# CMPUT 365: RL
# K-Armed Bandits

Rupam Mahmood

Jan 10, 2022

# Admin

- Due dates:
  - C1M1 practice quiz: Tomorrow
  - C1M1 graded notebook: This saturday

- Each week:
  - Focus on one Coursera module
  - Read the chapter
  - Watch the Coursera videos
  - Work on practice quiz (important for exams and written assignments)
  - Solve worksheet questions
  - Finish graded notebook

# Coursera videos: Video 1

- First introduction to problem of decision making

- Objects: actions, rewards, time steps or trials, values

写作 A 写作 R.

they are associate with some uncertainty

Can't appear as an input at the LHS 因为它不是一个单值，而是一个RV.

expectation. choosing a particular arm "a".

only takes a single value input, not a set value input. ⇒

$$q_\star(a) = E[R|A = a]$$

the reward of choosing a particular action "a".

Random variable A takes a particular outcome.

- Check [this extra video]() connecting bandits with probabilities we just learned

$$y = E[R|A=a] = \sum_r P(R=r|A=a)\,r$$

- 为了能够 decide which action is best, 就要 find the value of taking each action, 即 Action - Values. (也叫 Action - values function)

Defined the value of expecting an action as the expected reward: (goal: maximize the expected reward)

expected reward: $q_*(a) \doteq \mathbb{E}[R_t \mid A_t = a] \quad \forall a \in \{1, \dots, k\}$

is defined as

reward t

given selected action a

for each possible action 1 ~ k.

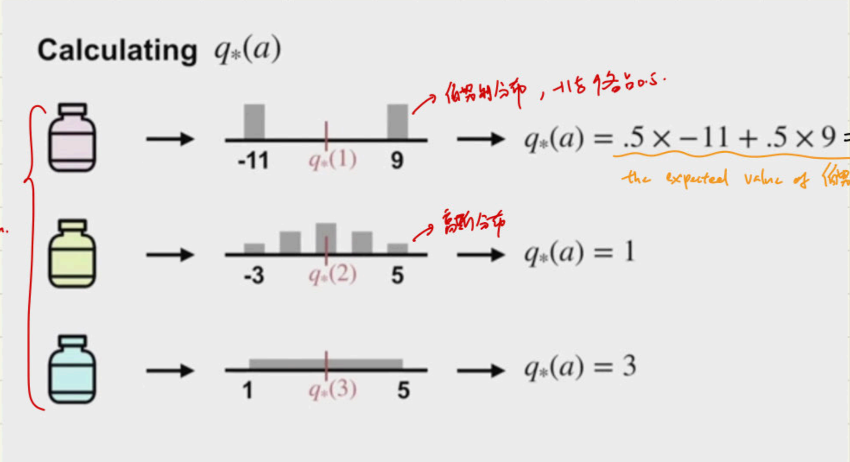
(Value of an action, not known)

$= \sum_r p(r|a)\, r$ ⇒ the sum of all possible rewards

(multiply possible reward with the probability of observing that reward).

If the agent (玩家) chose the action has the highest value, it achieve the goal. This procedure is called $\underset{a}{\text{Argmax}}\; q_*(a)$.

the mean of the distribution of each action.




伯努利分布, -11与9各占0.5.

the expected value of 伯努利 distribution

高斯分布

对应每一个 action 都有不同的 distribution.

# Video 2

- Estimating action values

The agent doesn't know $q$. But the agent knows $Q$. The agent wants to estimate $q$. ⇒

true quantity.

$$\underbrace{q_\star(a)}_{\text{value that we want to estimate.}} \doteq E[R|A=a] \approx \underbrace{Q_t(a)}_{\text{the estimator}} \doteq \boxed{\frac{\text{sum of rewards when } a \text{ taken prior to } t}{\text{number of times } a \text{ taken prior to } t}} = \frac{\sum_{i=1}^{t-1} R_i \cdot \mathbb{1}_{A_i=a}}{\sum_{i=1}^{t-1} \mathbb{1}_{A_i=a}}$$

a is constant.

Sample average.

the times

each time you get the different values.

随着每次尝试, estimator $Q(a)$ will get closer and closer to the true value $q_\star(a)$.

(the sample-average method to estimating $q_\star$).

(The estimator usually has the big letter. ∵它们 depend on $R_i$. giving a different value.)

- Why can't we choose actions greedily while estimating action values?

$$A_t \doteq \underset{a}{\operatorname{argmax}}\, \underline{Q_t(a)}$$

因为这里是个 random value.

∴ The greedy choice is bias in a sense.

shway to Everyone 1:37 PM

S It might be suboptimal but the agent does not know and it keeps choosing that suboptimal action greedily.

- If we had perfect estimates of action values, can we choose greedily?

