CS 4/5789 11 April 2022 Prof Sarah Dean

Lecture 20: Linear Contextual Bandits

1) Setting

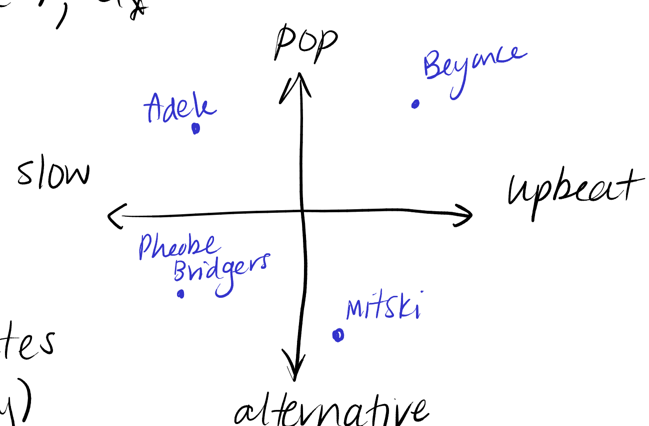
Our simplified MDP setting consists of:

- contexts $x \in \mathcal{X} \subseteq \mathbb{R}^d$
  drawn from distribution $\mathcal{D} \in \Delta(\mathcal{X})$ $x_t \sim \mathcal{D}$
- actions "arms" $a \in \mathcal{A} = \{1, \dots, K\}$
- rewards $r_t = r(x_t, a_t)$ with
  $\mathbb{E}[r(x,a)] = \mu_a(x) = \theta_a^\top x$ linear function
- Horizon T

Goal: find a policy $a_t = \pi(x_t)$ that acheives low regret.

$$R(T) = \sum_{t=1}^{T} \mathbb{E}_{x_t \sim \mathcal{D}} \Big[\underbrace{\max_a \theta_a^\top x_t}_{\mu_*(x),\ a_*} - \theta_{a_t}^\top x_t \Big]$$

Example: music recomentation
arms a are artists
$\theta_a \in \mathbb{R}^d$ represents attributes
$x \in \mathbb{R}^d$ represents a users affinity towards the attributes (observed from listening history)



Last lecture we considered an explore-then-commit algorithm for general function approximation/supervised learning

$$a_t = \arg\max_a \hat{y}_a(x_t) \quad \text{where}$$

$$\hat{y}_a = \arg\min_{y \in M} \sum_{i=1}^{N} (y(x_i^a) - r_i^a)^2$$

data collected during exploration phase.

Linear Regression

If we know that $y_a(x) = \theta_a^T x$ we can instatiate the general supervised learning framework with

$$M = \{ y(x) = \theta^T x \mid \theta \in \mathbb{R}^d \}$$

In this case the learning problem is equivalent to

$$\hat{\theta}_a = \arg\min_{\theta} \sum_{i=1}^{N} (\theta^T x_i^a - r_i^a)^2$$

We will sometimes drop the a subscript in these notes.

Lemma: As long as $(x_i)_{i=1}^N$ span $\mathbb{R}^d$,

$$\hat{\theta} = \Big(\underbrace{\sum_{i=1}^{N} x_i x_i^T}_{A}\Big)^{-1} \underbrace{\sum_{i=1}^{N} x_i r_i}_{b} = A^{-1} b$$

Proof:

$$\nabla_\theta \sum_{i=1}^{N} (\theta^T x_i - r_i)^2 = 2 \sum_{i=1}^{N} x_i (x_i^T \theta - r_i)$$

setting the gradient equal to zero,

$$\Big(\underbrace{\sum_{i=1}^{N} x_i x_i^T}_{A}\Big) \theta = \underbrace{\sum_{i=1}^{N} x_i r_i}_{b}$$

The matrix on the left hand side is invertible if $(x_i)_{i=1}^{N}$ span $\mathbb{R}^d$

(Why? Let $X = \begin{bmatrix} x_1^T \\ \vdots \\ x_N^T \end{bmatrix} \in \mathbb{R}^{N \times d}$. Then if x_i span $\mathbb{R}^d$, X has full row rank, $\text{rank}(X) = d$

$\sum_{i=1}^{N} x_i x_i^T = X^T X \in \mathbb{R}^{d \times d}$ is full rank because $\text{rank}(X^T X) = \text{rank}(X) = d$. Therefore it is invertible.)

$\square$

The matrix A is related to the <u>empirical covariance</u>

$$\Sigma = \mathbb{E}_{x \sim D}[x x^T] \qquad \hat{\Sigma} = \frac{1}{N} \sum_{i=1}^{N} x_i x_i^T$$

We can relate $A = N \hat{\Sigma}$.

2) Interactive Demo - Jyputer Notebook

3) LinUCB Algorithm

Recall that last lecture we wanted to estimate conditional errors $\mathbb{E}[(\hat{\mu}_a(x)-\mu_a(x))^2 \mid x]$. Using the structure of the linear regression problem, we can do this.

