Hierarchical modeling in education performance prediction

James Nguyen
Georgia Tech
Atlanta, GA, USA
cuong.nguyen@gatech.edu

ABSTRACT

Recent advances in Machine Learning and Data Mining enable new branches of educational research to thrive, particularly e-learning and data capturing through e-learning systems. [9] Academic performance research benefits strongly from this trend through better data and models. In this paper, I investigate **hierarchical models** - a flexible family of advanced parametric models - that could achieve not only high **out-of-sample prediction power** but also **meaningful confidence intervals** for **statistical inference**. Reviews and empirical comparisons versus traditional machine learning methods on an educational dataset are also provided in details.

ACM Classification Keywords

H.5.m. Information Interfaces and Presentation (e.g. HCI): Miscellaneous; J.4.m. Computer Applications: SOCIAL AND BEHAVIORAL SCIENCES; G.3.m. Mathematics of computing: Probability and Statistics - Multivariate statistics; I.5.1.m. Computing methodologies: Pattern Recognition - Models

Author Keywords

Parametric Models; Education Research; E-learning; Student Performance; Predictive Modeling; Statistical Inference; Bayesian computation; STAN; Hierarchical Models; MCMC; Confidence Interval; Prediction Interval

HISTORICAL REVIEW OF EDUCATION RESEARCH

The paper reviews the evolution of quantitative predictive methods in education research to motivate interest in advanced parametric methods.

Parametric Statistical Models

Traditionally, research in education and educational policies have applied similar quantitative methods to those used in the social sciences [5], because the main goal of these educational inquiries is to **establish causal effects**. **Parametric**, sparse models which allow ease of **hypothesis testing** therefore are both prevalent and important, while black-box, flexible predictive models are not popular. [13]

Some parametric models offer **prediction capability** with **confidence intervals (CIs)**. However, unlike causality testing, prediction only take secondary roles in traditional education research, so these models are not built to optimize for out-of-sample accuracy. [9].

Recent Machine learning methods

Advances in EdTech enable new branches of educational research to thrive, particularly **intelligent learning systems**. [9] New ways of generating data and richer features through e-learning systems enable researchers to build better predictive models. People apply more modern methods on these datasets; **ensembles, neural nets** and **SVMs** are the most common methods when optimizing for **out-of-sample accuracy**. [10], [1]

The body of work that apply predictive ML methods on educational data tends to *weight practicality more than interpretability*, and they do not focus on answering traditional inference questions. [9, 8]

RESEARCH GOAL AND METHODOLOGY

Could we build flexible predictive models that offer *not only high out-of-sample prediction power but also meaningful confidence intervals for statistical inference.*? A family of advanced parametric models that could satisfy this demand is **hierarchical models** - we will investigate their construction process and predictive power in this paper.

Hierarchical modeling in education research

Hierarchical models offer a balance between predictive accuracy and interpretability. ***Hierarchical models are particularly suitable for educational research, because education data are generally collected from a hierarchy.*** [11, 6]

For example, you can have performance scores for students in a state, at different districts, different schools within the same district, different class levels within the same school etc. Of course, if the data are generated in a true controlled fashion with balanced groups representation across all covariates, then we can do a pooled t-test, but these are survey or longitudinal data with known **biases** and **imbalances**. [7]

There are few studies focusing on *hierarchical models in academic performance* [14, 12], and there is no current study that focuses on building a highly predictive hierarchical model first before interpreting statistical significance.

Bayesian Techniques in Hierarchical modeling

Fitting hierarchical model is one of the major usage of **Bayesian techniques**. Previously, due to computational complexity of the sampling algorithms in these Bayesian techniques, building hierarchical models are hard. However, recent advances in hardware and software have opened up these techniques to a wider range of social science problems. [2]

DATA OVERVIEW

In coming sections, this research builds hierarchical models on a student performance dataset and compares those hierarchical models' predictive power against state-of-the-art tree-based models built on the same data.

The reference educational dataset in [1] is collected from a learning management system (LMS) using an activity tracker. This learning system monitors learning progress like reading an article or watching a training video.

The data consist of **480 student records** and **17 features**, with final performance as the desired predictive outcome - 1.

