Megerősítéses tanulás

segédlet a kurzushoz

February 16, 2012

1 Markov döntési folyamatok és a megerősítéses tanulás elemei

Környezet alatt egy rendezett $(\mathcal{S}, \mathcal{A}, \mathbf{P}, \mathbf{R})$ négyest értünk, ahol

- $\mathcal{S}$ tetszőleges halmaz, melyet *állapothalmaz*-nak, elemeit pedig *állapot*-oknak hívjuk.
- $\mathcal{A}$ tetszőleges halmaz, melyet *cselekvéshalmaz*-nak, elemeit pedig *cselekvés*-eknek hívjuk.
- $\mathbf{P} = \{\mathbf{P}(s,a)\}_{s\in\mathcal{S},a\in\mathcal{A}}$, ahol egy $s \in \mathcal{S}$ és $a \in \mathcal{A}$ esetén $\mathbf{P}(s,a)$ egy eloszlás az $\mathcal{S}$ halmaz felett. Diszkrét esetben használjuk a $\mathbf{P}(s,a,s') = \mathbb{P}[s_1(s,a) = s']$, úgynevezett *állapotátmenet valószínűség* értékeket is, ahol $s_1 \sim \mathbf{P}(s,a)$.
- $\mathbf{R} = \{\mathbf{R}(s,a,s')\}_{s,s'\in\mathcal{S},a\in\mathcal{A}}$, ahol egy $s, s' \in \mathcal{S}$ és $a \in \mathcal{A}$ esetén $\mathbf{R}(s,a,s')$ egy eloszlás az $\mathbb{R}$ halmaz felett; egy ilyen eloszlású változót nevezünk *jutalom*-nak.

Adott $\mathcal{S}$ állapothalmaz és $\mathcal{A}$ cselekvéshalmaz esetén egy $\pi : \mathcal{S} \to \mathcal{A}$ (esetleg véletlen) leképezést *stratégiá*-nak vagy *eljárásmód*-nak hívunk. Diszkrét esetben használjuk a $\pi(s,a) = \mathbb{P}[\pi(s) = a]$ jelölést, $s \in \mathcal{S}$, $a \in \mathcal{A}$.

Egy $(\mathcal{S}, \mathcal{A}, \mathbf{P}, \mathbf{R})$ környezet és egy $\pi : \mathcal{S} \to \mathcal{A}$ stratégia egy s_0 kezdőállapot esetén meghatároz egy $\{(s_t, a_t, r_t)\}_{t=0}^{T}$ úgynevezett *Markov döntési folyamat*-ot (amire az angol elnevezés, a Markov Decision Process alapján röviden *MDP*-ként is hivatkozunk), ahol

- $a_t \sim \pi(s_t)$,
- $s_{t+1} \sim \mathbf{P}_{s_t,a_t}$,
- $r_t \sim \mathbf{R}_{s_t,a_t,s_{t+1}}$,
- és teljesül a *Markov tulajdonság*, azaz hogy bármely $\sigma_0, \sigma_1, \ldots, \sigma_{t+1} \in \mathcal{S}$, $\alpha_0, \alpha_1, \ldots, \alpha_t \in \mathcal{A}$, $\rho_0, \rho_1, \ldots, \rho_t \in \mathbb{R}$ esetén

$$\mathbb{P}\left[s_{t+1} = \sigma_{t+1}, r_t = \rho_t \,\middle|\, s_0 = \sigma_0, a_0 = \alpha_0, r_0 = \rho_0, s_1 = \sigma_1, a_1 = \alpha_1, r_1 = \rho_1, \ldots, s_t = \sigma_t, a_t = \alpha_t\right] =$$
$$= \mathbb{P}\left[s_{t+1} = \sigma_{t+1}, r_t = \rho_t \,\middle|\, s_t = \sigma_t, a_t = \alpha_t\right] ,$$

