Warm-starting contextual bandits: robustly combining supervised and bandit feedback

Chicheng Zhang\textsuperscript{1}; Alekh Agarwal\textsuperscript{1}; Hal Daumé III\textsuperscript{1,2}; John Langford\textsuperscript{1}; Sahand Negahban\textsuperscript{3}

\textsuperscript{1}Microsoft Research, \textsuperscript{2}University of Maryland, \textsuperscript{3}Yale University
Warm-starting contextual bandits

• For timestep $t = 1, 2, \ldots T$:
  • Observe context $x_t$ with associated cost $c_t = (c_t(1), \ldots, c_t(K))$ from distribution $D$
  • Take an action $a_t \in \{1, \ldots K\}$
  • Receive cost $c_t(a_t) \in [0,1]$

• **Goal:** incur low cumulative cost: $\sum_{t=1}^{T} c_t(a_t)$
Warm-starting contextual bandits

• Receive warm-starting examples $S = \{(x, c)\} \sim W$

• For timestep $t = 1, 2, \ldots, T$:
  • Observe context $x_t$ with associated cost $c_t = (c_t(1), \ldots, c_t(K))$ from distribution $D$
  • Take an action $a_t \in \{1, \ldots, K\}$
  • Receive cost $c_t(a_t) \in [0, 1]$

• **Goal:** incur low cumulative cost: $\sum_{t=1}^{T} c_t(a_t)$
Warm-starting contextual bandits: motivation

• Some labeled examples often exist in applications, e.g.
  • News recommendation: editorial relevance annotations
  • Healthcare: historical medical records w/ prescribed treatments

• Leveraging historical data can reduce unsafe exploration
Warm-starting contextual bandits: motivation

• Some labeled examples often exist in applications, e.g.
  • News recommendation: editorial relevance annotations
  • Healthcare: historical medical records w/ prescribed treatments

• Leveraging historical data can reduce unsafe exploration

Key Challenge: $W$ may not be the same as $D$
  • Editors fail to capture users’ preferences
  • Medical record data from another population

How to utilize the warm-starting examples robustly and effectively?
Algorithm & performance guarantees

ARRoW-CB: iteratively finds the best relative weighting of warm-start and bandit examples to rapidly learn a good policy
Algorithm & performance guarantees

ARRoW-CB: iteratively finds the best relative weighting of warm-start and bandit examples to rapidly learn a good policy

• Theorem (informal):
Compared to algorithms that ignore $S$, * the regret of ARRoW-CB is
- never much worse (robustness)
- much smaller, if $W$ and $D$ are close enough, and $|S|$ is large enough

*S~W is the warm start data
Empirical evaluation

- 524 datasets from openml.org
- CDFs of normalized errors

Algorithm 1

Algorithm 2

% settings w/ error $\leq e$
Empirical evaluation

- 524 datasets from openml.org
- CDFs of normalized errors

- Moderate noise setting
- Algorithms:
  - ARRoW-CB,
  - Sup-Only,
  - Bandit-Only,
  - Sim-Bandit (uses both sources)
Empirical evaluation

- 524 datasets from openml.org
- CDFs of normalized errors

- Moderate noise setting
- Algorithms:
  - ARRoW-CB,
  - Sup-Only,
  - Bandit-Only,
  - Sim-Bandit (uses both sources)

Poster Thu #52