M.Sc. Research Project Proposal
Bill (Mufan) Li
Supervised by Professor Jeffrey Rosenthal
Department of Statistical Sciences
University of Toronto
Phone: (416)843-9179
Email: mufan.li@mail.utoronto.ca

Assessing GPA Measures From a Collaborative Filtering Perspective

Recent developments in machine learning have made significant contributions to a wide range of fields that are not traditionally considered data science. Notably the Netflix competition have attracted a collective effort in developing models that greatly improved prediction of movie ratings by different users, creating the best movie recommendation system at the time [2]. Similar to the Netflix rating data, student grades in different courses follow the same structure, allowing the application of the same machine learning techniques. This research project aims to apply these techniques to analyze the student grade dataset from [1], which contains complete transcripts of undergraduate students from a major Canadian University. Like predicting user ratings, we are able to predict the grades for courses. From the predictions, this project intends to analyze the effect of choosing easier courses on student grades, specifically by comparing the predicted grades of courses students did not take against the courses taken within the same program. By analyzing the variation in course difficulty, these results could potentially improve curriculum design for educational institutions and admission procedure for graduate programs.

Inference on missing data in the student grade dataset falls directly under a collaborative filtering (CF) problem, where a grade is only assigned for some matches of students and courses. The greatest difficulty of this type of problem is the sparsity of data, where each student can only take a small subset of courses, leaving majority of potential course grades missing. Additionally, data is not even distributed among courses, where some courses can have few attendance. That being said, matrix factorization and restricted Boltzmann machines techniques have been highly effective at collaborative filtering. This project intends to investigate these two techniques and the respective extensions with the student grade dataset. This proposal will provide a brief overview of two main techniques that will be investigated, as well as a discussion on several problems this research project intend to address.

The most basic matrix factorization (MF) method, known as singular value decomposition (SVD), decomposes a large matrix into a product of two low-rank matrices. Specifically for this problem, suppose there are M students and N courses, then we define $\mathbf{A} \in \mathbb{R}^{M\times N}$ as the matrix of grades, where A_{ij} corresponds to the grade of student i and course j. We then seek two low-rank matrices $\mathbf{U} \in \mathbb{R}^{M\times d}$ and $\mathbf{V} \in \mathbb{R}^{N\times d}$ such that

$$\mathbf{A} \approx \mathbf{U}\mathbf{V}^\top. \tag{1}$$

This is called a rank d approximation of $\mathbf{A}$. Note the value A_{ij} is the dot product of the ith row of $\mathbf{U}$ and the jth row of $\mathbf{V}$. The row vectors may be interpreted as features, i.e. row i of $\mathbf{U}$ contains all the information of student i. A common goal of optimization is to minimize the Frobenius norm of the difference for a rank d approximation $\left\|\mathbf{A} - \mathbf{U}\mathbf{V}^\top\right\|$, defined by the square root of the sum of squares of its entries [2]. In the case of collaborative filtering, the Frobenius norm is not computed for missing entries.

The greatest advantage of SVD is the simplicity of both implementation and inference, where model optimization requires only least square or gradient type methods, and each missing value A_{ij} is just the dot product of two row vectors of $\mathbf{U}$ and $\mathbf{V}$. However, one significant weakness is SVD cannot incorporate student or course specific information besides the grades. For example, a student's retaken or dropped courses cannot be easily used by SVD. Additionally, since the Frobenius norm cannot be computed for unknown values, SVD fails to optimize for regions where the data is sparse. One extension of MF is by defining a probabilistic distribution using the matrix product as the mean. Under the probabilistic MF setup, we can shift the mean of the distribution based on student and course specific information. In fact, [3] showed incorporating the probabilistic MF successfully resolves both problems for the Netflix dataset.

On the other hand, restricted Boltzmann machines (RBM) is a completely different approach to the problem. A RBM is a Markov random field in the form of a bipartite graph, where the joint probability follows a Boltzmann type distribution. The bipartite graph structure creates two layers without internal connections. One layer, called the visible layer, contain the observed values of grades for a specific student. These nodes are connected to the other layer, called the hidden layer, with symmetrical weighted connections.

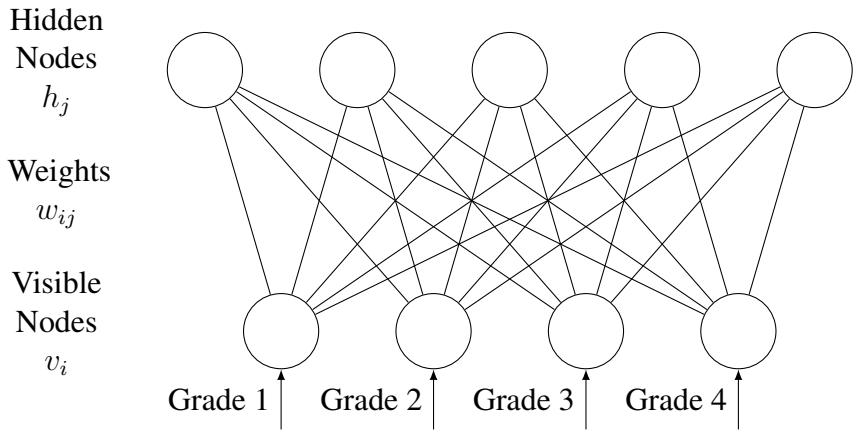

Figure 1: A restricted Boltzmann machine (RBM) with 4 courses and 5 hidden nodes for a specific student.