# Video 3

- Estimating action values incrementally

$$Q_n \doteq \frac{R_1 + R_2 + \cdots + R_{n-1}}{n-1}$$

Cameron Jen to Everyone 1:40 PM

CJ So Qn = Estimated value before choosing the action for the n-th time right ? yes!!

即 $N_t(a)$, the number of times that each action has been taken

$$Q_{n+1} = Q_n + \frac{1}{n}\Big[R_n - Q_n\Big]$$

finding the recursive update rule for sample average.

- Incremental learning is more generally applicable to both stationary and nonstationary problems

can be sth else.

$$\text{NewEstimate} \leftarrow \text{OldEstimate} + \text{StepSize}\Big[\text{Target} - \text{OldEstimate}\Big]$$

$$Q_{n+1} \doteq Q_n + \alpha\Big[R_n - Q_n\Big]$$

constant stepsize. (不变的). ⇒ we can solve the bandit problem change over time.

若设变化的 stepsize, 则其应为 $\frac{1}{t-1}$.

# Video 4 & 5

- (4) The exploration-exploitation tradeoff
- epsilon-greedy, as a simple method to balance exploration and exploitation

$$A \leftarrow \begin{cases} \arg\max_a Q(a) & \text{with probability } 1-\varepsilon \quad \text{(breaking ties randomly)} \\ \text{a random action} & \text{with probability } \varepsilon \end{cases}$$

- (5) Optimistic initial values encourage early exploration
- Limitation of optimistic initialization: not well suited for nonstationary problems

greedy 方法不一定会导致最优解!! 因为有可能 agent 会错误的喜欢某个 meal 然后只会选它，导致他根本没有发现其他 meal 的价值.

## Exploration versus Exploitation

∴ 该选择 Explore (尝试点一个菜) 还是 Exploit? (点你喜欢的菜).

- **Exploration** - **improve** knowledge for **long-term** benefit ⇒ get more accurate estimate to our value
- **Exploitation** - **exploit** knowledge for **short-term** benefit ⇒ might get more reward

- **How do we choose when to explore and when to exploit?**

Epslion : 继续 explore 的概率.

# Life of a bandit agent

## A simple bandit algorithm

Initialize, for $a = 1$ to $k$:

$$Q(a) \leftarrow 0$$
$$N(a) \leftarrow 0$$

Loop forever:

tell us what action we take and what reward we receive.

$$A \leftarrow \begin{cases} \arg\max_a Q(a) & \text{with probability } 1 - \varepsilon \quad \text{(breaking ties randomly)} \\ \text{a random action} & \text{with probability } \varepsilon \end{cases}$$
$$R \leftarrow bandit(A)$$
$$N(A) \leftarrow N(A) + 1$$
$$Q(A) \leftarrow Q(A) + \boxed{\frac{1}{N(A)}}\left[R - Q(A)\right]$$

A constant.

# Video 6

- Upper confidence bound action-selection uses uncertainty in the estimates to drive exploration

Mohamed Ahmed to Everyone 1:50 PM

MA For example issues with using bigger e is once you explored a lot and found the better actions, bigger e would still prefer to explore more

epsilon，可能是explore的概率. 是介于0到1之间的数.

$$A_t \doteq \underset{a}{\operatorname{argmax}} \left[ Q_t(a) + c\sqrt{\frac{\ln t}{N_t(a)}} \right]$$

if only $Q_t$, you choose the greedy action.

但现在加了这个项. (代表 select a random action).

- $N_t(a)$ being in the denominator decreases $a$'s uncertainty estimate after being chosen

- $\ln t$ in the numerator increases the uncertainty estimates of actions that are not chosen

- Over time, increases become smaller but remains unbounded

# Slido questions

- Action value: 1,
- Step size: 1,
- Epsilon-greedy: 1,

Recording

You are viewing Rupam Mahmood's screen View Options

View

365w22-bandits - Google Slides

365w22-bandits

Course: CMPUT 365 LEC B1 - W

https://app.sli.do/event/7BdTJeEnLpUdpZsnu9B4Ff/embed/polls/2f84b06b-22d9-47ee-9f68-3dbbe76ab725

Inbox - ashique@ua... University of Albert... ___evernote - Goog... Rupam's Website CMPUT 340 Winter ... CMPUT 365 Winter ... Dictionary.com | Me... Edmonton, Alberta ... describing the expe... diary - Google Docs Fellows Intranet Other Bookmarks

The action with the highest estimated value is called

optimal action

greedy action → the largest true value.

