STA4000H Research Course

Student: Mufan Li
Supervisor: Jeffery Rosenthal
Term: Winter 2016

Coursework Description

Recent developments in machine learning have made significant contributions to a wide range of fields that are not traditionally considered data science. In this research course, we intend to explore several of the machine learning techniques in applications to education.

Specifically, this research project aims to apply machine learning to analyze the student grade dataset from [2], which contains complete transcripts of undergraduate students from a major Canadian University. Similar to 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.

The project will focus on implementing three main methods of inference:

1. Matrix factorization (MF) [1] and if time permits probabilistic matrix factorization (PMF) [3, 4]
2. Restricted Boltzmann machines (RBM) [5]
3. Denoising auto-encoders (DAE) [6] and if time permits variational auto-encoders (VAE) [7]

The student is expected to evaluate each model's performance on the dataset, analyze the output results, and recommend policy changes to adjust for student behaviors in choosing courses. The student will be graded on weekly progress (40%) and a final paper with source code (60%).

References

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

[2] 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.

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

[4] Ruslan Salakhutdinov and Andriy Mnih. Bayesian probabilistic matrix factorization using markov chain monte carlo. In *Proceedings of the 25th international conference on Machine learning*, pages 880–887. ACM, 2008.

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

[6] Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, and Pierre-Antoine Manzagol. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *The Journal of Machine Learning Research*, 11:3371–3408, 2010.

[7] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.