TRƯỜNG ĐẠI HỌC BÁCH KHOA HÀ NỘI
VIỆN TOÁN ỨNG DỤNG VÀ TIN HỌC

ỨNG DỤNG HỌC TĂNG CƯỜNG VÀO ĐẦU TƯ CHỨNG KHOÁN

Giáng viên hướng dẫn: TS. Nguyễn Thị Ngọc Anh

Nhóm Sinh viên thực hiện:

Vũ Thị Ngọc

Trịnh Hoàng Đức

Đặng Đình Trung

Lớp: KSTN Toán Tin K60

HÀ NỘI - 06/2018

Mục lục

Chương 1

Bài toán

Chương 2

Phương pháp

2.1 Giới thiệu

Trong ngành khoa học máy tính, học tăng cường (Reinforcement Learning)là một lĩnh vực con của học máy (Machine Learning), nghiên cứu cách thức một tác tử (agent) trong một môi tường nên chọn thực hiện các hành động nào để có được phần thưởng có giá trị lớn nhất về lâu về dài. Các thuật toán học tăng cường cố gắng tìm một chiến lược ánh xạ từ không gian trạng thái tới không gian hành động mà agent nên chọn trong các trạng thái đó.

2.2 Quá trình quyết định Markov

Quá trình quyết định Markov (Markov Decision Processes, ký hiệu là MDP)cung cấp một nền tảng toán học cho việc mô hình hóa việc ra quyết

định trong các tình huống mà kết quả là một phần ngẫu niên, một phần dưới sự điều khiển của một người ra quyết định. MDP rất hữu dụng trong việc học một loạt các bài toán tối ưu hóa được giải quyết thông qua quy hoạch động và hocn tăng cường. MDP được biết đến sớm nhất vào những năm 1950 (cf. Bellman 1957). Một cốt lõi của nghiên cứu về quá trình ra quyết định Markov là từ kết quả của cuốn sách của Ronald A. Howard xuất bản năm 1960, $QuyhochØngvqutrnhMarkov$. Chúng được sử dụng trong rất nhiều các lĩnh vực khác nhau, bao gồm robot, điều khiển tự động,kinh tế, và chế tạo.

Chính xác hơn, một quá trình quyết định Markov là một quá trình điều khiển ngẫu nhiên thời gian rời rạc. Tại mỗi bước thời gian, quá trình này trong một vài trạng thái s, và người ra quyết định có thể chọn bất kỳ hành động a nào có hiệu lực trong trạng thái s. Quá trình này đáp ứng tại bước thời gian tiếp theo bằng cách di chuyển ngẫu nhiên vào một trạng thái mới s', và đưa ra cho người ra quyết định một phần thưởng tương ứng $R_a(s, s')$

Xác suất mà quá trình di chuyển vào trạng thái mới của nó s' bị ảnh hưởng bởi hành động được chọn. Đặc biệt, nó được đưa ra bởi hàm chuyển tiếp trạng thái $P_a(s, a')$. Do đó, trạng thái kế tiếp s' phụ thuộc vào trạng thái hiện tại s và a đã cho, lại độc lập có điều kiện với toàn bộ trạng thái và hành động trước đó. Nói cách khác, các trạng thái chuyển tiếp của một quá trình MDP thỏa mãn thuộc tính Markov.

Quá trình quyết định Markov là một phần mở rộng của chuỗi Markov; khác biệt là ở sự bổ sung của các hành động (cho phép lựa chọn) và phần thưởng (cho động cơ). Ngược lại, nếu chỉ có một hành động tồn tại cho mỗi

trạng thái và tất cả các phần thưởng là giống nhau (ví dụ: zero), một quá trình quyết định Markov làm giảm một chuỗi Markov.

