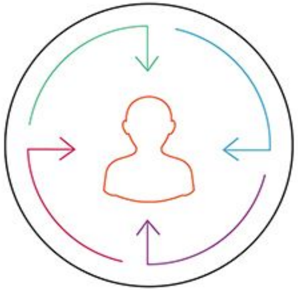
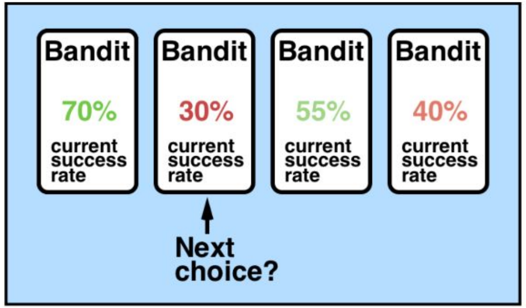
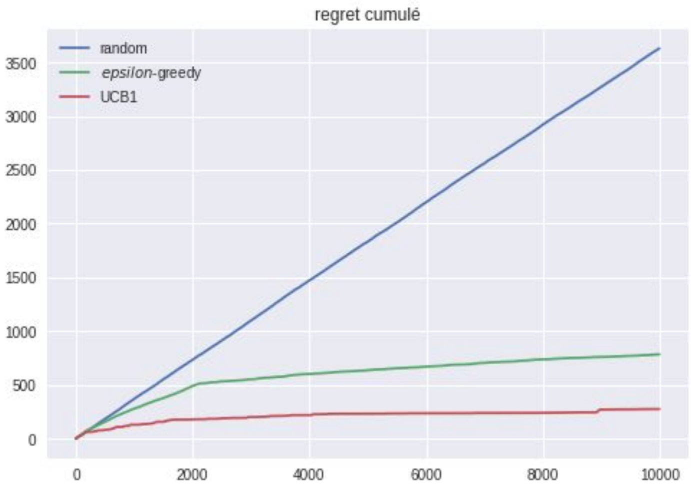
Collaborative Filtering
LinUCB for MovieLens dataset



Team: Chen Dang
Công Minh Đinh
Oskar Rynkiewicz

Bandit problem introduction





Goal: minimize the expected total regret

$$R(T) = \mathbb{E}_{I_t}\left[\sum_{t=1}^{T} r_{i^*,t}\right] - \mathbb{E}_{I_t}\left[\sum_{t=1}^{T} r_{i,t}\right]$$

Implemented UCB vs e-greedy vs random

LinUCB Disjoint

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

LinUCB Hybrid

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

Movielens dataset & Data processing

$$
\begin{array}{c|ccccccc}
users/items & i_1 & i_2 & \cdots & \cdots & \cdots & \cdots & i_m \\
\hline
u_1 & r_{1,1} & \cdots & r_{1,j} & \cdots & r_{1,k} & \cdots & r_{1,m} \\
u_2 & r_{2,1} & \cdots & r_{2,j} & \cdots & r_{2,k} & \cdots & r_{2,m} \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
u_i & r_{i,1} & \cdots & r_{i,j} & \cdots & r_{i,k} & \cdots & r_{i,m} \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
u_n & r_{n,1} & \cdots & r_{n,j} & \cdots & r_{n,k} & \cdots & r_{n,m}
\end{array}
$$

Dataset: Movielens 1M

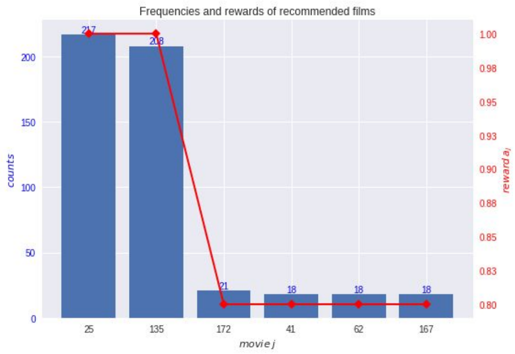
1 million votes of 6000 users on 4000 films

Pre-processing:

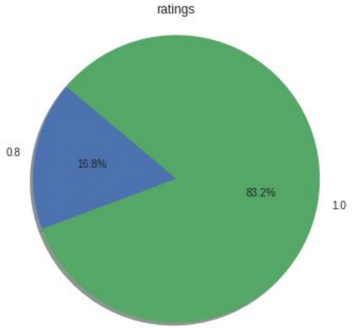
- Resized ratings between 0 and 1
- Selected the movies watched by more than 1000 users
- Selected users who watched more than 150 movies
- Selected 30 most important features

