Белманова једначина

Самообучавајући и адаптивни алгоритми

Милан Р. Рапаић

Департман за аутоматско управљање
Департман за рачунарство и аутоматику
Факултет техничких наука
Универзитет у Новом Саду
Нови Сад • Србија

1. новембар 2023.

Преглед

1. Одређивање вредности стања за дату политику у детерминистичком случају
2. Одређивање вредности стања у стохастичком случају
3. Одређивање најбољих вредности и најбољих политика

Белманова једначина и динамичко програмирање

Шта су и које проблеме решавају?

Белманова једначина

Различити облици Белманове једначине нам омогућавају да одредимо најбоље одлуке за свако стање Марковљевог процеса одлучивања. Свака политика одлучивања која задовољава Белманову једначину је најбоља (оптимална).

Динамичко програмирање

Иако се у литератури појам "динамичког програмирања" појављује у различитим контекстима, у оквиру овог предмета под овим појмом ћемо подразумевати све поступке за изналажење оптималне политике одлучивања решавањем Белманове једначине.

Одређивање вредности стања за дату политику у детерминистичком случају

1. Одређивање вредности стања за дату политику у детерминистичком случају
2. Одређивање вредности стања у стохастичком случају
3. Одређивање најбољих вредности и најбољих политика

Детерминистичка Белманова једначина

... односно кључна идеја на основу које ћемо је извести

$$g_k = \sum_{i=0}^{T} \gamma^i r_{k+i} = r_0 + \sum_{i=1}^{T} \gamma^i r_{k+i} = r_0 + \gamma \sum_{i=0}^{T} \gamma^i r_{k+1+i}$$

$$g_k = r_k + \gamma g_{k+1}$$

$$g_\pi(s) = r + \gamma g_\pi(s^+)$$

Рекурзивни израз за одређивање вредности стања (Детерминистички случају)

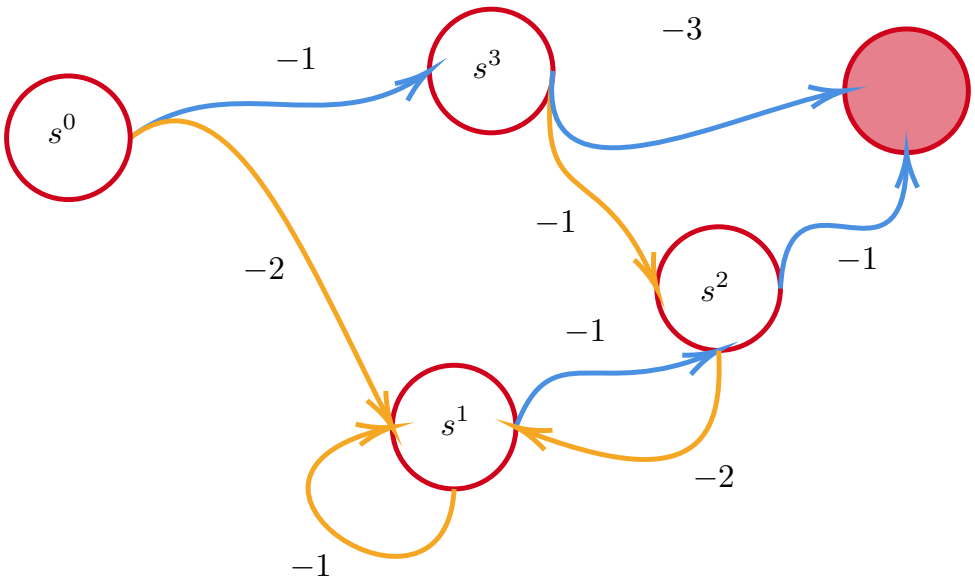
$$v_\pi(s) = h(s, \pi(s)) + \gamma v_\pi(f(s, \pi(s)))$$

Пример

Написати Белманову једначину и одредити вредности стања

ПОЛИТИКА

Увек користи плаву акцију.