MDP là một quá trình điều khiển ngẫu nhiên thời gian rời rạc. MDP là một tập 5 dữ liệu $< \mathbb{S}, \mathbb{A}, P, R, \gamma >$. Trong đó:

- $\mathbb{S}$: không gian các trạng thái
- $\mathbb{A}$: không gian các hành động
- $P : \mathbb{S} \times \mathbb{A} \times S \to \mathbb{R}$: hạt nhân chuyển tiếp Markov
- $R : \mathbb{S} \times \mathbb{A} \to \mathbb{R}$: hàm phần thưởng, $0 < \gamma < 1$ là hệ số chiết khấu

Giả sử rằng, tahi thời điểm t, trạng thái $S_t = s$ và agent có hành động $A_t = a$. Khi đó, xác suất của trạng thái $B \in \mathbb{S}$ tại thời điểm $t+1$ được cho bởi công thức:

$$P(s, a, B) = \mathbb{P}(S_{t+1} \in B | S_t = s, A_t = a) \ (1)$$

Sau quá trình này, agent nhận được một phần thưởng ngẫu nhiên là $R_(t+1)$. Hàm thưởng $R(s, a)$ là phần thưởng thu được khi thực hiện hành động a ở trang thái s

$$R(s, a) = \mathbb{E}[R_(t+1) | S_t = s, A_t = a] \ (2)$$

Tại bất kì bước thời gian nào, agent chọn hành động của nó theo một chính sách $\pi : \mathbb{S} \times A \to \mathbb{R}$ sao cho với mỗi $s \in \mathbb{S}, C \to \pi(s, C)$ là xác suất phân phối trên $(\mathbb{A}, A)$. Do đó, chính sách π và trạng thái ban đầu

$s_0 \in \mathbb{S}$ xác định chuỗi trạng thái - hành đồng - phần thưởng ngẫu nhiên $\{(S_t, A_t, R_{t+1})\}_{t \geq 0}$ với giá trị trên $\mathbb{S} \times \mathbb{A} \times \mathbb{R}$. Trong một không gian vô hạn,hiệu suất của agent thường được tính bằng tổng phần thưởng chiếu khấu thu được sau một chính sách là

$$G_t = \sum_{t=0}^{\infty} \gamma^t R_{t+k+1} \quad (3)$$

Vì phần thưởng này là ngẫu nhiên, agent xem xét giá trị kì vọng của nó, thưởng được gọi là hàm giá trị trạng thái

$$V_\pi(s) = \mathbb{E}_\pi[G_t | S_t = s] \quad (4)$$

Trong đó, chỉ số $\mathbb{E}_\pi$ chỉ ra rằng xá hành đồng được chọn theo chính sách π. Hàm giá trị trạng thái được đo tốt khi agent ở trong một trạng thái nhất định và tuân theo một chính sách nhất định. Tương tự, ta có hàm giá trị hành động

$$Q_\pi(s, a) = \mathbb{E}_\pi[G_t | S_t = s, A_t = a] \quad (5)$$

Ta có mối liên hệ giữa V_π và Q_π

$$V_\pi(s) = \int_A \pi(s, a) Q_\pi(s, a) da \quad (6)$$

Hầu như tất cả các thuật toán học tăng cường được thiết kế để tính các hàm giá trị này dựa trên các phương trình Bellman

$$V_\pi(s) = R_\pi(s) + \gamma T_\pi V_\pi(s) \quad (7)$$
$$Q_\pi(s, a) = R(s, a) + \gamma T_a V_\pi(s) \quad (8)$$

Khi chúng ta biểu diễn bở T_a thì toán tử chuyển với hành đồng a (chính sách π)

$$T_s F(s) = \mathbb{E}[F(S_{t+1}|S_t = s, A_t = a] = \int_S P(s, a, s')F(s')ds' \ (9)$$
$$T_\pi F(s) = \mathbb{E}_p i[F(S_{t+1}|S_t = s] = \int_{\mathbb{A}} \pi(s, a) \int_{\mathbb{S}} P(s, a, a')F(s')F(s')ds'da \ (10)$$