We keep track of

$$A_t^a = \sum_{k=1}^{t} x_k x_k^T \mathbb{1}\{a_k = a\}, \quad b_t^a = \sum_{k=1}^{t} x_k r_k \mathbb{1}\{a_k = a\}$$

$$\hat{\theta}_t^a = (A_t^a)^{-1} b_t^a$$

Alg: LinUCB

Initialize 0 mean & infinite confidence intervals

For $t = 1, \dots, T$:

$$a_t = \operatorname*{argmax}_a \ \hat{\theta}_t^{a\,T} x_t + \alpha \sqrt{x_t (A_t^a)^{-1} x_t}$$

update $\hat{\theta}^{a_t}$, b^{a_t}, A^{a_t}

Geometric Intuition:

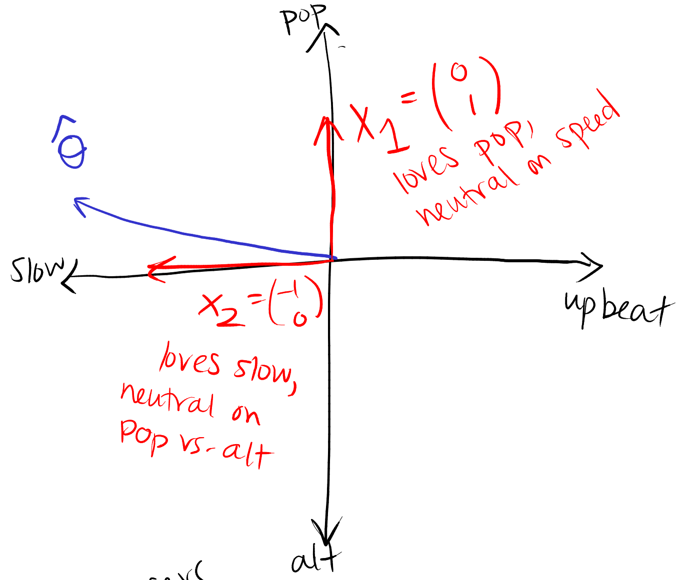
$$\underbrace{\hat{\theta}^T x}_{\text{large if } x \text{ and } \hat{\theta} \text{ are aligned}} + \alpha \underbrace{\sqrt{x^T A^{-1} x}}_{\text{large if } x \text{ is not aligned with much historical data}}$$

$$x^T A^{-1} x = x^T (N\hat{\Sigma})^{-1} x = \underbrace{\frac{1}{N}}_{\text{amount of data}} \underbrace{x^T \hat{\Sigma}^{-1} x}_{\text{alignment w/ data}}$$

ex- $\hat{\theta}^T x_2 > \hat{\theta}^T x_1$

$\hat{\theta}^T x_2 \rightarrow$ large alignment on speed

$\hat{\theta}^T x_1 \rightarrow$ less alignment on pop



ex- if previous data is

$(x_1, -x_1, x_1, x_1, -x_1, -x_2)$

most users so far are or neutral speed

then $A^{-1} = \begin{bmatrix} 1 & 0 \\ 0 & 1/5 \end{bmatrix}$

less sure on speed

pretty sure about pop

$\hat{\Sigma} = \begin{bmatrix} 1/6 & 0 \\ 0 & 5/6 \end{bmatrix}$

$\underbrace{x_1^T A^{-1} x_1}_{\text{more sure}} < \underbrace{x_2^T A^{-1} x_2}_{\text{less sure}}$

Statistical Explaination:

Claim: With high probability (over noisy rewards)

$$\theta_a^T x \leq \hat{\theta}_a^T x + \alpha \sqrt{x^T A_a^{-1} x}$$

where α depends on probability & variance of rewards

Lemma: (Chebychev's Inequality)

for a random variable u with $\mathbb{E}(u) = 0$,

$|u| \leq \beta \sqrt{\mathbb{E}(u^2)}$ with probability $1 - 1/\beta^2$

Proof: we will use chebychev's to show that w.h.p

$$|\underbrace{\hat{\theta}_a^T x - \theta_a^T x}_{u}| \leq \alpha \sqrt{\underbrace{x^T A^{-1} x}_{\mathbb{E}u^2}}$$

1) compute expectation. Define $w_i = r_i - \mathbb{E}[r_i]$ so $r_i = \theta_a^T x_i + w_i$.

$$\theta = \left(\sum_{i=1}^N x_i x_i^T\right)^{-1} \sum_{i=1}^N x_i (\theta_a^T x_i + w_i)$$

$$= \underbrace{\cancel{\left(\sum_{i=1}^N x_i x_i^T\right)^{-1} \sum_{i=1}^N x_i x_i^T}}_{I} \theta_a + \left(\sum_{i=1}^N x_i x_i^T\right)^{-1} \sum_{i=1}^N x_i w_i$$

$$\theta - \theta_a = A^{-1} \sum_{i=1}^N x_i w_i$$

therefore $\mathbb{E}((\theta - \theta_a)^T x) = A^{-1} \sum_{i=1}^N x_i \mathbb{E}[w_i] = 0$

2) compute variance

$$\mathbb{E}_{\theta}\left[((\theta - \theta_a)^T x)^2\right] = \mathbb{E}_{w_i}\left[x^T A^{-1} \sum_{i=1}^N x_i w_i \cdot \sum_{i=1}^N x_i^T w_i A^{-1} x\right]$$

$$= x^T A^{-1} \mathbb{E}\left[\sum_{\substack{i=1 \\ j=1}}^N x_i x_j^T \underbrace{w_i w_j}\right] A^{-1} x$$

The noise in rewards is iid so the expectation is 0 if $i \neq j$. Define σ^2 as variance of rewards.

$$= x^T A^{-1} \underbrace{\sum_{i=1}^N x_i x_i^T}_{=A!} \sigma^2 A^{-1} x$$

$$= \sigma^2 x^T A^{-1} x$$

Therefore, using Chebychev's, we have that w.p. $1 - 1/\beta^2$,

$$|\theta_a^T x - \hat{\theta}_a^T x| \le \underbrace{\beta\sigma}_{\alpha} \sqrt{x^T A^{-1} x}$$

Thus the upper bound of this confidence interval is $\hat{\theta}^T x + \alpha \sqrt{x^T A^{-1} x}$

$\square$