Matrix factorization (SVD): $M = \phi_u^T W \phi_a$

=> Feature user matrix and film (item) matrix

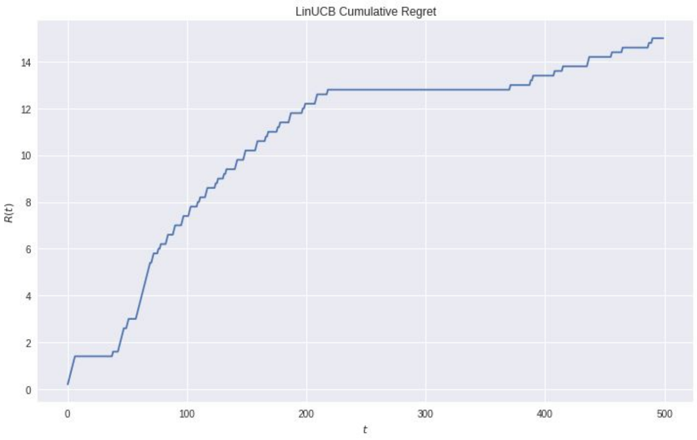


Most frequently recommended films

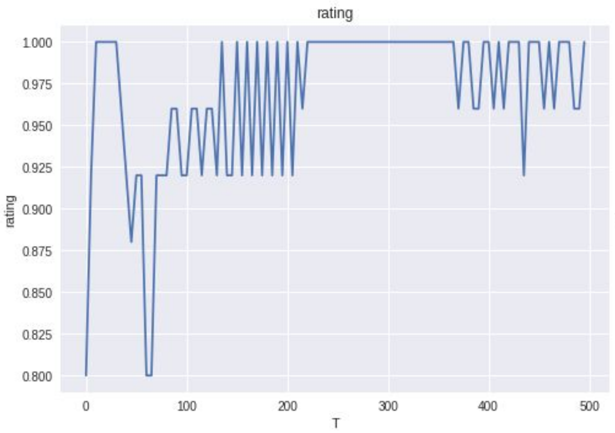


Rating percentages

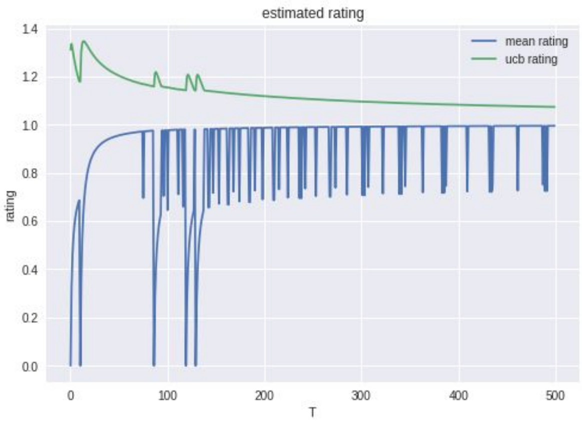
LinUCB Disjoint



Accumulated regret for one user