Các phương trình này có thể được viết dưới dạng phương trình điểm cố định tho một số giả thuyết về hàm thưởng, thừa nhận một giải pháp duy nhất theo định lý ánh xạ co. Mục tiêu của agent là chọn một chính sách π_* mà tối đa hóa được lợi nhuận ở tất cả các trạng thái có thể. Chính sách như vậy được gọi là tối ưu và các hàm giá trị (sửa lỗi) được gọi là hàm giá trị trạng thái tối ưu

$$V_*(s) = \sup_\pi V_\pi(s) \ (11)$$

và Optimal Action - Value Function

$$Q_*(s, a) = \sup_\pi Q_\pi(s, a) \ (12)$$

Các hàm giá trị tối ưu thỏa mãn phương trình Bellman sau

$$V_*(s) = \sup_a Q_*(s, a) = \sup_a R(s, a) + \gamma T_a V_*(s) \ (13)$$
$$Q_*(s, a) = R(s, a) + \gamma T_a V_*(s) \ (14)$$
$$= R(s, a) + \gamma \int_{\mathbb{S}} P(s, a, s') \sup_{a'} Q_*(s', a')ds$$

Một lần nữa, đây là những phương trình điểm cố định mà luôn tồn tại và duy nhất một giải pháp do định lý ánh xạ co. Với hàm giá trị trạng thái tối ưu Q_*. Một chính sách tối ưu thu được bằng cách chọn mỗi trạng thái hành động với giá trị lớn nhất Q_*

$$a_* = argsup_a Q_*(s, a) \ (15)$$

Chính sách tham lam này được xác định và chỉ phụ thuộc vào trạng thái hiện tại của hệ thống

Chương 3

Ứng dụng

3.1 Giới thiệu

Một cách tiếp cận tương đối mới với giao dịch tài chính là sử dụng thuật toán học máy để dự đoán giá cổ phiếu tăng hoặc giảm trước khi nó xả ra. Một nhà đầu tư tối ưu sẽ mua cổ phiếu trước khi giá của nó tăng và bán trước khi nó giảm.

Một nhà giao dịch tài chính sẽ sử dụng học tăng cường tiếp diễn. Thuật toán và các tham số lấy từ bài báo của Moody và Saffel. Thuật toán hướng tăng mà tối đa hóa giá tri hàm thưởng còn gọi là Sharpe. Việc chọn tham số tối ưu w cho nhà giao dịch, chúng ta cố gắng tận dụng lợi thế của thay đổi giá cổ phiếu. Một ví dụ về cách làm việc của nhà giao dịch, cả "thế giới thực" và ảo đều được minh họa trong phần cuối của chương.

3.2 Hàm mục tiêu: SHARPE

Với mỗi thời điểm của lợi nhuận đầu tư, Sharpe được xác định bởi công thức

$$S_T = \frac{Average(R_t)}{StandardDeviation(R_t)} \text{ với } t \text{ là số nguyên, } t = 1, 2, ..., \text{T}$$

Trong đó, R_t là lợi nhuận đầu tư cho giao dịch tại thời điểm t. Bằng trực giác, Sharpe thưởng cho các chiến lược đầu tư dựa vào xu hướng ít biến động để tạo ra lợi nhuận.

3.3 Hàm đầu tư

Nhà đầu tư sẽ cố gắng sao cho Sharpe đạt giá trị lớn nhất cho mỗi chuỗi thời gian nhất định. Với bài báo cáo natf, hàm đầu tư có công thức là:

$$F_t = tanh(w^T x_t)$$

trong đó M là số chuỗi thời gian đầu vào cho nhà đầu tư, tham số $w \in R^{M+2}$, vector đầu vào $x_t = [1, r_t, ..., r_{t-M}, F_{t-1}]$, và doanh thu $r_t = p_t - p_{t-1}$

