Co to jest Reinforcement Learning?

Tak w skrócie: mamy agenta (to nasz algorytm) oraz środowisko, w którym ten agent ma działać (w naszym przypadku zabijać zielone prosiaczki). Agent nie zna środowiska, nie dostaje od nas żadnych informacji/danych na ten temat. Wszystkie dane musi zebrać sam.

Interakcje agenta ze środowiskiem odbywają się zazwyczaj w dyskretnych krokach czasu i polegają na obserwowaniu przez agenta(można też mówić ucznia) kolejnych stanów środowiska oraz wykonywaniu wybranych zgodnie z jego obecną strategią decyzyjną akcji. Po wykonaniu akcji uczeń dostaje rzeczywistoliczbowe wartości wzmocnienia (albo nagrody), które są miarą oceny jego działania. Wykonanie akcji może spowodować zmianę stanu środowiska, np. W przypadku gry Angry Birds zbicie klocków, przewrócenie konstrukcji zbudowanej z klocków, zabicie świni.

Trzy podejścia do RL

1. Agent odruchowy (direct policy search) – uczy się polityki
2. Agent z funkcją użyteczności U – uczy się funkcji użyteczności U(s) i używa jej, aby wybierać akcje, które maksymalizują wartość oczekiwaną przyszłych nagród
3. Agent z funkcją Q – uczy się funkcji Q(s,a), która zwraca oczekiwaną użyteczność podjęcia danej akcji w danym stanie

Q-learning - agent nie musi posiadać modelu środowiska, ale przez to nie może wnioskować na więcej niż jeden ruch do przodu, bo nie wie w jakim stanie będzie.

Typy uczenia ze wzmocnieniem

- pasywne – uczumy się tylko użyteczności stanów funkcji U(s) lub użyteczności par stan-akcja: funkcja Q(s,a)
- aktywne – musimy również nay=uczyć się polityki (Co mam robić?). Konieczna jest eksploracja.

Proces decyzyjny Markova – jest to matematyczny model problemu uczenia się ze wzmocnieniem, a właściwe model środowiska. Proces decyzyjny Markova (Markov decision process - MDP) definiuje się jako czwórkę:

$$MDP = [X,A,\varrho,\delta]$$

gdzie

- X jest skończonym zbiorem stanów,
- A jest skończonym zbiorem akcji,
- ϱ jest funkcją nagrody/wzmocnienia,
- δ jest funkcją przejścia stanów.