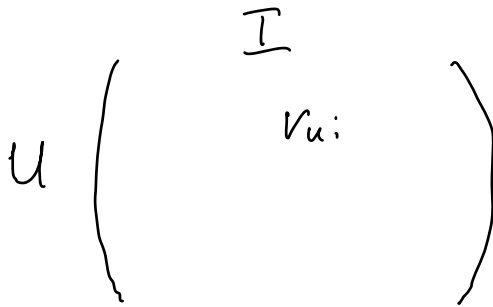


Рекомендательные системы.

U - Users

I - items

r_{ui} - оценка i -го товара пользов. u



r_{ui} :

1) Бинарные:

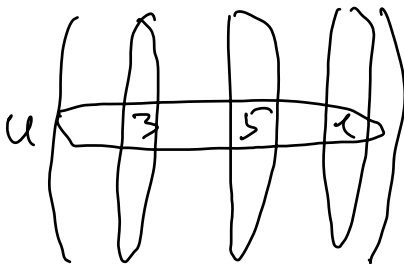
Купил/не купил
Посмотрел/не

2) Рейтинги

Оценка товара/фильма

Представим k товаров пользователям u
 $I(u)$, наиболее подходящие

r_{ui} для i , которые пользователи ещё
не купили.



User-based

Item-based

KNN

Комбинированная
формирование

1) User-based

- Ищем пользователей, похожих на u
$$P : \rho(u, p) > \alpha$$
- Смотрим товары с наибольшими оценками
- Выбираем из них k товаров, которые u не покупал

Item-based:

- Ищем товары, которые похожи на купленные пользователем
- $$I(u) = \{i \in I \mid \exists v_{u, i_0}, w_{i_0, i} > \alpha\}$$
- Выбираем k товаров с наибольшими p_i
$$p_i = \max_{i_0: \exists v_{u, i_0}} w_{i_0, i}$$

Для бинарных векторов не подходит корреляция Пирсона

- 1) Нужно хранить разреженную матрицу
- 2) Хороший бенчмарк, но можно лучше
- 3) Проблема холодного старта
- 4) Выбор метрики эвристика

Модель со скрытыми
переменными

$$p_u \in \mathbb{R}^d, \quad q_i \in \mathbb{R}^d$$

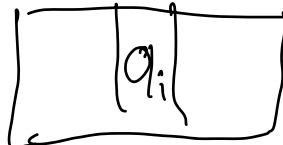
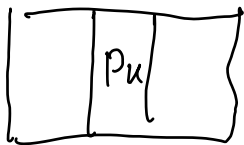
$$r_{ui} \approx \langle p_u, q_i \rangle$$

$$\sum_{u,i \in R} (r_{ui} - \bar{r}_u - \bar{r}_i - \langle p_u, q_i \rangle)^2 \rightarrow \min_{P, Q}$$

R — матрица рейтингов

R' — R с центрированными строками и столбцами

$$\|R' - P^T Q\|^2 \rightarrow \min_{P, Q} \quad (\otimes)$$



$$1) \quad \|R - 2P^T Q\|^2 \rightarrow \min_{P, Q}$$

$$2) \quad \sum_{(u,i) \in R} (r_{ui} - \bar{r}_u - \bar{r}_i - \langle p_u, q_i \rangle)^2 +$$

$$+ \lambda \sum_{u \in U} \|p_u\|^2 + \mu \sum_{i \in I} \|q_i\|^2$$

$$\rightarrow \min_{D, N}$$

LFM . Latent Factor Model

1) Случ. шум.

Выбираем пару (u, i)

$$P_{uk} := P_{uk} + \eta q_{ik} (r_{ui} - \bar{r}_u - \bar{r}_i - \langle p_u, q_i \rangle)$$

$$q_{ik} := q_{ik} + \eta P_{uk} (r_{ui} - \bar{r}_u - \bar{r}_i - \langle p_u, q_i \rangle)$$

2) ALS Alternating least squares

(*) Не получается по P, Q обновл.

Если зафиксировать P или Q , то получим.

