Machine Learning Approaches for Learning Analytics: Collaborative Filtering Or Regression With Experts?

Kangwook Lee
School of Electrical Engineering
KAIST
Daejeon, Korea
kw1jjang@kaist.ac.kr

Jichan Chung
School of Electrical Engineering
KAIST
Daejeon, Korea
jichan3751@kaist.ac.kr

Yeongmin Cha
Riiid, Inc.
Seoul, Korea
ymcha@riiid.co

Changho Suh
School of Electrical Engineering
KAIST
Daejeon, Korea
chsuh@kaist.ac.kr

Abstract

An intelligent learning analytics based on an enormous amount of education data is a key enabler of the next generation of education; among many tasks of the intelligent learning analytics, personalized prediction of test responses based on the record of each individual learner is of the utmost importance. In recent years, a variety of machine learning algorithms for predicting test outcomes have been proposed, and two of the most prominent approaches are collaborative filtering and logistic regression. Collaborative filtering is fully data-driven since it does not require any extra information other than test outcomes while logistic regression is applicable only when questions can be independently analyzed by experts. In this work, we first propose a new model for test responses, and propose a collaborative filtering algorithm with enhanced human-interpretability based on the new model. Then, we evaluate the prediction performance of these approaches using a large education data set, collected via mobile applications for English Language Learning. Our experimental results show that the fully data-driven collaborative filtering approach can predict test outcomes better than the logistic regression approach.

1 Introduction

Large-scale online education platforms such as Massive Open Online Courses (MOOCs) have become the source of an enormous amount of education data [8]. The goal of a modern intelligent learning analytics is to understand and optimize the learning progress of students using *big education data* with appropriate models [20]. Among many tasks of the intelligent learning analytics, personalized prediction of test responses based on the record of each individual learner is of the utmost importance.

A variety of machine learning algorithms have been proposed for data-driven learning analytics, and *collaborative filtering* has attracted much attention from researchers in recent years [2, 13]. Collaborative filtering is a popular technique for estimating preference (or taste) of users by discovering the implicit correlation between the revealed responses of users [21], and it is the core of the recommendation engines in a variety of applications, e.g., movie recommendation systems [1], product recommendation systems for e-commerce [14], and social networks [11], etc. The implicit correlation between the responses of different users can be captured by assuming latent variable associated

with each items, which we call *hidden concepts*. A learning analytics system is indeed a particular instance of recommendation systems: students and their learning progresses to the learning analytics systems are users and their preferences to the recommendation systems; and questions to the learning analytics systems are items to the recommendation systems. Recently, a few notable works [2, 13] have proposed novel collaborative filtering approaches for learning analytics, and the superior prediction performances of collaborative filtering-based approaches were reported. Further, such collaborative filtering approaches could predict test outcomes with test response data only, and do not necessarily require question analysis results, which is usually obtained through a lengthy, expensive process.

While collaborative filtering approaches are known to efficiently analyze both students and questions simultaneously, to the best of our knowledge, it has not been rigorously compared with other algorithms that can enjoy the benefit of question analysis results. More specifically, if every question is associated with relevant concepts by experts or teachers with appropriate weights, it is not clear whether one must resort to collaborative filtering-based algorithms: a simple regression that can effectively utilize question analysis results may yield more accurate student analysis results and hence superior prediction performance. Moreover, the existing collaborative filtering algorithms result in models with low human-interpretability.

This precisely sets the goal of this paper: *we conduct a systematic comparison of our collaborative filtering algorithm, which requires the response data of students only, and a simple regression algorithm, which requires both the response data of students and experts' analysis of questions.* For experiments, we prepare a set of TOEIC (Test Of English for International Communication) preparation questions, which are independently analyzed by English experts, and then collect responses from students via an online education application, called SantaTOEIC. Further, based on the variation of a popular item response theory (IRT) model called the multidimensional two-parameter logistic (M2PL) latent trait model, we propose a new collaborative filtering algorithm with enhanced human-interpretability, and compare the prediction performance of our collaborative filtering algorithm and that of logistic regression equipped with question analysis results.

Our preliminary results indicate that our collaborative filtering-based algorithm provides more accurate prediction than a regression algorithm even though the regression fully exploits the question analysis results. This implies that even when one is given with a careful analysis of the question set, a blind adoption of the given analysis results may result in a poor learning analytics system. We believe that such an unexpected observation is due to the inexactness of question analysis results since it is heavily subject to human errors and biases. It is also possible that the key concepts which experts believe useful for analyzing questions are not fully capturing the nature of questions, fortifying the need of fully data-driven learning analytics.

Related Works. Among many algorithms for collaborative filtering, matrix completion – a technique that can be used to fill a low rank matrix with unobserved entires [5–7] – has recently gained its popularity. For instance, the authors of [7] show that, under some mild conditions, one can reliably fill a square matrix of size n by n and of rank r if the number of observed entries is order of $nr\mathrm{polylog}(n)$. This can be done by filling the missing entries of the matrix so that the nuclear norm of the matrix is minimized. A similar result holds even when the observed entries are not exact but noisy. Including convex program approaches, many other efficient algorithms (e.g., spectral methods, non-convex algorithms, stochastic algorithms) have been proposed in the literature for matrix completion [4, 10, 19]. In this paper, we specifically use a stochastic gradient descent algorithm for matrix completion, proposed by [19], in order to solve our optimization problem.

