Workshop 10

COMP90051 Statistical Machine Learning

Semester 2, 2024

Learning Outcomes

By the end of this workshop you should be able to:

1. Be able to implement epsilon-greedy multi-armed bandits
2. Be able to implement upper confidence bound multi-armed bandits
3. Be familiar with offline evaluation of MABs
4. Develop intuition about exploitation vs. exploration

COMP90051 Statistical Machine Learning

Stochastic MAB setting

- Possible actions $\{1, \ldots, k\}$ called "arms"
  - Arm i has distribution P_i on bounded rewards with mean μ_i
- In round $t = 1 \ldots T$
  - Play action $i_t \in \{1, \ldots, k\}$ *(possibly randomly)*
  - Receive reward $R_{i_t}(t) \sim P_{i_t}$
- Goal: minimise cumulative regret
  - $\mu^* T - \sum_{t=1}^{T} E\left[R_{i_t}(t)\right]$

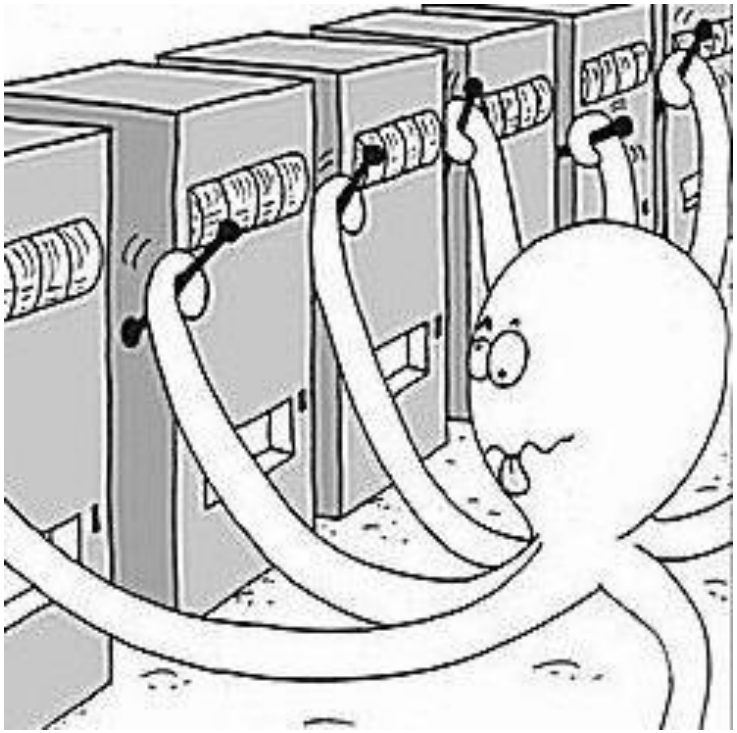
    Expected cumulative reward of bandit

    Best expected cumulative reward with hindsight

    where $\mu^* = \max_i \mu_i$
  - Intuition: Do as well as a rule that is simple but has knowledge of the future

Multi-armed bandits

each distribution
action has its reward



Open the mab/index.html for a simulation

$N_{t,k} = \sum_{I=1}^{t} \mathbb{1}[a_I = k]$ 计数k被选了多少遍

k: arm number
round t.

ε-Greedy exploit

: choose best arm with prob 1-ε otherwise

Worksheet 10

$|M|$ over n pulls

$\mu_i = \frac{n\mu + M_t}{n+1}$ new reward we observe