Dealbreaker: A Nonlinear Latent Variable Model for Educational Data

Andrew Lan SL29@RICE.EDU
Rice University

Tom Goldstein TOMG@CS.UMD.EDU
University of Maryland

Richard Baraniuk RICHB@RICE.EDU
Rice University

Christoph Studer STUDER@CORNELL.EDU
Cornell University

Abstract

Statistical models of student responses on assessment questions, such as those in homeworks and exams, enable educators and computer-based personalized learning systems to gain insights into students' knowledge using machine learning. Popular student-response models, including the Rasch model and item response theory models, represent the probability of a student answering a question correctly using an affine function of latent factors. While such models can accurately predict student responses, their ability to interpret the underlying knowledge structure (which is certainly nonlinear) is limited. In response, we develop a new, nonlinear latent variable model that we call the *dealbreaker* model, in which a student's success probability is determined by their *weakest* concept mastery. We develop efficient parameter inference algorithms for this model using novel methods for nonconvex optimization. We show that the dealbreaker model achieves comparable or better prediction performance as compared to affine models with real-world educational datasets. We further demonstrate that the parameters learned by the dealbreaker model are *interpretable*—they provide key insights into which concepts are critical (i.e., the "dealbreaker") to answering a question correctly. We conclude by reporting preliminary results for a movie-rating dataset, which illustrate the broader applicability of the dealbreaker model.

Proceedings of the 33^{rd} International Conference on Machine Learning, New York, NY, USA, 2016. JMLR: W&CP volume 48.

1. Introduction

A key problem in machine learning-based education is *student-response modeling*, i.e., developing principled statistical models that (i) accurately predict unobserved student responses to questions and (ii) identify the latent concepts that govern correct or incorrect responses. A wide range of student-response models have been proposed in the literature, including the Rasch (Rasch, 1993), item response theory (IRT) (Lord, 1980), knowledge tracing (Corbett & Anderson, 1994), and factor analysis-based models (Cen et al., 2006; Pavlik et al., 2009; Gong et al., 2010; Chi et al., 2011; Bergner et al., 2012; Lan et al., 2014b).

The Rasch model (Rasch, 1993) is simple yet effective for analyzing student-response data. This model characterizes the probability of a correct response as a function of two scalar parameters: the student's ability and the question's difficulty. The Rasch model lays the foundation for the IRT model (Lord, 1980), which features additional parameters characterizing the discrimination level of the questions across students and the effect of guessing. The multi-dimensional IRT (MIRT) model (Reckase, 2009) and the factor analysis-based models expand upon the IRT model by adding multi-dimensional ability and difficulty parameters (we refer to the model dimensions as "concepts").

1.1. Limits of Affine Student–Response Models

A key commonality of all the models described above is that they are *affine*—they characterize a student's probability of success on a question as an affine function of the student's knowledge on underlying concepts. While such models are simple and enable accurate prediction of unobserved student responses, they suffer from a key flaw known as the "explaining away" phenomenon (Wellman & Henrion, 1993): Affine models allow weak knowledge of

a concept to be erroneously covered up by strong knowledge of other potentially unrelated concepts. Affine models also fail to capture more complicated nonlinear dynamics underlying student responses. For instance, it may be impossible for a student to answer a question correctly without mastering a specific concept. Consider the situation where a student tries to solve the problem: "Simplify the expression $(5x^2 \sin^2 x + 5x^2 \cos^2 x + 10x)/(x + 2)$." Students that do not know the trigonometric identity $\sin^2 x + \cos^2 x = 1$, will by stymied, no matter how strong their knowledge of polynomial division. This kind of nonlinear "dealbreaker" property cannot be captured by an affine model.

Only limited progress has been made in nonlinear student-response models. For example, the deterministic inputs, noisy and-gate (DINA) model (de la Torre, 2011) posits that a student's probability of answering a question correctly depends on each specific combination of their binary-valued concept knowledge states (e.g., 101 means that the student has mastered Concepts 1 and 3, but not 2). While the DINA model enables the characterization of more complex response behavior, such as "students have to master both Concepts 1 and 3 in order to answer this question correctly," it remains an affine model, because the success probability is modeled as an affine function of the probabilities of the student being in each specific knowledge state. Moreover, the DINA model suffers from the fact that there can be up to 2^K possible knowledge patterns for each question involving K concepts; this prevents its use in domains that cover tens or more different concepts.

1.2. Contributions

In this paper, we develop a new statistical framework for student-response modeling, dubbed the *dealbreaker model*, that avoids the drawbacks of existing models. In the dealbreaker model, the probability of a student's success on a question depends only on their *weakest* concept mastery among all the concepts involved in that question and no others; this prevents the "explaining away" phenomenon. For the example question mentioned above, we say that not knowing the trigonometric identity $\sin^2(x) + \cos^2(x) = 1$ is the "dealbreaker" of the question.

To perform parameter inference for this non-affine model, we develop a novel, nonconvex optimization algorithm as well as a smooth approximation to the dealbreaker model that leads to even more efficient inference.

Using four distinct educational datasets, we demonstrate that the exact and approximate dealbreaker models achieve comparable or better prediction performance on unobserved student responses than state-of-the-art affine models (Rasch, MIRT, and DINA models). Moreover, we showcase the ability of our models to identify the key concept (the so-called "dealbreaker") that is needed to answer a question correctly. This new functionality could play a significant role in the modern, machine learning-based approach to personalized learning that has been identified as a national priority in the US (NAE, 2016). Going further, we report preliminary results for a movie rating dataset, which showcase the broader applicability and interpretability advantage of the dealbreaker model to domains outside of education.

2. The Dealbreaker Model

Let N be the total number of students and Q the total number of questions. Let $Y_{i,j}$ denote the binary-valued graded response of student j to question i, where $Y_{i,j} = 1$ denotes a correct response and $Y_{i,j} = 0$ an incorrect response. Note that some (or many) responses $Y_{i,j}$ may be unobserved or missing. Let K be the number of concepts underlying the questions in the dataset, where the concepts are the latent factors that control the probability of a correct answer. Let $C_{k,j}$ denote the knowledge mastery level of student j on concept k (Lan et al., 2014b). Also let $\mu_{i,k}$ denote the intrinsic difficulty of question i on concept k, which characterizes the level of knowledge required on this concept for a student to answer this question correctly.

