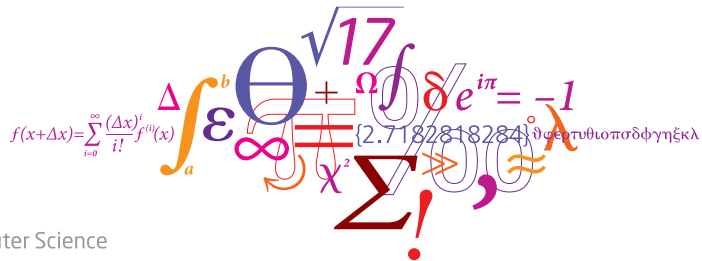
# 02525: Movie recommendation with collaborative filtering (Week 1)

Deena Francis

Postdoc, Section for Cognitive Systems (CogSys),

DTU Compute, Technical University of Denmark (DTU)

email: dfra@dtu.dk




DTU Compute

Department of Applied Mathematics and Computer Science

# Overview

## Week 1

- Introduction to recommender systems
- Introduction to collaborative filtering
- Similarity
- User based filtering
- Homework and Exercise 1

## Week 2

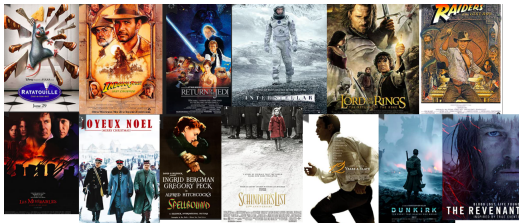
- Item based filtering
- Collaborative filtering - why it works, issues
- Evaluating the performance
- Analysis of top 50 IMDB movies recommendation
- Information about report
- Exercise 2

## Recommender systems

- Automated systems that can utilize the sheer volumes of data that are available in various forms to provide its users with meaningful predictions or recommendations.

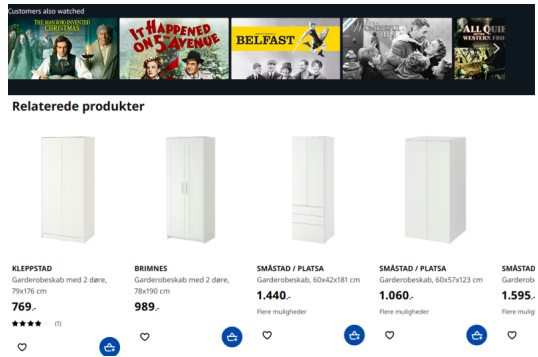
# Motivation

- An excess of choices: millions of books, movies, songs, ...




# Motivation

- An excess of choices: millions of books, movies, songs, ...
- A powerful tool - automation + quality
- Many applications



## Machine learning

Machine learning is a sub-field within computer science and artificial intelligence, which enables computers to *learn* without being explicitly programmed.

- Find patterns from data
- Make predictions

## Machine learning

Popular applications of machine learning:

- Computer vision (CV): facial recognition, medical imaging
- Robotics and reinforcement learning: self-driving cars, production robots, AI in computer games
- Audio: voice recognition, hearing aids and implants
- Natural language processing (NLP): text analysis, spam filtering, genome sequencing / bioinformatics
- General statistical modeling: finance, data-driven marketing, traffic, pharma and biotech
- **Recommender systems**: movies (Netflix), books (Amazon)

## A short discussion

How will I recommend a movie to a friend?

## A short discussion

How will I recommend a movie to a friend?

- Collect **data** about movies and ratings of people.
- Identify popular movies.
- Collect data about your friend's preferences in movies.

This is exactly what collaborative filtering does - with some math!

## Building blocks

- Data
- Algorithms

## Collaborative filtering

- Collaborative filtering is a collection of algorithms that predict *ratings* based on *similarities*.
- "Collaborative": data of users (people)
- "Filtering": choosing

## Assumptions

- Users like movies that other **similar users** like (user-based).
- Users like movies **similar** to those **movies** that they already like (item-based).

