

## COMP9417

# Machine Learning and Data Mining

## Assignment

## Topic: 3.4

# Movie recommender system using collaborative filtering

## Joel Lawrence (z3331029)

## Deepansh Singh (z5199370)

### August 2019

Introduction

The goal of this project is to determine whether a user will enjoy a movie or not. To do this, we examine the MovieLens Dataset which contains 100,000 movie ratings from users across a giant sample of movies.

Various methods will be tested to determine an overall “best” approach and then an examination of this method and refining it. Best will be measured by error when comparing our predicted result vs the test set.

Importantly, a predicted rating should aim to not only be accurate, but also be able to effectively determine enjoyment or not given the incredibly varied choices and tastes of people.

The project will work through each method and arrive at the Collaborative Filtering approach for recommender systems. The underlying assumption of the collaborative filtering approach is that if a person A has the same opinion as a person B on an issue, A is more likely to have B's opinion on a different issue than that of a randomly chosen person.

From a machine learning perspective, the idea is to find the most similar (cosine similarity method) users to your target user (Nearest Neighbours approach) and weight their ratings of an item as the prediction of the rating of this item for target user. There are quite a few limitations of this method. It doesn’t handle sparsity well when no one in the neighbourhood rated an item that is what you are trying to predict for target user. Also, it’s not computational efficient as the growth of the number of users and products.

# Implementation

### Data Curation

If your work was mostly implementation, focus on that. Otherwise briefly describe what you did.

### Coding Design and Packages used

Deepansh: explain ideas here please

# Experimentation

### Deepansh Experimentation on different approaches

All methods must be tested on some data, so these results should be included. Additionally, if this was a major focus, you will need to explain the work done and what was accomplished, for example on setting up the learning task, choice of evaluation, and so on. Detailed statistical analyses are probably outwith the scope of the project, so don't include these unless you are already very familiar with this kind of thing.

CF method and result here:

### Collaborative Filtering

Initially, the focus was on implementing Collaborative Filtering on the overall set to determine how it performed as an approach compared to our initial experimentation methods. It performed well in comparison. It had met our initial hypothesis of predicting results however we decided that it could still be improved.

The idea here was that outliers and people who were vastly different from the target user could be adversely affecting the target user’s prediction, even taking into account similarity weightings. To test this idea, we modified our Collaborative Filtering approach to only consider the Top-K most similar users when predicting a movie rating. The K value can represent any number less than the total number of users in the dataset.

As can be seen in the appendix in *Figure 1*, the result drastically improved from the overall when taking the Top-25 and Top-50 users for comparison, \*currently running a further analysis on the exact sweet spot\*

References

Should be there for algorithms used or other aspects of the work.

Copy-paste all algo’s here



# Appendix

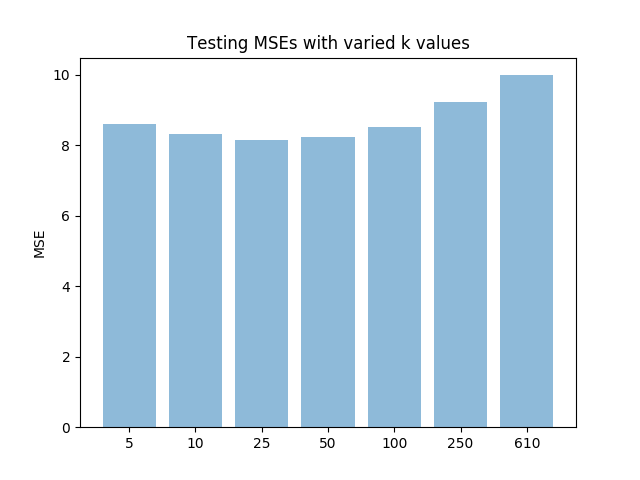


Figure : MSE error by varied K values

Should be used if you have a lot of experimental results. However, consider plotting graphs or using other visualizations like histograms to summarize a lot of results concisely.

Place any graphs here