The *hard dealbreaker model* represents the probability that student j answers question i correctly as follows:

$$\begin{aligned} p(Y_{i,j} = 1) &= \sigma\Big(\min_{k=1,\dots,K}(C_{k,j} - \mu_{i,k})\Big) \\ &= \min_{k=1,\dots,K} \sigma(C_{k,j} - \mu_{i,k})\,. \end{aligned} \tag{1}$$

Here, $\sigma(x)$ is a suitably-chosen link function that maps real values onto the success probability of a Bernoulli random variable in $[0, 1]$. Without loss of generality, we will exclusively use the *inverse logit link function* defined as $\sigma(x) = (1 + e^{-x})^{-1}$; hence, the second equality in (1) follows from the fact that $\sigma(x)$ is non-decreasing in x.

We will refer to $\min(\mathbf{x}) = \min_k x_k : \mathbb{R}^K \to \mathbb{R}$ as the $\min$ function, with the $\max$ function defined analogously. The $\min$ function is a non-smooth, non-convex function that makes parameter estimation a nontrivial task. As an alternative, we will also use the so-called *soft-min* function

$$f_\alpha(\mathbf{x}) = -\frac{1}{\alpha} \log \sum_{k=1,\dots,K} e^{-\alpha x_k},$$

which is a smooth approximation to the $\min$ function; the parameter $\alpha > 0$ determines the quality of the approximation (larger values correspond to tighter approximations). This soft-min approximation leads to the *soft dealbreaker model* for graded student responses:

$$p(Y_{i,j} = 1) = \sigma\left(-\frac{1}{\alpha} \log \sum_{k=1,\dots,K} e^{-\alpha(C_{k,j} - \mu_{i,k})}\right). \tag{2}$$

For $K = 1$, both dealbreaker models (1) and (2) coincide (trivially) with the classical Rasch model (Rasch, 1993).

Intuitively, the two dealbreaker models state that the probability of a student answering a question correctly depends only on their *weakest* concept mastery that is tested in the question. For example, suppose that geometry and algebra are both involved in a question. The dealbreaker model requires the student to have strong knowledge of *both* geometry and algebra in order to succeed with high probability. If they have strong knowledge of only geometry but not of algebra, then they are not likely to succeed—literally, algebra is a "dealbreaker" to their success on this question.

Remark 1. By defining

$$p(Y_{i,j} = 0) = \min_{k=1,\dots,K} \sigma(\mu_{i,k} - C_{k,j}) \tag{3}$$

instead of (1), we arrive at an alternative model, which we refer to as the hard deal*maker* model; analogously to (2), a soft-version can be derived. In contrast to the dealbreaker models, these dealmaker models imply that it is sufficient for student j to master *only one* concept $C_{k,j}$ to successfully answer question $Y_{i,j}$. In what follows, we will focus on the hard and soft dealbreaker models as they (i) better reflect educational scenarios and (ii) achieve superior prediction performance in our experiments on real-world educational datasets. Nevertheless, our proposed inference methods can easily be applied to the dealmaker model. We also note that the dealmaker model may be useful in the analysis of other datasets (e.g., to model single-issue politics in voting).

Remark 2. The two dealbreaker models (1) and (2), as well as the hard dealmaker model in (3), are only identifiable in their parameters $C_{k,j}$ and $\mu_{i,k}$ up to a constant offset in each concept, i.e., the model predictions remain unchanged if we add an arbitrary constant a_k to the parameters $C_{k,j}$, $\forall j$ and $\mu_{i,k}$, $\forall i$. Therefore, parameter estimation for these models is non-unique. We will alleviate this identifiability issue by regularizing the parameters $C_{k,j}$ and $\mu_{i,k}$ in Sec. 5.1.

3. Inference for the Hard Dealbreaker Model

We now develop a computationally efficient parameter inference algorithm for the hard dealbreaker model. We first outline the full algorithm, which employs the alternating direction method of multipliers (ADMM) framework (Boyd et al., 2011) for our nonconvex problem. We then detail the proximal operators that are required in our algorithm.

3.1. ADMM Algorithm

We formulate parameter estimation for the hard dealbreaker model as an optimization problem that minimizes the negative log-likelihood of the observed student responses. Let $\Omega_1 = \{(i,j) : Y_{i,j} = 1\}$, and $\Omega_0 = \{(i,j) : Y_{i,j} = 0\}$. The dealbreaker model decomposes into the form

$$\begin{aligned}\underset{C_{k,j},\mu_{i,k},\forall i,j,k}{\text{minimize}} & \sum_{(i,j)\in\Omega_1} -\log\sigma(\min_k(C_{k,j} - \mu_{i,k})) \\ & + \sum_{(i,j)\in\Omega_0} -\log\sigma(-\min_k(C_{k,j} - \mu_{i,k})) \\ = & \sum_{(i,j)\in\Omega_1} \max_k -\log\sigma(C_{k,j} - \mu_{i,k}) \\ & + \sum_{(i,j)\in\Omega_0} \min_k -\log\sigma(-(C_{k,j} - \mu_{i,k})).\end{aligned}$$

We have made use of the facts that $-\log(\sigma(x))$ is non-increasing in x and thus $p(Y_{i,j} = 0) = 1 - p(Y_{i,j} = 1) = 1 - \sigma(\min_k(C_{k,j} - \mu_{i,k})) = \sigma(\max_k -(C_{k,j} - \mu_{i,k}))$.

