Reinforcement Learning (RL)

Algorytm uczenia przez wzmacnianie jest w dużym uogólnieniu rekurencyjną procedurą zdobywania wiedzy metodą prób i błędów. Wyobraź sobie grę, której zasad nie znasz. Grasz, a po 50 ruchach sędzia mówi „przegrałeś”. To jest uczenie przez wzmocnienie.

Tworzymy naszego „ucznia/agenta” który próbuje grać i oczekuje na odpowiedz czy udało mu się osiągnąć cel czy nie. To pozwala mu na zmianę swojej strategii/posunięć celem zoptymalizowania jak najlepszego wyniku. Bez żadnej informacji ze środowiska uczeń/agent nie ma podstaw, aby decydować, który ruch wykonać:

Musi wiedzieć, że coś dobrego się stało, gdy wygrał albo wykonał dobry ruch.

Możemy go informować przez
nagrode (reward), wzmocnienie (reinforcement)

Cel to optymalna polityka(strategii gry).

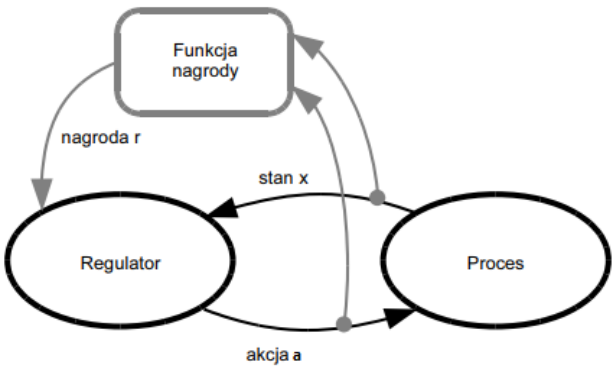
U podstaw uczenia się ze wzmocnieniem leżą dynamiczne interakcje ucznia/agenta ze środowiskiem, w którym działa, realizując swoje zadanie. Interakcje te odbywają się dyskretnych (na ogół) krokach czasu i polegają na obserwowaniu przez ucznia kolejnych *stanów* środowiska oraz wykonywaniu wybranych zgodnie z jego obecną *strategią* decyzyjną *akcji*. Po wykonaniu akcji uczeń otrzymuje rzeczywisto-liczbowe wartości *wzmocnienia* lub *nagrody*, które stanowią pewną miarę oceny jakości jego działania. Wykonanie akcji może również powodować zmianę stanu środowiska.

W każdym kroku czasu t:

1. obserwuj aktualny stan x_t
2. wybierz akcję a_t do wykonania w stanie x_t
3. wykonaj akcję a_t
4. obserwuj wzmocnienie r_t i następny stan x_{t+1}
5. ucz się na podstawie doświadczenia $\langle x_t, a_t, r_t, x_{t+1} \rangle$

A po ludzku można powiedzieć iż regulator/agent/uczeń wchodzi w interakcję z obiektem (procesem, środowiskiem) sterowania za pomocą trzech sygnałów:

- stanu x,
- sterowania (akcji) a
- nagrody (kosztu sterowania) r.



W każdym kroku algorytmu regulator obserwuje stan x_t obiektu, a następnie wykonuje akcję a_t, przeprowadzającą obiekt do następnego stanu x_{t+1},. Jednocześnie regulator otrzymuje sygnał wartościujący wykonaną akcję w postaci nagrody r_t. Po otrzymaniu nagrody regulator wykonuje kolejny krok algorytmu.

Środowisko

Środowisko pod wpływem wykonywanych przez ucznia/agenta akcji może zmieniać stany oraz dostarczać nagrody, stanowiące ocenę skuteczności działania ucznia. W uczeniu się ze wzmocnieniem dopuszcza się ***niepewność*** środowiska i zakłada się jego ***nieznajomość*** przez ucznia/agenta.

niepewność oznacza, że generowane pod wpływem wykonywanych akcji wzmocnienia i zmiany stanów mogą być stochastyczne (jak ktoś nie wie co to znaczy to pytać Rafała).

nieznajomość oznacza, że leżące u podstaw tych stochastycznych mechanizmów rozkłady prawdopodobieństwa nie są znane uczniowi/agentowi. Ponadto środowisko jest *niekontrolowalne*: uczeń/agent nie ma na te rozkłady prawdopodobieństwa żadnego wpływu.

