Contextual Bandits

Vipul Patil

Problem statement (theoretical)

Repeat:

1. Learner presented with context
2. Learner chooses an action
3. Learner observes reward (but only for chosen action)

Goal: learn to choose actions to maximize rewards

Problem Protocol: Contextual bandits

For each round $t \in [T]$:

1. algorithm observes a "context" x_t,
2. algorithm picks an arm a_t,
3. reward $r_t \in [0, 1]$ is realized.

- **Goal:** Maximize total reward: $\sum \mathbf{r_t}$
- Reward $\mathbf{r_t}$ in each round t depends both on the context $\mathbf{x_t}$ and the chosen action $\mathbf{a_t}$

Example

Improve user satisfaction by tailoring recommendations to specific user's need

- **Arms**: ads, movies, articles, or whatever is being recommended
- **Context**: user data and cookies, which can be utilized to predict their preferences.

Exploration versus Exploitation

- **Exploitation:** Pick choices that seem best based on past outcomes
- **Exploration:** Pick choices not yet tried out (or not tried enough)
- Exploitation has notions of “**being greedy**” and being “**short-sighted**”
- Too much Exploitation ⇒ Regret of missing unexplored “gems”
- Exploration has notions of “**gaining info**” and being “**long-sighted**”
- Too much Exploration ⇒ Regret of wasting time on “duds”

Examples

- **Restaurant Selection**

  Exploitation: Go to your favorite restaurant

  Exploration: Try a new restaurant

- **Online Banner Advertisement**

  Exploitation: Show the most successful advertisement

  Exploration: Show a new advertisement

- **Oil Drilling**

  Exploitation: Drill at the best known location

  Exploration: Drill at a new location

ϵ- Greedy/ Epoch-Greedy Algorithm

Explicit exploration and exploitation

On each round, choose action:

- According to “best” policy so far (with probability 1-ϵ)
- Uniformly at random (with probability ϵ)

Applications

Online advertising (personalized advertising):

- System aims to select the most relevant ad to display to a user based on their context (such as user demographics, browsing history, or current webpage content)

Healthcare Treatment Selection:

- Treatment decisions can be made based on patient characteristics and contextual factors.
- The bandit algorithm can learn from historical patient data to suggest the most suitable treatment options, considering individual patient needs and potential outcomes.

Dynamic Pricing:

- The system adapts pricing decisions based on contextual information such as customer segment, demand patterns, and market conditions.
- The bandit algorithm learns from past pricing experiments and user responses to optimize pricing for maximizing revenue or other desired objectives.