Multi Armed Bandits

Tuesday, 27 September 2022 12:03 PM

MAB

Conceptual problem used to choose next action to take when training model-free MDP algorithms

N actions labelled $X_1, X_2, \dots X_n$

Each play of action a at time t labelled $X_{a,t}$

Each action $X_1, X_2, \dots X_n$ follows a fixed probability distribution for rewards, and are independent from each other.

Each try of an action $X_{a,1}, X_{a,2}, \dots, X_{a,T}$ is is independent and identically distributed

However, we do not know the distribution or rewards for each action.

Goal: Maximise overall reward from applying a series of actions

We don't actually care about probability distributions, only care about expected reward for each action

$$E[X_a] = \frac{1}{N} \times \sum_{i=1}^{N} X_{a,i}$$

The more we play an action, the better estimate we have of $E[X]$

If we knew $E[X]$ for every action, best strategy is to always select action with highest $E[X]$

Since we don't have this information, we have to spend some actions trying different arms to learn their expected rewards.

Exploration vs Exploitation

Each try at an action should be to either gain more information (exploration)

Or to play action with best chance of maximising reward (exploitation)

More time spent exploring improves our estimate of which action is best, but means there is less time to use the information to maximise rewards

MAB strategies mostly have a parameter to determine this split

MAB Strategies

Epsilon Greedy

$\varepsilon \in [0,1]$ determines explore/exploit split

ε% of the time we randomly select an action to try (explore)

$1-\varepsilon$% of the time we perform action with highest expected reward so far.

After every action, observe reward and update expected reward $Q(a)$

$$Q(a) = Q(a) + \frac{1}{N(a)}[X_{a,k} - Q(a)]$$

Epsilon-Decreasing

$\varepsilon \in [0,1]$

$\alpha \in [0,1]$ decay of ε

Same as epsilon greedy with slight adjustment

After each action, update $\varepsilon \leftarrow \alpha \times \varepsilon$

Overtime, ε decreases to 0 (time spent exploring decreases)

Idea: As we get more information, no need to waste time exploring as much

larger α means ε decays slower

(spend more time exploring)

Softmax

$T > 0$ 'temperature' or effect of Q(a)

Assigns a probability of choosing each action

Weight of action $a = e^{\frac{Q(a)}{T}}$

Pr of choosing action $a = \frac{\text{weight of } a}{\text{Sum of all action weights}}$

larger value of T means probability of choosing each action is more uniform

Smaller value means probabilities weighted more by Q(a)

UCB 1 (Upper Confidence Bounds)

Next action chosen maximises the equation:

$$Q(a) + \sqrt{\frac{2 \ln t}{N(a)}}$$

t = number of rounds so far

Only works after every action is tried at least once

($N(a) > 0$)