CS145 Project Introduction

Movie Rating Predictions

Instructor: Yizhou Sun
TAs: Yunsheng Bai, Shengming Zhang

01/14/2019
Project Introduction

- Background & Motivation
- Project Task
- Dataset
- Evaluation
- Project Deadlines and Grading
Background

MovieLens Dataset

MovieLens helps you find movies you will like. Rate movies to build a custom taste profile, then MovieLens recommends other movies for you to watch.
Background Motivation

- Based on the movie watch history of a user (together with ratings), for an unwatched movie, can you predict the ratings that the user possibly gives?
- Recommendation System
  - Recommend potentially high-rating movies
- Multi-information
- Various types of data
- A good fit for our class
Based on the information of

- **Ratings**: UserId, MovieId, Rating, Timestamp
- **Tags**: UserId, MovieId, Tag, Timestamp
  - Tags are user-generated metadata
- **Movies**: MovieId, Title, Genres
  - Action, Adventure, Animation etc.
- **Links**: MovieId, IMDBId, TMDBIId
  - Helps you refer to additional information
- **Tag Genome**: MovieId, TagId, Relevance
  - Relevance indicates how strong a movie is associated with a tag.

We aim to:

Given an unseen UserId-MovieId pair, predict the rating.
Training part: Use training data to build your model

- Movie Genre/Tag Genome information
- User tag/rating history ~12M ratings

Validation part: Use validation data to evaluate your model

- user-movie pairs that were unseen in the training part
- ratings for the pairs ~4M ratings

Testing part: Get a score based on the testing result

- user-movie pairs that were unseen above ~4M tests
Dataset

For Our Course Project:

Based on the MovieLens Latest Datasets

The corresponding user IDs are re-hashed

Do not try to retrieve the original dataset and decode our hashing, you will end up wasting time, trust us :)
Submit your results to Kaggle!

Try your model on the Kaggle competition: https://www.kaggle.com/c/movieratepredictions

See your score on the leaderboard

Evaluation:

RMSE: Root Mean Squared Error

More details of the Kaggle competition will be released soon
Project Grading

- Midterm Report (5% of total)
- Final Report (20% of total)
- Performance on Kaggle
  - 50% of your final report score will be influenced by the performance of your project on the leaderboard (which is 10% of total).
Approximately 3 pages

Current progress about project, including
- Data processing and transformation
- Designed & tested models / methods

Discussion and future project plan
- Some conclusions and findings
- Analysis of current models and techniques
- Timeline of future project plan (around 4 weeks)

Details about midterm report guidelines will be released later!
No longer than 10 pages

Must include:
- Group member information
- Data selection and pre-processing
- Model and techniques
- Evaluation and conclusion
- *Current leaderboard rank and score
- References and credit (papers, other’s codes, maximum 1 page)
- Related work (maximum ½ page)
- Task distribution form and peer evaluation form

ACM paper format

Must NOT include:
- Background or too much description on given original datasets
- Any source code

Details about final report guidelines will be released later!
<table>
<thead>
<tr>
<th>Task</th>
<th>People</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data processing</td>
<td>Student A</td>
</tr>
<tr>
<td>Implementation: Algorithm 1</td>
<td>Student B, C</td>
</tr>
<tr>
<td>Implementation: Algorithm 2</td>
<td>Student B, D</td>
</tr>
<tr>
<td>Implementation: Algorithm 3</td>
<td>Student A, D</td>
</tr>
<tr>
<td>Writing final report</td>
<td>Student C</td>
</tr>
<tr>
<td>CRITERIA</td>
<td>John</td>
</tr>
<tr>
<td>--------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>Attendance at group meetings</td>
<td>4</td>
</tr>
<tr>
<td>Availability when needed</td>
<td>5</td>
</tr>
<tr>
<td>Highly contributed to writing and proof reading of the final report.</td>
<td>5</td>
</tr>
<tr>
<td>Reliability</td>
<td>5</td>
</tr>
<tr>
<td>Contributed ideas that were of high quality.</td>
<td>4</td>
</tr>
<tr>
<td>Approximately, the amount of time spent on this project was comparable to other group members.</td>
<td>5</td>
</tr>
<tr>
<td>Overall (Would you work with them again?)</td>
<td>5</td>
</tr>
</tbody>
</table>

Question:
Do you think some member in your group should be given a lower score than the group score? If yes, please list the name, and explain why.
Important Dates & Milestones

- **Jan. 18**: Group formation due (1% in total as participation)
- **Feb. 18**: Midterm project report due
- **Mar. 20**: Final project report due (together with all codes)
Thank you!

Enjoy “mining” and good luck!