Suppose the graph have N visible nodes and M hidden nodes, with each visible node denoted v_i, hidden nodes denoted h_j, weights between two nodes w_{ij}, b_i and a_j be bias parameters, and σ_i be the standard deviation of grades for each course. Here each visible node v_i represents the grade for course i, where a specific student is fixed. Let $\theta = \{w_{ij}, a_j, b_i, \sigma_i\}\ \forall i, j$, $\mathbf{v} = \{v_i\}\ \forall i$, and $\mathbf{h} = \{h_j\}\ \forall j$ denote the collections. Additionally, we let the hidden nodes only take on binary values, i.e. $v_i \in \mathbb{R}, h_j \in \{0, 1\}$. We can then define the energy function and the joint distribution for the graph:

$$\begin{aligned} E(\mathbf{v}, \mathbf{h}|\theta) &= \sum_{i=1}^{N} \frac{(b_i - v_i)^2}{2\sigma_i} - \sum_{i=1}^{N}\sum_{j=1}^{M} w_{ij} h_j \frac{v_i}{\sigma_i} - \sum_{i=j}^{M} a_j h_j \\ P(\mathbf{v}, \mathbf{h}|\theta) &= \frac{\exp\left[-E(\mathbf{v}, \mathbf{h}|\theta)\right]}{\mathcal{Z}} \end{aligned} \tag{2}$$

where $\mathcal{Z}$ is the partition function normalizing the distribution. After marginalizing over the hidden nodes $\mathbf{h}$, we can find the gradient of the likelihood function with respect to the parameters θ to perform steepest

descent optimization. Finding the gradient requires the use of Gibbs sampling, although [5] showed the approximate gradient after very few iterations of Gibbs sampling is sufficient for optimization.

To perform inference on a missing grade value, one simply include an additional "visible" node v_p, where the value is not known, but can be determined by the energy function:

$$P(v_p|\mathbf{v}) \propto \sum_{\mathbf{h}} \exp[-E(v_p, \mathbf{v}, \mathbf{h})] \tag{3}$$

Similar to probabilistic MF, RBM can also easily incorporate student specific information within the model. The RBM also has a natural extension into multiple layers called deep belief networks (DBN) [4]. This type of networks can represent much more complex relationships at a cost of more difficult optimization and expensive computation for inferencing.

Both MF and RBM have been proven successful in the Netflix competition [2], therefore these techniques are expected to perform well in other collaborative filtering type problems. In this project, we hope to address the following questions:

- Can the model correctly predict grades for future (or missing) courses given previous year's (or incomplete) grades?
- Does the current GPA measure appropriately represent academic abilities when students select easier courses to improve grades?

By leaving out some data during the optimization process, predictions on these data can be tested. This is commonly referred to the training data and test data in machine learning. With inference for missing grades, we can correctly assess the prediction power of the model, answering the first question. If grades are predicted for all courses that are available, we can estimate a "fair" grade for students as if all students have taken the same courses. This allows grading of student abilities on comparable grounds, regardless of which courses the students have taken. Finally, we can quantify the expected grade improvement due to the course selection by optimizing the choice of courses for all students in the dataset. This results in a hypothetical grade, comparing to the true grade answers the second question.

We intend to fully protect the privacy of students in the dataset during research by using completely anonymous data. The project will use the data to provide valuable insights to educational institutions, potentially improving curriculum design and fair grading.

References

[1] Michael A Bailey, Jeffrey S Rosenthal, and Albert H Yoon. Grades and incentives: assessing competing grade point average measures and postgraduate outcomes. *Studies in Higher Education*, (ahead-of-print):1–15, 2014.

[2] Andrey Feuerverger, Yu He, Shashi Khatri, et al. Statistical significance of the netflix challenge. *Statistical Science*, 27(2):202–231, 2012.

[3] Andriy Mnih and Ruslan Salakhutdinov. Probabilistic matrix factorization. In *Advances in neural information processing systems*, pages 1257–1264, 2007.

[4] Ruslan Salakhutdinov. *Learning deep generative models*. PhD thesis, University of Toronto, 2009.

[5] Ruslan Salakhutdinov, Andriy Mnih, and Geoffrey Hinton. Restricted boltzmann machines for collaborative filtering. In *Proceedings of the 24th international conference on Machine learning*, pages 791–798. ACM, 2007.