Since this optimization problem is non-convex, we seek an efficient approximate solution via the ADMM framework. Let $Z^k_{i,j} = C_{k,j} - \mu_{i,k}$ and rewrite the above problem as

$$\begin{aligned}\underset{C_{k,j},\mu_{i,k},\forall i,j,k}{\text{minimize}} & \sum_{(i,j)\in\Omega_1} \max_k -\log\sigma(Z^k_{i,j}) \\ & + \sum_{(i,j)\in\Omega_0} \min_k -\log\sigma(-Z^k_{i,j}), \\ \text{subject to} \quad & Z^k_{i,j} = C_{k,j} - \mu_{i,k}.\end{aligned}$$

The augmented Lagrangian for this problem is as follows:

$$\begin{aligned}\underset{C_{k,j},\mu_{i,k},\forall i,j,k}{\text{minimize}} & \sum_{(i,j)\in\Omega_1} \max_k -\log\sigma(Z^k_{i,j}) \\ & + \sum_{(i,j)\in\Omega_0} \min_k -\log\sigma(-Z^k_{i,j}) \\ & + \frac{\rho}{2}\sum_{i,j,k}(Z^k_{i,j} - C_{k,j} + \mu_{i,k} + \Lambda^k_{i,j})^2,\end{aligned}$$

where $\Lambda^k_{i,j}$ is the Lagrange multiplier for the constraint $Z^k_{i,j} = C_{k,j} - \mu_{i,k}$ and $\rho \geq 0$ is a (suitably chosen) scaling parameter.[1] We randomly initialize the variables $Z^k_{i,j}$, $C_{k,j}$, and $\mu_{i,k}$, $\forall\, i,j,k$ from the standard normal distribution, and initialize the Lagrange multipliers as $\Lambda^k_{i,j} = 0$, $\forall\, i,j,k$. We then iterate the following steps until convergence is reached (convergence of ADMM for non-convex problems is shown in (Li & Pong, 2015)).

Optimize over $Z^k_{i,j}$: For each index pair $(i,j) \in \Omega_1$, solve the following proximal problem:

$$\begin{aligned}\underset{Z^k_{i,j},\forall k}{\text{minimize}} \quad & \tfrac{1}{2}\textstyle\sum_k (Z^k_{i,j} - C_{k,j} + \mu_{i,k} + \Lambda^k_{i,j})^2 \\ & + \tfrac{1}{\rho}\max_k -\log\sigma(Z^k_{i,j}),\end{aligned}$$

[1] Note that we use the *scaled* augmented Lagrangian, in which the Lagrange multiplier appears inside of the least-squares penalty.

and for each index pair $(i,j) \in \Omega_0$, solve the following proximal problem:

$$\underset{Z^k_{i,j},\forall k}{\text{minimize}} \quad \tfrac{1}{2}\sum_k (Z^k_{i,j} - C_{k,j} + \mu_{i,k} + \Lambda^k_{i,j})^2 + \tfrac{1}{\rho}\min_k -\log\sigma(-Z^k_{i,j}).$$

The details of these two proximal problems are given in the next section.

Optimize over $C_{k,j}$: Solve the following problem:

$$\underset{C_{k,j}}{\text{minimize}} \quad \tfrac{1}{2}\sum_i (Z^k_{i,j} - C_{k,j} + \mu_{i,k} + \Lambda^k_{i,j})^2.$$

The closed-form solution is given by

$$\widehat{C}_{k,j} = \tfrac{1}{Q}\sum_i (Z^k_{i,j} + \mu_{i,k} + \Lambda^k_{i,j}).$$

Optimize over $\mu_{i,k}$: Solve the following problem:

$$\underset{\mu_{i,k}}{\text{minimize}} \quad \tfrac{1}{2}\sum_j (Z^k_{i,j} - C_{k,j} + \mu_{i,k} + \Lambda^k_{i,j})^2.$$

The closed-form solution is given by

$$\widehat{\mu}_{i,k} = \tfrac{1}{N}\sum_j (C_{k,j} - Z^k_{i,j} - \Lambda^k_{i,j}).$$

Update Lagrange multiplier: Compute

$$\hat{\Lambda}^k_{i,j} = \Lambda^k_{i,j} + Z^k_{i,j} - C_{k,j} + \mu_{i,k}, \forall\, i,j,k.$$

3.2. Proximal Operators

In the hard dealbreaker ADMM algorithm, we need to solve the following proximal problems:

$$\text{P}_{\max}: \quad \underset{\mathbf{x}}{\text{minimize}} \quad \tfrac{1}{2}\|\mathbf{y}-\mathbf{x}\|_2^2 + \max_k g(x_k),$$
$$\text{P}_{\min}: \quad \underset{\mathbf{x}}{\text{minimize}} \quad \tfrac{1}{2}\|\mathbf{y}-\mathbf{x}\|_2^2 + \min_k g(-x_k).$$

Here, $\mathbf{y} \in \mathbb{R}^K$ and $g(x) = \frac{1}{\rho}\log(1+e^{-x})$ is a non-increasing, non-negative convex function on $(-\infty,\infty)$. The following theorem characterizes the solution to $\text{P}_{\max}$.[2]

Theorem 1. *Assume that the entries in* $\mathbf{y}$ *are sorted in ascending order. Then, the solution to the proximal problem* P_{max} *is given by*

$$x_k = \begin{cases} \widehat{\tau} & \text{for } k = 1,\ldots,\widehat{K}, \\ y_k & \text{for } k = \widehat{K}+1,\ldots,K, \end{cases}$$

where $\widehat{K}$ *is the largest integer* M *such that*

$$My_M - \textstyle\sum_{k=1}^M y_k + g'(y_M) \le 0$$

and $\widehat{\tau}$ *is the solution to* $\widehat{K}\tau - \sum_{k=1}^{\widehat{K}} y_k + g'(\tau) = 0$.

[2]Our results apply to any non-increasing, differentiable function $g(\cdot)$, more general than the results in (Parikh & Boyd, 2014).

