# recommender systems

As a working example applying the various techniques of **Recommender Systems**, the Netflix Competition offering a $1 million prize to the team who could improve their existing **Recommender System** by **10%**, is used.

Assuming the following sample data:

$$\begin{array}{rc} & \begin{array}{cccc}\textbf{Alien} & \textbf{Bug's Life} & \textbf{Cars} & \textbf{Dark Night}\end{array} \\ \begin{array}{r}\textbf{Carmen}\\ \textbf{Joseph}\\ \textbf{Leonore}\\ \textbf{Esmerelda}\end{array} & \begin{bmatrix} 5 & - & 4 & 1 \\ 5 & 4 & - & - \\ 1 & 7 & - & 3 \\ 2 & 8 & 1 & - \end{bmatrix}\end{array}$$

The idea of a **Recommender System** in this context is to use the crowd's votes to complete entries.

There are two prominently used methods for Recommender Systems

- **User-Based Collaborative Filtering**
- **Item-Based Collaborative Filtering**

## User-Based Collaborative Filtering

Predict Carmen's rating for Alien using **similar user ratings**

$$\begin{array}{rc} & \begin{array}{cccc}\textbf{Alien} & \textbf{Bug's Life} & \textbf{Cars} & \textbf{Dark Night}\end{array} \\ \begin{array}{r}\textbf{Carmen}\\ \textbf{Joseph}\\ \textbf{Leonore}\\ \textbf{Esmerelda}\end{array} & \begin{bmatrix} ? & 4 & 4 & 1 \\ 5 & 4 & - & 2 \\ 1 & 1 & - & 3 \\ 2 & 3 & 1 & - \end{bmatrix}\end{array}$$

## Item-Based Collaborative Filtering

Predict Joseph's rating for Dark Night using **similar items' ratings**

$$\begin{array}{rc} & \begin{array}{cccc}\textbf{Alien} & \textbf{Bug's Life} & \textbf{Cars} & \textbf{Dark Night}\end{array} \\ \begin{array}{r}\textbf{Carmen}\\ \textbf{Joseph}\\ \textbf{Leonore}\\ \textbf{Esmerelda}\end{array} & \begin{bmatrix} 2 & - & 1 & 1 \\ 5 & 4 & 2 & ? \\ 4 & - & 3 & 3 \\ - & 4 & 1 & - \end{bmatrix}\end{array}$$

## User-Based Collaborative Filtering and Correlation

$$\hat{\boldsymbol{R}}(\textbf{Carmen}, \textbf{Alien}) = \bar{R}_{\text{Carmen}} + \frac{\sum_{\text{Users } U} sim(\textbf{Carmen}, U) \times \left(R_{U,\textbf{Alien}} - \bar{R}_U\right)}{\sum_{\text{Users } U} sim(\textbf{Carmen}, U)}$$

Where $\boldsymbol{sim} =$

$$sim(\textbf{Carmen}, U) = \frac{\sum_{\text{Movies } m}\left(R_{\text{Carmen},m} - \bar{R}_{\text{Carmen}}\right) - \left(R_{U,m} - \bar{R}_U\right)}{\sqrt{\sum_{\text{Movies } m}\left(R_{\text{Carmen},m} - \bar{R}_{\text{Carmen}}\right)^2}\sqrt{\sum_{\text{Movies } m}\left(R_{U,m} - \bar{R}_U\right)^2}}$$

$$\hat{R}(\textbf{Carmen}, \textbf{Alien}) = \bar{R}_{\text{Carmen}} + \frac{\sum_{\text{Users } U} sim(\textbf{Carmen}, U) \times (R_{U,\textbf{Alien}} - \bar{R}_U)}{\sum_{\text{Users } U} sim(\textbf{Carmen}, U)}$$

$\hat{R}(\textbf{Carmen}, \textbf{Alien})$ → The estimation of **Carmen's** rating of **Alien**, which she has not seen

$$\hat{R}(\textbf{Carmen}, \textbf{Alien}) = \bar{R}_{\text{Carmen}} + \frac{\sum_{\text{Users } U} sim(\textbf{Carmen}, U) \times (R_{U,\textbf{Alien}} - \bar{R}_U)}{\sum_{\text{Users } U} sim(\textbf{Carmen}, U)}$$

$\bar{R}_{\text{Carmen}}$ → **Carmen's** average rating → $\hat{R} > \bar{R}$ if it is expected that **Carmen** will enjoy **Alien**

$$\hat{R}(\textbf{Carmen}, \textbf{Alien}) = \bar{R}_{\text{Carmen}} + \frac{\sum_{\text{Users } U} sim(\textbf{Carmen}, U) \times (R_{U,\textbf{Alien}} - \bar{R}_U)}{\sum_{\text{Users } U} sim(\textbf{Carmen}, U)}$$

$\sum_{\text{Users } U} sim(\textbf{Carmen}, U) \times (R_{U,\textbf{Alien}} - \bar{R}_U)$ → All other **Users** are considered whether or not they rated **Alien** above or below the **User's** average rating. Each **User** is weighted by their similarity to **Carmen**; taking a weighted average of the **User's** votes.

$$\hat{R}(\textbf{Carmen}, \textbf{Alien}) = \bar{R}_{\text{Carmen}} + \frac{\sum_{\text{Users } U} sim(\textbf{Carmen}, U) \times (R_{U,\textbf{Alien}} - \bar{R}_U)}{\sum_{\text{Users } U} sim(\textbf{Carmen}, U)}$$

$\frac{\sum_{\text{Users } U} sim(\textbf{Carmen}, U) \times (R_{U,\textbf{Alien}} - \bar{R}_U)}{\sum_{\text{Users } U} sim(\textbf{Carmen}, U)}$ → The **Users** are weighted by how similar they are to **Carmen**. The **Users** who are more similar than others will be assigned larger weights; the denominator dictates all weights will sum to one, ultimately providing a prediction of **Carmen's** rating for **Alien**.

It is noted the sum over movies rated by both users term in the denominator:

$$\sqrt{\sum_{\text{Movies } m} \left(R_{\text{Carmen},m} - \bar{R}_{\text{Carmen}}\right)^2}$$

The above term is only summed over the movies that both users have rated. Therefore, the metric will favor users who have watched few movies that overlap with Carmen, but agree with Carmen on the ratings of movies they have both watched.

## **User**-Based Collaborative Filtering and Correlation

- A simple approach to apply and interpret within a model
- However, if there are many users & movies with few ratings, similarities will be inaccurate and estimates will likely be poor
- Does not take into account how similar **items** are with each other

## **Item**-Based Collaborative Filtering

The algorithm applied in **Item**-Based Collaborated Filtering is the same as **User**-Based. The correlations are used to weight the previous recommendations by how similar an **item** is to past items.

$$\begin{array}{r c} & \begin{array}{cccc} \textbf{Alien} & \textbf{Bug's Life} & \textbf{Cars} & \textbf{Dark Night} \end{array} \\ \begin{array}{r} \textbf{Carmen} \\ \textbf{Joseph} \\ \textbf{Leonore} \\ \textbf{Esmerelda} \end{array} & \begin{bmatrix} 2 & - & 1 & 1 \\ 5 & 4 & 2 & ? \\ 4 & - & 3 & 3 \\ - & 4 & 1 & - \end{bmatrix} \end{array}$$

## **Item**-Based Collaborative Filtering and Correlation

- Simple to use and is generally more robust in making predictions with **item-item** similarities opposed to **user-user** similarities
- Suffers from poor estimates when items are either not popular or do not contain much data
- Does not take into account how similar **users** are to the current one