To ostatnie założenie ma decydujące znaczenie na wytyczenie granicy między uczniem a środowiskiem: uczeń/agent ma wpływ na swoje własne mechanizmy działania, parametry itp., lecz nie ma wpływu na środowisko.

W jaki sposób się uczy

W najbardziej ogólnym przypadku możemy powiedzieć, że od ucznia/agenta oczekuje się nauczenia się strategii (czyli odwzorowania stanów na akcje do wykonania w tych stanach), która maksymalizuje pewne kryterium jakości za pomocą otrzymywanych przez niego nagród. Rodzaj tego kryterium decyduje o konkretnym typie uczenia się ze wzmocnieniem.

„Najciekawszy i najczęściej rozważany jest przypadek, kiedy uczeń ma maksymalizować swoje nagrody *długoterminowo*: dobra strategia niekoniecznie przynosi natychmiast wysokie nagrody, lecz jest opłacalna w dłuższym horyzoncie czasowym. Ten typ uczenia się ze wzmocnieniem wymaga uwzględnienia przez ucznia/agenta opóźnionych skutków wykonywanych przez niego akcji i określany jest mianem *uczenia się z opóźnionym wzmocnieniem* lub *uczenia się na podstawie opóźnionych nagród*. Stosowane wówczas algorytmy uczenia się rozwiązują tzw. **problem *temporalnego przypisania zasługi*** (*temporal credit assignment*), polegający na przypisaniu zasługi (bądź winy) za długoterminowe dochody ucznia/agenta jego poszczególnym akcjom, być może wykonanym wiele kroków przed faktycznym uzyskaniem tych dochodów.”

Efektywność ucznia/agenta często przyjmuje się sumę otrzymanych nagród.
Ucznia/Agenta rozpoczyna w czasie $t = 0$ i jego zadaniem jest maksymalizowanie sumy:

$$E\left[\sum_{t=0}^{\infty}\gamma^t r_t\right]$$

gdzie *współczynnik dyskontowania* $\gamma \in [0,1]$ reguluje względną ważność krótko- i długoterminowych nagród.

Procesy decyzyjne Markowa

Model matematyczny problemu uczenia przez wzmacnianie (model środowiska) przedstawia się jako proces decyzyjny Markowa (MDP), który wyrażamy wzorem:

$$MDP = \langle X, A, \vartheta, \delta \rangle$$

gdzie

- $X = \{x_1, x_2, \dots, x_n\}$ jest skończonym zbiorem stanów,
- $A = \{a_1, a_2, \dots, a_n\}$ jest skończonym zbiorem akcji,
- ϑ jest funkcją nagrody (wzmocnienia),
- δ jest funkcją przejścia stanów.

Dla każdej pary $\langle x, a \rangle \in X$ x A, mamy wartości (x to aktualny stan w czasie t [x_t], a to akcja w czasie t [a_t])

- $\vartheta(x, a)$ która jest zmienną losową oznaczającą nagrodę otrzymywaną po wykonaniu akcji a w stanie x
- $\delta(x, a)$ jest zmienną losową oznaczającą następny stan po wykonaniu akcji a w stanie x

Czyli:

$r_t = \vartheta(x_t, a_t)$, ($r_t$ to wartość wzmocnienia/nagrody w czasie t)
$x_{t+1} = \delta(x_t, a_t)$, ($x_{t+1}$ to następny stan)

Równanie Bellman'a:

$$U(s) = R(s) + \gamma \max_{a \epsilon A} \sum_{s'} U(s')\, P(s'|s, a)$$

Trzy podejścia do RL:

- **agent odruchowy (ang. direct policy search)**
  Uczy się polityki π : S → A

- **agent z funkcją użyteczności U**
  uczy się f. użyteczności U(s) i używa jej, aby wybierać akcje, które maksymalizują wartość oczekiwaną przyszłych nagród.

- **agent z funkcją Q**

Uczy się funkcji Q(s, a), która zwraca oczekiwaną użyteczność podjęcia danej akcji w danym stanie

Typy uczenia ze wzmocnieniem:

- **pasywne**
  Polityka π jest dana. Uczymy się tylko użyteczności stanów funkcja U(s) lub użyteczności par stan-akcja: funkcja Q(s, a)
- **aktywne**
  Musimy również nauczyć się polityki („co mam robić?”) Konieczna eksploracja.