Белманова једначина

$$v^0 = -1 + \gamma v^3$$

$$v^1 = -1 + \gamma v^2$$

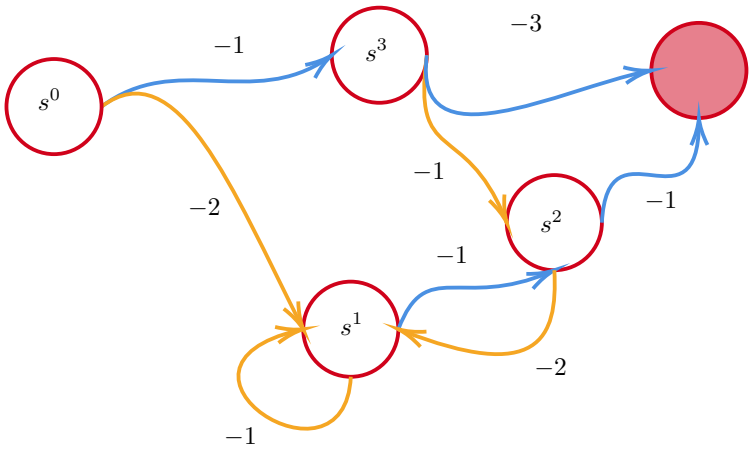
$$v^2 = -1 + \gamma v^{\text{term}}$$

$$v^3 = -3 + \gamma v^{\text{term}}$$

Уколико је политика одлучивања дата, Белманова једначина је увек систем линеарних једначина!

Одређивање вредности стања за дату политику

... решавањем линеарне матричне једначине



$$\underbrace{\begin{bmatrix} v^0 \\ v^1 \\ v^2 \\ v^3 \end{bmatrix}}_{\mathbf{v}} = \gamma \underbrace{\begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}}_{\mathbf{A}} \underbrace{\begin{bmatrix} v^0 \\ v^1 \\ v^2 \\ v^3 \end{bmatrix}}_{\mathbf{v}} + \gamma \underbrace{\begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \end{bmatrix} v^{\text{term}} + \begin{bmatrix} -1 \\ -1 \\ -1 \\ -3 \end{bmatrix}}_{\mathbf{r}}$$

Непосредно (директно) решење

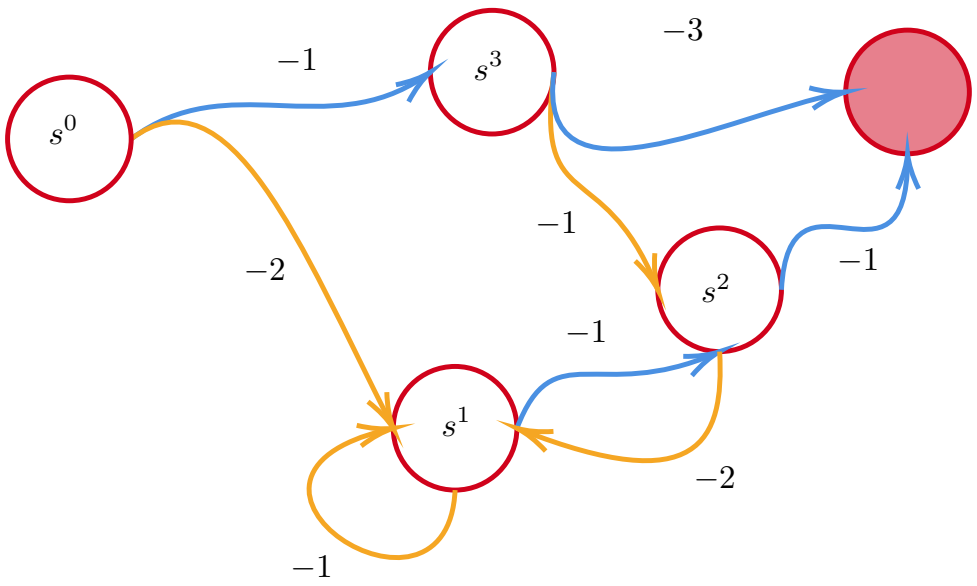
$$\mathbf{v} = (\mathbf{I} - \gamma \mathbf{A})^{-1} \mathbf{r}$$

