Collaborative Filtering

- Idea: rank items by score $s(j)$ $j = 1 \ldots M$ items, not personalised

- Notations

$$s(j) = \frac{\sum_{i \in \Omega_j} r_{ij}}{|\Omega_j|}$$

r_{ij} = rating user i gave item j

Ω_j = set of all users who rated item j

- Personalize the score

$$s(i,j) = \frac{\sum_{i' \in \Omega_j} r_{i'j}}{|\Omega_j|}$$

$i = 1 \ldots N$ number of users

i' is just index

- Ratings Matrix

r_{ij} = rating user i gave item j

$R_{N \times M}$ = user-item ranking matrix

- Relationship to NLP

$X(t,d)$ = # of times term t appears in document d

"gravity" really likes this document Einstein wrote

- Sparsity

Most entries $= 0$ in NLP

User-item matrix is sparse because most entries are empty

- Goal

most $r(i,j)$ don't exists

guess what the missing values are: $s(i,j) = \hat{r}(i,j)$

- Regression

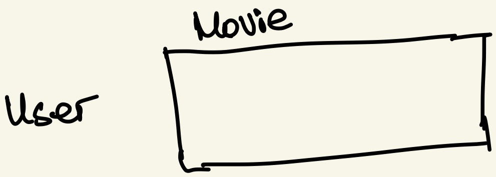
Because we are predicting real number

Objective: Mean Squared Error MSE

$$MSE = \frac{1}{|\Omega|} \sum_{i,j \in \Omega} (r_{ij} - \hat{r}_{ij})^2$$

Ω = set of pairs (i,j) where user i has rated item j

User - User Collaborative Filtering



Find users with high correlation

- Weighted Ratings

$$s(i,j) = \frac{\sum_{i' \in \Omega_j} w_{ii'} r_{i'j}}{\sum_{i' \in \Omega_j} w_{ii'}}$$

weight between user i and user i', large when users in aggreement

Users may be optimistic/pessimistic (3 for bad vs 1 for bad movie)

- Deviation

no absolute ratings → how much it diviates from own average

$dev(i,j) = r(i,j) - \bar{r}_i$ for a known rating

predicted:

$$\hat{dev}(i,j) = \frac{1}{|\Omega_j|} \sum_{i' \in \Omega_j} r(i',j) - \bar{r}_{i'}$$

→ deviations between item i and average for user

→ # rated item j

prediction: $s(i,j) = \bar{r}_i + \hat{dev}(i,j)$

- Combine

$$s(i,j) = \bar{r}_i + \sum_{i' \in \Omega_j} \frac{w_{ii'} \{ r_{i'j} - \bar{r}_{i'} \}}{\sum_{i' \in \Omega_j} |w_{ii'}|}$$

- Calculate weights

Pearson correlation coefficient

$$\rho_{xy} = \frac{\sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N} (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^{N} (y_i - \bar{y})^2}}$$

But not all data is present, matrix sparse

$$w_{ii'} = \frac{\sum_{j \in \Psi_{ii'}} (r_{ij} - \bar{r}_i)(r_{i'j} - \bar{r}_{i'})}{\sqrt{\sum_{j \in \Psi_i} (r_{ij} - \bar{r}_i)^2} \sqrt{\sum_{j \in \Psi_{i'}} (r_{i'j} - \bar{r}_{i'})^2}}$$

Ψ_i = set of movies that user i has rated

$\Psi_{ii'}$ = set both users have rated

$\Psi_{ii'} = \Psi_i \cap \Psi_{i'}$

- Cosine Similarity

$$\cos\theta = \frac{x^T y}{|x||y|}$$

- Neighbourhood

k nearest neighbours with highest weights

keep neighbours with high absolute correlation

Item-Item Collaborative Filtering

- Find users most like other users
- Look at matrix row-wise (2 similar users)

	PR	TR	NT
U1	4,5	5	4
U2	5	5	4,5
U3	1	2	0,5
U4	2	2	0,5

similar items for similar ratings

$$\frac{\sum_{i \in \Omega_{jj'}} (r_{ij} - \bar{r}_j)(r_{ij'} - \bar{r}_{j'})}{\sqrt{\sum_{i \in \Omega_j} (r_{ij} - \bar{r}_j)^2} \sqrt{\sum_{i \in \Omega_{j'}} (r_{ij'} - \bar{r}_{j'})^2}} = w_{jj'}$$

Ω_j = users who rated item j

$\Omega_{jj'}$ = users who rated item j and item j'

$\bar{r}_j$ = avg rating for item j

Item score

$$S(i,j) = \bar{r}_j + \frac{\sum_{j' \in \psi_i} w_{jj'}(r_{ij'} - \bar{r}_{j'})}{\sum_{j' \in \psi_i} |w_{jj'}|}$$

ψ_i = items user i has rated

choose similar items to the current

- Practical Differences

2 items have more users in common (more users to choose from)

item Based is faster $O(M^2 N)$ $N \gg M$

users