Features

The features are classified into three major categories:

- Demographic: gender and nationality etc. - Fig. 1, 2
- Academic: topic, grade level and section etc. - Fig. 3
- Behavioral: raised hand on class, visited resources, answering survey by parents, and school satisfaction - Fig 4

Reference tree-based models

Reference state-of-the-art tree-based models are taken from **Amrieh et. al.** [1], where the authors built learners on this data and show that ensemble models achieve strong out-of-sample prediction. I reproduced their results using **25pct** of data as hold-out test set with accuracy listed in table 1.

Shortcomings of tree-based/black-box methods

For inference, the predicted values from tree-based/black-box models do not provide useful information that aids decision making, e.g. confidence intervals and significance of contributing factors. [3, 8] E.g, by topic, the dataset in [1] is unbalanced (**Fig. 3**): **IT** has 95 students vs. **Maths** has 21 students, so performance prediction for a new Maths student should be a lot less certain than that of a new IT student. Meanwhile, none of the blackbox models can reflect such confidence adjustments based on their predictive values.

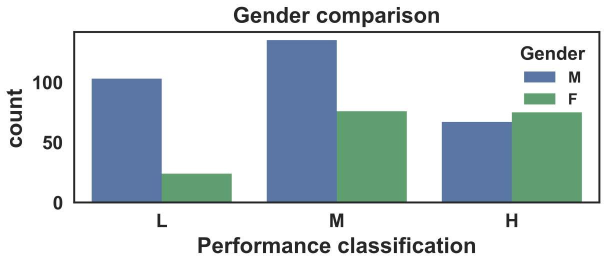

Figure 1. Women tend to perform better than men on average.

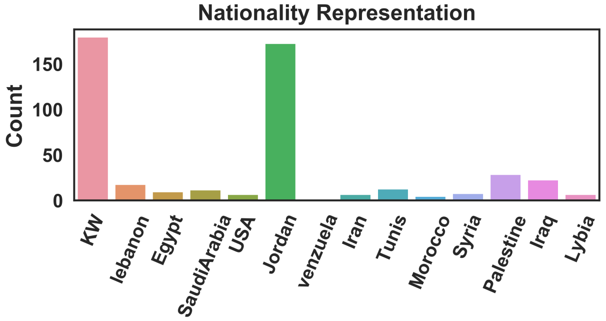

Figure 2. The students come from different origins: 179 are from Kuwait, 172 are from Jordan, and smaller groups < 30 from Palestine, Iraq, Lebanon, Tunis, Saudi Arabia, Egypt, Syria etc.

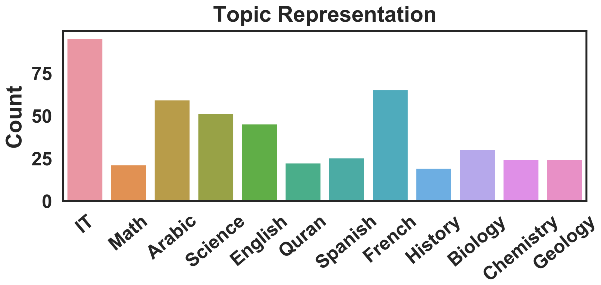

Figure 3. Multiple topics included - this feature is also unbalanced.

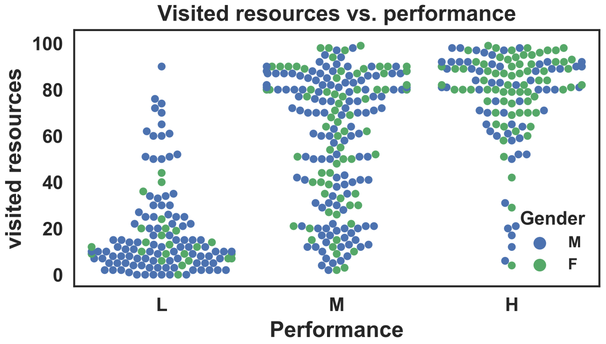

Figure 4. Tendency to visit resources is a *behavioral feature* - only made available by E-learning systems. This plot shows that students who received a lower grade (L) visited fever resources than students that scored a M or H grade. Additionally, women who received a high mark (H) almost exclusively visited a lot of the on-line resources.

Model	*Accuracy*
Decision Tree	0.77
Random Forest	0.83
Hierarchical Topic x Gender	**0.89**
Hierarchical Topic x Country	0.89

Table 1. Hierarchical models do better than both decision tree and random forest for out-of-sample prediction. The best hierarchical model (bold) accounts for groups interactions between Topics, Student Gender and responsible Parent Gender.