Корачно (итеративно) решење

$$\mathbf{v}^{k+1} = \gamma \mathbf{A} \mathbf{v}^k + \mathbf{r}\,, \quad \mathbf{v}^0 \in \text{rnd}$$

Одређивање вредности стања за дату политику

... решавањем уназад, почев од крајњег (терминалног) стања



- choose $v^{\text{term}} = 0$
- Наћи сва стања из ког се непосредно стиже до крајњег (s^2 and s^3)

$$v^2 = -1 + \gamma v^{\text{term}} = -1$$

$$v^3 = -3 + \gamma v^{\text{term}} = -3$$

- Наћи сва стања из којих се непосредно стиже до стања познате вредности (s^0 and s^1)

$$v^0 = -1 + \gamma v^3 = -3.7$$

$$v^1 = -1 + \gamma v^2 = -1.9$$

Белманова једначина

... за вредности акција у стањима при произвољној датој политици у детерминистичком случају

$$g_k = r_k + \gamma g_{k+1}$$

$$q_\pi(s_k, a_k) = r_k + \gamma q_\pi(s_{k+1}, a_{k+1})$$

Рекурзивна формула за одређивање вредности акција у стању (детерминистички случај)

$$q_\pi(s, a) = h(s, a) + \gamma q_\pi(f(s, a), \pi(f(s, a)))$$

Поново добијамо скуп линеарних једначина.
(Истина, у односу на већи број променљивих него у случају када тражимо вредности стања).

Одређивање вредности стања у стохастичком случају

1. Одређивање вредности стања за дату политику у детерминистичком случају
2. Одређивање вредности стања у стохастичком случају
3. Одређивање најбољих вредности и најбољих политика

Белманова једначина за одређивање вредности стања

... за дату политику одлучивања у стохастичком случају

$$g_\pi(s) = \mathbb{E}_\pi \left\{ \sum_{i=0}^{T} \gamma^i R_{k+i} \middle| S_0 = s \right\} = \mathbb{E}_\pi \left\{ R_0 + \gamma \sum_{i=1}^{T} \gamma^{i-1} R_i \middle| S_0 = s \right\}$$

$$g_\pi(s) = \sum_{a \in \mathcal{A}} \pi(a|s) \mathbb{E}_\pi \left\{ R_0 + \gamma \sum_{i=1}^{T} \gamma^{i-1} R_i \middle| S_0 = s, A_0 = a \right\}$$

$$g_\pi(s) = \sum_{a \in \mathcal{A}} \pi(a|s) \sum_{\substack{s^+ \in \mathcal{S} \\ r \in \mathcal{R}}} p(s^+, r|s, a) \mathbb{E}_\pi \left\{ R_0 + \gamma \sum_{i=1}^{T} \gamma^{i-1} R_i \middle| S_0 = s, A_0 = a, S_1 = s^+, R_0 = r \right\}$$

$$g_\pi(s) = \sum_{a \in \mathcal{A}} \pi(a|s) \sum_{\substack{s^+ \in \mathcal{S} \\ r \in \mathcal{R}}} p(s^+, r|s, a) \left[r + \gamma \mathbb{E}_\pi \left\{ \sum_{i=1}^{T} \gamma^{i-1} R_i \middle| S_1 = s^+ \right\} \right]$$

Рекурзиван израз за одређивање вредности стања

$$v_\pi(s) = \sum_{a \in \mathcal{A}} \pi(a|s) \sum_{\substack{s^+ \in \mathcal{S} \\ r \in \mathcal{R}}} p(s^+, r|s, a) \left[r + \gamma v_\pi(s^+) \right]$$

