Contextual Linear Bandit Problem & Applications

Feng Bi
Joon Sik Kim
Leiya Ma
Pengchuan Zhang

Contents

- Recap of last lecture
  - Multi-armed Bandit Problem / UCB1 Algorithm
- Application to News Article Recommendation
- Contextual Linear Bandit Problem
- LinUCB Algorithm
- Disjoint Model & General Model
- Derivation of LinUCB
- News Article Recommendation Revisited
- Conclusion

Last Lecture : Multi-armed Bandit Problem

- K actions (feature-free)
- Each action has an average reward (unknown): μ_k
- For t=1,…,T (unknown)
  - Choose an action a_t from {1, …, K} actions
  - Observe a random reward y_t, where y_t is bounded [0,1]
  - $E[y_t] = \mu_{a,t}$: Expected reward of action a_t
- Minimizing Regret: $$R = \sum_{t=1}^{T} [\mu^* - \mu_{a,t}]$$

Q. How to choose an action to minimize regret?

Last Lecture: UCB1 Algorithm

- At each iteration, choose the action with highest Upper Confidence Bound (UCB)

$$a_{t+1} = \arg\max_a \hat{\mu}_{a,t} + \sqrt{\frac{2 \ln t}{t_a}}$$

Exploitation Term   Exploration Term

- Regret Bound : with high probability, sublinear w.r.t T

#Actions

Time Horizon

$$R(T) = O\left(\frac{K}{\Delta} \ln T\right)$$

Gap between best & 2nd best

$\Delta = \mu_1 - \mu_2$

Application of
Multi-armed Bandit Problem?

News Article Recommendation

- Various algorithms for personalized recommendation
  - Collaborative filtering, Content-based filtering, Hybrid approaches
- But…
  - Web-based contents undergo frequent changes
  - Some users have no previous data to learn from (cold-start problem)
- Exploration vs. Exploitation
  - Need to gather more information about users with more trials
  - Optimize the article selection with past user experience

➔ Use multi-armed bandit setting

Article Recommendation in Feature-Free Bandit Setting

Users u_1 with age YOUNG and u_2 with age OLD

u_1

u_2

Retirement planning wishes vs. reality

The Player **Wizarding World of Harry Potter ride may conjure a new path for theme park rides**

Elon Musk: 198,000 Tesla Model 3 Orders Received in 24 Hours

Not tired yet: Warriors top Spurs for 72nd win, set up date with history

Article Recommendation in Feature-Free Bandit Setting

User u_1 with age YOUNG

u_1

Retirement planning wishes vs. reality

The Player **Wizarding World of Harry Potter ride may conjure a new path for theme park rides**

Elon Musk: 198,000 Tesla Model 3 Orders Received in 24 Hours

Not tired yet: Warriors top Spurs for 72nd win, set up date with history

Article Recommendation in Feature-Free Bandit Setting

User u_1 with age YOUNG

u_1

0.2	**Retirement planning wishes vs. reality**
0.4	The Player **Wizarding World of Harry Potter ride may conjure a new path for theme park rides**
0.5	**Elon Musk: 198,000 Tesla Model 3 Orders Received in 24 Hours**
0.8	**Not tired yet: Warriors top Spurs for 72nd win, set up date with history**

Article Recommendation in Feature-Free Bandit Setting

User u_1 with age YOUNG

u_1

0.2 **Retirement planning wishes vs. reality**

0.4 The Player **Wizarding World of Harry Potter ride may conjure a new path for theme park rides**

0.5 **Elon Musk: 198,000 Tesla Model 3 Orders Received in 24 Hours**

0.8 **Not tired yet: Warriors top Spurs for 72nd win, set up date with history**

Article Recommendation in Feature-Free Bandit Setting

User u_1 with age YOUNG

u_1

0.2 **Retirement planning wishes vs. reality**

0.4 The Player **Wizarding World of Harry Potter ride may conjure a new path for theme park rides**

0.5 **Elon Musk: 198,000 Tesla Model 3 Orders Received in 24 Hours**

0.8 **Not tired yet: Warriors top Spurs for 72nd win, set up date with history**

Article Recommendation in Feature-Free Bandit Setting

User u_2 with age OLD

u_2

0.2 **Retirement planning wishes vs. reality**

0.4 The Player **Wizarding World of Harry Potter ride may conjure a new path for theme park rides**

0.5 **Elon Musk: 198,000 Tesla Model 3 Orders Received in 24 Hours**

0.8 **Not tired yet: Warriors top Spurs for 72nd win, set up date with history**

Article Recommendation in Feature-Free Bandit Setting

User u_2 with age OLD

u_2

0.2 Retirement planning wishes vs. reality

0.4 The Player Wizarding World of Harry Potter ride may conjure a new path for theme park rides

0.5 Elon Musk: 198,000 Tesla Model 3 Orders Received in 24 Hours

0.8 Not tired yet: Warriors top Spurs for 72nd win, set up date with history

Article Recommendation in Feature-Free Bandit Setting

User u_2 with age OLD

???

u_2

0.2 **Retirement planning wishes vs. reality**

0.4 The Player **Wizarding World of Harry Potter ride may conjure a new path for theme park rides**

0.5 **Elon Musk: 198,000 Tesla Model 3 Orders Received in 24 Hours**

0.8 **Not tired yet: Warriors top Spurs for 72nd win, set up date with history**

Article Recommendation in Feature-Free Bandit Setting

User u_2 with age OLD

u_2

0.2 **Retirement planning wishes vs. reality**

0.4 **The Player Wizarding World of Harry Potter ride may conjure a new path for theme park rides**

0.5 **Elon Musk: 198,000 Tesla Model 3 Orders Received in 24 Hours**

0.8 **Not tired yet: Warriors top Spurs for 72nd win, set up date with history**

Contextual-Bandit Problem

- For t=1,…,T (unknown)
  - User u_t , set A_t of actions (a)
  - Feature vector (context) $\mathbf{x}_{t,a}$: summarizes both user u_t and action a
  - Based on previous results, choose a_t from A_t
  - Receive payoff r_{t,a_t}
  - Improve selection strategy with new observation set $(x_{t,a_t}, a_t, r_{t,a_t})$
- Minimizing Regret: $R(T) = \mathbf{E}\left[\sum_{t=1}^{T}\left(r_{t,a_t^*} - r_{t,a_t}\right)\right]$

Action with maximum expected payoff at time t

Difference?

- Contextual bandit problem becomes K-armed bandit problem when
  - The action set A_t is unchanged and contains K actions for all t
  - The user u_t (or the context) is the same for all t

- Also called context-free bandit problem

Article Recommendation in Feature-Free Bandit Setting

Users u_1 with age YOUNG and u_2 with age OLD

u_1

u_2

Retirement planning wishes vs. reality

The Player **Wizarding World of Harry Potter ride may conjure a new path for theme park rides**

Elon Musk: 198,000 Tesla Model 3 Orders Received in 24 Hours

Not tired yet: Warriors top Spurs for 72nd win, set up date with history

Article Recommendation in Contextual Linear Bandit Setting

Users u_1 with age YOUNG and u_2 with age OLD

$$\mathbf{x_1} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

$$\mathbf{x_2} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

θ_1 **Retirement planning wishes vs. reality**

θ_2 The Player **Wizarding World of Harry Potter ride may conjure a new path for theme park rides**

θ_3 **Elon Musk: 198,000 Tesla Model 3 Orders Received in 24 Hours**

θ_4 **Not tired yet: Warriors top Spurs for 72nd win, set up date with history**

Article Recommendation in Contextual Linear Bandit Setting

Linear Payoff = $x^T \boldsymbol{\theta}$

Users u_1 with age YOUNG and u_2 with age OLD

$$\mathbf{x_1} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

$$\mathbf{x_2} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

θ_1 **Retirement planning wishes vs. reality**

θ_2 The Player **Wizarding World of Harry Potter ride may conjure a new path for theme park rides**