Proof. The problem $\text{P}_{\max}$ is equivalent to

$$\underset{\mathbf{x}}{\text{minimize}} \quad \tfrac{1}{2}\|\mathbf{y}-\mathbf{x}\|_2^2 + t$$
$$\text{subject to} \quad g(x_k) \le t,\ \forall k.$$

The Karush-Kuhn-Tucker (KKT) conditions for this problem are as follows:

$$x_k - y_k + \gamma_k g'(x_k) = 0, \quad \forall k, \tag{4}$$
$$\textstyle\sum_k \gamma_k = 1, \tag{5}$$
$$\gamma_k(g(x_k) - t) = 0. \quad \forall k, \tag{6}$$

Here, γ_k is the non-negative Lagrange multiplier for the inequality constraint $g(x_k) \le t$. In the complimentary slackness condition (6), we have that if $\gamma_k = 0$, then $g(x_k) \le t$. In this case, the stationarity condition (4) gives $x_k = y_k$. On the other hand, if $\gamma_k > 0$, then $g(x_k) = t$, meaning that $x_k = g^{-1}(t) := \tau$. In this case, (4) leads to $x_k = y_k - \gamma_k g'(x_k) \ge y_k$, since $\gamma_k \ge 0$ and $g'(x_k) \le 0$ because $g(x_k)$ is non-increasing. As a consequence, we know that the solution to $\text{P}_{\max}$ is given by

$$x_k = \max\{y_k, \tau\} \tag{7}$$

for some constant τ. Hence, we need only find τ. Since $\frac{1}{2}\|\mathbf{y}-\mathbf{x}\|_2^2$ is non-decreasing and $\max_k g(x_k)$ is non-increasing as τ increases, we know that there will be a minimizer for τ. In order to find its value, we note that the analysis above gives

$$\gamma_k = \begin{cases} 0 & x_k = y_k, \\ \frac{y_k - \tau}{g'(\tau)} & x_k = \tau \ge y_k. \end{cases}$$

Together with the stationary condition for t (5), we have

$$\textstyle\sum_{k'} \frac{y_{k'} - \tau}{g'(\tau)} = 1 \iff \sum_{k'}(y_{k'} - \tau) - g'(\tau) = 0,$$

where k' corresponds to the indices in $\mathbf{x}$ that satisfy $x_k = \tau$. First, we need to identify these indices. By assumption, $y_1 \le \ldots \le y_K$. Then, we examine the value of $f(\tau) = \sum_{k'}(y_{k'} - \tau) - g'(\tau)$. Note that $f(\tau)$ is a non-increasing function of τ as both $\sum_{k'}(y_{k'} - \tau)$ and $-g'(\tau) = \frac{1}{\rho(1+e^{\tau})}$ are non-increasing functions of τ. To find the indices k', we check $f(\tau)$ for different values of τ:

$\tau < y_1$: $f(\tau) = -g'(\tau) > 0$, since we have $x_k = y_k$ for $k = 1,2,\ldots,K$ from (7).

$y_1 \le \tau < y_2$: $f(\tau) = y_1 - \tau - g'(\tau)$ since we have $x_1 = \tau$ and $x_k = y_k$ for $k = 2,\ldots,K$, giving $f(y_1) = y_1 - y_1 - g'(y_1)$.

$\vdots$

$\tau \ge y_K$: $f(\tau) = \sum_{k=1}^K y_k + K\tau - g'(\tau)$ since we have $x_1 = \ldots = x_K = \tau$, giving $f(y_K) = \sum_{k=1}^K y_k - Ky_K - g'(y_K)$.

According to the analysis above, the number of elements in $\mathbf{x}$ that are equal to τ is simply the largest integer M such that $My_M - \sum_{k=1}^{M} y_k + g'(y_M) \leq 0$.

Once we have found the integer $\widehat{K}$, the value of τ can be found by solving $f'(\tau) = \widehat{K}\tau - \sum_{k=1}^{\widehat{K}} y_k + g'(\tau) = 0$. We use Newton's method by initializing $\tau_0 = y_{\widehat{K}}$ and iteratively performing the following update:

$$\tau_{\ell+1} = \tau_\ell - \frac{\widehat{K}\tau_\ell - \sum_{k=1}^{\widehat{K}} y_k + g'(\tau_\ell)}{\widehat{K} + g''(\tau_\ell)}$$

until the sequence $\{\tau_\ell\}$ converges to $\widehat{\tau}$.

In summary, the solution of $\mathrm{P}_{\max}$ can be written as

$$x_k = \begin{cases} \widehat{\tau} & \text{for } k = 1, \ldots, \widehat{K}, \\ y_k & \text{for } k = \widehat{K}+1, \ldots, K. \end{cases}$$

We note that $\mathrm{P}_{\max}$ is a generalization of the proximal problem for the ℓ_∞-norm (Duchi et al., 2008; Studer et al., 2015), which corresponds to the special case of $g(x) = |x|$. □

The following theorem characterizes the solution to $\mathrm{P}_{\min}$.

Theorem 2. *Assume that the entries in* $\mathbf{y}$ *are sorted in ascending order. Then, the solution to the proximal problem* P_{min} *is given by*

$$x_k = \begin{cases} \widehat{\tau} & \textit{for } k = 1, \\ y_k & \textit{for } k = 2, \ldots, K, \end{cases}$$

where $\widehat{\tau}$ *is the solution to* $\tau - y_1 + g'(-\tau) = 0$.

Proof. In this case, the function $g(-x)$ is non-decreasing on $(-\infty, \infty)$. Therefore, the value of $\min_k g(-x_k)$ depends only on the smallest element in $\mathbf{x}$, and the other elements of $\mathbf{x}$ will simply be equal to their corresponding elements in $\mathbf{y}$. We need only solve for the smallest element; it will be given by the solution to the equation $\tau - y_1 + g'(-\tau) = 0$. In our algorithm, we use Newton's method, analogously to the one used to solve $\mathrm{P}_{\min}$ to find $\widehat{\tau}$. □

4. Inference for the Soft Dealbreaker Model

We now develop the inference algorithm for the soft dealbreaker model. As for the hard dealbreaker model, the inference problem minimizes the approximated negative log-likelihood (2) of the observed student responses

$$\begin{aligned} \underset{C_{k,j},\mu_{i,k},\forall i,j,k}{\text{minimize}} & \sum_{(i,j)\in\Omega_1} -\log\sigma(\min_k(C_{k,j} - \mu_{i,k})) \\ & + \sum_{(i,j)\in\Omega_0} -\log\sigma(-\min_k(C_{k,j} - \mu_{i,k})) \\ & \approx \sum_{(i,j)\in\Omega_1} -\log\sigma\Big(-\frac{1}{\alpha}\log\sum_k e^{-\alpha(C_{k,j}-\mu_{i,k})}\Big) \\ & + \sum_{(i,j)\in\Omega_0} -\log\sigma\Big(\frac{1}{\alpha}\log\sum_k e^{-\alpha(C_{k,j}-\mu_{i,k})}\Big), \end{aligned}$$

where $\alpha \geq 0$ controls how tight the soft-min approximates the hard-min function. Since the approximate negative log-likelihood function is smooth in the variables $C_{k,j}$ and $\mu_{i,k}$, we can use the fast adaptive shrinkage/thresholding algorithm (FASTA) framework (Goldstein et al., 2015) to efficiently find a locally optimal solution to this problem.