Lưu ý tằng r_t là khác nhau với giá trị của cổ phiếu giữa thời điểm hiện tại t và thời điểm trước đó. Vì vậy, r_t là lợi nhuận trong một phần cổ phiếu được mua ở thời gian $t - 1$

Ngoài ra, hàm $F_t \in [-1, 1]$ cho biết trường hợp đầu tư tại thời điểm t. Có 3 trường hợp có thể có: long, short, neutral

- Long khi $F_t > 0$. Trong trường hợp này, nhà đầu tư mua một cổ phiếu với giá p_t và hi vọng rằng nó tăng giá ở thời điểm $t + 1$

- Short khi $F_t < 0$. Ở trường hợp này, nhà đầu tư bán một cổ phiếu mà họ không sở hữu với giá p_t, với kì vọng nó có thể thực hiện giao dịch tại thời điểm $t+1$. Nếu giá cao hơn tại thời điểm $t+1$ thì nhà đầu tư bắt buộc phải mua ở giá cao hơn thời điểm $t + 1$ đẻ hoàn thành hợp đồng. Nếu giá ở thời điểm $t + 1$ thấp hơm thì nhà đầu tư thu được lợi nhuận.

- Neutral khi $F_t = 0$. Trong trường hợp này nhà đầu tư không được cũng không mất lợi nhuận.

Như vậy, F_t cho biết cổ phiếu tại thời điểm t. Đó là $n_t = \mu.F_t$, cổ phiếu được mua (long) hoặc bán (short) với μ là số lượng cổ phiếu tối đa cho mỗi giao dịch. Lợi nhuận ở thời điểm t phụ thuộc vào F_{t-1}:

$$R_t = \mu(F_{t-1}.r_t - \delta|F_t - F_{t-1}|$$

trong đó, δ là chi phí giao dịch tại thời điểm t. Nếu $F_t = F_{t-1}$ (không thay đổi đàu tư trong thời điểm này) thì sẽ không mất phí giao dịch. Nếu không sẽ mất phí tỷ lệ thuận với sự chênh lệch trong cổ phiếu nắm giữ.

Đầu tiên, $(\mu.F_{t-1}.r_t)$ là kết quả trả về từ quyết định đầu tư tại thời điểm $t-1$. Ví dụ nếu $\mu = 20$ cổ phiếu, quyết định mua một nửa mức tối đa cho phép ($F_{t-1} = 5$) và mỗi cổ phiếu tăng $r_t = 8$ đơn vị giá, kì hạn này sẽ là 80. tổng lợi nhuận thu được (bỏ qua các khoản phí giao dịch phát sinh ở thời điểm t)

3.4 Hướng tăng

Tối đa hóa Sharpe tuân theo hướng tăng. DDaaud tiên chúng ta xác định hàm thưởng bằng cách sử dụng công thức cơ bản của thống kê cho trung bình và phương sai.

Ta có:

$$S_r = \frac{E[R_t]}{\sqrt{E[R_t^2] - (E[R_t])^2}} = \frac{A}{\sqrt{B - A^2}}$$

Với $A = \frac{1}{T}\sum_{t=1}^{T} R_t$ và $B = \frac{1}{T}\sum_{t=1}^{T} R_t^2$

Sau đó chúng ta lấy đạo hàm của S_T sử dụng chuỗi:

$$\begin{aligned}
\frac{dS_T}{dw} &= \frac{d}{dw}\left\{\frac{A}{\sqrt{B-A^2}}\right\} = \frac{dS_T}{dA}.\frac{dA}{dw} + \frac{dS_T}{dB}.\frac{dB}{dw} \\
&= \sum_{t=1}^{T}\left\{\frac{dS_T}{dA}.\frac{dA}{dR_t} + \frac{dS_T}{dB}.\frac{dB}{dR_t}\right\}.\frac{dR_T}{dw} \\
&= \sum_{t=1}^{T}\left\{\frac{dS_T}{dA}.\frac{dA}{dR_t} + \frac{dS_T}{dB}.\frac{dB}{dR_t}\right\}.\left\{\frac{dR_t}{dF_t}.\frac{dF_t}{dw} + \frac{dR_t}{dF_{t-1}}\frac{dF_{t-1}}{dw}\right\}
\end{aligned}$$