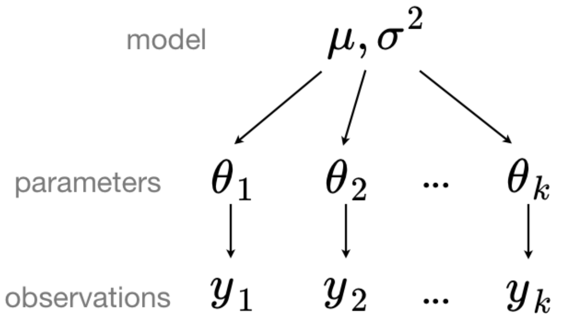

Figure 5. In a hierarchical model, we see data at both individual-level (y_t) and at group-level. Group-level parameters themselves (θ_i) are viewed as a sample from a distribution governed by hyper-parameters (μ and σ). Thus, we view θ_i as being neither entirely different or exactly the same - this is *partial pooling*.

MODEL BUILDING AND ASSUMPTIONS

This section walks through the parameterization of hierarchical models used, the model assumptions and their justifications. For a quick recap of hierarchical models and visualization, see Figure 5; a more detailed treatment is available in [4].

Hierarchy specifications

Figure 5 shows how we could view natural groups in the dataset such as Topics, Nationality or Gender as one level above the individual students. Each of these groups has its own average/parameter (θ_i), but the groups' averages are related to each others - **partial pooling**. Partial pooling is very useful in dealing with unbalanced data, where some categories have only a few samples and we must borrow inference/prediction from other groups' parameters - e.g. Topics in Figure 3.

Model parameterization

At the data level, we have the logistic regression:

$$Pr(y_i = 1) = logit^{-1}(X_i\theta_{j[i]}), \text{ for i=1,...,n} \quad (1)$$

Where **X** is the matrix of individual-level features and $j[i]$ indexes the Topic/Gender groups that student i belongs to.

The group-level coefficients θ_j follows another distribution:

$$\theta_j \,\tilde{}\, Gaussian(\mu, \sigma_\theta^2), \text{ for j=1,..., 48 (48 groups)} \quad (2)$$

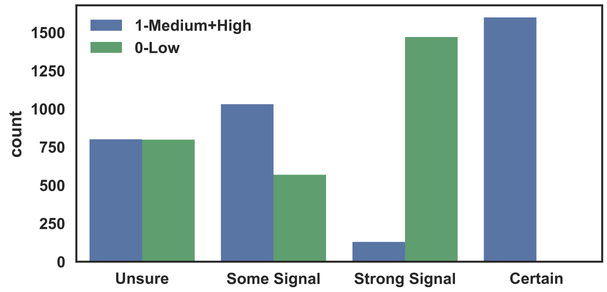

Figure 6. Uncertainty about out-of-sample predictions are demonstrated for 4 cases across the spectrum here. Since the model generates binary outcomes through MCMC simulation, if the split is 50-50, then the model is really not making any strong assertion - see the *Unsure* case on the left. From left to right, the splits of the simulations get more pronounced, meaning in those predictive samples, the model is more and more certain of its prediction.

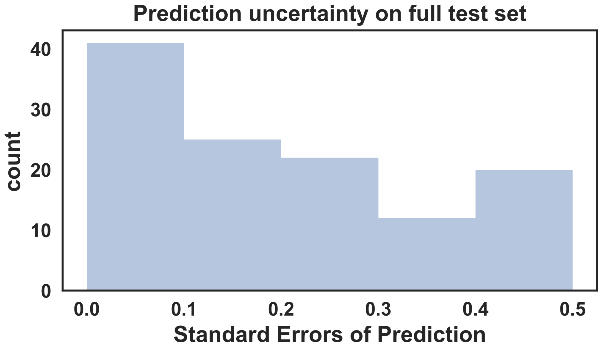

Figure 7. This graph shows the empirical uncertainty of all 120 out-of-sample predictions, measured by standard errors. Notice that about 17pct of these samples have really high uncertainties (>0.4), which means the model almost refuses to classify these samples, and the assignment of these predictions in to classes are almost purely numerical noises.

Where μ is the vector of group-level average per features and σ_θ^2 governs how much variability we expect in group-level coefficient θ_j. The model for θ_j in (2) allows us to include all 48 groups in the full model without worrying about collinearity, and this two-level parameterization gives the model the name "hierarchical".

