Contextual bandit models for personalized recommendation

Emilie Kaufmann

ALICIA meeting, May 12^{th}, 2014

Outline

1. Contextual bandit models
2. Algorithms for contextual linear bandits
3. A recommendation system with binary responses
4. Further challenges

Outline

Classical versus contextual bandits

Classical bandit model:

- K actions: action $a \leftrightarrow$ distribution ν_a with mean μ_a
- action $A_t \in \{1, \dots, K\}$ is chosen at time t
- rewards $r_t \sim \nu_{A_t}$ is observed:

$$r_t = \mu_{A_t} + \epsilon_t$$

(ϵ_t some centered noise)

- best action

$$a^* = \underset{a=1\dots K}{\text{argmax}}\ \mu_a$$

Classical versus contextual bandits

Classical bandit model:

- K actions: action $a \leftrightarrow$ distribution ν_a with mean μ_a
- action $A_t \in \{1, \dots, K\}$ is chosen at time t
- rewards $r_t \sim \nu_{A_t}$ is observed:

$$r_t = \mu_{A_t} + \epsilon_t$$

(ϵ_t some centered noise)

- best action

$$a^* = \underset{a=1\dots K}{\text{argmax}}\ \mu_a$$

Contextual bandit model :

- set $\mathcal{D}_t \subset \mathbb{R}^d$ of contexts *available at time t*
- context $x_t \in \mathcal{D}_t$ is chosen at time t
- reward r_t is observed:

$$r_t = f(x_t) + \epsilon_t$$

(f unknown function)

- best context *at time t*

$$x_t^* = \underset{x \in \mathcal{D}_t}{\text{argmax}}\ f(x)$$

Contextual bandits for recommendation

- At time t, a user (with some features) arrives on a website

Contextual bandits for recommendation

- At time t, a user (with some features) arrives on a website
- Several items/adds (with some features) could be presented to him

Contextual bandits for recommendation

- At time t, a user (with some features) arrives on a website
- Several items/adds (with some features) could be presented to him
- For each of these items a context $x \in \mathbb{R}^d$ is build according to features of the user, the item, the webpage on which it is displayed (etc.), forming the set $\mathcal{D}_t$

Contextual bandits for recommendation

- At time t, a user (with some features) arrives on a website
- Several items/adds (with some features) could be presented to him
- For each of these items a context $x \in \mathbb{R}^d$ is build according to features of the user, the item, the webpage on which it is displayed (etc.), forming the set $\mathcal{D}_t$
- if context x_t is chosen, the associated item is presented to the user

Contextual bandits for recommendation

- At time t, a user (with some features) arrives on a website
- Several items/adds (with some features) could be presented to him
- For each of these items a context $x \in \mathbb{R}^d$ is build according to features of the user, the item, the webpage on which it is displayed (etc.), forming the set $\mathcal{D}_t$
- if context x_t is chosen, the associated item is presented to the user
- a response (reward) r_t is collected, that depends on x_t:

$$r_t = f(x_t) + \epsilon_t$$

examples: time spent on the website, amount of money spent, binary response indicating a click or a conversion...

What assumptions on f?

$$r_t = f(x_t) + \epsilon_t$$

For some $\theta \in \mathbb{R}^d$, one can assume:

- $f(x) = \theta^T x$ (linear bandits)
- $f(x) = \mu(\theta^T x)$ (*generalized linear bandits*)

Measure of performance: the regret

Let $\theta \in \mathbb{R}^d$. For $x_t \in \mathcal{D}_t$ a context chosen at time t, one observes

$$r_t = \theta^T x_t + \epsilon_t$$

Classical MAB:

- *Pseudo regret* of an algorithm:

$$\mathcal{R}(T, \mathcal{A}) = \sum_{t=1}^{T} (\mu_{a^*} - \mu_{A_t})$$

- Known results: there exists algorithms s.t.