Item Response Theory (IRT) is a study of mathematical models for item analysis and student analysis. McKinley and Reckase propose the multidimensional two-parameter logistic (M2PL) latent trait model in [15]. The M2PL model assumes that there exist multiple factors affecting test responses. Further, each student is associated with a vector whose elements represent his/her understanding of different factors, and similarly each question is associated with a vector whose elements represent the correlation of different factors with the question. Berner et al. adopt this model for learning analytics, apply a collaborative filtering algorithm, and report the superior prediction performance of the proposed approach over the existing approaches [2]. Lan et al. propose a new algorithm that imposes sparsity on the question analysis results and hence provide human-interpretable question analysis results [13]. While the existing works have focused on developing a better model and a new algorithm for learning analytics, our goal is different: we would like to see whether other simple

algorithms, equipped with an external source of information such as question analysis results, can outperform purely data-driven approaches.

Organization. The rest of this paper is structured as follows. In Sec. 2, we present the our model for test responses, and define the relevant parameters. In Sec. 3, we first illustrate a logistic regression-based approach and our collaborative filtering-based approach. We describe the detailed description of the question set and the data collection process in Sec. 4, and present the experiment results in Sec. 5. We conclude the paper with a few interesting discussion topics and future research directions in Sec. 6.

2 Model

Consider an education system with n students and m questions. We assume that students' responses are generated according to the following variation of the M2PL latent trait model [15]. For each i, $1 \leq i \leq n$, student i is associated with an r-dimensional row vector $L_i \in [0,1]^{1\times r}$, where r is the upper bound on the number of *hidden concepts*. The j^{th} component of L_i represents the level of student i's understanding of the j^{th} hidden concept. For each i, $1 \leq i \leq m$, question i is associated with an r-dimensional row vector $R_i \in [0,1]^{1\times r}$, where the j^{th} component of R_i represents the fractional contribution of the j^{th} hidden concept to question i, and the sum of components of R_i is 1, i.e., $\sum_{j=1}^{r} R_i(j) = 1$. For notational simplicity, we define the student-concept matrix L and the question-concept matrix R: $L = [L_1^T, L_2^T, \ldots, L_n^T]^T \in [0,1]^{n\times r}$ and $R = [R_1^T, R_2^T, \ldots, R_m^T] \in [0,1]^{m\times r}$. For a pair of student index i and question index j, the level of student i's understanding of question j is defined as $X_{ij} = \sum_{k=1}^{r} L_i(k)R_j(k) = L_i R_j^T$. Further, we assume a non-linear mapping from the level of understanding to the probability of correct guess. More specifically, we use a logistic function for such a mapping: Given X_{ij}, the probability that student i correctly answers question j is defined as $\phi(X_{ij}) = \phi_a + \frac{1-\phi_a}{1+e^{-\phi_c(X_{ij}-\phi_b)}}$, where ϕ_a, ϕ_b, and ϕ_c are appropriately set, independently of questions or users. We define the understanding level matrix X and the probability of correct answer matrix P as follows: $X = [X_{ij}] \in [0,1]^{n\times m}$ and $P = [P_{ij}] \in [0,1]^{n\times m}$. Note that $X = LR^T$ and $P = \phi(X)$, where $\phi(\cdot)$ is applied component-wise. Finally, we assume that $Y_{ij} \in \{0,1\}$, which represents whether student i guessed the correct answer for question j (1) or not (0), follows a Bernoulli random distribution with success probability P_{ij}. We denote by Ω the set of student index-question index pairs for observed entries. Further, we denote by $\Omega_{i\star}$ and $\Omega_{\star j}$ the set of question indices attempted by student i and the set of indices of users who attempted question j, respectively. That is, $\Omega_{i\star} = \{j|(i,j) \in \Omega\}$, and $\Omega_{\star j} = \{i|(i,j) \in \Omega\}$.

One can note that the model described above cannot capture the inherent difficulty of problems. In order to resolve this issue, we introduce the following two auxiliary concepts: the $(r+1)^{\text{th}}$ concept is one that no one knows, and the $(r+2)^{\text{th}}$ concept is one that everyone knows. This can be imposed by setting $L_i(r+1) = 0$ and $L_i(r+2) = 1$ for all i. For $R_i(r+1)$ and $R_i(r+2)$, we treat them equally as the other hidden concepts, and hence our algorithm will find the estimates of them. In order to understand how the above normalization works, consider the following extreme cases. Imagine that for some j, $R_j(r+1) = 1$ and $R_j(k) = 0$ for all $k \in [r+2] \setminus \{r+1\}$. That is, question j consists of the $(r+1)^{\text{th}}$ concept, which is not known to everyone. Note that this can happen when the question is atypical, or uncommon words appear in the Language Learning question. One would like to model such cases in a way that every user, regardless of their background knowledges, will randomly guess the answer of the question. Note that this will be immediately achieved since we set $L_i(r+1) = 0$ for all i, and hence $X_i(j) = 0$. Similarly, if $R_j(r+2) = 1$ for some j, question j will be correctly answered by all users with probability one. We remark that our model with these auxiliary concepts can also be seen as an alternative form of the M3PL latent trait model since our model captures multidimensional item discrimination, item difficulty, as well as different guessing probability for each problem.

