Temporal multi-armed bandits

Tom Rochette <tom.rochette@coreteks.org>

July 24, 2025 — daae079c

1 Variations

- With preference for recently chosen arm
  - Preferable when recency is useful, for instance when learning new material
- With preference with least recently chosen arm
  - Preferable when arms that haven't been picked in a while are more likely to provide a better reward
- With a limited memory/maximum of arms
  - The MAB algorithm is given a fixed amount of arms it can remember so it has to decide which ones it keeps in memory
  - Once the limit is reached, it has to forget one arm in order to explore a new one and keep its state (number of pulls and received reward)
  - Preferable when the options might expire if not selected within a given timeframe

2 Notes

- In some cases where you have a lot of arms to choose from and very few evaluation for each of them (less than 5), with many more arms which have never been chosen, you need a strategy to pick from those arms
  - One option is to use contextual bandits and provide some information you may have about the options you've never explored (e.g., the number of people who picked the arm in the past, the average reward)
    - In my use case, I'm trying to decide which book to read next. We can translate the previous example as the number of people who have read the book and the average rating of the book