$$\mathbb{E}[\mathcal{R}(T, \mathcal{A})] \quad = \quad O(\sqrt{KT})$$

Measure of performance: the regret

Let $\theta \in \mathbb{R}^d$. For $x_t \in \mathcal{D}_t$ a context chosen at time t, one observes

$$r_t = \theta^T x_t + \epsilon_t$$

Classical MAB:

- *Pseudo regret* of an algorithm:

$$\mathcal{R}(T, \mathcal{A}) = \sum_{t=1}^{T} (\mu_{a^*} - \mu_{A_t})$$

- Known results: there exists algorithms s.t.

$$\mathbb{E}[\mathcal{R}(T, \mathcal{A})] = O(\sqrt{KT})$$

Contextual linear bandit model:

- *Pseudo regret* of an algorithm

$$\mathcal{R}(T, \mathcal{A}) = \sum_{t=1}^{T} (\theta^T x_t^* - \theta^T x_t)$$

- Known results: there exists algorithms with pseudo-regret of order

$$O(d\sqrt{T}) \quad \text{or} \quad O(\sqrt{dT \log(K)})$$

Measure of performance: the regret

Let $\theta \in \mathbb{R}^d$. For $x_t \in \mathcal{D}_t$ a context chosen at time t, one observes

$$r_t = \theta^T x_t + \epsilon_t$$

Classical MAB:

- *Pseudo regret* of an algorithm:

$$\mathcal{R}(T, \mathcal{A}) = \sum_{t=1}^{T} (\mu_{a^*} - \mu_{A_t})$$

- Known results: there exists algorithms s.t.

$$\mathbb{E}[\mathcal{R}(T, \mathcal{A})] = O(\sqrt{KT})$$

Contextual linear bandit model:

- *Pseudo regret* of an algorithm

$$\mathcal{R}(T, \mathcal{A}) = \sum_{t=1}^{T} (\theta^T x_t^* - \theta^T x_t)$$

- Known results: there exists algorithms with pseudo-regret of order

$$O(d\sqrt{T}) \quad \text{or} \quad O(\sqrt{dT \log(K)})$$

Question: How should x_t be chosen at each round to minimize regret?

Optimistic algorithms for linear bandits

- First step: Build a set of statistically plausible models, i.e. a **confidence region** for θ.

Optimistic algorithms for linear bandits

- First step: Build a set of statistically plausible models, i.e. a **confidence region** for θ.

$$R_t = X_t^T \theta + E_t,$$

with

$$X_t = \begin{pmatrix} x_1^T \\ x_2^T \\ . \\ x_t^T \end{pmatrix} \in \mathcal{M}_{t,d}(\mathbb{R}), \quad R_t = \begin{pmatrix} r_1 \\ r_2 \\ . \\ r_t \end{pmatrix} \in \mathbb{R}^t, \quad E_t = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ . \\ \epsilon_t \end{pmatrix} \in \mathbb{R}^t.$$

The regularized least-square estimate of θ at time t is

$$\hat{\theta}(t) = (B(t))^{-1} X_t^T R_t \quad \text{with} \quad B(t) = \lambda I_d + X_t^T X_t$$

For a suited exploration rate $\beta(t, \delta)$ (and assumptions on the noise),

$$C_t = \left\{ \theta' \in \mathbb{R}^d : ||\hat{\theta}(t) - \theta'||_{B(t)} \leq \beta(t, \delta) \right\}$$

satisfies $\mathbb{P}(\forall t \in \mathbb{N}, \theta \in C_t) \geq 1 - \delta$.