θ_3 **Elon Musk: 198,000 Tesla Model 3 Orders Received in 24 Hours**

θ_4 **Not tired yet: Warriors top Spurs for 72nd win, set up date with history**

Article Recommendation in Contextual Linear Bandit Setting

Linear Payoff = $x^T \boldsymbol{\theta}$

Users u_1 with age YOUNG and u_2 with age OLD

$$\mathbf{x_1} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

$$\mathbf{x_2} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

[0.1 , 0.6] **Retirement planning wishes vs. reality**

[0.5, 0.1] The Player **Wizarding World of Harry Potter ride may conjure a new path for theme park rides**

[0.6 , 0.1] **Elon Musk: 198,000 Tesla Model 3 Orders Received in 24 Hours**

[0.9 , 0.2] **Not tired yet: Warriors top Spurs for 72nd win, set up date with history**

Article Recommendation in Contextual Linear Bandit Setting

Linear Payoff = $x^T \boldsymbol{\theta}$

Users u_1 with age YOUNG and u_2 with age OLD

$$\mathbf{x}_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

$$\mathbf{x}_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

[0.1 , 0.6]	**Retirement planning wishes vs. reality**
[0.5, 0.1]	The Player **Wizarding World of Harry Potter ride may conjure a new path for theme park rides**
[0.6 , 0.1]	**Elon Musk: 198,000 Tesla Model 3 Orders Received in 24 Hours**
[0.9 , 0.2]	**Not tired yet: Warriors top Spurs for 72nd win, set up date with history**

LinUCB Algorithm

- Assumption: the payoff model is linear
  - Most Intuitive thought: Linear Model
  - Advantage: Confidence interval computed efficiently in closed form

- Tempting to apply UCB on general contextual bandit problems
  - asymptotic optimality
  - strong regret bound

➔ Called LinUCB algorithm.

Contextual Bandit

For each trail t=1,2,3..., T

1. Observe environment $x_{t,a} \in \mathbb{R}^d$, i.e. user u_t a set of actions $\mathcal{A}_t$ and both their features
2. Choose an arm $a_t \in \mathcal{A}$ based on previous trails an receive payoff r_{t,a_t}.
3. Improve arm selection strategy with new observation $(\mathrm{x}_{t,a_t}, a_t, r_{t,a_t})$

Example: News Recommendation

For each time the news page is loaded t=1,2,3..., T

1. Arms or actions are the articles, which can be shown to the user. The environment could be user and article information.
2. If the aricle is clicked $r_{t,a_t} = 1$ otherwise 0.
3. Improve new article selection

Minimize expected regret, i.e

$$R_A(T) = \mathbb{E}\left[\sum_{t=1}^{T} r_{t,a_t^*}\right] - \mathbb{E}\left[\sum_{t=1}^{T} r_{t,a_t}\right]$$

Two Models

- For convenience exposition, first describe simpler form
  - Disjoint linear model
- Then consider the general case
  - hybrid model

➔ LinUCB is a generic contextual bandit algorithm which applies to applications other than personalized news article recommendation.

Linear Disjoint Model

- We assume the expected payoff of an arm a is linear in its *d*-dimentional feature $x_{t,a}$ with some unknown coefficient vector θ_a^*; namely for all t,

$$\mathbf{E}[r_{t,a}|\mathbf{x}_{t,a}] = \mathbf{x}_{t,a}^\top \boldsymbol{\theta}_a^*.$$

- The model is called disjoint because the parameters are not shared among different arms.

“Disjoint” in example

Article Recommendation

Users u_1 with age YOUNG and u_2 with age OLD

u_1

u_2

Retirement planning wishes vs. reality

The Player **Wizarding World of Harry Potter ride may conjure a new path for theme park rides**

Elon Musk: 198,000 Tesla Model 3 Orders Received in 24 Hours

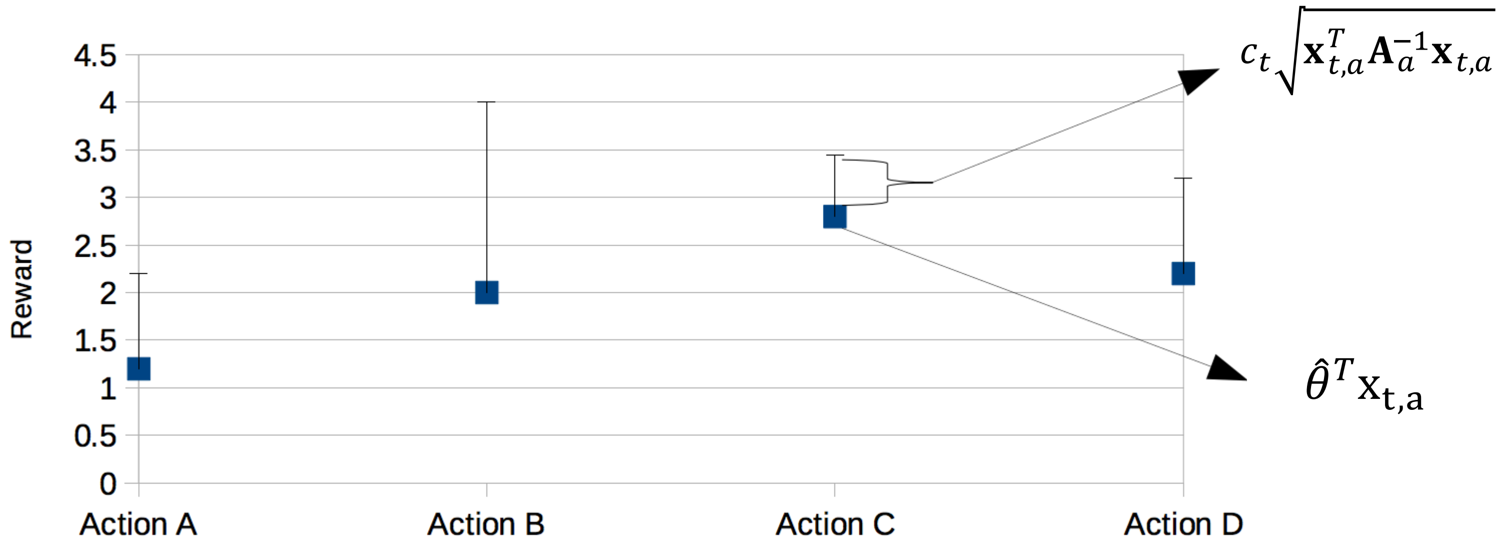
Not tired yet: Warriors top Spurs for 72nd win, set up date with history

Algorithm

Algorithm 1 LinUCB with disjoint linear models.

