Monte Carlo metódy
(Strojové učenie II)

M. Mach

Katedra kybernetiky a umelej inteligencie, FEI, TUKE

február 2021 - február 2025

Monte Carlo učenie posilňovaním

- Prístup založený na odhade hodnotových funkcií
- Model-free - nepotrebuje úplnú znalosť dynamiky prostredia, stačí skúsenosť s prostredím
  - skúsenosť = sekvencia stavov, akcií a odmien
- Interakcia s prostredím
  - skutočná
  - simulovaná - postačuje obmezený generatívny model (schopný generovať odmeny a prechody medzi stavmi)
- Obmedzenie na epizodické úlohy
- Inkrementálny v zmysle epizóda po epizóde (možnosť ale nie povinnosť)

Monte Carlo odhad $v_\pi(s)$

- Získaná skúsenosť vo forme epizodických sekvencií

$$S_0, A_0, R_1, S_1, A_1, R_2, S_2, A_2, \ldots, S_{T-1}, A_{T-1}, R_T, S_T$$

  - cieľom je získať odhad $v_\pi(s) = E_\pi[G_t | S_t = s]$
    - odhad pre stav v ktorom sekvencia začína
    - odhad pre všetky stavy v sekvencii (lepšie využitie dát)
- Založené na spriemerňovaní nezávislých kumulatívnych odmien (z nezávislých epizód)
  - očakávaná kumulatívna odmena je nahradená priemernou kumulatívnou hodnotou
  - so zvyšujúcim sa počtom vzoriek bude priemer konvergovať k očakávanej hodnote
- Viacnásobný výskyt kumulatívnej hodnoty v epizóde
  - prvá návšteva
  - každá návšteva

Výber akcií

- Hodnotová funkcia stavu sa dá použiť pre výber akcie (pre najlepšiu kombináciu R_{t+1} a S_{t+1})
  - odvodiť $q_\pi(s, a)$ - hľadanie do hĺbky 1

$$
\begin{aligned}
q_\pi(s, a) &= E_\pi[R_{t+1} + \gamma G_{t+1} | S_t = s, A_t = a] \\
&= E_\pi[R_{t+1} | S_t = s, A_t = a] + \\
&\quad \gamma E_\pi[G_{t+1} | S_t = s, A_t = a] \\
&= \sum_{r \in R} p(r|s, a) * r + \gamma \sum_{s' \in S} p(s'|s, a) * v_\pi(s')
\end{aligned}
$$

  - vybrať maximalizujúcu akciu
- Nemáme model dynamiky prostredia $p(s', r|s, a)$
  $\rightarrow$ pre výber akcií je nutné odhadovať $q_\pi(s, a)$ namiesto $v_\pi(s)$

Algoritmus First-visit MC odhadu q_π

Input: a policy π to be evaluated
Initialize:
$q(s, a) \in \mathcal{R}$, arbitrarily, for all $s \in S$, $a \in A$
$Returns(s, a) \leftarrow$ an empty list, for all $s \in S$, $a \in A$

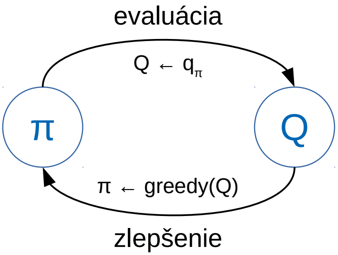
Loop forever (for each episode):
Generate an episode following π:
$S_0, A_0, R_1, S_1, A_1, R_2, \ldots, S_{T-1}, A_{T-1}, R_T$
$G \leftarrow 0$
Loop for each step of episode, $t = T-1, T-2, \ldots, 0$
$G \leftarrow \gamma G + R_{t+1}$
Unless S_t, A_t appears in $S_0, A_0, S_1, A_1, \ldots, S_{t-1}, A_{t-1}$:
Append G to $Returns(S_t, A_t)$
$q(S_t, A_t) \leftarrow average(Returns(S_t, A_t))$