We start by initializing the variables as for the hard dealbreaker model. To reduce the chance of getting stuck in a local optimum, we initialize α to a small positive value (e.g., $\alpha = 0.1$) that ensures smoothness of the initial objective function. We also initialize the stepsize s to a small positive value. Then, in each iteration, we perform the following steps until convergence is reached.

Gradient step on $C_{k,j}$ and $\mu_{i,k}$: Calculate the gradient of the cost function f with respect to $C_{k,j}$ and $\mu_{i,k}$ via

$$\frac{\partial f}{\partial C_{k,j}} = -\sum_{i:(i,j)\in\Omega_1} \frac{e^{-\alpha(C_{k,j}-\mu_{i,k})}}{u + u^{1-\frac{1}{\alpha}}} + \sum_{i:(i,j)\in\Omega_0} \frac{e^{-\alpha(C_{k,j}-\mu_{i,k})}}{u + u^{1+\frac{1}{\alpha}}},$$

$$\frac{\partial f}{\partial \mu_{i,k}} = \sum_{j:(i,j)\in\Omega_1} \frac{e^{-\alpha(C_{k,j}-\mu_{i,k})}}{u + u^{1-\frac{1}{\alpha}}} - \sum_{j:(i,j)\in\Omega_0} \frac{e^{-\alpha(C_{k,j}-\mu_{i,k})}}{u + u^{1+\frac{1}{\alpha}}},$$

where $u = \sum_{k'} e^{-\alpha(C_{k',j}-\mu_{i,k'})}$. Then, perform the gradient step with respect to each $C_{k,j}$ and $\mu_{i,k}$, $\forall\, i, j, k$, via

$$C_{k,j} \leftarrow C_{k,j} - s\frac{\partial f}{\partial C_{k,j}}, \quad \mu_{i,k} \leftarrow \mu_{i,k} - s\frac{\partial f}{\partial \mu_{i,k}},$$

and perform a backtracking line-search (Boyd & Vandenberghe, 2004) on s.

Stepsize s update: Adaptively select the stepsize s using the value of the variables from this iteration and the last iteration according to the Barzilai-Borwein rule (Barzilai & Borwein, 1988). This selection rule achieves faster empirical convergence than other methods, e.g., (Beck & Teboulle, 2009).

The steps above do not update the value of α, but in practice, we update the value of α using a rule inspired by the *continuation* method (Wen et al., 2010) in convex optimization. The procedure we use works as follows. First, we hold the value of α fixed and perform the above iterations until convergence. Then, we increase the value of α by multiplying it by a constant factor (e.g., 5), and run the iterations again by initializing them with the converged estimates of $C_{k,j}$ and $\mu_{i,k}$ from the previous iterations. We terminate the iterations until they converge for a large value of α (e.g., $\alpha = 20$). At this point, the large final value of α ensures that the soft min function closely approximates the true, non-smooth $\min$ function. We emphasize that this continuation approach also speeds up the numerical solver and

Model	Hard DB		Soft DB		DINA		3PL MIRT		Rasch	1-bit MC
K	3	6	3	6	3	6	3	6		
MT	0.798±0.016	0.796±0.017	0.801±0.013	0.799±0.012	0.770±0.012	0.775±0.017	0.673±0.024	0.723±0.020	0.795±0.016	**0.802±0.016**
UG	0.871±0.004	0.871±0.004	**0.875±0.004**	0.871±0.004	0.850±0.005	0.800±0.006	0.757±0.017	0.754±0.015	0.853±0.004	0.873±0.004
CE	**0.689±0.003**	0.685±0.003	0.685±0.003	0.682±0.004	0.684±0.003	0.641±0.004	0.533±0.006	0.558±0.007	0.686±0.004	0.688±0.005
edX	**0.929±0.001**	0.925±0.001	0.927±0.001	0.927±0.001	0.926±0.001	0.917±0.001	0.865±0.002	0.860±0.002	0.926±0.001	0.928±0.001

Table 1. Performance comparison in terms of the prediction accuracy (ACC) for the dealbreaker models (Hard DB and Soft DB) against the DINA, 3PL MIRT, and Rasch models, and also the 1-bit MC algorithm in (Lan et al., 2014a).

Model	Hard DB		Soft DB		DINA		3PL MIRT		Rasch	1-bit MC
K	3	6	3	6	3	6	3	6		
MT	**0.841±0.015**	0.839±0.015	0.840±0.018	0.839±0.018	0.784±0.023	0.730±0.021	0.646±0.027	0.690±0.024	0.839±0.017	0.838±0.019
UG	**0.832±0.004**	0.831±0.004	0.831±0.005	0.830±0.004	0.760±0.014	0.788±0.011	0.613±0.019	0.633±0.015	0.800±0.009	0.830±0.007
CE	0.744±0.004	0.744±0.004	0.748±0.004	0.746±0.003	**0.750±0.004**	0.679±0.005	0.524±0.009	0.560±0.007	0.747±0.004	0.747±0.004
edX	0.904±0.002	0.906±0.002	**0.912±0.002**	0.911±0.002	0.906±0.003	0.832±0.003	0.754±0.004	0.753±0.004	0.911±0.002	0.910±0.002

Table 2. Performance comparison in terms of the area under the receiver operating characteristic curve (AUC) for the dealbreaker models (Hard DB and Soft DB) against the DINA, 3PL MIRT, and Rasch models, and also the 1-bit MC algorithm in (Lan et al., 2014a).

reduces the chance that our method gets stuck in a bad local minimum, eventually improving the quality of our results.