Фикс Q :

$$P_u = \left(\sum_{i: r_{ui}} q_i q_i^T \right)^{-1} \left(\sum_{i: r_{ui}} r_{ui} q_i \right)$$

$d \times d$ $d \times 1$

$$q_i = \left(\sum_{u: r_{ui}} P_u P_u^T \right)^{-1} \sum_{u: r_{ui}} r_{ui} P_u$$

HALS : Hierarchical

Фиксируем всё кроме P_k и Q_k

$$P_k = Q_k \left(R - \sum_{s \neq k} P_s q_s^T \right)^T$$

$$q_k q_k^T$$

$$q_k = \frac{p_k (R - \sum_{s \neq k} p_s q_s^T)}{p_k p_k^T}$$

Латентная матрица

Implicit ALS

$$s_{ui} = \begin{cases} 1, & \exists r_{ui} \\ 0, & \text{иначе} \end{cases}$$

$$c_{ui} = 1 + \lambda r_{ui} \quad (\text{увеличивает } s_{ui})$$

$$\sum c_{ui} (s_{ui} - \bar{s}_u - \bar{s}_i - \langle p_u, q_i \rangle)^2 + \lambda \sum_{u \in U} \|p_u\|^2 + \mu \sum_{i \in I} \|q_i\|^2 \rightarrow \min_{P, Q}$$

Факторизационные
матрицы

$$x \in R^d$$

$$d(x) = w_0 + \sum_{j=1}^d w_j x_j + \sum_{j_1=1}^d \sum_{j_2=j_1+1}^d w_{j_1 j_2} x_{j_1} x_{j_2}$$

$$\frac{d(d-1)}{2} + d + 1$$

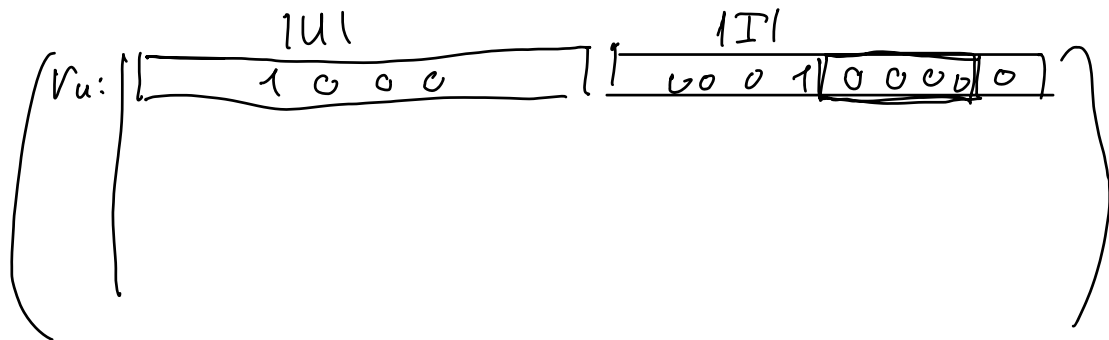
$$w_{j_1 j_2} \approx \langle v_{j_1} v_{j_2} \rangle$$

$$d(x) = w_0 + \sum_{j=1}^d w_j x_j + \sum_{j_1=1}^d \sum_{j_2=j_1+1}^d \langle v_{j_1} v_{j_2} \rangle x_{j_1} x_{j_2}$$

$$v_i \in \mathbb{R}^r \quad dr + d + 1$$

Обобщение (*)

v_{ui} описывается $|U| + |I|$ признаками



$$v_{ui} \approx w_0 + w_u + w_i + \langle v_u, v_i \rangle \quad (*)$$

MC MC - Monte-Carlo Markov Chain

FFM - field-aware factor machine

Комбинированный подход

$$\sum_{i \in I: \exists v_{ui}} (\langle w_u, q_i \rangle - v_{ui})^2 \rightarrow \min_{w_u}$$

Метрики

1) Предсказание рейтингов

MSE, MAE, ...

2) Предсказание событий

F-score ROC-AUC, log-loss ...

$R_u(k)$ - рейтинг к товарам для u
 L_u - товары, для которых
произошло событие

- $\text{hitrate}@k = [R_u(k) \cap L_u \neq \emptyset]$

- $\text{precision}@k = \frac{|R_u(k) \cap L_u|}{|R_u(k)|}$

- $\text{recall}@k = \frac{|R_u(k) \cap L_u|}{|L_u|}$

$$P(\Theta | \text{data}) = \frac{P(\text{data} \cap \Theta)}{P(\Theta)} = \frac{\overset{\text{Likelihood}}{P(\text{data} | \Theta)} \cdot \underset{\text{prior}}{P(\Theta)}}{P(\Theta)}$$