Several hierarchies types and depths are explored, but for this data, I found a simple 1-level hierarchy of **Topic x Student Gender x Parent Gender** interaction groups perform the best.

Model performance

One caveat for this dataset is that we actually need **2** logistic models specified by Equation **(1)** above, because we must classify **Low vs. Medium/High** and **High vs. Medium+Low** separately then combine the predictions for a 3-outcome target.

The out-of-sample accuracy of the hierarchical models outperform reference tree-based models in [1] - shown in Table 1. Not only that, this type of models can also make use of the **partial pooling** concept to predict outcomes for observations from a new group - by using the group average μ [4].

Prediction confidence

These hierarchical models are fitted using MCMC methods, so a distinctive advantage over black-box models is that each out-of-sample prediction comes with a whole distribution, which allows researchers to assess how confident the model is about a particular point estimate. Figures 6 and 7 look deeper into different cases where the confidence varies.

The ability of this model to allow statistical inference on predictive cases are very useful, because it could aid users to **not** make a decision. For example, if the model is used to detect cheating or predict adverse drug reaction, refraining from making quick actions on uncertain predictions are almost always better than acting quickly on numerical noises (see Figure 7).

CONCLUSION

This paper demonstrates that it is possible to build advanced parametric models - **hierarchical models** in particular - that can outperform popular machine learning methods in prediction accuracy. Furthermore, hierarchical models also provides meaningful confidence intervals for statistical inference and decision making, which makes them well applicable to education research.

For future works, and extensions of the models built here with better assumptions and more informative priors could achieve even higher accuracy. Better methodology of model comparison and selection could also be further investigated. Alternatively, deeper inference into the fitted parameters could shed lights on the variability and differences of the groups in the hierarchy, which contributes further domain-specific insights to the educational research community.

REFERENCES

1. Elaf Abu Amrieh, Thair Hamtini, and Ibrahim Aljarah. 2016. Mining educational data to predict StudentâĂŹs academic performance using ensemble methods. *International Journal of Database Theory and Application* 9, 8 (2016), 119–136.
2. Bob Carpenter, Andrew Gelman, Matthew D Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. 2017. Stan: A probabilistic programming language. *Journal of statistical software* 76, 1 (2017).
3. Bradley Efron and Trevor Hastie. 2016. *Computer age statistical inference*. Vol. 5. Cambridge University Press.
4. A Gelman, JB Carlin, HS Stern, and DB Rubin. Bayesian data analysis,3rd Edition. (????).
5. Andrew Gelman and Jeronimo Cortina. 2009. *A quantitative tour of the social sciences*. Cambridge University Press.
6. David Kaplan. 2016. Causal inference with large-scale assessments in education from a Bayesian perspective: A review and synthesis. *Large-Scale Assessments in Education* 4, 1 (2016), 7.
7. Christoph König and Rens van de Schoot. 2017. Bayesian statistics in educational research: A look at the current state of affairs. *Educational Review* (2017), 1–24.
8. Yannick Meier, Jie Xu, Onur Atan, and Mihaela Van der Schaar. 2016. Predicting grades. *IEEE Transactions on Signal Processing* 64, 4 (2016), 959–972.
9. Cristóbal Romero and Sebastián Ventura. 2010. Educational data mining: a review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 40, 6 (2010), 601–618.
10. Amirah Mohamed Shahiri, Wahidah Husain, and others. 2015. A review on predicting student's performance using data mining techniques. *Procedia Computer Science* 72 (2015), 414–422.
11. Anders Skrondal and Sophia Rabe-Hesketh. 2009. Prediction in multilevel generalized linear models. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 172, 3 (2009), 659–687.
12. Bidya Raj Subedi, Nancy Reese, and Randy Powell. 2015. Measuring teacher effectiveness through hierarchical linear models: Exploring predictors of student achievement and truancy. *Journal of Education and Training Studies* 3, 2 (2015), 34–43.
13. Timothy Teo. 2014. *Handbook of quantitative methods for educational research*. Springer Science & Business Media.
14. Liang-Ting Tsai and Chih-Chien Yang. 2015. Hierarchical effects of school-, classroom-, and student-level factors on the science performance of eighth-grade Taiwanese students. *International Journal of Science Education* 37, 8 (2015), 1166–1181.