3 Algorithms for Learning Analytics

In this section, we first describe an approach that can be used when one is given the question-concept matrix R in addition to the test response data set Y in Sec. 3.1. In Sec. 3.2, we describe an algorithm that does not require such extra analysis: it instead learns the question-concept matrix R from the test response data set Y_Ω.

3.1 A Logistic Regression-Based Algorithm with R

Given the question-concept matrix R and observed responses Y_Ω, one can use logistic regression [9] to estimate the student-concept matrix L. That is, for each i we solve the following logistic regression problem:

$$\min_{L_i} \sum_{j \in \Omega_{i\star}} [-Y_{ij} \log(P_{ij}) - (1 - Y_{ij}) \log(1 - P_{ij})] \tag{1}$$
$$\text{s.t. } 0 \leq L_{ij} \leq 1, \ \sum_j L_{ij} = 1, \ P_{ij} = L_i R_j^T.$$

We then construct L by concatenating L_i's, and use the constructed L to find P for predicting student responses.

3.2 A Collaborative Filtering Algorithm without R

We now describe a collaborative filtering algorithm that estimates X, each element of which represents a student's level of understanding of a question, using the test response data Y only. More specifically, our algorithm finds the maximum likelihood (ML) estimator of X given a set of observation $\{Y_{ij}\}_{(i,j)\in\Omega}$. Equivalently, the ML estimator can be found by solving a minimization problem whose objective function is the negative of the log likelihood of the observed entries. In order to encourage X to have a low rank solution, we also add to the objective function the nuclear norm regularization term [7]. That is, one would like to solve the following optimization problem:

$$\min_{L,R} \sum_{(i,j)\in\Omega} [-Y_{ij} \log(P_{ij}) - (1 - Y_{ij}) \log(1 - P_{ij})] + \mu \|LR^T\|_* \tag{P1}$$
$$\text{s.t. } 0 \leq L_{ij} \leq 1, \ 0 \leq R_{ij} \leq 1, \ P = LR^T, \ \sum_j L_{ij} = 1, \ \forall i.$$

We approximate the optimization problem (P1) with (P2) by replacing the nuclear norm of X with the sum of the squared Frobenius norms of factor matrices L and R. This approximation is based on the following property of nuclear norm [18]: the nuclear norm of a matrix X is equal to the minimum sum of the squared Frobenius norms of factor matrices L and R such that $X = LR^T$.

$$\min_{L,R} \sum_{(i,j)\in\Omega} [-Y_{ij} \log(P_{ij}) - (1 - Y_{ij}) \log(1 - P_{ij})] + \frac{\mu}{2} \left(\|L\|_F^2 + \|R\|_F^2 \right) \tag{P2}$$
$$\text{s.t. } 0 \leq L_{ij} \leq 1, \ 0 \leq R_{ij} \leq 1, \ P = LR^T, \ \sum_j L_{ij} = 1, \ \forall i.$$

Indeed, any local minimum of the above optimization problem (P2) is known to match that of the global minimum of the original problem (P1) under mild conditions, and this agreement can be further certified by checking rank deficiency of the factored matrices L and R [19].

As a specific choice of an algorithm, we use the projected Stochastic Gradient Descent (SGD) method. That is, starting with randomly initialized $L^{(0)}$ and $R^{(0)}$, we iteratively find a sequence of $L^{(k)}$ and $R^{(k)}$ as follows:

$$L_{i_k}^{(k+1)} = \Pi_{P_L} \left(\left(1 - \frac{\mu_1 \alpha_k}{|\Omega_{i_k\star}|} \right) L_{i_k}^{(t)} - \alpha_k \frac{\phi_c \left(Y_{i_k j_k} - \phi(L_{i_k} R_{j_k}^T) \right)}{\phi(L_{i_k} R_{j_k}^T)(1 + e^{-\phi_c (L_{i_k} R_{j_k}^T - \phi_b)})} R_{j_k}^{(t)} \right), \tag{2}$$

$$R_{j_k}^{(k+1)} = \Pi_{P_R} \left(\left(1 - \frac{\mu_1 \alpha_k}{|\Omega_{\star j_k}|} \right) R_{j_k}^{(t)} - \alpha_k \frac{\phi_c \left(Y_{i_k j_k} - \phi(L_{i_k} R_{j_k}^T) \right)}{\phi(L_{i_k} R_{j_k}^T)(1 + e^{-\phi_c (L_{i_k} R_{j_k}^T - \phi_b)})} L_{i_k}^{(t)} \right), \tag{3}$$

where the index pair (i_k, j_k) is chosen uniformly at random from Ω for k^{th} iteration, and $\Pi_{P_L}(\cdot)$ and $\Pi_{P_R}(\cdot)$ are projections of a vector onto the spaces of feasible L's and R's, respectively. The projected SGD is known to converge to a globally optimal solution when the objective function and the regularization terms (including those induced by constraints) are convex [16]. The objective function of (P2), however, is non-convex, so we run the above algorithm multiple times with different initialization points.