5. Experiments

We now demonstrate the prediction performance of the dealbreaker model on unobserved student responses using four real-world educational datasets. We furthermore showcase the interpretability of the dealbreaker model by visualizing the "dealbreaker" concept for each question. In addition, we use a movie rating dataset to show that the dealbreaker model can be applied to other datasets outside of education.

5.1. Predicting Unobserved Student Responses

We compare the dealbreaker models against three state-of-the-art student-response models: the DINA model (de la Torre, 2011), the 3PL multi-dimensional item response theory (3PL MIRT) model (Reckase, 2009), and the Rasch model (Rasch, 1993). We also include a comparison against the 1-bit matrix completion (1-bit MC) algorithm proposed in (Lan et al., 2014a) and analyzed in (Davenport et al., 2014). The following four datasets are used.

MT: $N = 99$ students answering $Q = 34$ questions in a high-school algebra test administered in Amazon's Mechanical Turk (Amazon, 2016); 100% of the responses are observed.

UG: $N = 92$ students answering $Q = 203$ questions in an undergraduate course on introduction to computer engineering; 99.5% of the responses are observed.

CE: $N = 1567$ students answering $Q = 60$ questions in a college entrance exam; 70.7% of the responses are observed.

edX: $N = 6403$ students answering $Q = 197$ questions in a massive open online course (MOOC) on signals and systems; 15.0% of the responses are observed.

Experimental setup: To reduce the identifiability issue of the dealbreaker model, we add the regularization term $\frac{\lambda}{2}(\sum_{k,j} C_{k,j}^2 + \sum_{i,k} \mu_{i,k}^2)$ to the cost functions of both the hard and soft dealbreaker optimization problems and select the parameter λ using cross-validation. In each cross-validation run, we randomly leave out 20% of the student responses in the dataset (the "unobserved" data) and train the algorithms on the rest of the responses before testing their prediction performance on the unobserved data. We repeat each experiment 20 times with different random partitions of the dataset.

For the Rasch model and the MIRT model, we perform inference using the R MIRT package (Chalmers, 2012). The DINA model is implemented as detailed in (de la Torre, 2009; 2011). For the MIRT model, the DINA model, and both dealbreaker models, we use $K \in \{3, 6\}$ concepts.

We evaluate the prediction performance on the unobserved student responses of each model using two different metrics: (i) prediction accuracy (ACC), which is simply the portion of correct predictions, and (ii) area under the receiver operating characteristic curve (AUC) of the resulting binary classifier (Jin & Ling, 2005). Both metrics take on values in $[0, 1]$, with large values indicating better prediction performance.

Results and discussion: Tables 1 and 2 show the average performance of each algorithm on each dataset using each metric over 20 random splits of the data. With only two exceptions, we see that both dealbreaker models slightly outperform the other educational models in terms of prediction accuracy (ACC) and achieve slightly better or comparable performance with the Rasch model in terms of AUC. Moreover, the performances of the dealbreaker models and the Rasch model are very close to each other and much better than the DINA model and the 3PL MIRT model. The performance of the dealbreaker models is comparable to the 1-bit MC algorithm, whose parameters admit no interpretability.

	MT-DB	MT-Rasch
Soft DB	0.799	0.810
Rasch	0.775	0.808

Table 3. Comparison of the dealbreaker (DB) model against the Rasch model in terms of ACC when both models are fitted separately on subsets of the MT dataset. The dealbeaker model performs well on both subsets while the Rasch model does not perform well on questions with diverse response patterns across students.

Note that the results shown in Tables 1 and 2 correspond to the prediction performance over the *entire* dataset. We now compare the prediction performance of the dealbreaker models against other models (the Rasch model, in particular, as it is the best performing educational baseline algorithm) on different questions. Towards this end, we fit the soft dealbreaker model (with $K = 3$ concepts) and the Rasch model on the MT dataset and analyze their prediction performance on each question separately.

The top 5 questions that the dealbreaker model performs best on have average scores (portion of students with correct answers) of $67\%, 63\%, 67\%, 74\%, 70\%$, while the top-5 questions that the Rasch model predicts best on have average scores of $83\%, 11\%, 95\%, 87\%, 99\%$. Thus, we conclude that the Rasch model excels at very easy and very hard questions, while the dealbreaker model excels at questions with more diverse response patterns across students.

To further validate our observation, we divide the MT dataset with $Q = 34$ questions into two smaller, separate datasets, each with $Q = 17$ questions. One of them consists of questions on which the dealbreaker model outperforms the Rasch model (labeled MT-DB) in the prediction experiments above (using the entire dataset), and the other consists of questions on which the Rasch model outperforms the dealbreaker model (labeled MT-Rasch). We then repeat prediction experiments on these two small datasets *separately*.

Table 3 shows the performance of each algorithm on each small dataset using the ACC metric. We see that the dealbreaker model performs well on both small datasets, while the Rasch model's performance deteriorates significantly on the subset of the questions in the MT-DB dataset. These results support our observation that the simplicity of the Rasch model is best suited for questions with uniform response patterns across students (i.e., very easy or very hard questions), whereas the dealbreaker model is better suited for questions having more complex concept-understanding requirements for students to achieve success.

We emphasize that the soft dealbreaker model enables more efficient parameter inference compared to the hard dealbreaker model. For example, a single run of our Python code for the soft dealbreaker model with the UG dataset with 92 students and 203 questions takes only 10 s compared to 30 s for the hard dealbreaker model on an Intel i7 laptop with a 2.8 GHz CPU and 8 GB memory.

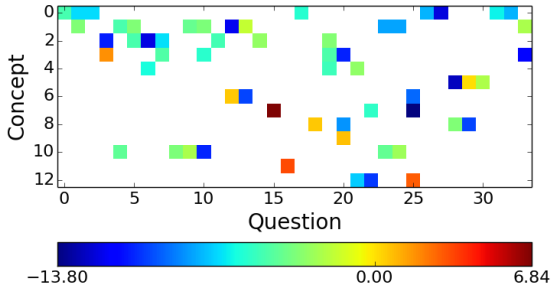

Figure 1. Visualization of the estimated question difficulty parameters $\mu_{i,k}$. "Warm" colors mean that the question requires the learners to have high knowledge on those concepts. For questions testing multiple concepts, we can see that the estimated difficulty parameters clearly show which concept is the "dealbreaker."