## Data - Ratings matrix

- N = Number of unique users/individuals
- M = Number of unique movies

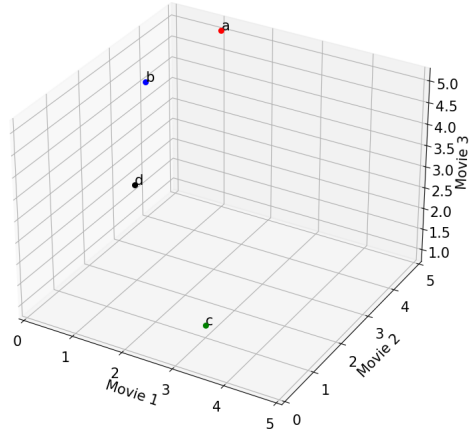
$$\mathbf{R} = \begin{bmatrix} R_{1,1} & R_{1,2} & ... & R_{1,M} \\ R_{2,1} & R_{2,2} & ... & R_{2,M} \\ \vdots & \vdots & \vdots & \vdots \\ R_{N,1} & R_{N,2} & ... & R_{N,M} \end{bmatrix}$$

The values of $\mathbf{R}$ are ratings of the users for each movie, usually in a scale of 0 to 5.

# Similarity

| | Movie 1 | Movie 2 | Movie 3 |
|---|---|---|---|
| a | 1 | 5 | 5 |
| b | 0 | 4 | 4 |
| c | 3 | 1 | 1 |
| d | 1 | 2 | 3 |

- Vectorized form of ratings of users a, b, c, d: $\mathbf{R}_a, \mathbf{R}_b, \mathbf{R}_c, \mathbf{R}_d$ respectively.
- For example: $\mathbf{R}_a = [1, 5, 5]$

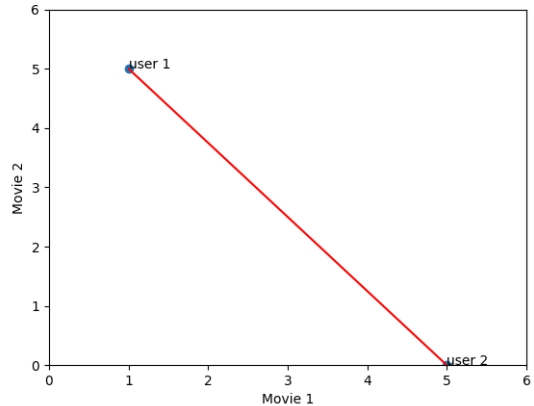


# Similarity - Euclidean similarity

For any two **general** users: a and b.

We compute the distance:

$$D_{Euclidean}(a,b) = \sqrt{\left(\sum_{i=1}^{M}(R_{a,i} - R_{b,i})^2\right)}$$



## Similarity - Euclidean similarity

and then define similarity:

$$S_{Euclidean}(a, b) = \frac{1}{1 + D_{Euclidean}(a, b)}$$

# Similarity - Pearson's similarity

- Pearson's Correlation Coefficient (p).

$$S_{Pearson}(a,b) = \frac{\frac{1}{M}\sum_{i=1}^{M}(R_{a,i} - \bar{R}_a)(R_{b,i} - \bar{R}_b)}{\sqrt{\frac{1}{M}\sum_{i=1}^{M}(R_{a,i} - \bar{R}_a)^2} \cdot \sqrt{\frac{1}{M}\sum_{i=1}^{M}(R_{b,i} - \bar{R}_b)^2}}$$

- A measure of **linear correlation** between two sets of data.