Aproximácia optimálnej politiky

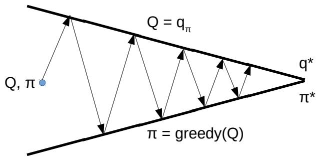
$$\pi_0 \xrightarrow{E} q_{\pi_0} \xrightarrow{Z} \pi_1 \xrightarrow{E} q_{\pi_1} \xrightarrow{Z} \pi_2 \xrightarrow{E} \dots \xrightarrow{Z} \pi_* \xrightarrow{E} q_*$$

- Zlepšenie politiky: $\pi(s) = \arg\max_a q(s,a)$

$$q_{\pi_k}(s, \pi_{k+1}(s)) = q_{\pi_k}(s, \arg\max_a q_{\pi_k}(s,a)) = \max_a q_{\pi_k}(s,a) \geq q_{\pi_k}(s,a)$$



Aproximácie - konvergencia k optimálnym hodnotám



Predpoklady garancie konvergencie

- Počet epizód pre vyhodnocovanie politiky
  - hodnoty sa iba asymptoticky blížia skutočným hodnotám
  - dostatočná aproximácia skutočných hodnôt
    - konvergencia až po nejaký stupeň aproximácie
    - potrebných mnoho sekvencií (epizód)
    - použiteľné iba pre malé úlohy
  - pohyb smerom k skutočným hodnotám
    - rezignácia na úplné vyhodnotenie politiky pred jej zlepšením
    - vyhodnotenie politiky obmedzené na jednu epizódu
    - zlepšenie iba pre tie dvojice (s, a), ktoré sa v epizóde vyskytli
- Preskúmanie všetkých párov (s, a)
  - aby bolo možné zohľadniť príspevok voľby každej akcie
  - aby bolo možné porovnávať alternatívy

On-policy a off-policy vyhodnocovanie a učenie

- Dva základné prístupy
  - on-policy metódy sú zamerané na politiku použitú pri tvorbe sekvencií $\rightarrow$ jedna politika
  - off-policy metódy sú zamerané na politiku, ktorá je rozdielna od politiky použitej pri tvorbe sekvencií $\rightarrow$ dve politiky
- Oblasť použitia
  - vyhodnocovanie danej politiky
  - iteračné vylepšovanie danej politiky

Použitie politiky π

- Problém s použitou politikou π
  - niektoré kombinácie stav-akcia sa nebudú v sekvenciách vyskytovať vôbec (žiadny odhad) alebo iba príliš zriedkavo (nespoľahlivý odhad)
  - extrém - deterministická politika v každom stave vyberie iba 1 akciu
  - problém so zachovávaním explorácie
- Riešenie
  - exploračné štarty - iba obmedzená použiteľnosť
    - epizóda štartuje v zadanej kombinácii stav-akcia, pokračuje už podľa politiky
    - pravdepodobnosť páru stav-akcia začínať sekvenciu je nenulová
  - stochastické politiky
    - $p(a|s) > 0$ pre všetky stavy a k nim príslušné akcie

Mäkké politiky

- Mäkká politika
  - $\pi(a|s) > 0$ pre všetky $s \in S$ a $a \in A(s)$
  - môže byť posúvaná stále viac k deterministickej optimálnej politike
  - ϵ-mäkká politika $\pi(a|s) \geq \frac{\epsilon}{|A(s)|}$
- ϵ-greedy politika
  - je príkladom ϵ-mäkkej politiky
  - výber akcie
    - väčšinu času sa vyberá akcia maximalizujúca hodnotovú funkciu akcie (pravdepodobnosť $1 - \epsilon$)
    - občas sa zvolí náhodne vybratá akcia (pravdepodobnosť ϵ)
  - realizovaná ako náhodný výber podľa pravdepodobností
    $$\pi(a|s) = \begin{cases} 1 - \epsilon + \frac{\epsilon}{|A(s)|} & a = \arg\max_a q(s, a) \\ \frac{\epsilon}{|A(s)|} & inak \end{cases}$$
  - ak viac maximalizujúcich akcií - vysoká pravdepodobnosť sa pridelí (náhodne) iba jednej z nich