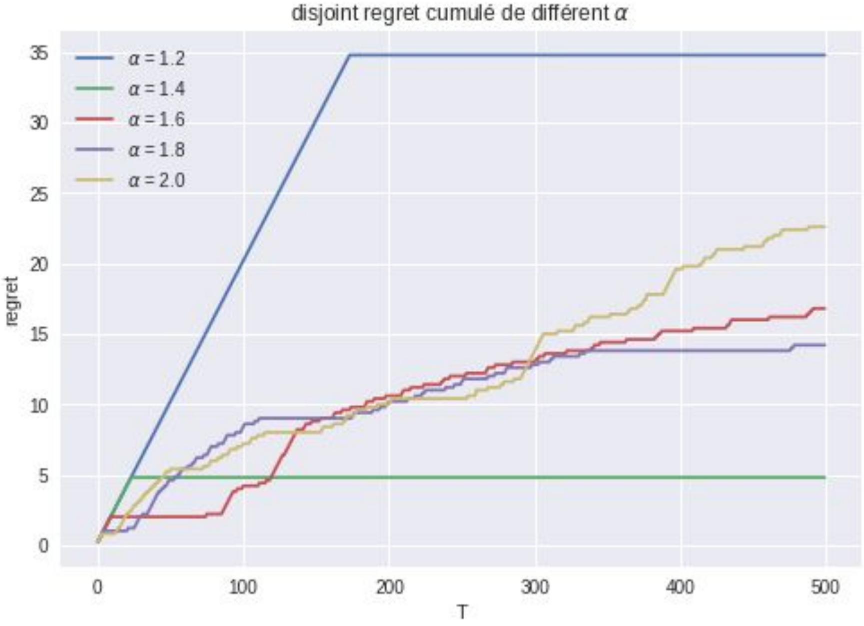


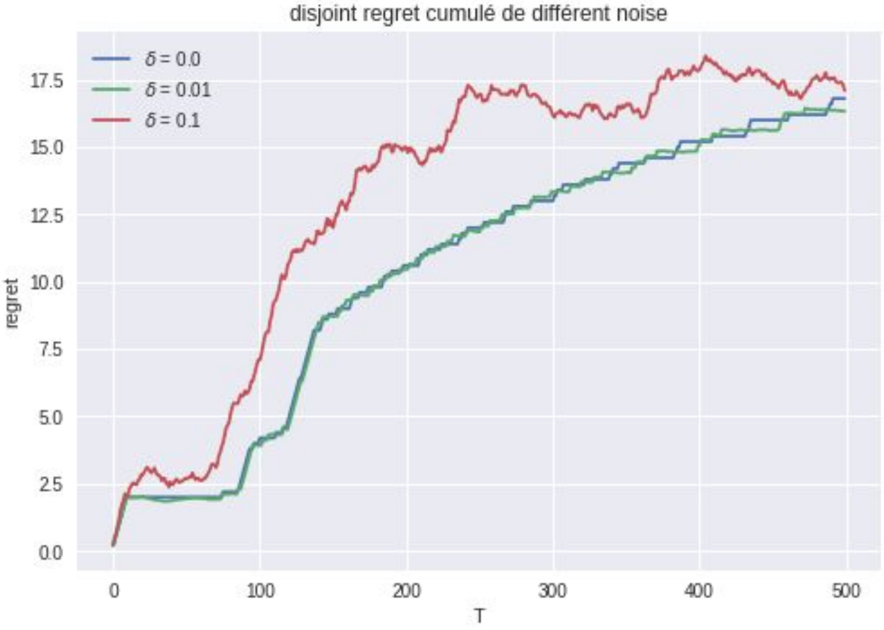
Average movies rating in iterations



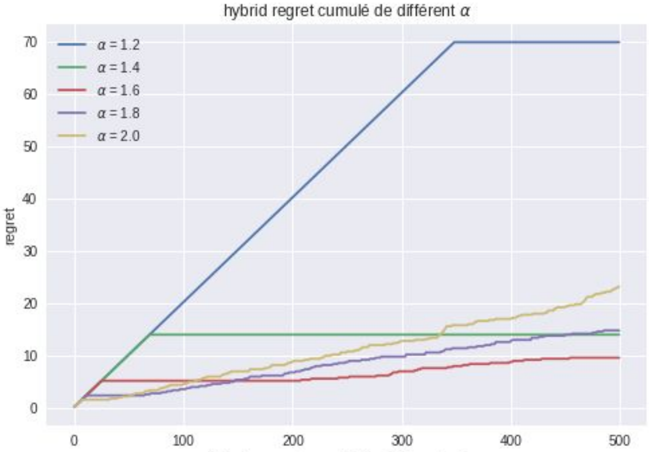
Estimated movies rating

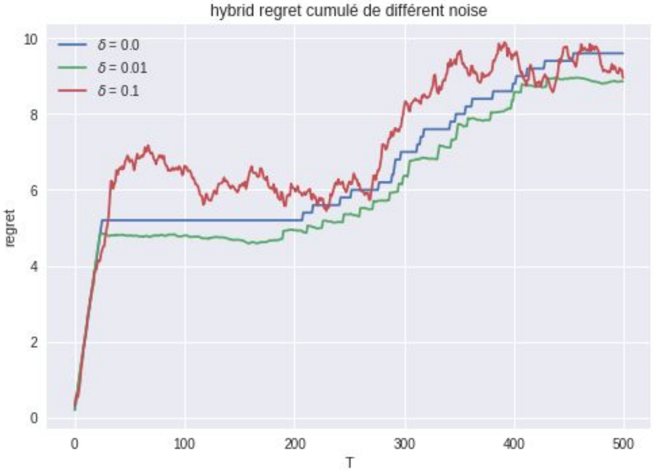
LinUCB Disjoint - Parameters

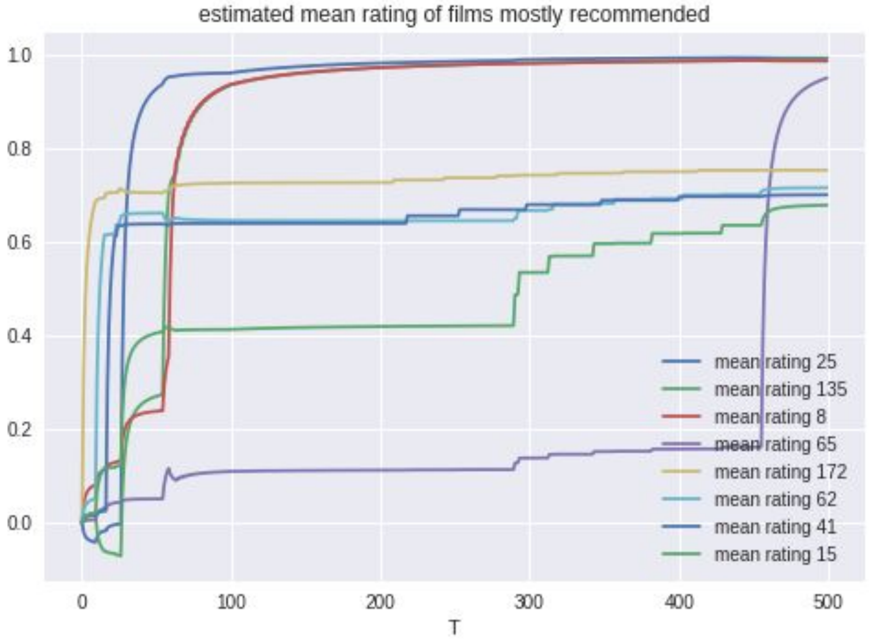