Inputs: $c_t \in \mathbb{R}_+$
for $t = 1, 2, 3, \ldots, T$ **do**
  Observe features of all arms $a \in \mathcal{A}_t$: $\mathbf{x}_{t,a} \in \mathbb{R}^d$
  for all $a \in \mathcal{A}_t$ **do**
    if a is new **then**
      $\mathbf{A}_a \leftarrow \mathbf{I}_d$ (d-dimensional identity matrix)
      $\mathbf{b}_a \leftarrow \mathbf{0}_{d \times 1}$ (d-dimensional zero vector)
    end if
    $\hat{\boldsymbol{\theta}}_a \leftarrow \mathbf{A}_a^{-1} \mathbf{b}_a$
    $p_{t,a} \leftarrow \hat{\boldsymbol{\theta}}_a^\top \mathbf{x}_{t,a} + c_t \sqrt{\mathbf{x}_{t,a}^T \mathbf{A}_a^{-1} \mathbf{x}_{t,a}}$
  end for
  Choose arm $a_t = \arg\max_{a \in \mathcal{A}_t} p_{t,a}$ with ties broken arbitrarily, and observe a real-valued payoff r_t
  $\mathbf{A}_{a_t} \leftarrow \mathbf{A}_{a_t} + \mathbf{x}_{t,a_t} \mathbf{x}_{t,a_t}^\top$
  $\mathbf{b}_{a_t} \leftarrow \mathbf{b}_{a_t} + r_t \mathbf{x}_{t,a_t}$
end for

Visualization Representation



Feature-free bandit v.s. linear bandit

Feature-free bandit

- $\mathbb{E}[r_{t,a}|x_{t,a}] = \mu_a^*$.
- μ_a^* is not known a priori.
- Confidence interval $C_{t,a}$

$$\{\mu_a : \frac{|\mu_a - \bar{\mu}_{t,a}|}{1/\sqrt{n_{t,a}}} \leq \sqrt{2\log t}\}$$

-

$$a_t = \arg\max_{a \in \{1,\ldots,K\}} \max_{\mu_a \in C_{t,a}} \mu_a$$

$$= \arg\max_{a \in \{1,\ldots,K\}} \bar{\mu}_{t,a} + \sqrt{\frac{2\log t}{n_{t,a}}}$$

Feature-free bandit v.s. linear bandit

Feature-free bandit

- $\mathbb{E}[r_{t,a}|x_{t,a}] = \mu_a^*$.
- μ_a^* is not known a priori.
- Confidence interval $C_{t,a}$

$$\{\mu_a : \frac{|\mu_a - \bar{\mu}_{t,a}|}{1/\sqrt{n_{t,a}}} \leq \sqrt{2\log t}\}$$

- $$a_t = \arg\max_{a\in\{1,\ldots,K\}} \max_{\mu_a\in C_{t,a}} \mu_a$$
$$= \arg\max_{a\in\{1,\ldots,K\}} \bar{\mu}_{t,a} + \sqrt{\frac{2\log t}{n_{t,a}}}$$

Linear bandit

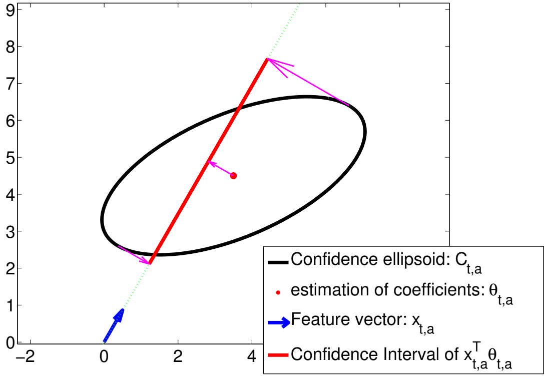
- $\mathbb{E}[r_{t,a}|x_{t,a}] = x_{t,a}^T\theta_a^*$.
- θ_a^* is not known a priori.
- Confidence ellipsiod $C_{t,a}$

$$\{\theta_a : \|\theta_a - \hat{\theta}_{t,a}\|_{A_{t,a}} \leq c_t\}$$

where $\|x\|_A \equiv \sqrt{x^TAx}$.

- $$a_t = \arg\max_{a\in\{1,\ldots,K\}} \max_{\theta_a\in C_{t,a}} x_{t,a}^T\theta_a$$
$$= \arg\max_{a\in\{1,\ldots,K\}} x_{t,a}^T\hat{\theta}_{t,a} + c_t\sqrt{x_{t,a}^TA_{t,a}^{-1}x_{t,a}}$$

Confidence ellipsiod $C_{t,a} = \{\theta_a : ||\theta_a - \hat{\theta}_{t,a}||_{A_{t,a}} \le c_t\}$



$$x_{t,a}^T \hat{\theta}_{t,a} + c_t \sqrt{x_{t,a}^T A_{t,a}^{-1} x_{t,a}} = \max_{\theta_a} \quad x_{t,a}^T \theta_a$$

$$\text{s.t.} \quad (\theta_a - \hat{\theta}_{t,a})^T A_{t,a} (\theta_a - \hat{\theta}_{t,a}) \le c_t$$

Feature-free bandit = Linear bandit with $x_{t,a} \equiv 1, \theta_a = \mu_a$

Feature-free bandit

- $\mathbb{E}[r_{t,a}|x_{t,a}] = 1^T \mu_a^*$.
- μ_a^* is not known a priori.
- Confidence interval $C_{t,a}$

  $$\{\mu_a : ||\mu_a - \bar{\mu}_{t,a}||_{n_{t,a}} \leq \sqrt{2 \log t}\}$$

  where $||\mu||_n \equiv \sqrt{\mu^T n \mu}$.
- $$a_t = \arg\max_{a \in \{1,\ldots,K\}} \max_{\mu_a \in C_{t,a}} 1^T \mu_a$$

  $$= \arg\max_{a \in \{1,\ldots,K\}} 1^T \bar{\mu}_{t,a} + \sqrt{\frac{2 \log t}{n_{t,a}}}$$

Linear bandit

- $\mathbb{E}[r_{t,a}|x_{t,a}] = x_{t,a}^T \theta_a^*$.
- θ_a^* is not known a priori.
- Confidence ellipsiod $C_{t,a}$

  $$\{\theta_a : ||\theta_a - \hat{\theta}_{t,a}||_{A_{t,a}} \leq c_t\}$$

  where $||x||_A \equiv \sqrt{x^T A x}$.
- $$a_t = \arg\max_{a \in \{1,\ldots,K\}} \max_{\theta_a \in C_{t,a}} x_{t,a}^T \theta_a$$

  $$= \arg\max_{a \in \{1,\ldots,K\}} x_{t,a}^T \hat{\theta}_{t,a} + c_t \sqrt{x_{t,a}^T A_{t,a}^{-1} x_{t,a}}$$

A Bayesian approach to derive $\hat{\theta}_{t,a}$ and $A_{t,a}^{-1}$

- Gaussian prior $p_0(\theta_a) \sim \mathcal{N}(0, \lambda I_d)$.

A Bayesian approach to derive $\hat{\theta}_{t,a}$ and $A_{t,a}^{-1}$

- Gaussian prior $p_0(\theta_a) \sim \mathcal{N}(0, \lambda I_d)$.
- $n_{t,a}$ noisy measurements: $\mathbf{y}_{t,a} \sim \mathcal{N}(\mathbf{D}_{t,a}\theta_a, I_{n_{t,a}})$.

$$\begin{bmatrix} \vdots \\ \mathbf{y}_{t,a}(i) \\ \vdots \end{bmatrix} = \begin{bmatrix} \vdots \\ x_{i,a}^T \\ \vdots \end{bmatrix} \theta_a + \begin{bmatrix} \vdots \\ \eta_{i,a} \\ \vdots \end{bmatrix}$$

A Bayesian approach to derive $\hat{\theta}_{t,a}$ and $A_{t,a}^{-1}$

- Gaussian prior $p_0(\theta_a) \sim \mathcal{N}(0, \lambda I_d)$.