Algoritmus MC on-policy odhadu π_*

On-policy first-visit MC control (for ε-soft policies), estimates $\pi \approx \pi_*$

Algorithm parameter: small $\varepsilon > 0$

Initialize:

$\pi \leftarrow$ an arbitrary ε-soft policy

$Q(s,a) \in \mathbb{R}$ (arbitrarily), for all $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$

$Returns(s,a) \leftarrow$ empty list, for all $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$

Repeat forever (for each episode):

Generate an episode following π: $S_0, A_0, R_1, \ldots, S_{T-1}, A_{T-1}, R_T$

$G \leftarrow 0$

Loop for each step of episode, $t = T-1, T-2, \ldots, 0$:

$G \leftarrow \gamma G + R_{t+1}$

Unless the pair S_t, A_t appears in $S_0, A_0, S_1, A_1 \ldots, S_{t-1}, A_{t-1}$:

Append G to $Returns(S_t, A_t)$

$Q(S_t, A_t) \leftarrow \text{average}(Returns(S_t, A_t))$

$A^* \leftarrow \arg\max_a Q(S_t, a)$ (with ties broken arbitrarily)

For all $a \in \mathcal{A}(S_t)$:

$$\pi(a|S_t) \leftarrow \begin{cases} 1 - \varepsilon + \varepsilon/|\mathcal{A}(S_t)| & \text{if } a = A^* \\ \varepsilon/|\mathcal{A}(S_t)| & \text{if } a \neq A^* \end{cases}$$

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

Konvergencia pri použití ϵ-greedy politiky

- Ak π' je učená ϵ-greedy politika a π je ϵ-mäkká politika, tak $\pi' \geq \pi$. Dôkaz podľa TZP:

$$
\begin{aligned}
v_{\pi'}(s) &= \sum_a \pi'(a|s) q_\pi(s,a) \\
&= \frac{\epsilon}{\mid A(s) \mid} \sum_a q_\pi(s,a) + (1-\epsilon) \max_a q_\pi(s,a) \\
&\geq \frac{\epsilon}{\mid A(s) \mid} \sum_a q_\pi(s,a) + (1-\epsilon) \sum_a \frac{\pi(a|s) - \frac{\epsilon}{|A(s)|}}{1-\epsilon} q_\pi(s,a) \\
&= \frac{\epsilon}{\mid A(s) \mid} \sum_a q_\pi(s,a) - \frac{\epsilon}{\mid A(s) \mid} \sum_a q_\pi(s,a) \\
&\quad + \sum_a \pi(a|s) q_\pi(s,a) \\
&= v_\pi(s)
\end{aligned}
$$

- Dá sa získať najlepšia spomedzi ϵ-mäkkých politík

Off-policy prístup

- On-policy dilema - učiť optimálnu politiku avšak nutnosť použiť neoptimálnu exploračnú politiku
  - on-policy prístup naučí iba near-optimal politiku
- Použitie dvoch politík pre riešenie dilemy
  - cieľová politika π (typicky deterministická)
  - exploračná (behaviour) politika b (môže byť ϵ-mäkká)
- Pokrytie politík: $\pi(a|s) > 0 \rightarrow b(a|s) > 0$
  - ak výber $\pi(s) \neq b(s)$ tak b musí byť v s stochastická
- Výhody off-policy prístupu
  - všeobecnejší prístup (zahŕňa on-policy)
  - širšie použitie (učenie z pozorovania iných, znovupoužitie skúseností zo starších politík, vyhodnocovanie rôznych politík na rovnakých dátach)
- Nevýhody off-policy prístupu
  - pomalšia konvergencia