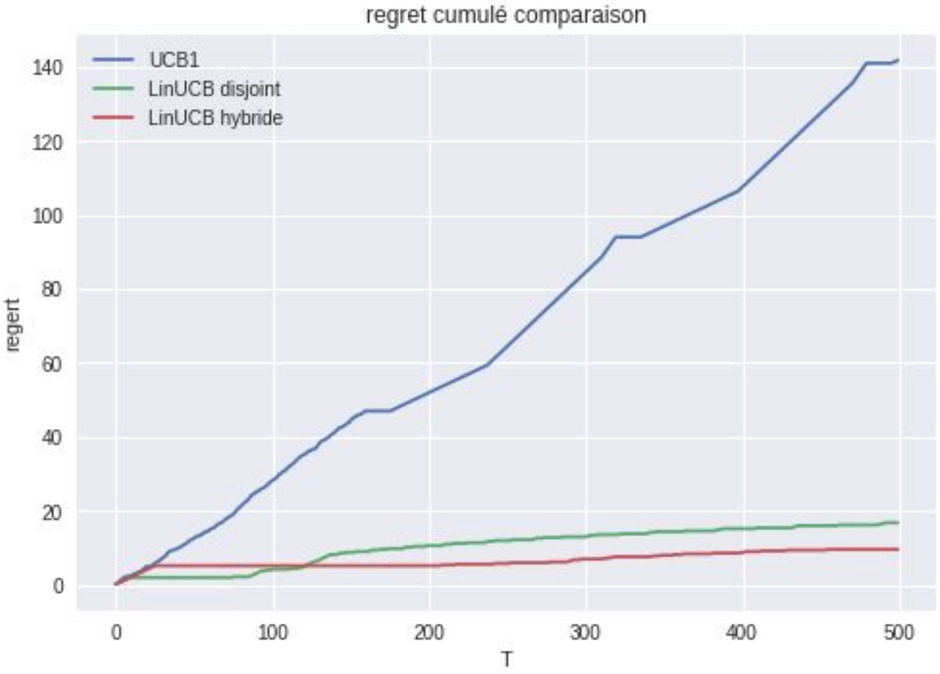
LinUCB Hybrid - Parameters







Method comparison and conclusions



- LinUCB is much better, especially LinUCB hybrid
- Choice of parameters for the model is important, as we noted that they can drastically impact the results
- Expanding the context with other information would be worthwhile