CS 4/5789 30 March 2022 Prof Sarah Dean

Lecture 19: Contextual Bandits

1) Motivation – slides

2) Formal Setting

Simplified RL setting with simplified version of state: context. Unlike states, contexts are memoryless. They are drawn from a fixed distribution independent of previous context & actions.

$\mathcal{X}$: a set of contexts x

$\mathcal{A} = \{1, \dots, K\}$ a set of discrete actions

$\mathcal{D} \in \Delta(\mathcal{X})$: context distrubution $x_t \overset{iid}{\sim} \mathcal{D}$

$r: \mathcal{X} \times \mathcal{A} \to \Delta(\mathbb{R})$ noisy reward $r_t \sim r(x_t, a_t)$, depends on context & action $\mathbb{E}[r(x,a)] = \mu_a(x)$

T: integer time horizon

Notice that the context distribution $\mathcal{D}$ in some sense subsumes the transition probabilities P and the initial distribution μ_0

The actions should depend on the observed context, therefore policy $\pi: \mathcal{X} \to \mathcal{A}$ (or stochastic $\pi: \mathcal{X} \times \mathcal{A} \to [0,1]$)

The optimal policy maximizes cumulative reward at each step:

$$\pi^*(x) = \max_a \mu_a(x)$$

Q: why is it sufficient to consider each timestep independently? How is this different from full MDP?

Goal: Minimize expected Regret (in terms of cumulative reward)

$$R(T) = \sum_{t=1}^{T} \mathbb{E}_{x_t \sim \mathcal{D}}\Big[\underbrace{\max_a \mu_a(x_t)}_{\mu_{a^*}(x_t)} - \mu_{a_t}(x_t)\Big]$$

3) Naive (Tabular) Approach

Suppose there are M discrete possible contexts.

IDEA: run a separate MAB for each different context.

Instead of computing mean-per-arm $\hat{y}_a$, compute mean-per-arm-and-context

$$\hat{y}_a(x) = \sum_{k=1}^{t} r_t \mathbb{1}\{a_t = a\}\mathbb{1}\{x_t = x\} \Big/ \underbrace{\sum_{k=1}^{t} \mathbb{1}\{a_t=a\}\mathbb{1}\{x_t=x\}}_{\text{number of times } a \text{ pulled in context } x}$$

Algorithm: Explore-then-Commit with context:

For $t = 1, 2, \ldots, T$:

  Observe x_t

  If $\exists a$ s.t. # times a pulled in context $x_t \leq N$ } exploration

    $a_t = a$

  else: } exploitation

    $a_t = \arg\max_a \hat{y}_a(x_t)$

(exploration and exploitation: context dependent)

Algorithm: UCB with contexts:

for $t = 1, 2, \ldots, T$:

  pull $a_t = \arg\max_a \hat{y}_t^a(x_t) + \sqrt{\dfrac{\log(KTM/\delta)}{N_t^a(x_t)}}$

Keep track of confidence intervals for arm-context pairs.

This is similar to a classic MAB but with $M \cdot K$ arms.
(M: # contexts, K: #arms)

We can show similar regret bounds as previous lectures where $K \leftarrow MK$.

In some sense we are searching over all possible $M \cdot K$ policies (vs. K actions).

4) Function Approximation

In reality, contexts include many peices of information (e.g. demographic information, recent browsing behaviour, etc) and the number of discrete contexts may be very large! We may never see the exact same context twice!

However, it is also likely that correlations exist between similar contexts

ex: user 1: { gender: F, age: 22, major: CS } $= x_1$
user 2: { gender: M, age 21, major: econ } $= x_2$
user 3: { gender: F, age: 21, major: econ } $= x_3$

Information about user 1 & user 2 should help us predict for user 3.

Instead of estimating $\hat{\mu}_a(x)$ by counting, we can use function approximation:

$$\hat{\mu}_a(x) = \underset{\mu \in M}{\operatorname{argmin}} \sum_{k=1}^{t} (\mu(x_k) - r_k)^2 \mathbb{1}\{a_k = a\} \Big/ \sum_{k=1}^{t} \mathbb{1}\{a_k = a\}$$

function class

Q: How to get confidence intervals on $\hat{\mu}_a(x)$?
A: supervised learning guarantees:

<u>Lemma</u>: for $x_i \overset{iid}{\sim} D$, $\mathbb{E}[y_i] = f_*(x_i)$ for $f_* \in \mathcal{F}$,

$$\hat{f} = \underset{f \in \mathcal{F}}{\operatorname{argmin}} \sum_{i=1}^{N} (\hat{f}(x_i) - y_i)^2$$

we have that, with high probability,

$$\mathbb{E}_{x \sim D}\left[|\hat{f}(x) - f_*(x)| \right] \lesssim \sqrt{\frac{c_{\mathcal{F}}}{N}}$$

Algorithm: Explore-then-commit w/ fn. approx

1) pull each arm N times and record $\{\{x_i^a, r_i^a\}_{i=1}^N\}_{a=1}^K$

Estimate $\hat{y}_a(x) = \underset{y \in M}{\operatorname{argmin}} \sum_{i=1}^{N} (y(x_i^a) - r_i^a)^2$

2) For $t = NK+1, \dots, T$ pull $a_t = \underset{a}{\operatorname{argmax}}\ \hat{y}_a(x_t)$

Regret analysis:

$$R(T) \leq \underbrace{NK}_{\text{from explore}} + \sum_{NK+1}^{T} \mathbb{E}_{x_t \sim D}\left[y_{a^*}(x_t) - y_{a_t}(x_t)\right]$$

$$\mathbb{E}_{x_t \sim D}\left[y_{a^*}(x_t) - y_{a_t}(x_t)\right] = \mathbb{E}_{x_t \sim D}\Big[\left(y_{a^*}(x_t) - \hat{y}_{a^*}(x_t)\right) + \overbrace{\left(\hat{y}_{a^*}(x_t) - \hat{y}_{a_t}(x_t)\right)}^{\leq 0 \text{ since } a_t \text{ argmax}} + \left(\hat{y}_{a_t}(x_t) - y_{a_t}(x_t)\right)\Big]$$

$$\lesssim 2\sqrt{\frac{C_M}{N}}$$

Then $R(T) \lesssim NK + 2T\sqrt{\frac{C_M}{N}}$ very similar to non-contextual!

Set $N = \left(\frac{T}{2K}\sqrt{C_M}\right)^{2/3}$

so that $R(T) \lesssim T^{2/3}(KC_M)^{1/3}$

What about UCB type algorithm? Good confidence intervals require knowing conditional expected error

$$\mathbb{E}\left[|y_a(x) - \hat{y}_a(x)| \,\middle|\, x\right]$$

Next lecture: Linear contextual bandits & LinUCB

$$y_a(x) = \theta_a^T x$$