Vzorkovanie podľa dôležitosti

$$S_t, A_t, S_{t+1}, A_{t+1}, S_{t+2}, \ldots, S_{T-1}, A_{T-1}, S_T$$

- Pravdepodobnosť výskytu sekvencie pri politike π
  $$P[A_t, S_{t+1}, A_{t+1}, \ldots, S_T | S_t]$$
  $$= \pi(A_t|S_t)p(S_{t+1}|S_t, A_t)\pi(A_{t+1}|S_{t+1}) \ldots p(S_T|S_{T-1}, A_{T-1})$$
  $$= \prod_{k=t}^{T-1} \pi(A_k|S_k)p(S_{k+1}|S_k, A_k)$$
- Pravdepodobnosť výskytu sekvencie pri politike b
  $$P[A_t, S_{t+1}, A_{t+1}, \ldots, S_T | S_t] = \prod_{k=t}^{T-1} b(A_k|S_k)p(S_{k+1}|S_k, A_k)$$
- Pomer pravdepodobností
  $$W_t = \frac{\prod_{k=t}^{T-1} \pi(A_k|S_k)p(S_{k+1}|S_k,A_k)}{\prod_{k=t}^{T-1} b(A_k|S_k)p(S_{k+1}|S_k,A_k)} = \frac{\prod_{k=t}^{T-1} \pi(A_k|S_k)}{\prod_{k=t}^{T-1} b(A_k|S_k)} = \prod_{k=t}^{T-1} \frac{\pi(A_k|S_k)}{b(A_k|S_k)}$$
- Určenie $q(s, a)$ podľa b a π zo sekvencie generovanej podľa b
  $$q_b(s, a) = E_b[G_t|S_t = s, A_t = a]$$
  $$q_\pi(s, a) = E_\pi[W_t * G_t|S_t = s, A_t = a]$$

Off-policy odhad π_*

- Použité politiky
  - cieľová: deterministická $\pi(s) \leftarrow \arg\max_a q(s, a)$
  - exploračná: stochastická ϵ-mäkká
- Relatívna pravdepodobnosť vykonania kroku v sekvencii $\frac{\pi(A_t|S_t)}{b(A_t|S_t)}$ môže byť
  - 0 (lebo $\pi(A_t|S_t) = 0$) - exploračná politika zvolila iný ako maximalizujúci krok
  - $\frac{1}{b(A_t|S_t)}$ (lebo $\pi(A_t|S_t) = 1$) - exploračná politika zvolila maximalizujúci krok
- Nevýhody
  - algoritmus sa učí iba z koncových častí sekvencií epizód (učí sa od konca iba do posledného non-greedy výberu) $\rightarrow$ preferencia všetkých návštev voči prvej návšteve
  - pomalé učenie pri dlhých epizódach (najmä pre páry na začiatku epizód)

Algoritmus MC off-policy odhadu π_*

Initialize, for all $s \in S$, $a \in A$:

$q(s, a) \in \mathcal{R}$, arbitrarily,

$Returns(s, a) \leftarrow$ an empty list

$\pi(s) \leftarrow argmax_a q(s, a)$

Loop forever (for each episode):

Generate an episode following b (any soft policy)

$S_0, A_0, R_1, S_1, A_1, R_2, \ldots, S_{T-1}, A_{T-1}, R_T$

$G \leftarrow 0,\ W \leftarrow 1$

Loop for each step of episode, $t = T-1, T-2, \ldots, 0$

$G \leftarrow \gamma G + R_{t+1}$

Append (W, G) to $Returns(S_t, A_t)$

$q(S_t, A_t) \leftarrow average(Returns(S_t, A_t))$

$\pi(S_t) \leftarrow argmax_a q(S_t, a)$

If $A_t \neq \pi(S_t)$ then exit inner Loop (proceed to next episode)

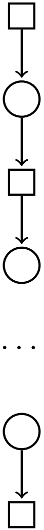
$W \leftarrow W * \frac{1}{b(A_t|S_t)}$

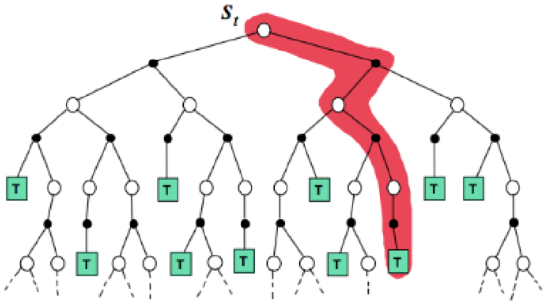
Typy vzorkovania podľa dôležitosti

$$S_{11}, \ldots, s_{t1}, a_{t1}, \ldots, S_{T1} \quad \ldots \quad S_{1n}, \ldots, s_{tn}, a_{tn}, \ldots, S_{Tn}$$