- $n_{t,a}$ noisy measurements: $\mathbf{y}_{t,a} \sim \mathcal{N}(\mathbf{D}_{t,a}\theta_a, I_{n_{t,a}})$.

$$\begin{bmatrix} \vdots \\ \mathbf{y}_{t,a}(i) \\ \vdots \end{bmatrix} = \begin{bmatrix} \vdots \\ x_{i,a}^T \\ \vdots \end{bmatrix} \theta_a + \begin{bmatrix} \vdots \\ \eta_{i,a} \\ \vdots \end{bmatrix}$$

- Posterior distribution $p_{t,a}(\theta_a) \sim \mathcal{N}(\hat{\theta}_{t,a}, A_{t,a}^{-1})$.

$$\hat{\theta}_{t,a} = (\mathbf{D}_{t,a}^T \mathbf{D}_{t,a} + \frac{1}{\lambda} I_d)^{-1} \underbrace{\mathbf{D}_{t,a}^T \mathbf{y}_{t,a}}_{\mathbf{b_{t,a}}},$$

$$A_{t,a} = \mathbf{D}_{t,a}^T \mathbf{D}_{t,a} + \frac{1}{\lambda} I_d \,.$$

A Bayesian approach to derive $\hat{\theta}_{t,a}$ and $A_{t,a}^{-1}$

- Gaussian prior $p_0(\theta_a) \sim \mathcal{N}(0, \lambda I_d)$.
- $n_{t,a}$ noisy measurements: $\mathbf{y}_{t,a} \sim \mathcal{N}(\mathbf{D}_{t,a}\theta_a, I_{n_{t,a}})$.

$$\begin{bmatrix} \vdots \\ \mathbf{y}_{t,a}(i) \\ \vdots \end{bmatrix} = \begin{bmatrix} \vdots \\ x_{i,a}^T \\ \vdots \end{bmatrix} \theta_a + \begin{bmatrix} \vdots \\ \eta_{i,a} \\ \vdots \end{bmatrix}$$

- Posterior distribution $p_{t,a}(\theta_a) \sim \mathcal{N}(\hat{\theta}_{t,a}, A_{t,a}^{-1})$.

$$\hat{\theta}_{t,a} = (\mathbf{D}_{t,a}^T \mathbf{D}_{t,a} + \frac{1}{\lambda} I_d)^{-1} \underbrace{\mathbf{D}_{t,a}^T \mathbf{y}_{t,a}}_{\mathbf{b_{t,a}}},$$

$$A_{t,a} = \mathbf{D}_{t,a}^T \mathbf{D}_{t,a} + \frac{1}{\lambda} I_d .$$

- The reward $x_{t,a}^T \theta_a \sim \mathcal{N}(x_{t,a}^T \hat{\theta}_{t,a}, x_{t,a}^T A_{t,a}^{-1} x_{t,a})$. The upper confidence bound (UCB) is $x_{t,a}^T \hat{\theta}_{t,a} + c_t \sqrt{x_{t,a}^T A_{t,a}^{-1} x_{t,a}}$.

A general form of linear bandit

Disjoint linear model

- $\mathbb{E}[r_{t,a}|x_{t,a}] = x_{t,a}^T \theta_a^*$.
- $$a_t = \arg\max_{a \in \{1,\ldots,K\}} \max_{\theta_a \in C_{t,a}} x_{t,a}^T \theta_a$$

A general linear model

- $\mathbb{E}[r_t|x_t] = x_t^T \theta^*$.
- $$x_t = \arg\max_{x \in \mathcal{A}_t} \max_{\theta \in C_t} x^T \theta$$

A general form of linear bandit

Disjoint linear model

- $\mathbb{E}[r_{t,a}|x_{t,a}] = x_{t,a}^T \theta_a^*$.
-

$$a_t = \arg\max_{a \in \{1,\ldots,K\}} \max_{\theta_a \in C_{t,a}} x_{t,a}^T \theta_a$$

A general linear model

- $\mathbb{E}[r_t|x_t] = x_t^T \theta^*$.
-

$$x_t = \arg\max_{x \in \mathcal{A}_t} \max_{\theta \in C_t} x^T \theta$$

$$\theta = \begin{bmatrix} \theta_1 \\ \vdots \\ \theta_a \\ \vdots \\ \theta_K \end{bmatrix}, \quad \mathcal{A}_t = \left\{ \begin{bmatrix} \vdots \\ 0 \\ x_{t,a} \\ 0 \\ \vdots \end{bmatrix} : a = 1, 2, \ldots, K \right\}$$

A hybrid linear model

A hybrid linear model

- $\mathbb{E}[r_{t,a}|z_{t,a}, x_{t,a}] = z_{t,a}^T\beta^* + x_{t,a}^T\theta_a^*$.
-

$$a_t = \arg\max_{a\in\{1,\dots,K\}} \max_{\beta,\theta_a\in C_t} z_{t,a}^T\beta + x_{t,a}^T\theta_a$$

A general linear model

- $\mathbb{E}[r_t|x_t] = x_t^T\theta^*$.
-

$$x_t = \arg\max_{x\in\mathcal{A}_t} \max_{\theta\in C_t} x^T\theta$$

$$\theta = \begin{bmatrix} \beta \\ \theta_1 \\ \vdots \\ \theta_a \\ \vdots \\ \theta_K \end{bmatrix}, \quad \mathcal{A}_t = \left\{ \begin{bmatrix} z_{t,a} \\ \vdots \\ 0 \\ x_{t,a} \\ 0 \\ \vdots \end{bmatrix} : a = 1, 2, \dots, K \right\}$$

A general form of linear bandit, continued

Disjoint linear model

$$a_t = \arg\max_{a \in \{1,\ldots,K\}} \max_{\theta_a \in C_{t,a}} x_{t,a}^T \theta_a$$

-

$$C_{t,a} = \{\theta_a : ||\theta_a - \hat{\theta}_{t,a}||_{A_{t,a}} \leq c_t\}$$

-

$$\hat{\theta}_{t,a} = (\mathbf{D}_{t,a}^T \mathbf{D}_{t,a} + \frac{1}{\lambda} I_d)^{-1} \mathbf{D}_{t,a}^T \mathbf{y}_{t,a} \, ,$$

$$A_{t,a} = \mathbf{D}_{t,a}^T \mathbf{D}_{t,a} + \frac{1}{\lambda} I_d \, .$$

A general form of linear bandit, continued

Disjoint linear model

$$a_t = \arg\max_{a \in \{1,\ldots,K\}} \max_{\theta_a \in C_{t,a}} x_{t,a}^T \theta_a$$

- $$C_{t,a} = \{\theta_a : \|\theta_a - \hat{\theta}_{t,a}\|_{A_{t,a}} \le c_t\}$$

- $$\hat{\theta}_{t,a} = (\mathbf{D}_{t,a}^T \mathbf{D}_{t,a} + \frac{1}{\lambda} I_d)^{-1} \mathbf{D}_{t,a}^T \mathbf{y}_{t,a} ,$$
  $$A_{t,a} = \mathbf{D}_{t,a}^T \mathbf{D}_{t,a} + \frac{1}{\lambda} I_d .$$

A general linear model

$$x_t = \arg\max_{x \in \mathcal{A}_t} \max_{\theta \in C_t} x^T \theta$$

- $$C_t = \{\theta : \|\theta - \hat{\theta}_t\|_{A_t} \le c_t\}$$

- $$\hat{\theta}_t = (\mathbf{D}_t^T \mathbf{D}_t + \frac{1}{\lambda} I_d)^{-1} \mathbf{D}_t^T \mathbf{y}_t ,$$
  $$A_t = \mathbf{D}_t^T \mathbf{D}_t + \frac{1}{\lambda} I_d .$$

An $O(d\sqrt{T})$ regret bound

Theorem (Theorem 2 + Theorem 3 in APS_2011)

Assume that

1. *The measurement noise η_t is independent of everything and is σ-sub-Gaussian for some $\sigma > 0$, i.e., $\mathbb{E}[e^{\lambda\eta_t}] \le \exp(\frac{\lambda^2\sigma^2}{2})$ for all $\lambda \in \mathbf{R}$.*
2. *For all t and all $x \in \mathcal{A}_t$, $x^T\theta^* \in [-1, 1]$.*