4 Experiment Setup and Data Set

4.1 Question Pool

In order to conduct experiments comparing the prediction performance of different approaches for learning analytics, we first created a pool of TOEIC (Test Of English for International Communication) preparation questions. TOEIC is a test of English for international communication, and each test is composed of 150 multiple-choice questions, divided into 7 parts. Among the 7 parts of TOEIC, we focused on Part 5 questions, with which students are asked to fill a blank with a grammatically correct word or phrase, and Part 6 questions, with which students are asked to read a short paragraph and to fill a blank with a word or phrase that is both grammatically correct as well as consistent with the rest of the given paragraph.

We first created the question pool of 4202 Part 5/6 TOEIC questions, and then had *every* question in the question set *analyzed* by English experts as follows. English experts first investigated the question pool, and then came up with a set of 69 English concepts, which they considered useful and necessary for describing the questions in the question pool. They then tagged each question with up to 6 relevant concepts. [1] More specifically, a total of 15 experts were employed to tag the questions with relevant concepts, and each question was randomly assigned to at least two experts. We develop an online tagging system where the experts were able to individually work on the assigned questions. In order to reduce systematic bias between experts, we chose to reveal the first reviewers response to a question to the second reviewer of the question. That is, the second reviewer's job was to adjust the response of the first reviewer. [2] Fig. 1 shows the histogram of the total number of appearances of the English concepts in our finalized data set. We observe that a few popular concepts are tagged much more frequently than the other English concepts.

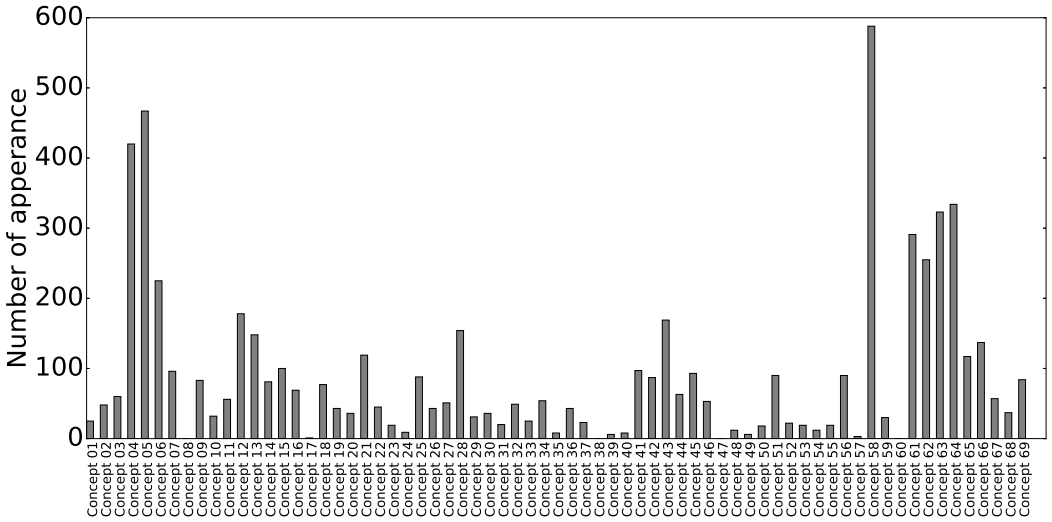


Figure 1: **The total number of appearances of the** 69 **English concepts.** This bar chart shows the total number of appearances of the 69 English concepts in the filtered data set. Note that each of the 1933 questions is associated with up to 6 English concepts. In this bar chart, the gray bars show that the number of appearances of the concepts as a tag. We note that there are 4 concepts that are not tagged at all in our filtered dataset.

Using this question analysis results, we construct a question-concept matrix R, which will be used for the logistic regression approach: if question i is associated with concept $k_1, k_2, \ldots, k_{\mathcal{I}_i}$, we set $R_i = \frac{1}{|\mathcal{I}_i|} \sum_{j=1}^{|\mathcal{I}_i|} e_{k_j}$, where e_i is the i^{th} unit vector.

[1] In addition to the simple tagging relations, the experts rated how relevant each concept is to each question with an integer relevance score, and also identified the most important concept for each question. However, we observe that using this additional information did not improve the performance of the logistic regression-based algorithm, so we chose not use this relevance score.

[2] We did not measure inter-rater reliability (IRR) since the responses of different experts were not independent under our sequential rating scheme.

4.2 Data Collection and Preprocessing

With the TOEIC question data set we prepared, we have collected a large response data set via an online TOEIC education platform, *Riiid SantaTOEIC*. From 1/1/2016 to 8/10/2016, a total of 106612 students had signed up for SantaTOEIC through its iOS and Android applications. A screenshot of the running Android application is shown in Fig. 2.