5.2. Visualizing the Dealbreaker Model

We now demonstrate the parameter interpretability afforded by the dealbreaker model using the MT dataset.

Experimental setup: The MT dataset comes with 13 domain-expert provided tags (or labels) on every question, which summarize the tested concepts. We use these tags as information on the underlying knowledge structure of the dataset and set K equal to the number of unique tags, letting each tag correspond to a unique concept. For each question, we only estimate the difficulty parameters of the concepts that it is associated with, and set the difficulty parameters of the other concepts to $\mu_{i,k} = -\infty$ so that they cannot be chosen as the minimum element in the min function on $C_{k,j} - \mu_{i,k}$ in the dealbreaker model.

Results and discussions: Figure 1 visualizes the estimated parameters $\mu_{i,k}$ for the MT dataset. Each grid cell in the figure represents the difficulty of a question with respect to a particular concept; "warm" colors (positive values) mean that the question requires high knowledge of a concept, "cold" colors (negative values) mean that the question requires only a moderate level of knowledge on a concept, and white means that a concept is not tested in the question.

Now we take a closer look at the questions that involve multiple concepts. For example, Question 3 corresponds to

$$\text{If } \tfrac{3x}{7} - \tfrac{9}{8} = -5, \text{ then } x = ?$$

The question tags are "Solving equations" and "Fractions," and the estimated question concept difficulties show that "Fractions" is the dealbreaker in this question. This matches with the observation that the key to answering this question correctly is to understand fractions, while the part that involves equation solving is relatively straightforward. As another example, Question 20 in this dataset is:

$$\text{Compute } \lim_{x \to 0} \tfrac{1}{x} \sin(x).$$

	soft DB	Rasch	1-bit MC
ACC	0.710 ± 0.003	0.689 ± 0.004	$\mathbf{0.718 \pm 0.002}$
AUC	0.775 ± 0.004	0.730 ± 0.005	$\mathbf{0.779 \pm 0.003}$

Table 4. Prediction performance of the dealbreaker model, the Rasch model, and 1-bit MC on the MovieLens dataset.

The question tags are "Fractions," "Trigonometry," and "Limits," and the estimated question concept difficulties show that "Limits" is the dealbreaker here, in agreement with the fact that the key to solving this question is to have a good knowledge on limits (more precisely, l'Hôpital's rule), while the fraction and trigonometry concepts needed to answer this question are less critical.

These examples highlight the advantage of the nonlinear dealbreaker model over affine models, since it can identify the most critical concepts involved in a question (e.g., in the widely used Q-matrix model (Barnes, 2005), every concept involved in the question is treated as though it contributes equally to the students' success probability). This information could enable a machine learning-based intelligent tutoring system to generate more targeted feedback for remediation or when a student asks for a hint on a question.

5.3. Interpreting Movie Ratings

To demonstrate the broader applicability of the dealbreaker model to domains outside of education, we perform inference on the "MovieLens 100k" dataset (Herlocker et al., 1999) consisting of the integer-valued (1-to-5) ratings of $N = 943$ users on $Q = 1682$ movies. To evaluate the performance of the dealbreaker model, we convert the entries into binary values using the approach proposed in (Davenport et al., 2014), i.e., we compare each entry to the average rating across the entire dataset (1 and 0 implies above and below average, respectively). We perform a prediction experiment as in Sec. 5.1 and compare the performance of the soft dealbreaker model with $K = 19$ (using the provided 19 genres with the genre labels of each movie) to the Rasch model and to the 1-bit matrix completion algorithm (1-bit MC) as proposed in (Davenport et al., 2014).

Table 4 shows the average prediction performance on the MovieLens dataset on both the ACC and AUC error metrics over 20 random splits of the dataset. Note that although the 1-bit MC algorithm slightly outperforms the soft dealbreaker model in terms of prediction performance, it offers virtually no interpretability of its model parameters.

We now report some interesting observations made by interpreting the estimated dealbreaker model parameters. The movies "Pretty Woman," "Sabrina," and "While You Were Sleeping," all have "Comedy" and "Romance" as their genres, with "Romantic" being the dealbreaker for these movies. "Romance" is the dealbreaker in all of these movies as they have large negative μ values on the "Comedy" genre (i.e., even users who do not particularly favor comedy would not dislike these movies) and large positive values on the "Romance" genre (i.e., users who do not favor romance would dislike these movies). On the contrary, "Bram Stoker's Dracula" has both "Horror" and "Romance" genres but with "Horror" as its dealbreaker—users who dislike horror movies are much less likely to enjoy it than users who dislike romantic movies.

Another interesting observation is that most of the highly rated movies (e.g., "Fargo," "Forrest Gump," and "Star Wars") cover many genres yet have no particular dealbreaker (i.e, they have large negative μ values for all involved genres). This implies that, even if a user does not like some of the genres, they may still like these movies. We feel that these preliminary results are encouraging, since they highlight the advantage of the nonlinear dealbreaker model for collaborative filtering applications as compared to affine models that excel in prediction but lack interpretability.

6. Conclusions

We have developed the dealbreaker model for analyzing students' responses to questions. Our model is nonlinear and characterizes the probability of a student's success on a question as a function of their *weakest* concept mastery, i.e., the "dealbreaker." This model helps us to gain deep insights into the knowledge structure of questions and to identify the key factors behind student response patterns on different questions. We have developed two inference algorithms for estimating the parameters of the hard and soft versions of the dealbreaker model, and have shown that they achieve excellent prediction performance on unobserved student responses, while enabling human interpretability of the estimated parameters. In addition, an application of the dealbreaker model to a movie rating dataset has shown that it provides an advantage compared to affine models in terms of interpretability of the model parameters.