Then, for any $\delta > 0$, with probability at least $1 - \delta$, for all $t \ge 0$,

1. *θ^* lies in the confidence ellipsoid*

$$C_t = \left\{\theta : \|\theta - \hat{\theta}_t\|_{A_t} \le c_t := \sigma\sqrt{\log\det A_t + d\log\lambda + 2\log\frac{1}{\delta}} + \frac{\|\theta^*\|}{\sqrt{\lambda}}\right\}$$

2. *The regret of the linUCB algorithm satisfies*

$$R_t = \underbrace{\sqrt{8t}}_{\text{I}}\ \underbrace{\sqrt{\log\det A_t + d\log\lambda}}_{\text{II}}\ \underbrace{\left(\sigma\sqrt{\log\det A_t + d\log\lambda + 2\log\frac{1}{\delta}} + \frac{\|\theta^*\|}{\sqrt{\lambda}}\right)}_{\text{III}:c_t}$$

An $O(d\sqrt{T})$ regret bound

Theorem (Theorem 2 + Theorem 3 in APS_2011)

Assume that

1. *The measurement noise η_t is independent of everything and is σ-sub-Gaussian for some $\sigma > 0$, i.e., $\mathbb{E}[e^{\lambda\eta_t}] \leq \exp(\frac{\lambda^2\sigma^2}{2})$ for all $\lambda \in \mathbf{R}$.*
2. *For all t and all $x \in \mathcal{A}_t$, $x^T\theta^* \in [-1, 1]$.*

Then, for any $\delta > 0$, with probability at least $1 - \delta$, for all $t \geq 0$,

1. *θ^* lies in the confidence ellipsoid*

$$C_t = \left\{\theta : \|\theta - \hat{\theta}_t\|_{A_t} \leq c_t := \sigma\sqrt{\log\det A_t + d\log\lambda + 2\log\frac{1}{\delta}} + \frac{\|\theta^*\|}{\sqrt{\lambda}}\right\}$$

2. *The regret of the linUCB algorithm satisfies*

$$R_t = \underbrace{\sqrt{8t}}_{\text{I}}\underbrace{\sqrt{\log\det A_t + d\log\lambda}}_{\text{II}}\underbrace{\left(\sigma\sqrt{\log\det A_t + d\log\lambda + 2\log\frac{1}{\delta}} + \frac{\|\theta^*\|}{\sqrt{\lambda}}\right)}_{\text{III}:c_t}$$

Lemma (Determinant-Trace Inequality, Lemma 10 in APS_2011)

If for all $t \geq 0$, $\|x_t\|_2 \leq L$ then

$$\log\det A_t \leq d\log(\frac{1}{\lambda} + \frac{tL^2}{d})$$

.

The ♥ of the proof

We consider the high probability event $\theta^* \in C_t$ for all $t \geq 0$.

$$
\begin{aligned}
r_t &= \langle x_t^*, \theta^* \rangle - \langle x_t, \theta^* \rangle && x_t, \tilde{\theta}_t = \arg\max_{x \in \mathcal{A}_t} \max_{\theta \in C_t} \langle x, \theta \rangle \\
&\leq \langle x_t, \tilde{\theta}_t \rangle - \langle x_t, \theta^* \rangle && \theta^* \in C_t \\
&= \langle x_t, \tilde{\theta}_t - \theta^* \rangle \\
&= \langle x_t, \hat{\theta}_t - \theta^* \rangle + \langle x_t, \tilde{\theta}_t - \hat{\theta}_t \rangle \\
&\leq \|x_t\|_{A_t^{-1}} \|\hat{\theta}_t - \theta^*\|_{A_t} + \|x_t\|_{A_t^{-1}} \|\tilde{\theta}_t - \hat{\theta}_t\|_{A_t} && \text{Cauchy-Schwarz} \\
&\leq 2c_t \|x_t\|_{A_t^{-1}} && \theta^*, \tilde{\theta}_t \in C_t = \{\theta : \|\theta - \hat{\theta}_t\|_{A_t} \leq c_t\}
\end{aligned}
$$

The ♥ of the proof

We consider the high probability event $\theta^* \in C_t$ for all $t \geq 0$.

$$\begin{aligned}
r_t &= \langle x_t^*, \theta^* \rangle - \langle x_t, \theta^* \rangle \qquad x_t, \tilde{\theta}_t = \arg\max_{x \in \mathcal{A}_t} \max_{\theta \in C_t} \langle x, \theta \rangle \\
&\leq \langle x_t, \tilde{\theta}_t \rangle - \langle x_t, \theta^* \rangle \qquad \theta^* \in C_t \\
&= \langle x_t, \tilde{\theta}_t - \theta^* \rangle \\
&= \langle x_t, \hat{\theta}_t - \theta^* \rangle + \langle x_t, \tilde{\theta}_t - \hat{\theta}_t \rangle \\
&\leq \|x_t\|_{A_t^{-1}} \|\hat{\theta}_t - \theta^*\|_{A_t} + \|x_t\|_{A_t^{-1}} \|\tilde{\theta}_t - \hat{\theta}_t\|_{A_t} \quad \text{Cauchy-Schwarz} \\
&\leq 2c_t \|x_t\|_{A_t^{-1}} \qquad \theta^*, \tilde{\theta}_t \in C_t = \{\theta : \|\theta - \hat{\theta}_t\|_{A_t} \leq c_t\}
\end{aligned}$$

Since $x^T \theta^* \in [-1, 1]$ for all $x \in \mathcal{A}_t$, then we have $r_t \leq 2$. Therefore,

$$r_t \leq \min\{2c_t \|x_t\|_{A_t^{-1}}, 2\} \leq 2c_t \min\{\|x_t\|_{A_t^{-1}}, 1\}$$

The ♥ of the proof, continued

$$r_t^2 \leq 4c_t^2 \min\{\|x_t\|^2_{A_t^{-1}}, 1\} \tag{1}$$

The ♥ of the proof, continued

$$r_t^2 \leq 4c_t^2 \min\{\|x_t\|^2_{A_t^{-1}}, 1\} \tag{1}$$

Consider the regret $R_T \equiv \sum_{t=1}^T r_t$,

$$R_T \leq \sqrt{T \sum_{t=1}^T r_t^2} \underbrace{\leq}_{\text{By (1)}} \sqrt{T \sum_{t=1}^T 4c_t^2 \min\{\|x_t\|^2_{A_t^{-1}}, 1\}}$$

$$\leq 2 \underbrace{\sqrt{T}}_{\text{I}} \underbrace{c_T}_{\text{III}} \underbrace{\sqrt{\sum_{t=1}^T \min\{\|x_t\|^2_{A_t^{-1}}, 1\}}}_{\text{II}}$$

c_t is monotonically increasing.

The ♥ of the proof, continued

$$r_t^2 \leq 4c_t^2 \min\{\|x_t\|^2_{A_t^{-1}}, 1\} \tag{1}$$

Consider the regret $R_T \equiv \sum_{t=1}^T r_t$,

$$R_T \leq \sqrt{T \sum_{t=1}^T r_t^2} \underbrace{\leq}_{\text{By (1)}} \sqrt{T \sum_{t=1}^T 4c_t^2 \min\{\|x_t\|^2_{A_t^{-1}}, 1\}}$$

$$\leq 2 \underbrace{\sqrt{T}}_{\text{I}} \underbrace{c_T}_{\text{III}} \underbrace{\sqrt{\sum_{t=1}^T \min\{\|x_t\|^2_{A_t^{-1}}, 1\}}}_{\text{II}} \qquad c_t \text{ is monotonically increasing.}$$

Since $x \leq 2\log(1+x)$ for $x \in [0,1]$, we have

$$\sum_{t=1}^T \min\{\|x_t\|^2_{A_t^{-1}}, 1\} \leq 2 \sum_{t=1}^T \log(1 + \|x_t\|^2_{A_t^{-1}}) = 2(\log\det A_t + d\log\lambda).$$

The last equality is proved in Lemma 11 in APS_2011.