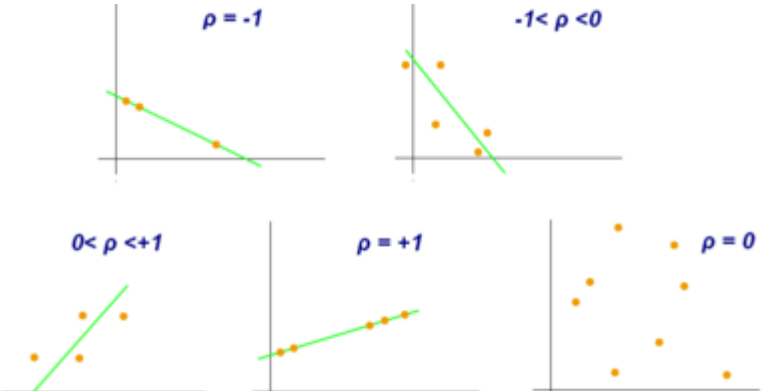
Here

$$\bar{R}_a = \frac{1}{M}\sum_{i=1}^{M} R_{a,i}$$

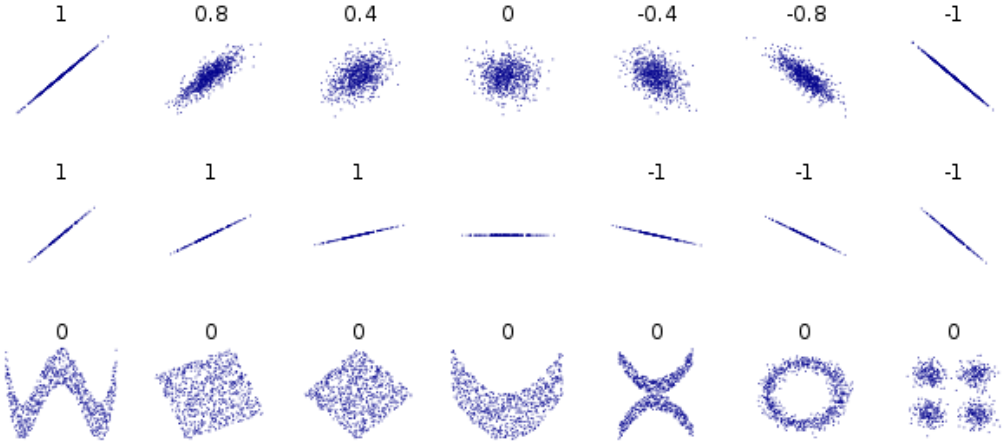
$$\bar{R}_b = \frac{1}{M}\sum_{i=1}^{M} R_{b,i}$$

(the mean or average ratings)

# Pearson's similarity



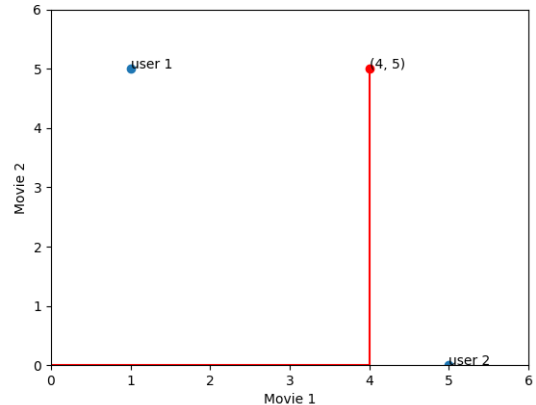
# Pearson's similarity



# Manhattan similarity

$$D_{Manhattan} = \sum_{i=1}^{M} |R_{a,i} - R_{b,i}|$$

$$S_{Manhattan}(a, b) = \frac{1}{1 + D_{Manhattan}}$$



## Similarity measures

Consider two users $a$ and $b$.

- Manhattan similarity:

$$S_{Manhattan}(a,b) = \frac{1}{1 + \sum_{i=1}^{M} |R_{a,i} - R_{b,i}|}$$

- Euclidean similarity:

$$S_{Euclidean}(a,b) = \frac{1}{1 + \sqrt{\sum_{i=1}^{M} (R_{a,i} - R_{b,i})^2}}$$

- Pearson's similarity:

$$S_{Pearson}(a,b) = \frac{\frac{1}{M} \sum_{i=1}^{M} (R_{a,i} - \bar{R}_a)(R_{b,i} - \bar{R}_b)}{\sqrt{\frac{1}{M} \sum_{i=1}^{M} (R_{a,i} - \bar{R}_a)^2} \cdot \sqrt{\frac{1}{M} \sum_{i=1}^{M} (R_{b,i} - \bar{R}_b)^2}}$$