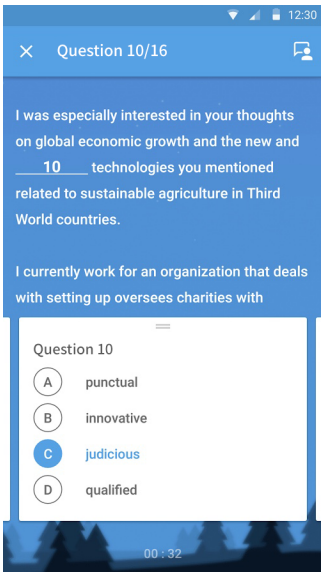


Figure 2: **A screenshot of SantaTOEIC, a Riiid's TOEIC preparation mobile application.** From 1/1/2016 to 8/10/2016, a total of 106612 students had signed up for SantaTOEIC. We preprocess the raw response data set of 13902274 responses in order to obtain a high quality data set. The final data set we use for the experiments consists of 4202 questions, 106612 students, and 13902274 responses.

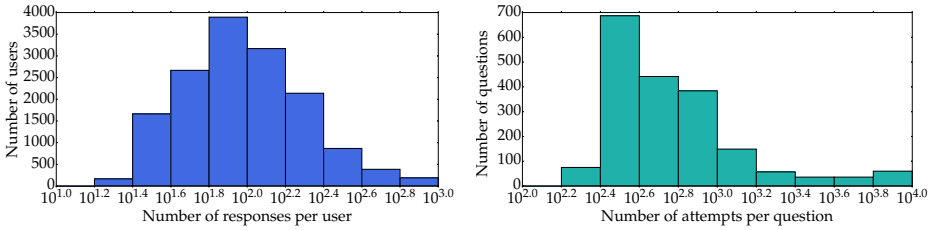
During the data collection period, a total of 13902274 responses had been collected. In order to obtain a high quality data set, we preprocess the raw data set as follows. We first removed the students who had attempted less than 30 questions during the observation period or had spent less than 3 seconds for more than or equal to 95% of their attempts. Then, we filtered out students whose correct answer rate is less than or equal to 30%. After we obtained the refined set of students, we filter out the questions that are responded less than 400 distinct students. The rationale behind these filtering conditions is that those students not satisfying the above conditions are likely to be ones who simply want to try out and explore the mobile applications for fun.

After we applied the aforementioned filtering process, we obtained the final data set consisting of $|\Omega| = 1920085$ responses of $n = 15137$ students on $m = 1933$ questions. Note that the density of the observation matrix is 0.0656 or about 6.5%.

4.3 A Quick Look At The Data Set

In this section, we provide a quick overview of the finalized data set. Fig. 3 shows two histograms: the histogram of the number of questions solved by each student and the histogram of the number of attempts for each question. Note that the minimum number of questions solved by each student is no less than 30 according to the preprocessing step.

Plotted in Fig. 4 is the usage pattern of students in terms of the number of responses collected in each month. We would like to remark that using the entire data set for training is not always the best thing to do. This is because our model is not capturing the changes in student performance, and hence a data set collected over a long time period might not fit well with our model. In our experiments, we observed that using the entire data set still gives the best performance with the validation set but this might not be the case in general. For more discussion about this issue, we refer the readers to Sec. 6.3.



(a) **Student Histogram** (b) **Question Histogram**

Figure 3: **Histograms.** Shown on the left is the histogram of the number of questions solved by each student, and shown on the right is the histogram of the number of attempts for each question.

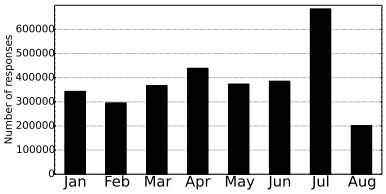


Figure 4: **Monthly Usage Pattern.** Plotted in the figure is the monthly usage pattern of the students.

5 Experiment Results

5.1 Algorithm Implementation and Specification

We first divide the entire data set into the training set (90%) and the test set (10%). All the experiment results reported in this section are with respect to the test set. The learning analytics algorithms described in Sec. 3 are implemented in Python. We then conduct a heuristic optimization for finding the optimal hyper-parameters such as the regularization parameter μ, the sequence of step sizes, the rank of the matrix X, and etc. More precisely, we randomly choose the hyper-parameters, and conduct a grid search for one of the hyper-parameters, where the performance of the algorithm is measured with respect to the validation set, a random partition of the training set. By repeating the above procedure, we chose $r = 4$ and $\mu = 1$; for the step size, we begin with $\alpha = 0.1$ and reduce the step size by a multiplicative factor of $10^{0.5}$ whenever the validation score stops decreasing for 3 epochs in a row. For $\phi(\cdot)$, we use $\phi_a = 0.25,\ \phi_b = 0.5$ and $\phi_c = 10$ for the logistic function, and the rational behind these choices is that one can correctly guess the answer of a question without knowing anything about a 4-choice question with probability at least 0.25. We plot the logistic function with these parameters in Fig. 5.