Algorithm Evaluation

How do we evaluate the performance of a recommendation algorithm?

- Can we just run the algorithm on “live” data?

- Build a simulator to model the bandit process, evaluate the algorithm based on the simulated data?

Algorithm Evaluation

How do we evaluate the performance of a recommendation algorithm?

- Can we just run the algorithm on "live" data?

  Difficult logistically.

- Build a simulator to model the bandit process, evaluate the algorithm based on the simulated data? May introduce bias from the simulator.

- Yahoo! Today Module! (random article)

Algorithm Evaluation

Inputs: $T > 0$, algorithm π, stream of events

$h_0 = 0, \;\; R_0 = 0$

for $t = 1, 2, ..., T$ do

  repeat

    Get next event $(x_1, ..., x_K, a, r_a)$

  until $\pi(h_{t-1}, (x_1, ..., x_K)) = a$

  $h_t \leftarrow \text{add}(h_{t-1}, (x_1, ..., x_K, a, r_a))$

  $R_t \leftarrow R_{t-1} + r_a$

end for

Output R_t / T

Algorithm Evaluation

Inputs: $T > 0$, algorithm π, stream of events

$h_0 = 0, \;\; R_0 = 0$

for $t = 1, 2, ..., T$ do

  repeat

    Get next event $(x_1, ..., x_K, a, r_a)$

  until $\pi(h_{t-1}, (x_1, ..., x_K)) = a$

  $h_t \leftarrow \text{add}(h_{t-1}, (x_1, ..., x_K, a, r_a))$

  $R_t \leftarrow R_{t-1} + r_a$

end for

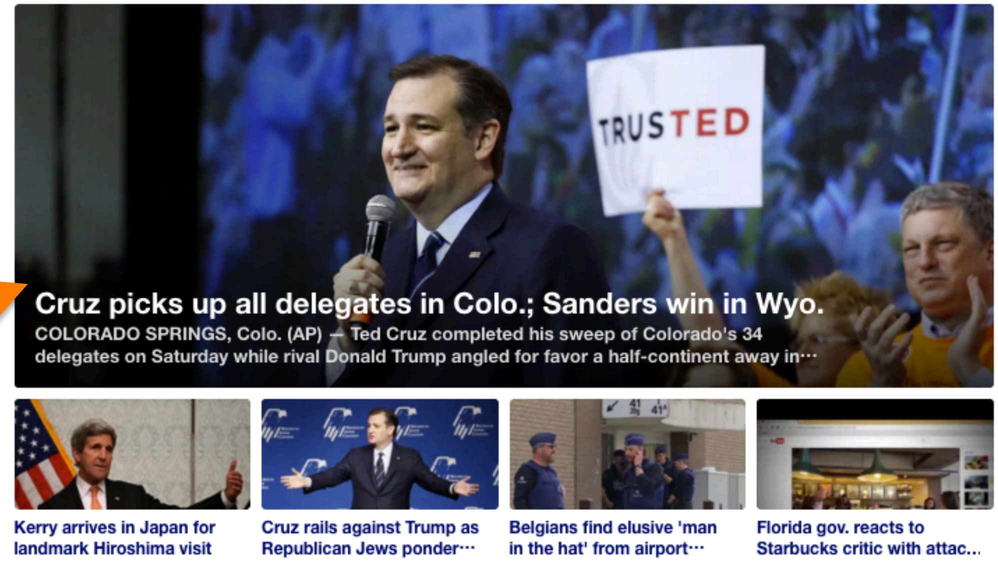
Output R_t / T

No bias! About T events are accepted from TK trails.

Algorithm Evaluation (data collection)

Yahoo News …

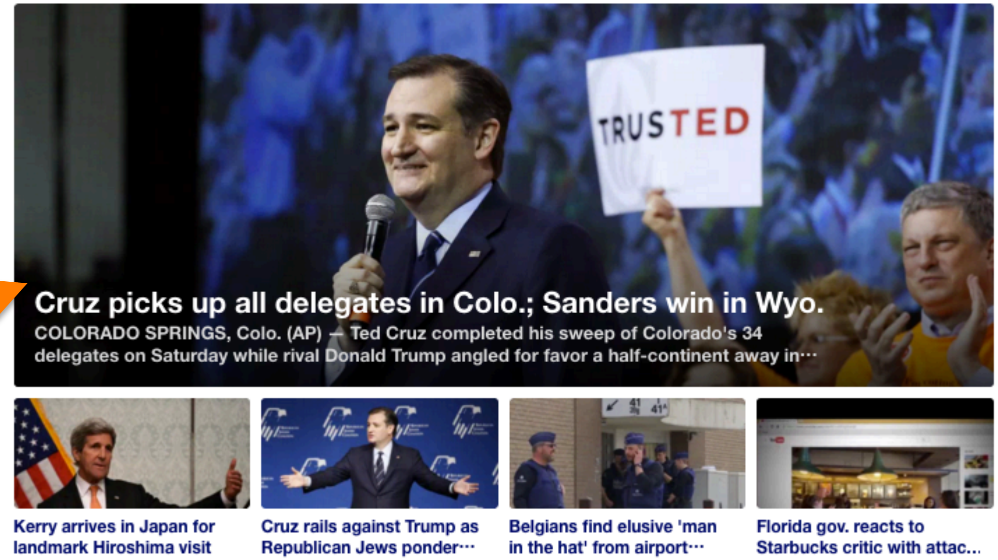
Randomly shoot user an article *a* as highlighted news.



Algorithm Evaluation (data collection)

Yahoo News …

Randomly shoot user an article *a* as highlighted news.



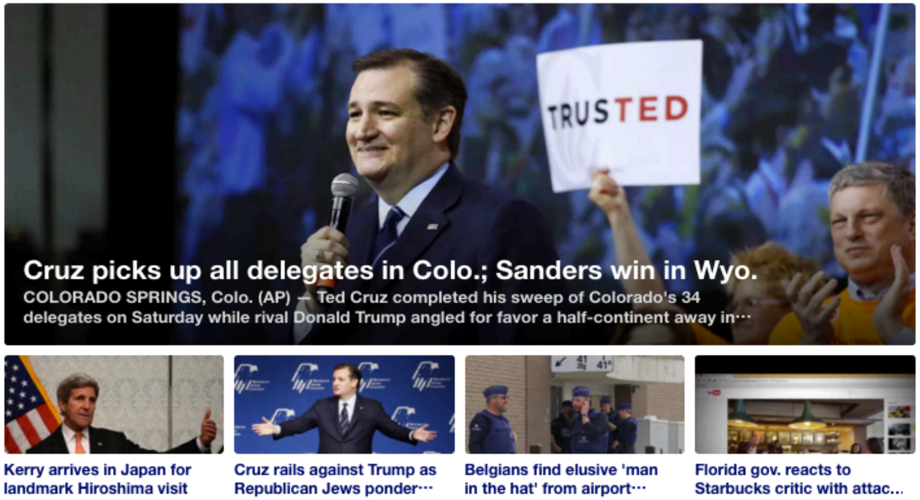
This event contains article a, feature vector X, and response r = 1/0 .

Accept this event iff algorithm predicts the same article a.

Algorithm Evaluation (construct features)

In Yahoo's data base, either a user or article is depicted by hundreds raw features.