Send

Voting as Anonymous

Acceptable Use - Privacy Policy

Talking:

From Sajjad Kavyani Baghbaderani...
greedy

Unmute Start Video 50 Participants Chat Share Screen Record Reactions Leave

Recording

You are viewing Rupam Mahmood's screen    View Options

365w22-bandits - Google Slides | 365w22-bandits | 365w22-bandits | Course: CMPUT 365 LEC B1 - W

https://app.sli.do/event/7BdTJeEnLpUdpZsnu9B4Ff/embed/polls/3c151163-b7e9-411b-8c2b-a2eaa566511f

Inbox - ashique@ua... | University of Albert... | ___evernote - Goog... | Rupam's Website | CMPUT 340 Winter ... | CMPUT 365 Winter ... | Dictionary.com | Me... | Edmonton, Alberta ... | describing the expe... | diary - Google Docs | Fellows Intranet | Other Bookmarks

A good step size for a non-stationary task is

- 1/t ⇒ 若选这个, 则 step size will be smaller and smaller. 且 Earlier values get more weight. 最后的结果必须是 the sample average $Q = \frac{R_1 + \dots + R_{n-1}}{n-1}$
- ✓ constant ⇒ you will be able to estimate the change from the beginning to the end

Send

Voting as Anonymous

Acceptable Use - Privacy Policy

Talking: Rupam Mahm

Unmute | Start Video | 50 Participants | Chat | Share Screen | Record | Reactions | Leave

Recording

You are viewing Rupam Mahmood's screen

View Options

View

365w22-bandits - Google Slides | 365w22-bandits | 365w22-bandits | 365w22-bandits | Course: CMPUT 365 LEC B1 - W

https://app.sli.do/event/7BdTJeEnLpUdpZsnu9B4Ff/embed/polls/85a2e82b-e83f-4803-8a4d-f6dbc2e57395

Inbox - ashique@ua... | University of Albert... | ___evernote - Goog... | Rupam's Website | CMPUT 340 Winter ... | CMPUT 365 Winter ... | Dictionary.com | Me... | Edmonton, Alberta ... | describing the expe... | diary - Google Docs | Fellows Intranet | Other Bookmarks

Is it possible that in an epsilon-greedy method, the action taken by the agent randomly could be the greedy action?

- [x] Yes ⇒ 有 $1-\varepsilon$ 的机率 是 $\arg\max_a Q(a)$. ∴ 当有 $\varepsilon$ 的机率选到 a random action 时, 该 action 依然有可能是它指在 $\arg\max_a Q(a)$ 里的.
- [ ] No ∴ probability of greedy action $= (1-\varepsilon) + \dfrac{\varepsilon}{\#\text{ of actions}}$

Send

Voting as Anonymous

Acceptable Use - Privacy Policy

Talking: Rupam Mahm

Unmute | Start Video | 49 Participants | Chat | Share Screen | Record | Reactions | Leave band-rev.mp4

Bandit problems are a sequence of two experiments.

3:37 PM Sun Jan 24 53%

Bandit Review

Untitled Notebook | https:%2Farmahm... | https:%2Farmahm... | https:%2Farmahm... | Bandit Review

# Bandits review

✓ What is the experiment? => the process the agent choosing actions.

✓ What are the outcomes?

✓ What are the random variables involved?

在Bandit问题中, we have an agent choosing actions and an environment providing rewards.

有two actions: $a \longrightarrow 1, 2$

gives rewards:

$b \longrightarrow 2, 3$

则此时experiment 为 {a, b}. 可写为 Random Variable : A for actions.

∴ 有 $P(A=a)$ 与 $P(A=b)$.

outcomes 为 {1, 2, 3}, the probability is depend on the action that was chosen. 写为 random variable R. ∴ 有 $P(R=1 \mid A=a)$

probability of R=1 base on what action was taken. 如 pick action b, 即 A=b, 则 $P(R=1 \mid A=b) = 0$

∴ 该 experiment 的 outcome 为 {(a,1), (a,2), (b,2), (b,3)}.

when we talk about probabilities here, we talk about probabilities with joint event: $P(A=a, R=1)$

$= P(R=1 \mid A=a)\, P(A=a)$

comes from the reward (R) process.

comes from the action process.

0:31 / 4:13

# Contextual Bandits

In Bandit problems, the agent chose from n actions, and one of them gives the highest reward: highest reward doesn't depend on the context.

Actions: O O O ⇒ The action that gives the highest reward

但 the best action depends on the context!!

如：给苹果，则 the best action is the first one : $\boxed{O}$ O O

给香蕉，则 the best action is the second one : O $\boxed{O}$ O

The context 可用一个 R.V. 来表示，称 $S$ as state. Depending on the state, the agent chose an action $A$ and receive the reward $R$.

∴ The whole experiment is: $P(\underbrace{S=s}_{\text{先让 state being presented}},\ \underbrace{A=a}_{\text{然后 an action being chosen by the agent}},\ \underbrace{R=r}_{\text{最后 a reward is received}})$ ⇒ a composition of 3 different events.

can be computed using the condition probability definition, 有:

$$\underbrace{P(R=r \mid S=s, A=a)}_{\text{the reward depends on the state and the action}} \times \underbrace{P(A=a \mid S=s)}_{\text{The action depends on the state}} \times \underbrace{P(S=s)}_{\text{probability of the environment choosing this state}}$$