26. Uczenie ze wzmocnieniem.

Nie jest uczeniem indukcyjnym czyli generalizacją dużej liczby przykładów. Nie jest to inna forma uczenia się pojęć. Nie jest to uczenie z nauczycielem. Jest to uczenie się umiejętności (tzw. wiedzy proceduralnej).
Źródłem informacji trenującej jest krytyk, nie będący nauczycielem. W uczeniu nadzorowanym nauczyciel podaje informację trenującą w formie gotowej odpowiedzi (np. klasy obiektu). Z kolei w przypadku uczenia ze wzmocnieniem uczeń najpierw daje odpowiedź na podstawie dotychczasowej strategii, a następnie krytyk ją ocenia.
Krytyk jest częścią środowiska, w odróżnieniu od ucznia, który środowiska nie zna, nie kontroluje i nie jest go pewnym. Uczeń poznaje środowisko jedynie za pośrednictwem wykonywanych przez siebie akcji.
Proces uczenia podzielony jest na kroki wykonywane w dyskretnych przedziałach czasowych, a każdy z kroków składa się z następujących etapów:
- obserwacja aktualnego stanu x(t)
- wybór akcji na podstawie strategii (a = PI(x(t))), strategia dla podanego stanu zwraca akcję
- wykonanie akcji, a następnie zaobserwowanie otrzymanej nagrody/wzmocnienia r(t)
- obserwacja stanu x(t+1) po wykonaniu akcji
- zmiana strategii na podstawie < x(t), a(t), r(t), x(t+1) >

Zmiana strategii powinna powodować naukę celowego zachowania. Cel wyznaczany jest otrzymywanymi nagrodami, a dążenie do niego wiąże się z maksymalizacją kryterium sumy nagród (ważonej). Maksymalizacja ta może być długo lub krótkoterminowa, odpowiednio gdy uczeń pamięta wszystkie poprzednie nagrody lub tylko kilka ostatnich.
Poprawianie strategii może następować po każdym kroku (tryb uczenia inkrementacyjny) lub po epoce uczenia (tryb epokowy).
Formalnym modelem środowiska jest proces decyzyjny Markowa (procesy stochastyczne w środowisku). Rozwiązaniem problemu uczenia ze wzmocnieniem jest poznanie prawdopodobieństw i wyznaczenie optymalnej strategii na tej podstawie - wartościowanie strategii.
Używa się w tym celu metod:
• programowania dynamicznego - metoda analityczna, wymaga pełnej wiedzy o procesie,
daje w wyniku strategię optymalną, rozwiązanie za pomocą równań Bellmana
• różnic czasowych TD - metody numeryczne, szybsze i prostsze, wynik przybliżony, algorytmy AHC, Q-learning, SARSA.
Dziedziny, w których możliwości zastosowań uczenia się ze wzmocnieniem wydają się w świetle dotychczasowych prac najbardziej obiecujące, to przede wszystkim:
• inteligentne sterowanie optymalne,
• uczące się roboty,
• gry planszowe,
• optymalizacja kombinatoryczna i szeregowanie.
Do najbardziej spektakularnych przykładów należy użycie uczenia się ze wzmocnieniem w
połączeniu z reprezentacją funkcji wartości za pomocą sieci neuronowej do gry w trik-traka
(backgammon): uzyskany w ten sposób program na podstawie własnej gry (ze sobą) doszedł do mistrzostwa (należy do kilku najlepszych graczy na świecie).

Uczenie się ze wzmocnieniem (reinforcement learning, RL) jest pod wieloma względami odmienne od innych form maszynowego uczenia się. W przeciwieństwie do klasyfikacji i regresji, jego celem nie jest aproksymowanie pewnego nieznanego odwzorowania przez generalizację na podstawie zbioru przykładów trenujących, chociaż wewnątrz systemów uczących się ze wzmocnieniem możemy łatwo odkryć wykorzystanie aproksymatorów. Jednak systemowi uczącemu się ze wzmocnieniem nie są dostarczane żadne przykłady trenujące, a jedynie wartościująca informacja trenująca, oceniające jego dotychczasową skuteczność.

U podstaw uczenia się ze wzmocnieniem leżą dynamiczne interakcje ucznia 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 rzeczywistoliczbowe 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 $\langle x_t, a_t, r_t, x_{t+1} \rangle$.