- $\mathcal{T}(s,a)$ - množina uvažovaných výskytov páru (s,a) v množine sekvencií
- Obyčajné vzorkovanie podľa dôležitosti
$$q(s,a) = \frac{\sum_{i \in \mathcal{T}(s,a)} W_i * G_i}{|\mathcal{T}(s,a)|}$$
  - bez odchýlky: $|\,\mathcal{T}(s,a)\,| = 1 \rightarrow q(s,a) = q_\pi(s,a)$
  - variancia nie je ohraničená
- Vážené vzorkovanie podľa dôležitosti
$$q(s,a) = \frac{\sum_{i \in \mathcal{T}(s,a)} W_i * G_i}{\sum_{i \in \mathcal{T}(s)} W_i}$$
  - odchýlka: $|\,\mathcal{T}(s,a)\,| = 1 \rightarrow q(s,a) = G_1 = q_b(s,a)$
    - odchýlka konverguje asymptoticky k nule
  - ohraničená variancia

Backup diagram





©/*lilianweng.github.io*

Inkrementálne určovanie priemeru

- Náhrada opakovaného spriemerňovania inkrementálnym updatom (pre vážené vzorkovanie)

$$
\begin{aligned}
q_n &= \frac{W_1 G_1 + \ldots + W_n G_n}{W_1 + \ldots + W_n} \\
&= \frac{\frac{W_1 G_1 + \ldots + W_{n-1} G_{n-1}}{W_1 + \ldots + W_{n-1}} (W_1 + \ldots + W_{n-1}) + W_n G_n}{W_1 + \ldots + W_n} \\
&= \frac{q_{n-1}(W_1 + \ldots + W_{n-1}) + W_n q_{n-1} - W_n q_{n-1} + W_n G_n}{W_1 + \ldots + W_n} \\
&= q_{n-1} + \frac{W_n (G_n - q_{n-1})}{W_1 + \ldots + W_n} \\
&= q_{n-1} + \frac{W_n}{C_n} (G_n - q_{n-1}) \\
C_n &= C_{n-1} + W_n
\end{aligned}
$$

- Inicializácia: $C_0 = 0$, $q_0 =$ ľubovoľná hodnota
- Poradie updatu: $C_n = \ldots$, $q_n = \ldots$

Algoritmus MC off-policy odhadu π_*

Off-policy MC control, for estimating $\pi \approx \pi_*$

Initialize, for all $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$:
- $Q(s,a) \in \mathbb{R}$ (arbitrarily)
- $C(s,a) \leftarrow 0$
- $\pi(s) \leftarrow \arg\max_a Q(s,a)$ (with ties broken consistently)

Loop forever (for each episode):
- $b \leftarrow$ any soft policy
- Generate an episode using b: $S_0, A_0, R_1, \ldots, S_{T-1}, A_{T-1}, R_T$
- $G \leftarrow 0$
- $W \leftarrow 1$
- Loop for each step of episode, $t = T-1, T-2, \ldots, 0$:
  - $G \leftarrow \gamma G + R_{t+1}$
  - $C(S_t, A_t) \leftarrow C(S_t, A_t) + W$
  - $Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \frac{W}{C(S_t, A_t)} [G - Q(S_t, A_t)]$
  - $\pi(S_t) \leftarrow \arg\max_a Q(S_t, a)$ (with ties broken consistently)
  - If $A_t \neq \pi(S_t)$ then exit inner Loop (proceed to next episode)
  - $W \leftarrow W \frac{1}{b(A_t|S_t)}$

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