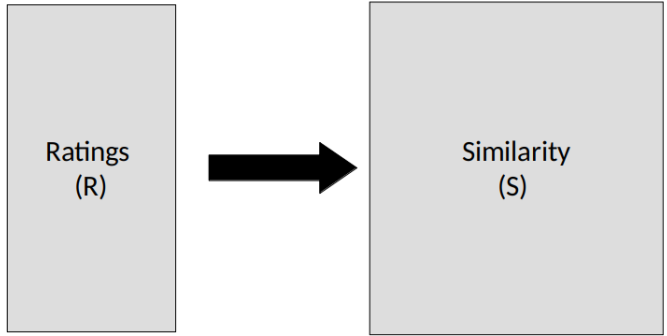
# Similarity matrix

$$\mathbf{S} = \begin{bmatrix} S_{1,1} & S_{1,2} & ... & S_{1,N} \\ S_{2,1} & S_{2,2} & ... & S_{2,N} \\ \vdots & \vdots & \vdots & \vdots \\ S_{N,1} & S_{N,2} & ... & S_{N,N} \end{bmatrix}$$

- It captures the similarity between all users.

# Similarity matrix

- How is $\mathbf{S}$ computed?




It is computed from the ratings matrix and a chosen similarity measure.

## A small exercise

Given,

$$\mathbf{R} = \begin{bmatrix} 2 & 1 & 1 \\ 2 & 1 & 5 \\ 5 & 3 & 1 \end{bmatrix}$$

1. Compute the similarity matrix using Manhattan similarity measure for the above ratings matrix.
2. Who is user 3's nearest neighbor (top most similar)?

## A small exercise - solution

$$\mathbf{R} = \begin{bmatrix} 2 & 1 & 1 \\ 2 & 1 & 5 \\ 5 & 3 & 1 \end{bmatrix}$$

1. Compute the Manhattan similarity matrix for the above ratings matrix.
   - We have 3 users, call them 1, 2 and 3.
   - First we compute the similarity between them using Manhattan similarity measure.

$$S_{1,2} = \frac{1}{1 + \sum_{i=1}^{3} |R_{1,i} - R_{2,i}|} = \frac{1}{1 + |2-2| + |1-1| + |1-5|} = \frac{1}{5}$$

$$S_{2,3} = \frac{1}{1 + \sum_{i=1}^{3} |R_{2,i} - R_{3,i}|} = \frac{1}{1 + |2-5| + |1-3| + |5-1|} = \frac{1}{10}$$

$$S_{1,3} = \frac{1}{1 + \sum_{i=1}^{3} |R_{1,i} - R_{3,i}|} = \frac{1}{1 + |2-5| + |1-3| + |1-1|} = \frac{1}{6}$$

# A small exercise - solution

$$\mathbf{S} = \begin{bmatrix} 1 & \frac{1}{5} & \frac{1}{6} \\ \frac{1}{5} & 1 & \frac{1}{10} \\ \frac{1}{6} & \frac{1}{10} & 1 \end{bmatrix}$$

2. Who is user 3's nearest neighbor (top most similar)?
   - User 1

## Generalized nearest neighbors

- We looked at the nearest neighbor for a user.
- In general, we can find top $k$ neighbors for any user.

For example:

$$\mathbf{S} = \begin{bmatrix} 1 & 0.76 & 0.19 & 0.84 \\ 0.76 & 1 & 0.3 & 0.28 \\ 0.19 & 0.3 & 1 & 0.69 \\ 0.84 & 0.28 & 0.69 & 1 \end{bmatrix}$$

Top $k = 2$ neighbors of user 4 are: users 1 and 3.

KNN(user 4) $= \{1, 3\}$

## Prediction

**Goal**: Compute predicted rating of movie $m$ by user $a$ denoted as $P$ using collaborative filtering with $k$ nearest neighbors.

## Prediction

There are a few ways to compute this prediction $P$ for a user $a$ and movie $m$.