$t = 0, 1, \ldots, T$. Ezen folyamat segítségével meghatározható az adott π stratégia mellett egy $s \in \mathcal{S}$ állapot $\gamma \in (0,1)$ diszkont-rátával leszámított összjutalmának V^π várható értéke, $V^\pi(s) = \mathbb{E}\left[r_0 + \gamma \cdot \sum_{t=0}^{\infty} \gamma^t \cdot r_{t+1} \,\middle|\, s_0 = s\right]$. Ez a V^π az *értékfüggvény*. Hasonlóképpen definiálható a Q^π *cselekvés-érték függvény*, ahol $Q^\pi(s,a) = \mathbb{E}\left[r_0 + \gamma \cdot \sum_{t=0}^{\infty} \gamma^t \cdot r_{t+1} \,\middle|\, s_0 = s, a_0 = a\right]$, $s \in \mathcal{S}$, $a \in \mathcal{A}$.

2 Bellmann-egyenlet

Egy $(\mathcal{S}, \mathcal{A}, \mathbf{P}, \mathbf{R})$ környezet és egy $\pi : \mathcal{S} \to \mathcal{A}$ stratégia meghatározta $\{(s_t, a_t, r_t)\}_{t=0}^{T}$ MDP tetszőleges $s \in \mathcal{S}$ esetén teljesíti az úgynevezett *Bellmann-egyenlet*-et:

$$V^\pi(s) = \mathbb{E}\left[r_0 + \gamma \cdot \sum_{t=0}^{\infty} \gamma^t \cdot r_{t+1} \,\middle|\, s_0 = s \right] = \qquad \text{(BE-MC)}$$

$$= \mathbb{E}\left[r_0 + \gamma \cdot V^\pi(s_1) \middle| s_0 = s \right] \ . \qquad \text{(BE-TD)}$$

Diszkrét esetben a (BE-TD) egyenlőséget kifejtve a következő összefüggéshez jutunk:

$$V^\pi(s) = \sum_{a \in \mathcal{A}} \pi(s,a) \sum_{s' \in \mathcal{S}} \mathbf{P}(s,a,s') \cdot \Big(\mathbb{E}\left[r_0 \mid s_0 = s, a_0 = a, s_1 = s' \right] + \gamma \cdot V^\pi(s') \Big) \qquad \text{(BE-DP)}$$

Jelölje $B_{\mathcal{S}} = \{V \in \mathbb{R}^{\mathcal{S}} : \|V\|_\infty < \infty\}$ az $\mathcal{S}$ fölött értelmezett korlátos függvények halmazát. Legyen, ahogy fönt is, valamely $s_0 \in \mathcal{S}$ esetén $a_0 \sim \pi(s_0)$ és $s_1 \sim \mathbf{P}_{s_0,a_0}$. Definiáljuk segítségükkel a $T^\pi : B_{\mathcal{S}} \to B_{\mathcal{S}}$, $V \mapsto T^\pi V$, leképezést, ahol tetszőleges $s \in \mathcal{S}$ állapotra

$$(T^\pi V)(s) = \mathbb{E}\left[r_0 + \gamma \cdot V(s_1) \mid s_0 = s \right] \ .$$

T^π használatával a (BE-TD) egyenlőséget a következőképpen fogalmazhatjuk át.

tétel 2.1. *V^π fixpontja T^π-nek, azaz $T^\pi V^\pi = V^\pi$.*

T^π egy másik fontos tulajdonsága a következő.

tétel 2.2. *T γ-kontrakció, azaz tetszőleges V és $\hat{V}$ esetén $\|T^\pi V - T^\pi \hat{V}\|_\infty \leq \gamma \cdot \|V - \hat{V}\|_\infty$.*