Белманова једначина за одређивање акција у стањима

... за дату политику одлучивања у стохастичком случају

$$q_\pi(s,a) = \mathbb{E}_\pi \left\{ R_0 + \gamma \sum_{i=1}^{T} \gamma^{i-1} R_i \middle| S_0 = s, A_0 = a \right\}$$

$$q_\pi(s,a) = \sum_{\substack{s^+ \in \mathcal{S} \\ r \in \mathcal{R}}} p(s^+, r|s,a) \mathbb{E}_\pi \left\{ R_0 + \gamma \sum_{i=1}^{T} \gamma^{i-1} R_i \middle| S_0 = s, A_0 = a, S_1 = s^+, R_0 = r \right\}$$

$$q_\pi(s,a) = \sum_{\substack{s^+ \in \mathcal{S} \\ r \in \mathcal{R}}} p(s^+, r|s,a) \left[r + \gamma \mathbb{E}_\pi \left\{ \sum_{i=1}^{T} \gamma^{i-1} R_i \middle| S_1 = s^+ \right\} \right]$$

$$q_\pi(s,a) = \sum_{\substack{s^+ \in \mathcal{S} \\ r \in \mathcal{R}}} p(s^+, r|s,a) \left[r + \gamma \sum_{a^+ \in \mathcal{A}} \pi(a^+|s^+) \mathbb{E}_\pi \left\{ \sum_{i=1}^{T} \gamma^{i-1} R_i \middle| S_1 = s^+, A_1 = a^+ \right\} \right]$$

Рекурзиван израз за одређивање вредности акција у стањима

$$q_\pi(s,a) = \sum_{\substack{s^+ \in \mathcal{S} \\ r \in \mathcal{R}}} p(s^+, r|s,a) \left[r + \gamma \sum_{a^+ \in \mathcal{A}} \pi(a^+|s^+) q_\pi(s^+, a^+) \right]$$

Одређивање најбољих вредности и најбољих политика

1. Одређивање вредности стања за дату политику у детерминистичком случају
2. Одређивање вредности стања у стохастичком случају
3. Одређивање најбољих вредности и најбољих политика

Белманова једначина

... за вредности стања при оптималној политици у детерминистичком случају

Пошто је оптимална политика само једна од могућих политика, важи рекурзивни образац за срачунавање v_π уз смену $\pi \equiv \pi^*$

$$v_{\pi^*}(s) = h(s, \pi^*(s)) + \gamma v_{\pi^*}(f(s, \pi^*(s)))$$

Пошто је v^* оптимално, десна страна мора бити максимална над скупом свих могућих политика

$$v_{\pi^*}(s) = \max_{\pi \in \mathscr{P}} \{g(s, \pi(s)) + \gamma v_{\pi^*}(f(s, \pi(s)))\}$$

Белманова једначина за одређивање вредности стања

$$v^*(s) = \max_{a \in \mathcal{A}} \{g(s, a) + \gamma v^*(f(s, a))\}$$

Белманова једначина

... за вредности стања при оптималној политици у стохастичком случају

Пошто је оптимална политика само једна од могућих политика, важи рекурзивни образац за срачунавање v_π уз смену $\pi \equiv \pi^*$

$$v_{\pi^*}(s) = \sum_{a \in \mathcal{A}} \pi^*(a|s) \sum_{\substack{s^+ \in \mathcal{S} \\ r \in \mathcal{R}}} p(s^+, r|s, a) \left[r + \gamma v_{\pi^*}(s^+) \right]$$

Пошто је v^* оптимално, десна страна мора бити максимална над скупом свих могућих политика

$$v_{\pi^*}(s) = \max_{\pi \in \mathscr{P}} \sum_{a \in \mathcal{A}} \pi^*(a|s) \sum_{\substack{s^+ \in \mathcal{S} \\ r \in \mathcal{R}}} p(s^+, r|s, a) \left[r + \gamma v_{\pi^*}(s^+) \right]$$

