Loop forever (for each episode):

Initialize S (first state of episode)

$I \leftarrow 1$

Loop while S is not terminal forever (for each time step)

$\delta \leftarrow R - \bar{R} + \gamma \hat{v}(S', \mathbf{w}) - \hat{v}(S, \mathbf{w})$

$\bar{R} \leftarrow \bar{R} + \alpha^{\bar{R}} \delta$

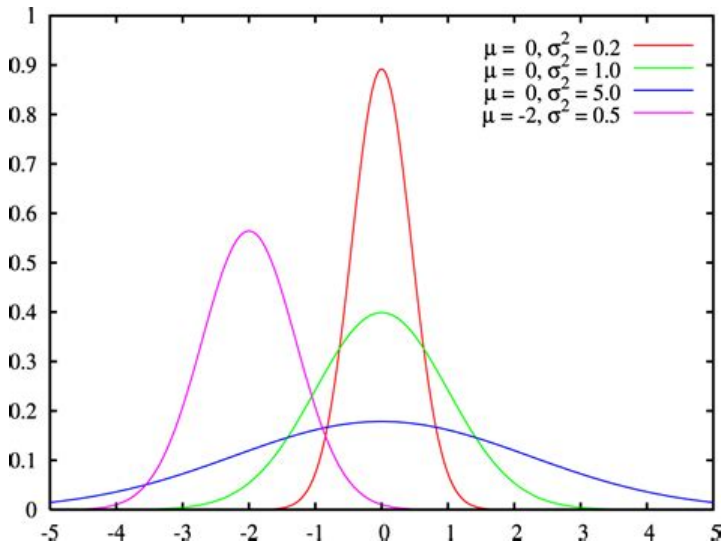
$\theta \leftarrow \theta + \alpha^{\theta} I \delta \nabla ln \pi(A, S, \theta)$

$I \leftarrow \gamma I$

Spojité akcie

- Akcie sú reálne skaláry
- Parametrizácia politiky
  - spojitá distribúcia pravdepodobnosti pre každý stav
  - distribúcia daná parametrickou hustotou pravdepodobnosti (napr. $N(\mu, \sigma)$)

    $\pi(a|s, \bar{\theta}) = \frac{1}{\sigma(s,\bar{\theta})\sqrt{2\pi}} \exp\left(-\frac{(a-\mu(s,\bar{\theta}))^2}{2\sigma(s,\bar{\theta})^2}\right)$
  - učia sa parametre hustoty pravdepodobnosti



- $\bar{\theta} = [\bar{\theta_\mu}, \bar{\theta_\sigma}]^T$
- Lineárne aproximátory
  - $\mu(s, \bar{\theta}) = (\bar{\theta_\mu})^T \bar{x}_\mu(s)$
  - $\sigma(s, \bar{\theta}) = \exp\left((\bar{\theta_\sigma})^T \bar{x}_\sigma(s)\right)$ (odchýlka musí byť kladná)

Výhody metód aproximujúcich politiku

- Výhody
  - lepšie konvergenčné vlastnosti
  - efektívne pre mnohorozmerný alebo spojitý priestor akcií
  - vedia učiť stochastické politiky s vhodnou úrovňou explorácie, blížiace sa deterministickým politikám
  - pre niektoré problémy je jednoduchšie parametricky reprezentovať politiku než hodnotové funkcie
- Nevýhody
  - typicky konvergujú k lokálnemu a nie globálnemu optimu
  - vyhodnotenie politiky je typicky neefektívne