Optimistic algorithms for linear bandits

- Second step: Acts as if the model were the one leading to the best possible outcome among all the statistically plausible models

$$x_{t+1} = \underset{x \in \mathcal{D}_{t+1}}{\text{argmax}} \max_{\theta' \in C_t} x^T \theta'.$$

Optimistic algorithms for linear bandits

- Second step: Acts as if the model were the one leading to the best possible outcome among all the statistically plausible models

$$x_{t+1} = \underset{x \in \mathcal{D}_{t+1}}{\text{argmax}} \max_{\theta' \in C_t} x^T \theta'.$$

With the above region C_t, this rewrites

$$x_{t+1} = \underset{x \in \mathcal{D}_{t+1}}{\text{argmax}} \left[\hat{\theta}(t)^T x + ||x||_{B(t)^{-1}} \beta(t, \delta) \right].$$

Examples: The OFUL algorithm [Abbassi-Yadkori et al. 11], with

$$\beta(t, \delta) \simeq \sqrt{d \log \left(\frac{Ct}{\delta \lambda} \right)}$$

satisfies $\mathbb{P}\left(\mathcal{R}(T, \text{OFUL}) = \tilde{O}(d\sqrt{T}) \right) \geq 1 - \delta.$

Similar algorithms:[Dani et al. 08][Rusmeviechentong and Tsitsiklis 10], [Chu et al 11](Lin-UCB)

A Bayesian view on linear bandits

The model is still

$$\begin{aligned} r_t &= x_t^T\theta + \epsilon_t \\ R_t &= X_t^T\theta + E_t. \end{aligned}$$

Assume that the noise is Gaussian $\epsilon_t \sim \mathcal{N}\left(0, \sigma^2\right)$ and that θ is drawn from some prior distribution:

$$\theta \sim \mathcal{N}\left(0, \kappa^2 I_d\right)$$

The posterior distribution on θ is given by

$$p(\theta | X_t, R_t) = \mathcal{N}\left(\hat{\theta}(t), \Sigma_t\right)$$

with

$$\left\{ \begin{array}{lcl} \hat{\theta}(t) & = & (B(t))^{-1} X_t^T R_t \quad \text{with} \quad B(t) = \frac{\sigma^2}{\kappa^2} I_d + X_t^T X_t \\ \Sigma_t & = & \sigma^2 (B(t))^{-1}. \end{array} \right.$$

$\hat{\theta}(t)$ is the regularized least-square estimate with $\lambda = \frac{\sigma^2}{\kappa^2}$.

Bayes-UCB for contextualized linear bandit

Bayes-UCB is a Bayesian, optimistic, algorithm, originally designed for bandits with independent arms ([Kaufmann et al. 2012]).
For $x \in \mathcal{D}_{t+1}$, the posterior distribution on $\theta^T x$ is

$$\pi_x(t) = \mathcal{N}\left(x^T\hat{\theta}(t), \sqrt{x^T\Sigma_t x}\right).$$

If $|\mathcal{D}_t| \leq K$, Bayes-UCB chooses at time $t+1$ the context

$$\begin{aligned} x_{t+1} &= \underset{x\in\mathcal{D}_{t+1}}{\text{argmax}}\ Q\left(1 - \frac{6\delta}{\pi^2 K t^2}, \pi_x(t)\right) \\ x_{t+1} &= \underset{x\in\mathcal{D}_{t+1}}{\text{argmax}}\ x^T\hat{\theta}(t) + ||x||_{\Sigma_t} Q\left(1 - \frac{6\delta}{\pi^2 K t^2}; \mathcal{N}(0,1)\right) \end{aligned}$$

with $Q(\alpha, \pi)$ the quantile of order α of the distribution π.
One can show that $\mathbb{P}\left(\mathcal{R}(T, \text{Bayes-UCB}) = \tilde{O}\left(\sqrt{dT\log(K)}\right)\right) \geq 1 - \delta$, *under the Bayesian probabilistic model*.