Белманова једначина за одређивање вредности стања

$$v^*(s) = \max_{a \in \mathcal{A}} \sum_{\substack{s^+ \in \mathcal{S} \\ r \in \mathcal{R}}} p(s^+, r|s, a) \left[r + \gamma v^*(s^+) \right]$$

Белманова једначина

... за вредности акција у стањима при оптималној политици у детерминистичком случају

Пошто је оптимална политика само једна од могућих политика, важи рекурзивни образац за срачунавање q_π уз смену $\pi \equiv \pi^*$

$$q_{\pi^*}(s,a) = h(s,a) + \gamma q_{\pi^*}(f(s,a), \pi^*(f(s,a)))$$

Пошто је q^* оптимално, десна страна мора бити максимална над скупом свих могућих политика

$$q_{\pi^*}(s,a) = \max_{\pi \in \mathscr{P}} \{h(s,a) + \gamma q_\pi(f(s,a), \pi(f(s,a)))\}$$

Белманова једначина за одређивање вредности акција у стањима

$$q^*(s,a) = h(s,a) + \gamma \max_{a^+ \in \mathcal{A}} q^*(f(s,a), a^+)$$

Белманова једначинан

... за вредности акција у стањима при оптималној политици у стохастичком случају

Пошто је оптимална политика само једна од могућих политика, важи рекурзивни образац за срачунавање q_π уз смену $\pi \equiv \pi^*$

$$q_{\pi^*}(s,a) = \sum_{\substack{s^+ \in \mathscr{S} \\ r \in \mathcal{R}}} p(s^+, r|s,a) \left[r + \gamma \sum_{a^+ \in \mathcal{A}} \pi^*(a^+|s^+) q_{\pi^*}(s^+, a^+) \right]$$

Пошто је q^* оптимално, десна страна мора бити максимална над скупом свих могућих политика

$$q_{\pi^*}(s,a) = \sum_{\substack{s^+ \in \mathcal{S} \\ r \in \mathcal{R}}} p(s^+, r|s,a) \left[r + \gamma \max_{\pi \in \mathscr{P}} \sum_{a^+ \in \mathcal{A}} \pi^*(a^+|s^+) q_{\pi^*}(s^+, a^+) \right]$$

Белманова једначина за одређивање вредности акција у стањима

$$q^*(s,a) = \sum_{\substack{s^+ \in \mathcal{S} \\ r \in \mathcal{R}}} p(s^+, r|s,a) \left[r + \gamma \max_{a^+ \in \mathcal{A}} q^*(s^+, a^+) \right]$$

Белманова једначина

... у свим својим прелепим облицима :)

Белманова једначина за вредности стања

$$v^*(s) = \max_{a \in \mathcal{A}} \{g(s,a) + \gamma v^*(f(s,a))\}$$

$$v^*(s) = \max_{a \in \mathcal{A}} \sum_{\substack{s^+ \in \mathcal{S} \\ r \in \mathcal{R}}} p(s^+, r|s,a) \left[r + \gamma v^*(s^+)\right]$$

Белманова једначина за вредности акција у стањима

$$q^*(s,a) = h(s,a) + \gamma \max_{a^+ \in \mathcal{A}} q^*(f(s,a), a^+)$$

$$q^*(s,a) = \sum_{\substack{s^+ \in \mathcal{S} \\ r \in \mathcal{R}}} p(s^+, r|s,a) \left[r + \gamma \max_{a^+ \in \mathcal{A}} q^*(s^+, a^+)\right]$$

Услед операције тражења максимума (оператора `max`), Белманова једначина за одређивање оптималних вредности представља скуп нелинеарних једначина. Непосредно решење ове једначине, у затвореном облику није могуће, чак ни принципски.

Further Reading

For further reading please consult [Sutton and Barto, 2018], **Chapter 4**.

References I


Sutton, R. S. and Barto, A. G. (2018).
Reinforcement Learning: An Introduction.
The MIT Press, second edition.