# Leveraging Bayesian Networks for Movie Recommendation System

Chenshu Liu, Tianqi Li

June 1, 2023

## Abstract

This project aims to develop a robust movie recommendation system by using Bayesian Networks as a our model. The goal is to predict user ratings for movies and provide personalized recommendations based on these predictions. The system will need to handle missing rating data, which is a common challenge in real-world scenarios. The technical work involved will comprise constructing a Bayesian Network using the dataset as a knowledge base, where users, movies, genres, and ratings are represented as nodes. We will use known ratings and genres in the datasets to learn the parameters of the Bayesian Network in order to predict missing ratings. The project will use multiple metrics such as Root Mean Squared Error (RMSE) and R-squared to provide a comprehensive evaluation of the system's performance. This project could contribute to the field of recommender systems by offering a probabilistic and context-aware approach to movie recommendations.

## 1 Group Introduction

- Chenshu Liu is currently a first-year master's student in Bioengineering focusing on integrating deep learning techniques in biomedical and bioengineering topics. His current research involves a Human Computer Interface (HCI) project using single-channel Electromyography (EMG) signal and a computerized tomography (CT) image super-resolution project. He received his bachelor's degrees in Neuroscience and Statistics at UCLA.
- Tianqi Li is currently a first-year master's student in Computer Science. His primary areas of interest lie in the domains of backend engineering, full-stack engineering, and machine learning. He received his bachelor's degree in Computer Science and Mathematics from UW-Madison.

## 2 Motivation and Introduction

With the blooming variety and potency of media platforms, information sharing has became unprecedentedly efficient, which enabled huge opportunities in personalized information feed that are tailored to user's needs [1]. The day-to-day digital content has made recommender systems essential for creating a great user experiences as well as for the targeted advertisement and promotions.

In the context of movie recommendations, these systems face multiple challenges such as data sparsity and context-specific variables. Though there has been numerous attempts of using traditional machine learning or deep learning algorithms to predict user preferences based on prior data and has generated encouraging result, those techniques haven't considered the underlying convoluted interaction of the variables of the topic being investigated. On the other hand, Bayesian network is a versatile method in modeling systems with complex

relationships through joint probabilities [1]. Given that different people have different taste of movies, in order to generate more personalized movie recommendations, our project is motivated by devising a good and robust recommender system to address the aforementioned issues by leveraging Bayesian Networks. This approach aims to capture the dependencies between users, movies, and movie genres to enhance the predictive accuracy of movie ratings. We expect that this project could potentially fit into a bigger picture and contribute to the improvement of the user experience in the digital content domain.

Dataset titled MovieLens 25M collected and distributed by GroupLens, a research lab in the Department of Computer Science and Engineering at the Univeristy of Minnesota, Twin Cities, was chosen for the purpose of this project. The MovieLens 25M dataset contains 25 million ratings of 62,000 movies rated by 162,000 users, which was collected between January 09, 1995 and November 21, 2019 [2]. The dataset also includes an aspect called the genome tagging, which was developed by Vig et al., that offers an alternative tagging technique that allows users to interactively tag the subject which resolves the dilemma of tags being binary and discrete [3]. Overall, the MovieLens dataset contains the following variables that could potentially be considered as nodes for the Bayesian Network model of the project: movie genre, time of rating, genome tagging.

# 3 Course Relevance

Our project is directly related to several key topics covered in our class, such as Bayesian Networks and Propositional Logic. The use of Bayesian Networks is the backbone of this project. We will construct a Bayesian Network to predict the users ratings for our recommendation system, which reflects the principles of probabilistic reasoning and graphical models covered in the class. Propositional Logic plays an important role in the interpretation each node or attribute in the dataset can be viewed as a propositional variable, and the edges represent logical implications.

# 4 Background

After data exploration, we have determined several preliminary nodes for the Bayesian network, including the users, movies, genres, and ratings. This variables are candidates for the variables that we will later use to construct the Bayesian network model for movie rating prediction. Once the structure of the Bayesian network is determined, we can construct probability tables for the variables in each node and determine joint probabilities.

Further, we will use the knowledge learned in class about propositional logic and implementation of probabilistic programming techniques to generate preliminary Bayesian network models. At the same time, we need to decide on whether to use collaborative or content-based filtering for the recommender system. Collaborative filtering, which is the aggregation of a large population of users' preference to predict the preference of another user, which is not efficient in generating highly personalized suggestions. On the other hand, content-based filtering, which focuses on the subject's previous preferences, requires large amount of data to be satisfactory [1]. Thus, we need to consider the personalize-ability vs. data-intensive trade-off when constructing the model.

# 5 Technical Contribution

## 5.1 Objective

In this project, the technical contribution lies in the development and evaluation of a Bayesian network-based model for predicting user ratings for movies they haven't watched. The objective is to leverage the user's past movie ratings and movie genres to make predictions about their ratings for unseen movies. We chose to evaluate the performance of Bayesian network on the task movie rating prediction given the movie, genre, and evaluator. Our assumption is that Bayesian network can have a better performance than other traditional machine learning algorithms in the prediction task, because Bayesian networks models the uncertainty

and dependencies between variables using probability theory, which makes it capable of solving problems where uncertainty and probabilistic reasoning are important, such as decision support systems of movie rating. At the same time, since we are dealing with a really dynamic knowledge domain: movie rating, where new movies are being released overtime, and new ratings are being generated by people along the way, the scope of the knowledge base is increasing. Since Bayesian network support incremental learning which allows the model to be updated as new data becomes available. This adaptability is valuable in dynamic environments where the data distribution or relationships may change over time. It enables the model to evolve and improve its predictions based on new evidence.