Thompson Sampling for contextual linear bandit

Thompson Sampling (TS) heuristic: Draw a model from the current posterior distribution and act optimally in this sampled model.

$$\begin{aligned} \tilde{\theta}(t) &\sim \mathcal{N}\left(\hat{\theta}(t), \Sigma_t\right) \\ x_{t+1} &= \underset{x \in \mathcal{D}_{t+1}}{\operatorname{argmax}} \ \tilde{\theta}(t)^T x \end{aligned}$$

(draw each context acording to its posterior probability of being optimal)

From [Russo, Van Roy 2014] it can be shown that

$$\begin{aligned} \mathbb{E}\left[\mathcal{R}(T, \mathsf{TS})\right] &= \tilde{O}\left(d\sqrt{T}\right) \\ \text{if } |\mathcal{D}_t| \leq K, \ \mathbb{E}\left[\mathcal{R}(T, \mathsf{TS})\right] &= \tilde{O}\left(\sqrt{dT\log(K)}\right), \end{aligned}$$

up to logarithmic factors in T, and with the expectation taken under the *Bayesian model*. ([Agrawal, Goyal 2013] give first frequentist guarantees).

In practice

1 Thompson Sampling versus optimistic algorithms

$$\text{Opt.}: \quad x_{t+1} = \underset{x \in \mathcal{D}_{t+1}}{\text{argmax}} \left[\hat{\theta}(t)^T x + ||x||_{B(t)^{-1}} \beta(t,\delta)\right] \quad (1)$$

$$\text{TS}: \quad x_{t+1} = \underset{x \in \mathcal{D}_{t+1}}{\text{argmax}} \ \tilde{\theta}(t)^T x, \quad \text{with } \tilde{\theta}(t) \sim \mathcal{N}\left(\hat{\theta}(t), \sigma^2 B(t)^{-1}\right) \quad (2)$$

Both algorithms require to store the matrix

$$B(t) = \lambda I_d + X_t^T X_t = \lambda I_d + \sum_{s=1}^{t} x_s x_s^T$$

and compute its inverse at each round (which is costly)

In practice

① Thompson Sampling versus optimistic algorithms

$$\text{Opt.}: \quad x_{t+1} = \underset{x \in \mathcal{D}_{t+1}}{\text{argmax}} \left[\hat{\theta}(t)^T x + ||x||_{B(t)^{-1}} \beta(t,\delta)\right] \quad (1)$$

$$\text{TS}: \quad x_{t+1} = \underset{x \in \mathcal{D}_{t+1}}{\text{argmax}} \ \tilde{\theta}(t)^T x, \quad \text{with } \tilde{\theta}(t) \sim \mathcal{N}\left(\hat{\theta}(t), \sigma^2 B(t)^{-1}\right) \quad (2)$$

Both algorithms require to store the matrix

$$B(t) = \lambda I_d + X_t^T X_t = \lambda I_d + \sum_{s=1}^{t} x_s x_s^T$$

and compute its inverse at each round (which is costly)

② Variant with batch updates: $\theta(\hat{t})$ and $B(t)^{-1}$ remain constant for several rounds, and the context is still chosen according to (1) or (2)

Case study

We present elements from a paper by Chapelle et al. (2014):

Simple and scalable response prediction for display advertising

Case study

We present elements from a paper by Chapelle et al. (2014):

Simple and scalable response prediction for display advertising

(1) The features used:

- the set $\mathcal{D}_t$ contains sparse binary entries
- their are built by concatenating categorial features from user/add/campaign/website (and conjunctions of these features)
- to reduce the dimension, a 'hashing trick' can be used

The model: logistic regression

2 Goal: maximize the number of clicks or conversions (i.e. a target event happens).

Responses $r_t \in \{-1, 1\}$ depending on whether the target event happens

The model: logistic regression

(2) Goal: maximize the number of clicks or conversions
(i.e. a target event happens).

