뉴스기사를 제공하는 플랫폼 → 접속하는 유저에게 가장 관심이 있을법한 뉴스 제공
→ 유저가 해당 기사 클릭

Contextual, 20대 유저에겐 시사/정치보단 스포츠, 생활, 문화와 관련된 기사를 노출시키는 것이 유리

→ 성별, 연령대가 사전에 관측가능하다면 이에 대응하는 최적의 의사결정을 하는 알고리즘 개발

Regret function

$$E[Rew(ALG)] = \sum_{t=1}^{T} \mu(a_t|x_t)$$

최적의 reward를 얻는 정책 $\pi(x^*) = \max_{a \in A} \mu(a|x)$

$$R(T) = Rew(\pi^*) - Rew(ALG)$$

$x_{t,a} \in R^d$, t round의 a arm에 대한 d차원 Context,

Context vector의 element의 강도를 arm별로 나타내는 coefficient vector를 θ_a

x_t Context에서 a arm을 선택했을 때 기대보상

$$E[r_t|x_{t,a}] = x_{t,a}^T \theta_a$$

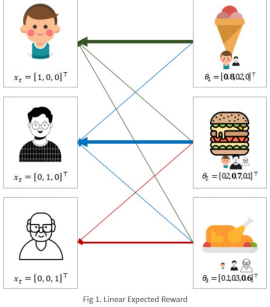

Fig 1. Linear Expected Reward

ex) t round에 10대에게 음식 추천, 기대보상

아이스크림 : $[1,0,0]\theta_1 = 0.8$ ✓최선의 선택

햄버거 : $[1,0,0]\theta_2 = 0.2$

치킨 : $[1,0,0]\theta_3 = 0.1$

핵심가정 : Context와 Reward 사이에는 선형 관계가 존재함

$$a_t = \arg\max_a \left(x_{t,a}^T \hat{\theta}_a + \alpha \sqrt{x_{t,a}^T A_a^{-1} x_{t,a}} \right) \qquad \alpha = 1 + \sqrt{\frac{\ln(2/\delta)}{2}}$$

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

벤딧 파라미터 ← $\theta_a = (D_a^T D_a + I_d)^{-1} D_a^T C_a$

Ridge Regression으로 추정

↳ 영향을 미치지 않는 특성에 대하여 0에 가까운 가중치 부여