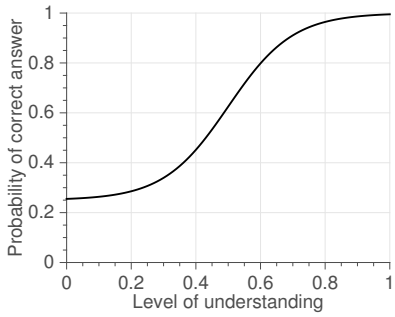
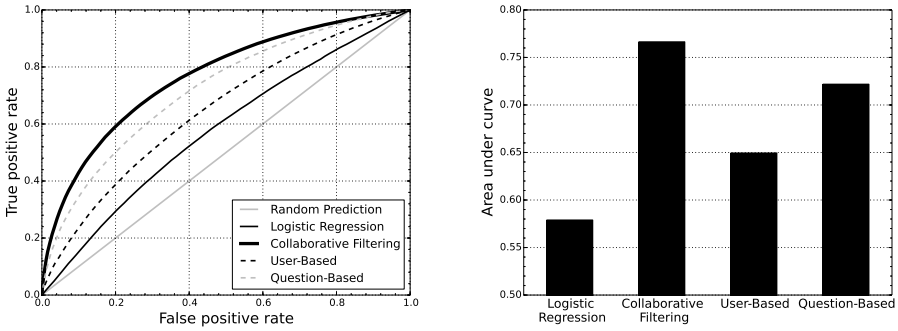


Figure 5: **The logistic function.** The logistic function $\phi(x) = \phi_a + \frac{1-\phi_a}{1+e^{-\phi_c(x-\phi_b)}}$ is used to model a non-linear mapping from the level of understanding to the probability of correct answer. The above plot shows the logistic function with parameters $\phi_a = 0.25,\ \phi_b = 0.5$ and $\phi_c = 10$.



(a) **ROC curves.** (b) **AUC.**

Figure 6: **Prediction performance of the various approaches for learning analytics.**

In addition to the approaches described in Sec. 3, we also implement the student-based prediction algorithm and the question-based prediction algorithm. The student-based prediction algorithm simply estimates the one-dimensional level of each student from the fraction of correct responses of the student in the training set, and similarly the question-based prediction estimates the one-dimensional difficulty of each problem from the fraction of correct responses for the problem in the training set.

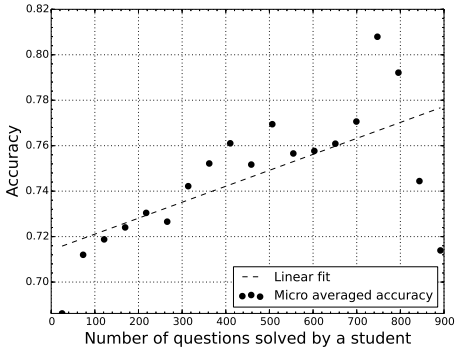
5.2 Prediction Performance

We plot the prediction performance of various approaches with respect to the test set in Fig. 6. More precisely, we train the models using the training set and measure the prediction (classification) performance of various models. A prediction outcome for a data point is called a true positive (negative) if the predictor correctly guessed that the student will respond to the question with a correct (wrong) answer. Similarly, a prediction outcome is a false positive (negative) if the predictor made a wrong guess that the student will respond to the question with a correct (wrong) answer. We denote the number of true positives, false positives, true negatives, and false negatives by tp, fp, tn, and fn, respectively. A few useful performance measures of a classifier are as follows: the accuracy is the fraction of correct predictions or $\frac{\mathrm{tp}+\mathrm{tn}}{\mathrm{tp}+\mathrm{fp}+\mathrm{tn}+\mathrm{fn}}$; the true positive rate is the fraction of true positives among condition positives or $\mathrm{tpr} = \frac{\mathrm{tp}}{\mathrm{tp}+\mathrm{fn}}$; and the false positive rate is the fraction of false positives among condition negatives or $\mathrm{fpr} = \frac{\mathrm{fp}}{\mathrm{fp}+\mathrm{tn}}$.

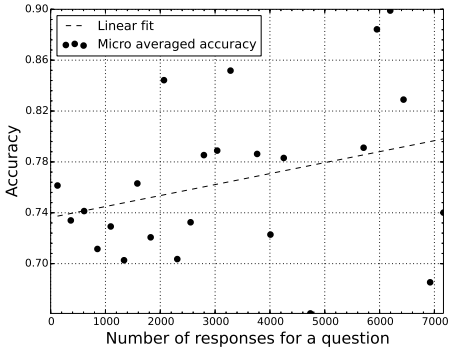
A receiver operating characteristic (ROC) curve is a way of illustrating the performance of a binary classifier: for classification threshold $\theta \in [0, 1]$, the ROC curve is a collection of pairs $(\mathrm{fpr}(\theta), \mathrm{tpr}(\theta))$. Note that the ROC curve of a random predictor is a line segment connecting $(0, 0)$ and $(1, 1)$, and that of a perfect predictor is line segments connecting $(0, 0)$, $(0, 1)$, and $(1, 1)$. Thus, the area under curve (AUC) of a ROC curve can represent the classification performance of a predictor: the larger the AUC is, the better the prediction performance is.