Responses $r_t \in \{-1, 1\}$ depending on whether the target event happens

(3) Model used: Logistic regression

$$\mathbb{P}(r_t = 1 | x_t, \theta) = \frac{1}{1 + \exp(-\theta^T x_t)}$$

The model: logistic regression

2 Goal: maximize the number of clicks or conversions (i.e. a target event happens).

Responses $r_t \in \{-1, 1\}$ depending on whether the target event happens

3 Model used: Logistic regression

$$\mathbb{P}(r_t = 1 | x_t, \theta) = \frac{1}{1 + \exp(-\theta^T x_t)}$$

4 Response prediction based on a training set $T = (x_i, r_i)_{1 \leq i \leq n}$

$$\hat{\theta} = \underset{w \in \mathbb{R}^d}{\text{argmin}} \ \frac{\lambda}{2} ||w||^2 + \sum_{i=1}^{n} \log(1 + \exp(-r_i w^T x_i))$$

(regularized maximum likelihood estimator)

A Bayesian view on logistic regression

Training set $T = (x_i, r_i)_{1\leq i\leq n}$, model

$$\mathbb{P}(r_t = 1 | x_t, \theta) = \frac{1}{1+\exp(-\theta^T x_t)}$$

If $\theta \sim \mathcal{N}\left(0, \frac{1}{\lambda} I_d\right)$, the posterior distribution $p(\theta|T)$ has no close form expression ($\neq$ linear case), but can be approximated, using a Laplace approximation, by

$$p(\theta|T) \sim \mathcal{N}\left(m, \text{Diag}(q_i^{-1})\right)$$

with

$$\begin{aligned} m &= \underset{w\in\mathbb{R}^d}{\text{argmin}} \ \frac{\lambda}{2}||w||^2 + \sum_{i=1}^{n} \log(1+\exp(-r_i w^T x_i)) &= \hat{\theta} \\ q_i &= \sum_{j=1}^{n} x_{j,i}^2 p_j(1-p_j) \ \text{ with } \ p_j = (1+\exp(-m^T x_j))^{-1} \end{aligned}$$

The posterior mean is the previously proposed estimator of θ.

Updates of the model and Thompson Sampling

(5) Regularized logistic regression with batch updates: The Bayesian interpretation allow for an easy sequential update of the model.

Initialization:

$m = 0$ and $q_i = \lambda$ for $i = 1 \dots d$

(corresponding to the prior distribution $\mathcal{N}\left(m, \text{Diag}(q_i^{-1})\right)$)

For $t = 1 \dots T$,

- Get a new batch of training data $(x_j, r_j)_{1 \leq j \leq n}$
- $\mathcal{N}\left(m, \text{Diag}(q_i^{-1})\right)$ is the new posterior distribution (obtained with Laplace approximation)

$$m \leftarrow \underset{w \in \mathbb{R}^d}{\text{argmin}} \ \frac{1}{2} \sum_{i=1}^{d} q_i (w_i - m_i)^2 \ + \sum_{i=1}^{n} \log(1 + \exp(-r_i w^T x_i))$$

Updates of the model and Thompson Sampling

(5) Regularized logistic regression with batch updates: The Bayesian interpretation allow for an easy sequential update of the model.

Initialization:
$m = 0$ and $q_i = \lambda$ for $i = 1 \dots d$
(corresponding to the prior distribution $\mathcal{N}\left(m, \text{Diag}(q_i^{-1})\right)$)

For $t = 1 \dots T$,

- Get a new batch of training data $(x_j, r_j)_{1 \leq j \leq n}$
- $\mathcal{N}\left(m, \text{Diag}(q_i^{-1})\right)$ is the new posterior distribution (obtained with Laplace approximation)

$$m \leftarrow \underset{w \in \mathbb{R}^d}{\text{argmin}} \ \frac{1}{2} \sum_{i=1}^{d} q_i (w_i - m_i)^2 \ + \sum_{i=1}^{n} \log(1 + \exp(-r_i w^T x_i))$$

Updates of the model and Thompson Sampling

(6) Thompson Sampling to obtain a new batch of data The current posterior is $\mathcal{N}\left(m, \text{Diag}(q_i^{-1})\right)$. For $t = 1 \dots n$