Đạo hàm từng phần, kết quả trả về được hàm:

$\frac{dR_t}{dF_t}$

$= \frac{d}{dF_t}\left\{\mu\left(F_{t-1}.r_t - \delta\left|F_t - F_{t-1}\right|\right)\right\}$

$= \frac{d}{dF_t}\left\{-\mu.\delta.\left|F_t - F_{t-1}\right|\right\}$

$$= \begin{cases} -\mu.\delta & F_t - F_{t-1} > 0 \\ \mu.\delta & F_t - F_{t-1} < 0 \end{cases}$$

$= -\mu\delta.sign\left(F_t - F_{t-1}\right)$

$\frac{dR_t}{dF_{t-1}}$

$= \frac{d}{dF_{t-1}}\left\{\mu\left(F_{t-1}.r_t - \delta\left|F_t - F_{t-1}\right|\right)\right\}$

$= \mu.r_t - \frac{d}{dF_{t-1}}\left\{-\mu.\delta.\left|F_t - F_{t-1}\right|\right\}$

$$= \begin{cases} \mu.\delta & F_t - F_{t-1} > 0 \\ -\mu.\delta & F_t - F_{t-1} < 0 \end{cases}$$
$$= \mu.r_t + \mu\delta.sign\left(F_t - F_{t-1}\right)$$

Khi đó, đạo hàm từng phần dF_t/dw và dF_{t-1}/dw được tính như sau:

$$\frac{dF_t}{dw}$$
$$= \frac{d}{dw}\left\{tanh(w^T x_t)\right\}$$
$$= \left(1 - tanh\left(w^T x_t\right)^2\right).\frac{d}{dw}.\left\{w^T x_t\right\}$$
$$= \left(1 - tanh\left(w^T x_t\right)^2\right)\left\{x_t + w_{M+2}\frac{dF_{t-1}}{dw}\right\}$$

Lưu ý rằng đạo hàm dF_t/dw có tính hồi quy và phụ thuộc vào tất cả các giá trị trước đó của nó. Điều này nghĩa là để tính được tham số thì ta phải có giá trị dF_t/dw từ chuỗi thời gian ban đầu. Bởi vì dữ liệu cổ phiếu trong khoảng 1000 - 2000 mẫu, điều này làm chậm hướng tăng nhưng vẫn tính toán được. Một cách khác là sử dụng học trực tuyến và ước lượng dF_t/dw chỉ sử dụng điều kiện dF_{t-1}/dw giúp cho thuật toán hướng tăng như trong bài báo của Moody và Saffell. Tuy nhiên, trong bài báo cáo này sử dụng các công thức chính xác ở trên.

Khi điều kiện dS_t/dw được tính toán, khối lượng ta cập nhật tuân theo nguyên tắc hướng tăng: $w_{t+1} = w_t + \rho.dS_T/dw$. Quá trình đươc lặp lại với số lần lặp N_e với N_e được chọn đảm bảo rằng Sharpe hội tụ.

3.5 Thuật toán

Chương 4

Kết luận

Danh mục công trình công bố của tác giả

[1] Vũ Thu Thảo, Nguyễn Ngọc Doanh, Nguyễn Nhị Gia Vinh, Nguyễn Thị

Ngọc Anh, *"Mô hình hóa ảnh hưởng phân bố không gian của hoa diệt rầy*

tới sự phát triển của rầy nâu hại lúa", Tạp chí Khoa học và Công nghệ -

Đại học Thái Nguyên, 2015, pp.119-123.