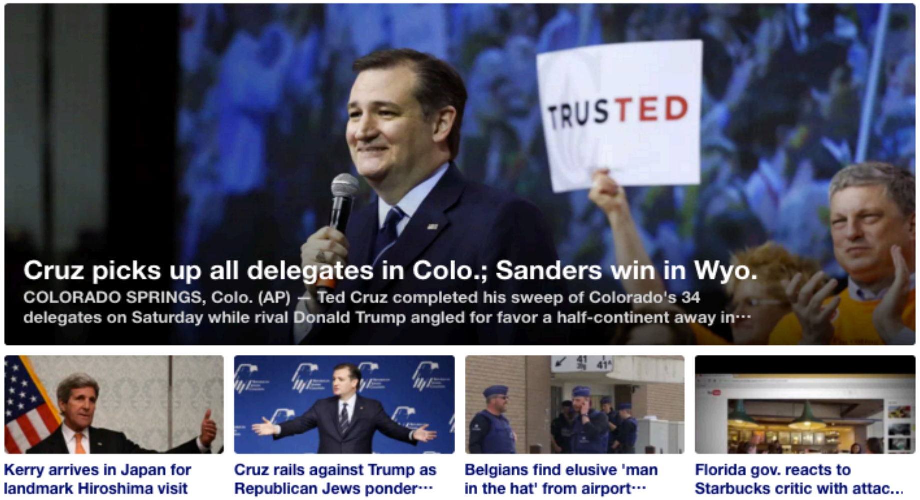
Need to reduce the feature dimensions.



Algorithm Evaluation (construct features)

In Yahoo's data base, either a user or article is depicted by hundreds raw features.

Need to reduce the feature dimensions.



Raw ϕ_a

Article	Long	Domestic	Tech	Politics	...
ϕ_a	1	1	0	1	...

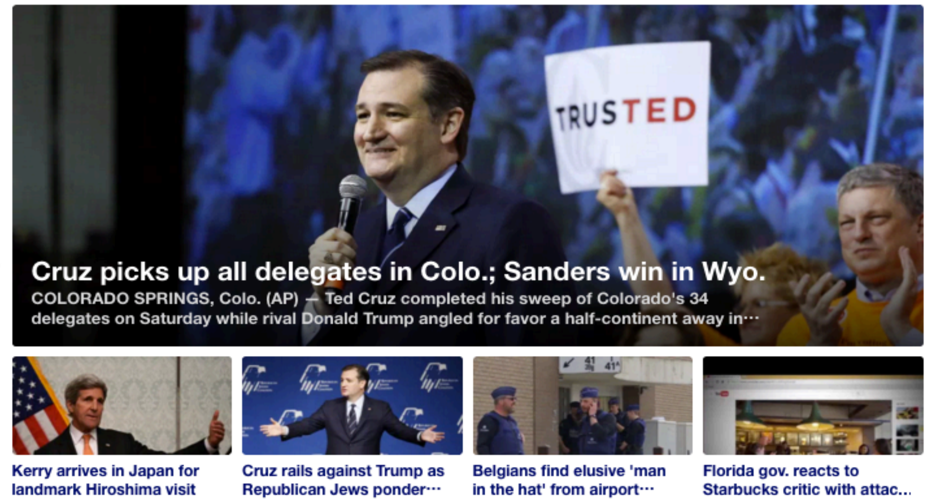
Raw ϕ_u

User	Gender	Age>20	Age>40	Student	...
ϕ_u	1	1	0	1	...

Algorithm Evaluation (construct features)

In Yahoo's data base, either a user or article is depicted by hundreds raw features.

Need to reduce the feature dimensions.



Raw ϕ_a

Article	Long	Domestic	Tech	Politics	...
ϕ_a	1	1	0	1	...

Raw ϕ_u

User	Gender	Age>20	Age>40	Student	...
ϕ_u	1	1	0	1	...

Suppose there's a weight matrix W, st. the probability of user clicking on article a is:

$$P = \phi_u^T W \phi_a$$

Algorithm Evaluation (construct features)

$$P = \phi_u^T W \phi_a$$

Logistic Regression to get *W*

K-means method to find clusters/groups of the users.

$\psi_u \equiv \phi_u^T W$ (Cluster this projected feature vector.)

Algorithm Evaluation (construct features)

$$P = \phi_u^T W \phi_a$$

Logistic Regression to get *W*

K-means method to find clusters/groups of the users.

$\psi_u \equiv \phi_u^T W$ (Cluster this projected feature vector.)

- The constructed feature vector for a user would be the possibilities of being in different groups.
  Denote as: $x_{t,a}$
- The same procedure can be applied to the article and get the constructed feature vector.

- Disjointed LinUCB, use $x_{t,a}$ as input data.
- Hybrid model, the outer product of constructed user and article feature vectors is also included as global features.

Algorithm Evaluation (construct features)

For example, we have binary raw feature vectors:

$$\phi_a = (1, 1, 0, 1, 1, ...) \qquad \phi_u = (1, 1, 0, 1, 0, ...)$$

Algorithm Evaluation (construct features)

For example, we have binary raw feature vectors for user and article:

$\phi_a = (1, 1, 0, 1, 1, ...)$ $\phi_u = (1, 1, 0, 1, 0, ...)$

- After clustering:

User, $x_{t,a}$

Group	A	B	C	D	E
Membership	0.2	0.1	0.35	0.3	0.05

Article

Group	1	2	3	4	5
Membership	0.7	0.05	0.15	0.05	0.05

Algorithm Evaluation (construct features)

For example, we have binary raw feature vectors for user and article:

$$\phi_a = (1, 1, 0, 1, 1, \ldots) \qquad \phi_u = (1, 1, 0, 1, 0, \ldots)$$

- After clustering:

User, $x_{t,a}$

Group	A	B	C	D	E
Membership	0.2	0.1	0.35	0.3	0.05

Article

Group	1	2	3	4	5
Membership	0.7	0.05	0.15	0.05	0.05

- The <u>outer product</u> of the two constructed feature vectors is: $z_{t,a}$ (25 dimensional here)