Bizonyítás: Az állítás egyszerűen következik abból, hogy a várhatóérték linearitása miatt bármely $s \in \mathcal{S}$ állapotra

$$\begin{aligned}(T^\pi V)(s) - (T^\pi \hat{V})(s) &= \mathbb{E}\left[r_0 + \gamma \cdot V(s_1) \middle| s_0 = s \right] - \mathbb{E}\left[r_0 + \gamma \cdot \hat{V}(s_1) \middle| s_0 = s \right] = \\ &= \gamma \cdot \mathbb{E}\left[V(s_1) - \hat{V}(s_1) \middle| s_0 = s \right] = \\ &\leq \gamma \cdot \|V - \hat{V}\|_\infty \ .\end{aligned}$$

□

Ez utóbbi eredményt kombinálva az előzővel és a Banach-féle fixpont-tétellel kapjuk:

következmény 2.3. *T^π-nek pontosan egy fixpontja létezik, és az V^π. Ezen felül tetszőleges $V_0 = V \in B_{\mathcal{S}}$ és $V_{t+1} = T^\pi V_t$, $t = 0, 1, \ldots$ esetén $\|V_k - V^\pi\|_\infty = O(\gamma^k)$.*

A föntiekre épül a (BE-DP) egyenlőségből kiinduló *iterált stratégia-kiértékelés* algoritmus ([1], 4.1 szekció, 92. o.). Kapcsolódó példák: [1] 3.8 példa (3.7 szekció, 71. o.) és 4.1 példa (4.1 szekció, 92.o.).

3 Bellmann-féle optimális operátor

Legyen $(\mathcal{S}, \mathcal{A}, \mathbf{P}, \mathbf{R})$ egy környezet, legyen valamely $s_0 \in \mathcal{S}$ és $a_0 \in \mathcal{A}$ esetén $s_1 = s_1(s_0, a_0) \sim \mathbf{P}_{s_0,a_0}$. és legyen $T^* : B_{\mathcal{S}} \to B_{\mathcal{S}}$, $V \mapsto T^*V$, ahol tetszőleges $s \in \mathcal{S}$ állapotra

$$(T^*V)(s) = \max_{a\in\mathcal{A}} \mathbb{E}\left[r_0 + \gamma \cdot V(s_1) \middle| s_0 = s, a_0 = a\right] =$$
$$= \max_{a\in\mathcal{A}} \sum_{s'\in\mathcal{S}} \mathbf{P}(s,a,s') \cdot \left(\mathbb{E}\left[r_0 \middle| s_0 = s, a_0 = a, s_1 = s'\right] + \gamma \cdot V(s') \right) .$$

megjegyzés 3.1. *Tetszőleges $V : \mathcal{S} \to \mathbb{R}$ értékfüggveny esetén egy π stratégia mohó stratégia a V-re vonatkozóan, ha $\pi(s) = \operatorname{argmax}_{a\in\mathcal{A}} \mathbb{E}\left[r_0 + \gamma \cdot V(s_1) \mid s_0 = s, a_0 = a\right]$. Ekkor $T^\pi V = T^*V$.*

tétel 3.2. *T^* γ-kontrakció*

Bizonyítás: Az állítás könnyen adódik abból, hogy tetszőleges $V, V' \in B_{\mathcal{S}}$ és $s \in \mathcal{S}$ esetén, egy $\mathcal{A}$-beli α-t úgy választva, hogy $\max_{a\in\mathcal{A}} \mathbb{E}[r_0 + \gamma \cdot V(s_1)|s_0 = s, a_0 = a] = \mathbb{E}[r_0 + \gamma \cdot V(s_1)|s_0 = s, a_0 = \alpha]$,

$$\max_{a\in\mathcal{A}} \mathbb{E}[r_0 + \gamma \cdot V'(s_1)|s_0 = s, a_0 = a] \geq \mathbb{E}[r_0 + \gamma \cdot V'(s_1)|s_0 = s, a_0 = \alpha] \geq$$
$$\geq \mathbb{E}[r_0 + \gamma \cdot V(s_1)|s_0 = s, a_0 = \alpha] - \gamma \cdot \|V - V'\|_\infty .$$

