Week 9 - Lecture 42

Stochastic Contextual Bandits

$t = 1, 2, \ldots T$

Actions:

1. For every user have a context at every time step
2. $\therefore x_t$ = context or side-information
3. $x_t \in C \subset \mathbb{R}^d$
4. $r_t = R(x_t, I_t)$ ← reward

where, $I_t \in [K]$ is the action chosen based on x_t.

Reward depends on context & arm-played

Actually reward also has a noise component.

$$r_t = R(x_t, I_t) + \eta_t$$ ← η_t is conditionally sub-gaussian

$R : C \times [K] \to \mathbb{R}$ R is reward function

↳ is a function from context, arm giving reward

For the learner the reward is unknown.

Thus goal is to figure out the function R.

If η_t is σ- subgaussian, then:

$$\mathbb{E}\left[e^{\lambda \eta_t}\right] \leq e^{\lambda^2 \sigma^2/2}$$

But η_t is <u>conditionally σ-subgaussian on $\mathcal{F}_t$</u>

$$\mathbb{E}\left[e^{\lambda \eta_t} \mid \mathcal{F}_t\right] \leq e^{\lambda^2 \sigma^2/2}$$

$$\mathcal{F}_t = \sigma\left(x_1, I_1, x_2, I_2 \ldots . x_t, I_t\right)$$

↑ σ-algebra generated by observations (context, action pair)

σ-subgaussian will have zero mean

$\therefore$

$$\mathbb{E}\left[r_t \mid \mathcal{F}_t\right] = R(x_t, I_t)$$

The policy would be mapping from context to arm

$$\pi : \mathcal{C} \longrightarrow [K]$$

The context are generated here (assume) stochastical in an i.i.d. fashion according to some common distribution, that would be revealed to the learner.

$\therefore \quad X_1, X_2, \ldots \ldots X_T \leftarrow$ randomly generated

Say a particular realization is:

$$x_1, x_2 - \; - \; - \; x_T$$

$$\sum_{t=1}^{T} R(x_t, a_t^*) \leftarrow \text{Best total reward}$$

where

$$a_t^* = \underset{a \in [K]}{\operatorname{argmax}} \; R(x_t, a)$$

The learner does not know R $\therefore$ Regret of a policy would be:

$$R_T(\pi) = \mathbb{E}\left[\sum_{t=1}^{T} R(X_t, a_t^*) - \sum_{t=1}^{T} R(X_t, I_t) \right]$$

↑ randomness over contexts & arms played

Assume that:

$$R(x, a) = \langle \underbrace{\Psi(x, a)}_{\text{known feature map } \Psi(x,a) \in \mathbb{R}^d \; \forall (x,a)}, \; \underbrace{\theta^*}_{\text{unknown } \theta^* \in \mathbb{R}^d} \rangle$$

} does not depend on context