- Hybrid model: $E[r_{t,a}|x_{t,a}] = z_{t,a}^T \beta^* + x_{t,a}^T \theta_a^*$

Remove this term will get back to disjointed LinUCB

Algorithm Evaluation

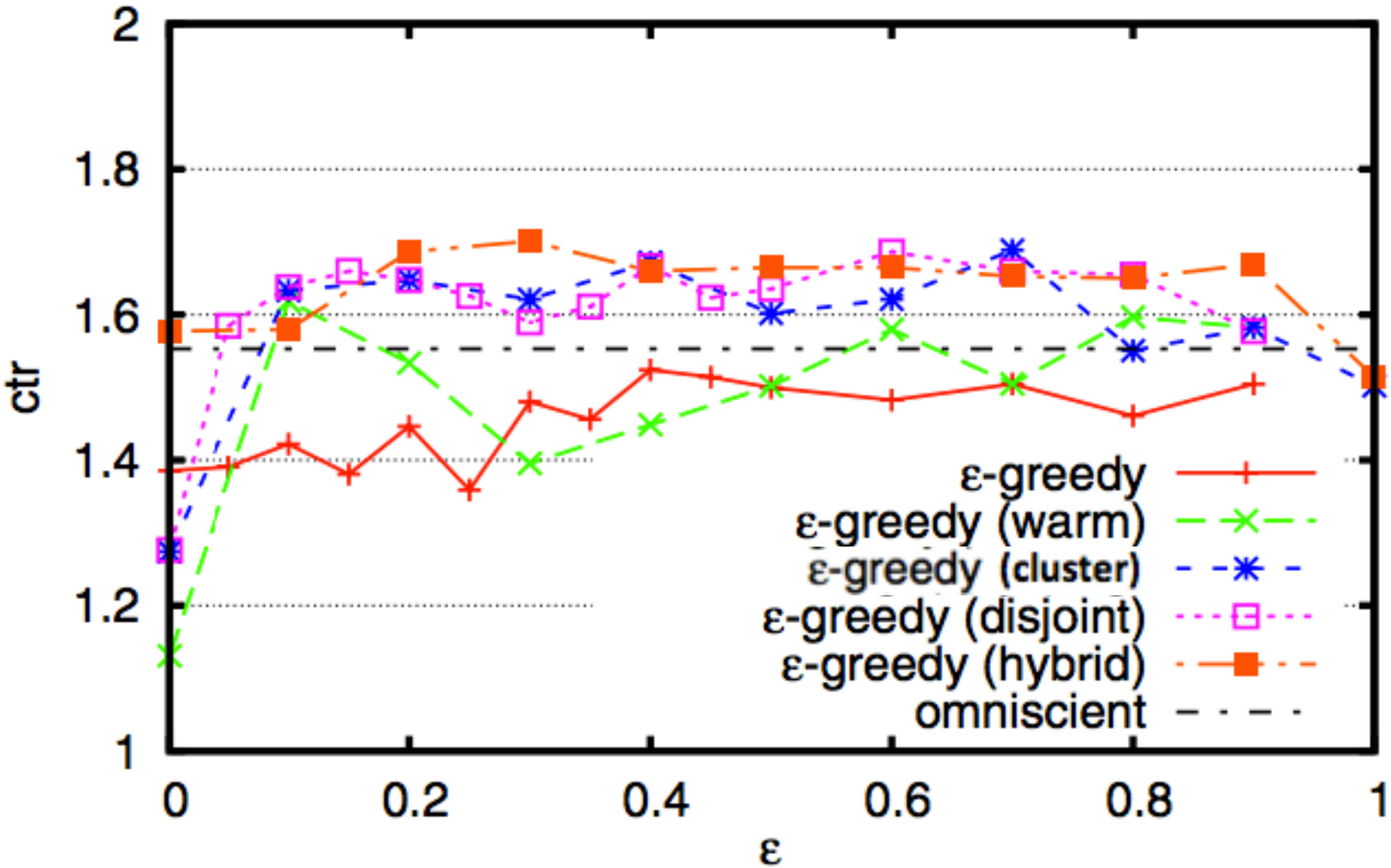
Without utilizing features:

- Purely Random
- E-greedy
- UCB
- Omniscient

Algorithms with features:

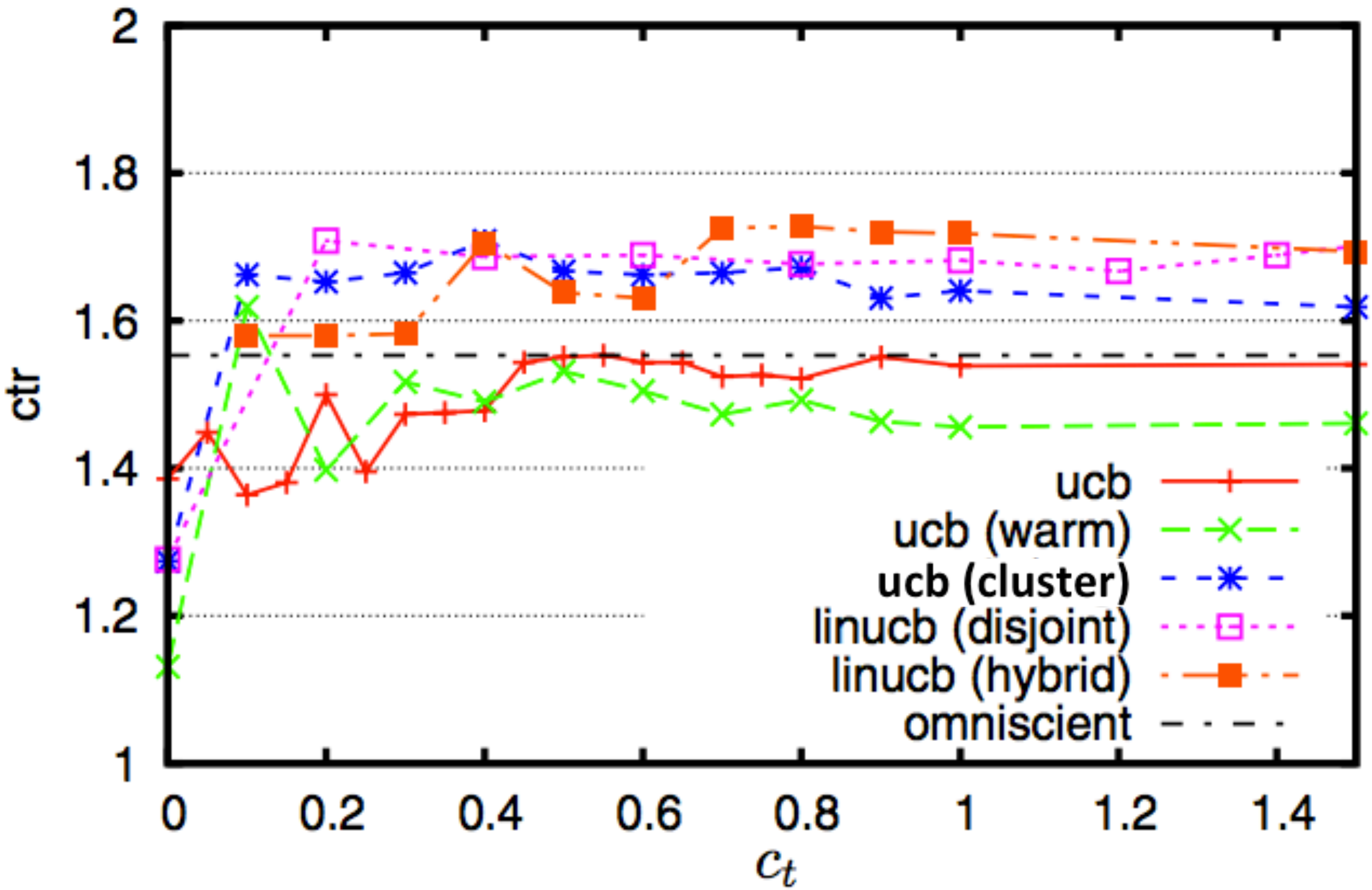
- E-greedy (run on different clusters)
- E-greedy (hybrid, epoch greedy)
- UCB (cluster)
- LinUCB (disjoint)
- LinUCB (hybrid)

Algorithm Evaluation



(CTR normalized by the random recommendation CTR)

Algorithm Evaluation



(CTR normalized by the random recommendation CTR)

Conclusion

For multi-armed bandits problem

- UCB algorithm without feature has regret bound:

$$R_T = O(\frac{K}{\epsilon} \ln T)$$

- LinUCB using feature vectors has regret bound:

$$R_T = O(D\sqrt{T})$$

- Evaluate using Yahoo Front Page Today Module data.
- Introducing contextual information (features) to the recommendation algorithm, the CTR (reward) has been improved by about 10%.

References

- Li, Lihong, et al. "A contextual-bandit approach to personalized news article recommendation." *Proceedings of the 19th international conference on World wide web*. ACM, 2010.
- Abbasi-Yadkori, Yasin, Dávid Pál, and Csaba Szepesvári. "Improved algorithms for linear stochastic bandits." *Advances in Neural Information Processing Systems*. 2011.
- Auer, Peter, Nicolo Cesa-Bianchi, and Paul Fischer. "Finite-time analysis of the multiarmed bandit problem." *Machine learning* 47.2-3 (2002): 235-256.
- Auer, Peter. "Using confidence bounds for exploitation-exploration trade-offs." *The Journal of Machine Learning Research* 3 (2003): 397-422.