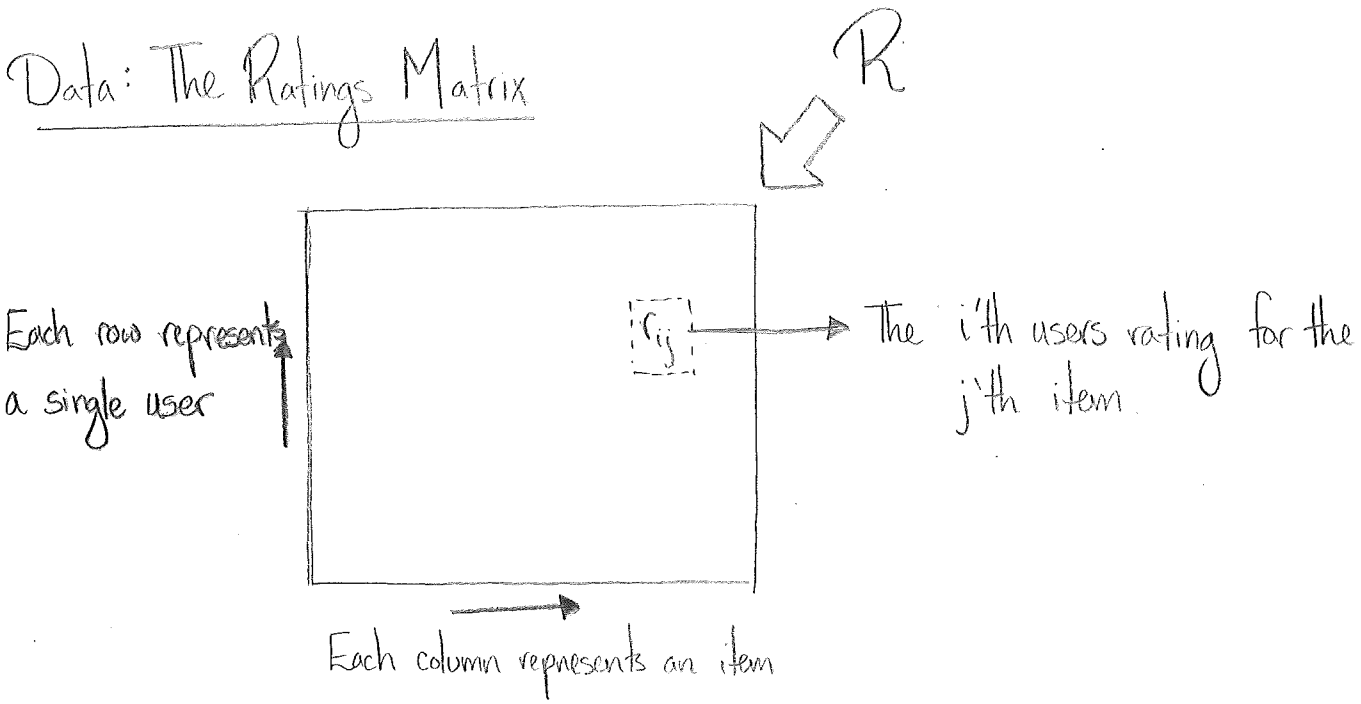
Reccomendation Systems

Morning: Similarity based / Collaborative Filtering

Afternoon: Matrix Factorization

Data: The Ratings Matrix



Usually, most entries are missing, since most users have not rated most items!

Ratings can be explicit or implicit

Explicit: User supplies ratings for items

Implicit: User consumes selected items, measure of consumption is taken as a rating.

Goal: Predict missing ratings!

r_{ij}: Actual rating of user i for item j

$\hat{r}_{ij}$: Predicted rating of user i for item j.

Similarity Based

User - User Hypothesis:

Similar users tend to give similar ratings to a single product.

Item - Item Hypothesis:

A single user will tend to give similar ratings to similar products.

Need: A way to measure similarity between users and/or items.

Then: We can predict ratings as a weighted average of the actual ratings

- of similar users (for a fixed item)
- of similar products (for a fixed user)

User - User:

$$\hat{r}_{ij} = \frac{\sum_{\text{users } u \text{ that have rated item } j} \overbrace{sim(i,u)}^{\text{user-user similarity}} r_{uj}}{\sum_{u} sim(i,u)}$$

↑ rating for fixed item j.

Item - Item:

$$\hat{r}_{ij} = \frac{\sum_{\text{items } t \text{ that are rated by user } i} \overbrace{sim(j,t)}^{\text{item-item similarity}} r_{it}}{\sum_{t} sim(j,t)}$$

Discussions:

1. What are the possible benefits/drawbacks of both approaches.
2. How may we make these calculations more efficient?

Similarity Measures:

Requirements:

1. $\text{sim}(a, b)$ is between (inclusive) zero and one.
2. $\text{sim}(a, b) = 0$ means "a and b are not at all similar."
3. $\text{sim}(a, b) = 1$ means "a and b are as similar as possible."

Similarity between users / items is based off the rows (user-user) or columns (item-item) of the rating matrix.

Cosine Similarity:

$$\cos\left(\theta_{\vec{a},\vec{b}}\right) = \frac{\vec{a}}{|\vec{a}|} \cdot \frac{\vec{b}}{|\vec{b}|}$$

Notation

$\vec{a}, \vec{b}$: rows (user-user) or columns (item-item) of the rating matrix R.

$$\text{cos-sim}\left(\vec{a}, \vec{b}\right) = \frac{1}{2} + \frac{1}{2}\cos\left(\theta_{\vec{a},\vec{b}}\right)$$

Pearson-Correlation Similarity

$$\text{corr}\left(\vec{a}, \vec{b}\right) = \frac{\text{cov}(\vec{a}, \vec{b})}{\text{sd}(\vec{a})\,\text{sd}(\vec{b})}$$

$$\text{corr-sim}\left(\vec{a}, \vec{b}\right) = \frac{1}{2} + \frac{1}{2}\,\text{corr}\left(\vec{a}, \vec{b}\right)$$

Discussion:

When are these large (≈ 1)

When are these small (≈ 0)

What should we do with missing ratings when computing similarity?

Jaccard-Similarity

R must be binary: user consumed item or not?

$$\text{jacc-sim}\left(\vec{a}, \vec{b}\right) = \frac{\#\text{ of items consumed by both a and b}}{\#\text{ of items consumed by either a or b}}$$