In Fig. 6a, the ROC curves for the various approaches are shown. We can observe that the ROC of the collaborative filtering-based approach is above those of the other approaches. We also compare the AUC of the approaches in Fig. 6b: the collaborative filtering-based approach achieves the highest AUC of 0.766, while the logistic regression approach achieves the AUC of 0.579. More surprisingly, we observe that even simple student-based and question-based approaches achieve higher AUC than the logistic regression approach, which fully exploit the question analysis provided by experts. This result implies that even though one is given question analysis results provided by experts or teachers, it is better to use fully data-driven approaches such as the collaborative filtering-based approach.

We conjecture that the poor performance of the logistic regression approach is due to the property of the R matrix given by the question analysis results. During the process of question analysis, the experts came up with too many different concepts, while they assigned weights to a few number of concepts per question. Since the number of model parameters to be estimated increases linearly in the number of concepts, we believe that the logistic regression model is overfit due to the insufficiency of data. Indeed, we observed an improved performance of the logistic-regression based



(a) Prediction accuracy as a function of the total number of questions solved by a user



(b) Prediction accuracy as a function of the total number of responses for a question

Figure 7: **Prediction performance as a function of the number of per-student, per-question observations.**

approach when we reduced the dimension of the matrix R using non-negative matrix factorization. Even though the improved performances were still not comparable with those of the other performances but it is an interesting open question whether a similar approach can further improve the performance of the logistic regression approach.

5.3 Correlation between the number of responses and prediction accuracy

In Sec. 4, we observed that in our data set, the number of responses per student and the number of responses per question widely vary. One important question is that how the prediction performance changes when the number of questions per student (or per question) increases. In order to answer this question, we first bin the students according to the number of questions submitted by them, and then micro average the prediction accuracy of the students within the same bin. Similarly, we bin the questions according to the number of responses, and then micro average the prediction accuracy of the responses for the questions within the same bin. Plotted in Fig. 7 are micro-averaged accuracy as a function of the number of questions solved by students and that as a function of the number of responses for questions. We use the bin size of 50 for Fig. 7a, and the bin size of 250 for Fig. 7b. From Fig. 7a, we can observe that the predicted accuracy of a student's responses linearly increases as the number of questions submitted by the student increases. Similarly, the predicted accuracy of the responses for a question linearly increases as the number of responses for the question increases. This observation can help design the right number of responses per student or per question at which an intelligent learning analytics system starts providing prediction results with high confidence.

6 Discussion and Conclusion

In this paper, we conducted an empirical evaluation of the collaborative filtering-based approach and the regression approach for learning analytics. We collected the data set via mobile applications for English test preparation exams, post-processed it, and trained our models using a matrix completion algorithm. Even though the regression approach is equipped with the question analysis results prepared by experts, we observed a superior prediction performance of the collaborative filtering-based approach, justifying the superiority of fully data-driven approaches. Our results are still premature since it is not clear whether such a predominance will persist even under more delicate models that can capture the real world better. We conclude the paper by discussing a few interesting open problems and some aspects of our current model that are subject to improvements.

6.1 Human-interpretability and hidden concepts

The normalization constraints on L and R in our proposed model provide an enhanced human-interpretability of the trained model. This is because while the rows of L and R of the vanilla model can have sign ambiguity, our model does not. That is, when $L_{i_1}(k) > L_{i_2}(k)$, student i_1 has a

better understanding of hidden concept k than student i_2. Similarly, $R_{j_1}(k) > R_{j_2}(k)$ implies that hidden concept k is more crucial when one solves question j_1 than when one solves question j_2. We believe that using such monotonicity can provide a way to study how hidden concepts can be interpreted in terms of the concepts that human know. For instance, if $R_j = e_k$ for some j and k, this implies that question j is a representative question of hidden concept k. By inspecting the content of question k, one may be able to understand how hidden concept k can be interpreted using human knowledge. Hence, we believe that our model with new normalization constraints can allow for a better understanding of hidden concepts, eventually making huge impacts on the way we create education content and teach students.

6.2 More than binary responses: Incorporation of other forms of data

While we used the binary response data only, the actual response data set contains several additional sources of side information such as the option chosen by students, the options marked wrong by students, the time taken to respond to a question, and etc. By incorporating the other forms of data with a more complicated model, one may be able to obtain better estimates of students and questions, and hence to provide superior prediction performance as well as personalized learning of a better quality. For instance, the nominal response model (NRM) proposed in [3] can model the probability of students responding to a certain option of a question. Ning et al. propose a new model for option responses with human-interpretable outputs, and show that the new model fits better with real world data as well [17]. It is an interesting future direction to study how one can apply a similar collaborative filtering-approach under such models capturing option responses.