## 5.2 Implementation

The technical merit of this project includes the following components.

### 5.2.1 Bayesian Network Model

This project employs a Bayesian network using the 'pgmpy' library to capture the dependencies between the user's rating, user ID, and movie genres. The model represents the relationships and conditional dependencies among these variables, allowing for the prediction of the user's ratings for movies they haven't seen.

### 5.2.2 Data Preprocessing

The project involves preprocessing the movie ratings dataset and merging it with the movie metadata dataset to obtain relevant information about the movies, such as the genres for different movies. Then data is transformed into a suitable format for the Bayesian network model by encoding genre presence with binary indicators.

### 5.2.3 Training and Evaluation

The Bayesian network model is trained on the user's past ratings data. The model is trained using the Bayesian estimator, which estimates the parameters of the network based on the available data. The performance of the model is evaluated using the root mean squared error (RMSE) and R-Square metrics.

## 5.3 Progress and Anticipated Challenges

The project has made progress in implementing the Bayesian network model for rating prediction based on user preferences and genres. The code successfully trains the model on the user's historical ratings and makes predictions for unseen movies. Some anticipated challenges include:

1. Fine-tune: The structure of Bayesian network model (e.g., how edges should be constructed) require tuning for optimal performance. Exploring different edge settings and conducting cross-validation can be challenging but necessary to improve the model's accuracy and generalization.

2. Generalization to Diverse Movie Genres: The model relies on movie genres as a key factor in predicting ratings. However, generalizing the model to handle a diverse range of movie genres, including rare or niche genres for some outlier movies, could pose challenges in terms of data representation and model adaptation which has a direct impact on the prediction accuracy.

# 6 Measures of Success

## 6.1 Baseline:

The baseline is to build and train a Bayesian Network that captures our knowledge base and predicts the ratings. If we can train the network and use it to make non-random predictions of missing ratings, we will consider that as a baseline level of success.

## 6.2 Medium:

We will evaluate the system using standard metrics such as precision, recall, and F1-score. We will also explore vertical analysis scheme that cross-compare the Bayesian Network model with the performance of other traditional machine learning

models. If the recommendation system can achieve satisfactory performance according to these metrics comparable or outperform other ML models (such as decision trees), we will consider it an unqualified success.

### 6.3 Stretch:

Given extra time, we might consider extending the complexity of our model to include more attributes (variables) such as movie tags and directors. At the same time, we may try to include some other variables, such as the genomics tags, provided by the MovieLens dataset, and aggregate other related datasets to further construct the knowledge graph. For example, given that the demographic information of the reviewers were not disclosed in the MovieLens dataset, we can incorporate. We could also explore the application of more advanced techniques for parameter learning and inference to achieve an even higher accuracy.

## 7 Planning and Timeline

### 7.1 Week 6-7: Model Design and Implementation

During this period, both members will collaboratively pre-process the datasets provided by Grouplens Research, including aggregating data tables using shared variables, data engineering of variables, and formatting. And then develop and implement the Bayesian Network model to achieve the baseline.

### 7.2 Week 8: Model Evaluation and Tuning

Member 1 will set up the metrics and evaluate the model performance. Member 2 will assist with this process, conduct necessary model tuning, and start drafting the project report. Then both members will work together aiming at the medium objective.

### 7.3 Week 9-10: Report drafting and stretch challenge

Both members will work on drafting the report. If there is extra time, we will work on the stretch objective. As the project concludes, both members will finalize the report.

## References


[1] Lin Yu-Chu, Yuusuke Kawakita, Etsuko Suzuki, and Haruhisa Ichikawa. "Personalized Clothing-Recommendation System based on a Modified Bayesian Network". In: *2012 IEEE/IPSJ 12th International Symposium on Applications and the Internet*. IEEE, July 2012, pp. 414–417. ISBN: 978-1-4673-2001-6 978-0-7695-4737-4. DOI: 10.1109/SAINT.2012.75. URL: http://ieeexplore.ieee.org/document/6305322/.

[2] F. Maxwell Harper and Joseph A. Konstan. "The MovieLens Datasets: History and Context". In: *ACM Transactions on Interactive Intelligent Systems (TiiS)* 5.4 (2015), 19:1–19:19. DOI: 10.1145/2827872. URL: https://doi.org/10.1145/2827872.

[3] Jesse Vig, Shilad Sen, and John Riedl. "The Tag Genome: Encoding Community Knowledge to Support Novel Interaction". In: *ACM Transactions on Interactive Intelligent Systems* 2.3 (Sept. 2012), pp. 1–44. ISSN: 2160-6455, 2160-6463. DOI: 10.1145/2362394.2362395. URL: https://dl.acm.org/doi/10.1145/2362394.2362395.