- Average prediction

$$P_{avg}(a,m) = \frac{1}{K} \sum_{b \in KNN(a)} R_{b,m}$$

- Weighted average prediction

$$P_{wavg}(a,m) = \sum_{b \in KNN(a)} w_{a,b} R_{b,m}$$

- Weighted average corrected

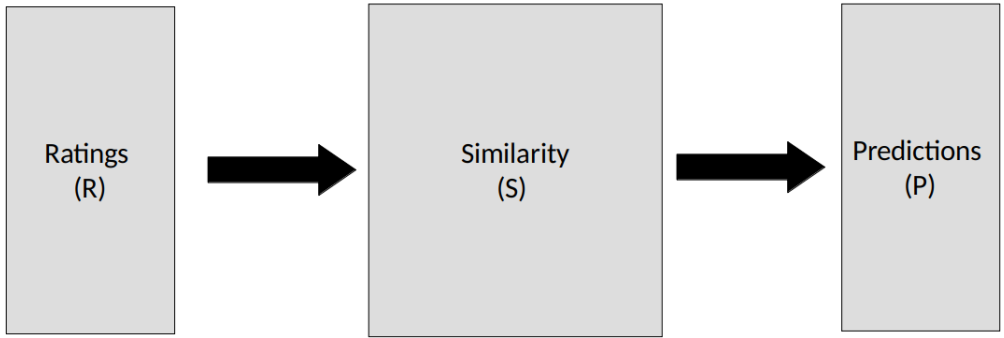
$$P_{wavg-corrected}(a,m) = \bar{R}_a + \sum_{b \in KNN(a)} w_{a,b}(R_{b,m} - \bar{R}_b)$$

where

$$w_{a,b} = \frac{S_{a,b}}{\sum_{b \in KNN(a)} S_{a,b}}$$

## Discussion

1. Can you describe in words what the predictions $P$ mean?


Ratings
(R)
Similarity
(S)
Predictions
(P)

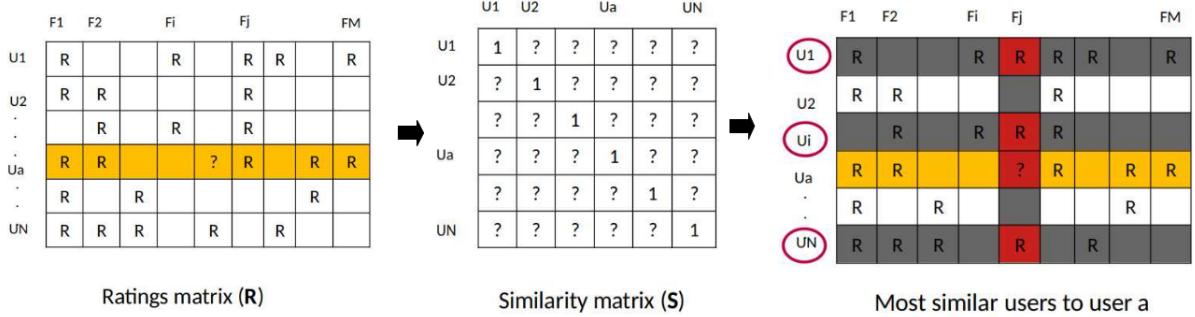
# The recommendation algorithm

1. Get the ratings matrix $\mathbf{R}$ of $N$ users and $M$ movies.
2. Compute the similarity matrix $\mathbf{S}$ using a similarity measure.
3. Find the $k$ nearest neighbors of user $a$.
4. Calculate predictions for all unseen movies for user $a$.
5. Recommend user $a$ his/her top $l$ movies ($l = \{1, 2, ..., M\}$).

## User-based collaborative filtering

- The similarity matrix is computed between users.
- What happens when new user is added?

# User-based collaborative filtering




Ratings matrix (**R**)

Similarity matrix (**S**)

Most similar users to user a

Prediction and recommendations

# Homework

- You are asked to fill out a form of movie recommendations in this link: Top50-IMDB-movies
- Please rate only the movies you have seen.
- I will create a recommended movie for each person next week.

**Rating scheme to use:**

- 0: Bad
- 1: I wouldn't watch it again, but I could recommend it to people I don't like.
- 2: I might not watch it again, but maybe there are others who like it. Maybe I will give this movie a few years.
- 3: Nice movie, but missing a little more of what I like.
- 4: Really good movie! Definitely something I want to see again and recommend.
- 5: Legendary movie! Shall we see it again?

## Exercises

- Exercises are given in the file *02525-Movie-recommendations-Exercise1.pdf* (the file can be found on Learn).
- You should do exercise 1 given in the document today (3:00 to 5:00 pm)
- Please go in and fill in the form by Sunday at 23:59.
  Link: Top50-IMDB-movies

  Then I will get some predictions for next time.

Thanks and see you next Thursday!