HUMAN MACHINE CO-LEARNING

Xiaojin Zhu
Bryan Gibson
KwangSung Jun
University of Wisconsin-Madison, USA
Guide human learning by machine learning

- Human is the boss
- Multi-Armed Bandit testbed
- Suggestions, more suggestions, and reverse psychology
- Speculations
General Brewster (PI): “Mr. Chairman, I need to make myself very clear. If we uplink now, Skynet will be in control of your military.”

“But you'll be in control of Skynet, right?”

“(pause) That is correct, sir.”

“Then do it.”
Human-Machine Co-Learning: Learning when human is the boss

- Not active learning: Human is not oracle
- Not computer tutoring: Machine does not know the world either
- Two learning systems interact. Goal: maximally help the human learner
To make things concrete…

- **Example:**
  - **World** = Multi-Armed Bandit (Whistler Restaurant Problem)
  - **Human** = user
  - **Machine learning** = smartphone

- **Demo**

\[ \text{reward} \sim P_A \quad \text{reward} \sim P_B \]
The truth

Mean $\mu_A = 35.2$

Mean $\mu_B = 50.5$
Let $x_1, \ldots, x_n$ be the rewards received in $n$ trials

Regret $n \mu^* - \sum_{i=1}^{n} x_i$ where $\mu^* = \max(\mu_A, \mu_B)$

Per-trial regret $\mu^* - \frac{1}{n} \sum_{i=1}^{n} x_i$

There is a rich literature in machine learning on optimal MAB strategies

- e.g., UCB1
Initialization: play each arm once

Repeat:

- Play arm \( \text{arg} \max_j x_j + \sqrt{\frac{2 \ln n}{n_j}} \)

- \( x_j \) is the average reward from arm \( j \) so far
- \( n_j \) is the number of times arm \( j \) has been played
- \( n \) is the overall number of plays

Regret \( O(\ln n) \)
UCB1-tuned

- Empirical enhancement
- Play arm
  \[
  \text{arg max } x_j + \sqrt{\ln n \min\left(\frac{1}{4}, V_j(n_j)\right)}
  \]
- Upper variance bound for arm \( j \) which is played \( s \) times in \( t \) trials:
  \[
  V_j(s) = \left(\frac{1}{s} \sum_{r=1}^{s} x_{jr}^2\right)^{-2} x_{js} - x_{js} + \sqrt{\frac{2 \ln t}{s}}
  \]
Machine good

- UCB1-tuned performance, averaged over 5000 sessions. Each session has 29 trials. Each trial has length 150.
There is also a rich psychology literature on human sub-optimal performance on MAB [e.g., Daw, O’Doherty, Dayan, Seymour, & Dolan 06; Lee, Zhang, Munro, & Steyvers 09; Acuna & Schrater 08]

Psychology experiment
- 28 undergrads
- 150 pulls each
Human bad

**per-trial regret**

![Graph showing per-trial regret](chart1)

**best-arm percentage**

![Graph showing best-arm percentage](chart2)
Co-Learning in MAB

Q: how can machine help?

- Machine suggests
- Human obeys, ignores, vetoes

UCB1-tune, something else?

Pull arm

Machine sees

reward

Human
Idea 1: Giving suggestions

- Demo

Your total score is: 132

I suggest you play machine B

[Agree] [Disagree]
Idea 1: Giving suggestions

- "Human": 28 subjects, "S": 27 subjects
Idea 2: Giving detailed suggestions

- Demo

Your total score is: 101

You have played machine A (B respectively) 2 (1) times, the sample mean is 48 (5), while the upper confidence bound of the true mean can be as high as 85 (57). I suggest you play machine A.

[Agree] [Disagree]
Idea 2: Giving detailed suggestions

Idea 3: Reverse psychology

- Let’s model humans
  - $A_i$: “agree” or “disagree” at iteration $i$
  - $x_i$: reward at iteration $i$
  - $S_i$: machine suggestion at iteration $i$

\[
P(A_i \mid A_{1:i-1}, x_{1:i-1}, S_{1:i-1}) \\approx P(A_i \mid A_{i-1})
\]

- Let $M_i$ be the true intention of UCB1

\[
S_i = \begin{cases} 
M_i & \text{if } P(A_i \mid A_{i-1}) \geq 1/2 \\
\neg M_i & \text{otherwise}
\end{cases}
\]
Idea 3: Reverse psychology

- Demo (always disagree)

  Your total score is: 1091

I suggest you play machine A

[Agree] [Disagree]
Idea 3: Reverse psychology

Speculations

- Multi-Armed Bandit with trembling hands?
- RL?
- Ethics
- What if humans do better than machines?
- Synergy?