There are a number of avenues for future work. Clearly, the performance of the dealbreaker model on a variety of educational datasets as well as on a movie rating dataset (especially in terms of interpretability) provides a call-to-action for the exploration of other nonlinear models. Adding extra functionality to the dealbreaker model also appears promising. For example, it is often the case that each question only covers a small number of concepts out of many (i.e., concept usage is *sparse*) (Lan et al., 2014b). Enforcing such a sparsity property on a dealbreaker model is challenging when there are no a-priori question labels available for the dataset. Furthermore, it is possible to extend the dealbreaker model from modeling binary data to ordinal data (e.g., the actual ratings in collaborative filtering applications), which may further improve the performance of the dealbreaker model on other applications.

References

Amazon. Amazon's Mechanical Turk. online: https://www.mturk.com/mturk/welcome, 2016.

Barnes, T. The Q-matrix method: Mining student response data for knowledge. In *Proc. AAAI Workshop on Educational Data Mining*, pp. 1–8, July 2005.

Barzilai, J. and Borwein, J. M. Two-point step size gradient methods. *IMA J. Numerical Analysis*, 8(1):141–148, Jan. 1988.

Beck, A. and Teboulle, M. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Science*, 2(1):183–202, Mar. 2009.

Bergner, Y., Droschler, S., Kortemeyer, G., Rayyan, S., Seaton, D., and Pritchard, D. Model-based collaborative filtering analysis of student response data: Machine-learning item response theory. In *Proc. 5th Intl. Conf. on Educational Data Mining*, pp. 95–102, June 2012.

Boyd, S. and Vandenberghe, L. *Convex Optimization*. Cambridge University Press, 2004.

Boyd, S., Parikh, N., Chu, E., Peleato, B., and Eckstein, J. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122, Jan. 2011.

Cen, H., Koedinger, K. R., and Junker, B. Learning factors analysis – A general method for cognitive model evaluation and improvement. In *Proc. 8th Intl. Conf. on Intelligent Tutoring Systems*, pp. 164–175, June 2006.

Chalmers, R. P. MIRT: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, 48(6):1–29, May 2012.

Chi, M., Koedinger, K. R., Gordon, G., and Jordan, P. Instructional factors analysis: A cognitive model for multiple instructional interventions. In *Proc. 4th Intl. Conf. on Educational Data Mining*, pp. 61–70, July 2011.

Corbett, A. T. and Anderson, J. R. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-adapted Interaction*, 4(4):253–278, Dec. 1994.

Davenport, M. A., Plan, Y., van den Berg, E., and Wootters, M. 1-bit matrix completion. *Information and Inference*, 3(3):189–223, Sep. 2014.

de la Torre, J. DINA model and parameter estimation: A didactic. *J. Educational and Behavioral Statistics*, 34(1): 115–130, Mar. 2009.

de la Torre, J. The generalized DINA model framework. *Psychometrika*, 76(2):179–199, Apr. 2011.

Duchi, J., Shalev-Shwartz, S., Singer, Y., and Chandra, T. Efficient projections onto the $\ell 1$-ball for learning in high dimensions. In *Proc. 25th Intl. Conf. on Machine Learning*, pp. 272–279, Helsinki, Finland, July 2008.

Goldstein, T., Studer, C., and Baraniuk, R. G. FASTA: A generalized implementation of forward-backward splitting. *arXiv eprint: 1501.04979*, Jan. 2015.

Gong, Y., Beck, J. E., and Heffernan, N. T. Comparing knowledge tracing and performance factor analysis by using multiple model fitting procedures. In *Proc. 12th Intl. Conf. on Intelligent Tutoring Systems*, pp. 35–44, June 2010.

Herlocker, J. L., Konstan, J. A., Borchers, A., and Riedl, J. An algorithmic framework for performing collaborative filtering. In *Proc. 22nd Annual Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval*, pp. 230–237, Aug. 1999.

Jin, H. and Ling, C. X. Using AUC and accuracy in evaluating learning algorithms. *IEEE Trans. Knowledge and Data Engineering*, 17(3):299–310, Mar. 2005.

Lan, A. S., Studer, C., and Baraniuk, R. G. Matrix recovery from quantized and corrupted measurements. In *IEEE Intl. Conf. on Acoustics, Speech and Signal Processing*, pp. 4973–4977, May 2014a.

Lan, A. S., Waters, A. E., Studer, C., and Baraniuk, R. G. Sparse factor analysis for learning and content analytics. *J. Machine Learning Research*, 15:1959–2008, June 2014b.

Li, G. and Pong, T. K. Global convergence of splitting methods for nonconvex composite optimization. *SIAM J. Optimization*, 25(4):2434–2460, Dec. 2015.

Lord, F. *Applications of Item Response Theory to Practical Testing Problems*. Erlbaum Associates, 1980.

NAE. NAE grand challenges: Advanced personalized learning. online: http://www.engineeringchallenges.org/challenges/learning.aspx, 2016.

Parikh, N. and Boyd, S. P. Proximal algorithms. *Foundations and Trends in Optimization*, 1(3):127–239, Jan. 2014.

Pavlik, P. I., Cen, H., and Koedinger, K. R. Learning factors transfer analysis: Using learning curve analysis to automatically generate domain models. In *Proc. 2nd Intl. Conf. on Educational Data Mining*, pp. 121–130, July 2009.

Rasch, G. *Probabilistic Models for Some Intelligence and Attainment Tests*. MESA Press, 1993.

Reckase, M. D. *Multidimensional Item Response Theory*. Springer, 2009.

Studer, C., Goldstein, T., Yin, W., and Baraniuk, R. G. Democratic representations. *arXiv eprint: 1401.3420*, Apr. 2015.

Wellman, P. P. and Henrion, M. Explaining "explaining away". *IEEE Trans. Pattern Analysis and Machine Intelligence*, 15(3):287–292, Mar. 1993.

Wen, Z., Yin, W., Goldfarb, D., and Zhang, Y. A fast algorithm for sparse reconstruction based on shrinkage, subspace optimization, and continuation. *SIAM J. Scientific Computing*, 32(4):1832–1857, June 2010.