□

Definiáljuk a V^* értékfüggvényt a következőképpen: $V^*(s) = \max_\pi V^\pi(s)$, $s \in \mathcal{S}$, és egy adott s állapotra jelöljön π^s mindig egy olyan stratégiát, melyre éppen $V^*(s) = V^{\pi^s}(s)$. Legyen π^* egy (determinisztikus!) mohó stratégia a V^*-ra vonatkozóan, azaz $\pi^*(s) = \operatorname{argmax}_{a\in\mathcal{A}} \mathbb{E}\left[r_0 + \gamma \cdot V^*(s_1) \mid s_0 = s, a_0 = a\right]$.

$$\begin{aligned}
\mathbb{E}\left[r_0 + \gamma \cdot V^{\pi^*}(s_1) \mid s_0 = s, a_0 = \pi^*(s)\right] &= V^{\pi^*}(s) \leq \\
&\leq V^*(s) = \\
&= V^{\pi^s}(s) \overset{\text{(BE-TD)}}{=} \\
&= \mathbb{E}\left[r_0 + \gamma \cdot V^{\pi^s}(s_1) \,\middle|\, s_0 = s, a_0 = \pi^s(s)\right] \leq \\
&\leq \mathbb{E}\left[r_0 + \gamma \cdot V^*(s_1) \mid s_0 = s, a_0 = \pi^s(s)\right] \leq \\
&\leq \max_{a\in\mathcal{A}} \mathbb{E}\left[r_0 + \gamma \cdot V^*(s_1) \mid s_0 = s, a_0 = a\right] = \\
&= \mathbb{E}\left[r_0 + \gamma \cdot V^*(s_1) \mid s_0 = s, a_0 = \pi^*(s)\right] ,
\end{aligned}$$

következésképpen $0 \leq V^*(s) - V^{\pi^*}(s) \leq \gamma \cdot \mathbb{E}\left[V^*(s_1) - V^{\pi^*}(s_1) \,\middle|\, s_0 = s, a_0 = \pi^*(s)\right] \leq \gamma \cdot \|V^* - V^{\pi^*}\|_\infty$. Supremumot véve kapjuk, hogy $\|V^* - V^{\pi^*}\|_\infty \leq \gamma \cdot \|V^* - V^{\pi^*}\|_\infty$, ami miatt $\|V^* - V^{\pi^*}\|_\infty = 0$ (tekintve, hogy $\gamma < 1$), vagy $V^* \equiv V^{\pi^*}$. Azaz a fenti egyenlőtlenségek mind egyenlőségek, és

$$V^*(s) = V^{\pi^*}(s) = \max_{a\in\mathcal{A}} \mathbb{E}\left[r_0 + \gamma \cdot V^*(s_1) \,\middle|\, s_0 = s, a_0 = a\right] = (T^*V^*)(s) . \qquad \text{(BE-DP}^*\text{)}$$

Ez a fenti tétellel, valamint a Banach-féle fixponttétellel együtt adja, hogy:

következmény 3.3. *T^*-nak pontosan egy fixpontja létezik, és az V^*. Ezen felül tetszőleges $V_0 = V \in B_{\mathcal{S}}$ és $V_{t+1} = T^*V_t$, $t = 0, 1, \ldots$ esetén $\|V_k - V^*\|_\infty = O(\gamma^k)$.*

Tehát létezik *optimális stratégia*, és az π^*. A föntiekre épül a (BE-DP*) egyenlőségből kiinduló *stratégia iteráció* ([1], 4.3 szekció, 98. o.) valamint az *érték-iteráció* algoritmus ([1], 4.4 szekció, 102. o.). Kapcsolódó példák: [1] 3.12 példa (3.8 szekció, 78. o.) és 4.1 példa (4.1 szekció, 94.o.).

References

[1] R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, 1998.