- a new user arrives at time t
- form the set $\mathcal{D}_t$ of contexts corresponding to the different items that can be recommended to him
- sample a vector from the current (approximate) posterior

$$\tilde{\theta}(t) \sim \mathcal{N}\left(m, \text{Diag}(q_i^{-1})\right)$$

- choose the context x_t that maximize the probability of positive response in this sampled model

$$x_t = \arg\max_{x \in \mathcal{D}_t} \frac{1}{1 + \exp(-\tilde{\theta}(t)^T x)} = \arg\max_{x \in \mathcal{D}_t} \tilde{\theta}(t)^T x$$

- recommend the associated item and get response r_t

A new batch $T = (x_t, r_t)_{1 \leq t \leq n}$ is obtained

Comments

- We explained how Thompson Sampling could be implemented in an example of generalized linear bandit model (based on logistic regression)
- It should be compared to optimistic algorithms for generalized linear bandits (presented by [Filippi et al. 2010])
- Thompson Sampling in the linear and logistic model is essentially the same algorithm

$$x_{t+1} = \underset{x \in \mathcal{D}_{t+1}}{\text{argmax}}\ \tilde{\theta}(t)^T x$$

but with $\tilde{\theta}(t)$ being sampled from a different posterior distribution. How do they compare in practice?

Comments

- We explained how Thompson Sampling could be implemented in an example of generalized linear bandit model (based on logistic regression)
- It should be compared to optimistic algorithms for generalized linear bandits (presented by [Filippi et al. 2010])
- Thompson Sampling in the linear and logistic model is essentially the same algorithm

$$x_{t+1} = \underset{x \in \mathcal{D}_{t+1}}{\text{argmax}} \; \tilde{\theta}(t)^T x$$

  but with $\tilde{\theta}(t)$ being sampled from a different posterior distribution. How do they compare in practice?
- A crucial part of the design of the recommendation system relies on the way the contexts are built

Further challenges

Some more involved contextual bandit models are currently been developped to face new challenges in recommendation systems:

- Recommendation of more than one item
  Example: [Yue, Guestrin 2011]
  Linear Submodular Bandits and their application to Diversified Retrieval
- Bandits with budget constraints (each item/add can be shown a limited number of time only)
  Example: [Badanidiyuru et al. 2014]
  Resourceful Contextual Bandits

References: 1/2

- Abbasi-Yadkori, Pal, Szepesvari, *Improved algorithms for linear bandits*, NIPS 2011
- Agrawal, Goyal, *Thompson Sampling for Contextual Bandits with Linear Payoffs*, ICML 2013
- Badanidiyuru, Langford, Slivkins, *Resourceful Contextual Bandits*, COLT 2014
- Chapelle, Manavoglu, Rosales, *Simple and scalable response prediction for display advertising*, ACM Transaction on Intelligent Systems and Technology, 2014
- Chu et al., *Contextual Bandits with Linear Payoff Functions*, AISTATS 2011
- Dani, Hayes, Kakade, *Stochastic Linear Optimization under Bandit Feedback*, NIPS 2008
- Filippi, Cappé, Garivier, Szepesvari *Parametric Bandits : The Generalized Linear case*, NIPS 2010

References: 2/2

- Kaufmann, Cappé, Garivier, *On Bayesian Upper Confidence Bounds for Bandits Problems*, AISTATS 2012
- Krause, Ong, *Contextual Gaussian Process Bandit Optimization*, NIPS 2011
- Rusmevichientong,Tsitsiklis, *Linearly Parametrized Bandits*, Mathematics of Operation Research, 2010
- Russo, Van Roy, *Learning to Optimize via Posterior Sampling*, Mathematics of Operation Research, 2014
- Valko, Korda, Munos, Cristinini, *Finite-time analysis of kernelized contextual bandits*, UAI 2013
- Yue, Guestrin, *Linear Submodular Bandits and their application to diversified retrieval*, NIPS 2011