6.3 Time-varying L

The generative model for responses implicitly assumes that the level of students' understanding is *time-invariant*. If the data set is collected over a long time period during which student's level of understanding is likely to fluctuate, such an assumption may totally fail, and the estimated L will be close to the time average of L. In practice, one is usually interested in predicting future responses, not unseen responses of the past, and the estimate of time average of L is hardly useful for predicting the future responses of students.

If one is given with an enormous amount of data, there is a simple fix: one can simply throw out the old responses and use the responses collected over a short time period only: a time-invariant model for students' understanding will fit better for a shorter range of time. The number of responses in the dataset, however, decreases when one reduces the data collection period, possibly deteriorating the prediction performance.

Time-variant response models can help resolve this issue. For instance, the authors of [12] have proposed a time-variant model for learning analytics capturing the time varying level of understanding of learners. We believe that such a time-variant model can take advantage of a large amount of data without compromising the fitness of the model.

6.4 Sparsity of R

It is reasonable to believe that among many concepts only a few are required to correctly answer a question but our collaborative filtering algorithm (usually) results in a dense question-concept matrix (R). [3] Therefore, imposing sparsity on R can potentially allow for a better model and hence an improved prediction performance. In [13], the authors propose a collaborative filtering that can find a sparse question-concept matrix R by incorporating the ℓ_1 regularization term into the objective function of the optimization problem. The authors observe a superior prediction performance of their proposed sparse model compared with the non-sparse model proposed in [2]. Inspired by this observation, we also measured the performance of the variation of our algorithm where the ℓ_1 regularization term is incorporated but we did not observe an improvement in prediction performance with our data set. Even though we could not observe an improvement in prediction performance with our data set, we believe that the sparse models, capturing the natural sparsity of R, will result in more accurate estimates in general.

[3] Indeed, we observe that our normalization often results in a sparse question-concept matrix but we do not know how to explain this phenomenon.

References

[1] Robert M Bell and Yehuda Koren. Lessons from the netflix prize challenge. *ACM SIGKDD Explorations Newsletter*, 9(2):75–79, 2007.

[2] Yoav Bergner, Stefan Droschler, Gerd Kortemeyer, Saif Rayyan, Daniel Seaton, and David E Pritchard. Model-based collaborative filtering analysis of student response data: Machine-learning item response theory. *International Educational Data Mining Society*, 2012.

[3] R Darrell Bock. Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, 37(1):29–51, 1972.

[4] Jian-Feng Cai, Emmanuel J Candès, and Zuowei Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization*, 20(4):1956–1982, 2010.

[5] Emmanuel J Candes and Yaniv Plan. Matrix completion with noise. *Proceedings of the IEEE*, 98(6):925–936, 2010.

[6] Emmanuel J Candès and Benjamin Recht. Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 9(6):717–772, 2009.

[7] Emmanuel J Candès and Terence Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory*, 56(5):2053–2080, 2010.

[8] John Daniel. Making sense of moocs: Musings in a maze of myth, paradox and possibility. *Journal of interactive Media in education*, 2012(3), 2012.

[9] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics Springer, Berlin, 2001.

[10] Prateek Jain, Praneeth Netrapalli, and Sujay Sanghavi. Low-rank matrix completion using alternating minimization. In *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*, pages 665–674. ACM, 2013.

[11] Ioannis Konstas, Vassilios Stathopoulos, and Joemon M Jose. On social networks and collaborative recommendation. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 195–202. ACM, 2009.

[12] Andrew S Lan, Christoph Studer, and Richard G Baraniuk. Time-varying learning and content analytics via sparse factor analysis. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 452–461. ACM, 2014.

[13] Andrew S Lan, Andrew E Waters, Christoph Studer, and Richard G Baraniuk. Sparse factor analysis for learning and content analytics. *Journal of Machine Learning Research*, 15(1):1959–2008, 2014.

[14] Greg Linden, Brent Smith, and Jeremy York. Amazon. com recommendations: Item-to-item collaborative filtering. *IEEE Internet computing*, 7(1):76–80, 2003.

[15] Robert L. McKinley and Mark D. Reckase. Maxlog: A computer program for the estimation of the parameters of a multidimensional logistic model. *Behavior Research Methods & Instrumentation*, 15(3):389–390, 1983.

[16] Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization*, 19(4):1574–1609, 2009.

[17] Ryan Ning, Andrew E Waters, Christoph Studer, and Richard G Baraniuk. Sprite: A response model for multiple choice testing. *arXiv preprint arXiv:1501.02844*, 2015.

[18] Benjamin Recht, Maryam Fazel, and Pablo A Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review*, 52(3):471–501, 2010.

[19] Benjamin Recht and Christopher Ré. Parallel stochastic gradient algorithms for large-scale matrix completion. *Mathematical Programming Computation*, 5(2):201–226, 2013.

[20] George Siemens. What are learning analytics? `http://www.elearnspace.org/blog/2010/08/25/what-are-learning-analytics/`, 2010. Accessed: 2016-09-30.

[21] Xiaoyuan Su and Taghi M Khoshgoftaar. A survey of collaborative filtering techniques. *Advances in artificial intelligence*, 2009:4, 2009.