β^3-IRT: A New Item Response Model and its Applications

Yu Chen
University of Bristol

Telmo Silva Filho
Universidade Federal de Pernambuco

Ricardo B. C. Prudêncio
Universidade Federal de Pernambuco

Tom Diethe
Amazon

Peter Flach
University of Bristol & Alan Turing Institute

Abstract

Item Response Theory (IRT) aims to assess latent abilities of respondents based on the correctness of their answers in aptitude test items with different difficulty levels. In this paper, we propose the β^3-IRT model, which models continuous responses and can generate a much enriched family of Item Characteristic Curves. In experiments we applied the proposed model to data from an online exam platform, and show our model outperforms a more standard 2PL-ND model on all datasets. Furthermore, we show how to apply β^3-IRT to assess the ability of machine learning classifiers. This novel application results in a new metric for evaluating the quality of the classifier's probability estimates, based on the inferred difficulty and discrimination of data instances.

1 INTRODUCTION

IRT is widely adopted in psychometrics for estimating human latent ability in tests. Unlike classical test theory, which is concerned with performance at the test level, IRT focuses on the items, modelling responses given by respondents with different abilities to items of different difficulties, both measured in a known scale [Embretson and Reise, 2013]. The concept of item depends on the application, and can represent for instance test questions, judgements or choices in exams. In practice, IRT models estimate latent difficulties of the items and the latent abilities of the respondents based on observed responses in a test, and have been commonly applied to assess performance of students in exams.

Recently, IRT was adopted to analyse Machine Learning (ML) classification tasks [Martínez-Plumed et al., 2016]. Here, items correspond to instances in a dataset, while respondents are classifiers. The responses are the outcomes of classifiers on the test instances (right or wrong decisions collected in a cross-validation experiment). Our work is different in two aspects: 1) we propose a new IRT model for continuous responses; 2) we applied the proposed IRT to predicted probabilities instead of binary responses.

For each item (instance), an Item Characteristic Curve (ICC) is estimated, which is a logistic function that returns the probability of a correct response for the item based on the respondent ability. This ICC is determined by two item parameters: difficulty, which is the location parameter of the logistic function; and discrimination, which affects the slope of the ICC. Despite the useful insights [Martínez-Plumed et al., 2016] and applicability to other AI contexts, binary IRT models are limited when the techniques of interest return continuous responses (e.g. class probabilities). Continuous IRT models have been developed and applied in psychometrics [Noel and Dauvier, 2007] but have limitations when applied in this context: limited interpretability since abilities and difficulties are expressed as real numbers; and limited flexibility since ICCs are limited to logistic functions.

In this paper, we propose a novel IRT model called β^3-IRT to addresses these limitations by means of a new parameterisation of IRT models such that: a) the resulting ICCs are not limited to logistic curves, different shapes can be obtained depending on the item parameters, which allows more flexibility when fitting responses for different items; b) abilities and difficulties are in the $[0, 1]$ range, which gives a unified scale for easier interpretation and evaluation.

We first applied the β^3-IRT model in a case study to predict students' responses in an online education platform. Out model outperforms the 2PL-ND model [Noel and Dauvier, 2007] on all datasets. In a second case study, β^3-IRT was used to fit class probabilities estimated by classifiers in binary classification tasks. The experimental results show the ability inferred by the model allows to

Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS) 2019, Naha, Okinawa, Japan. PMLR: Volume 89.

evaluate probability estimation of classifiers in an instance-wise manner with the aid of item parameters (difficulty and discrimination). Hence we show that the β^3-IRT model is useful not only in the more traditional applications associated with IRT, but also in the ML context outlined above.

Our contributions can be summarised as follows:

- we propose a new model for IRT with richer ICC shapes, which is more versatile than existing models;
- we demonstrate empirically that this model has better predictive power;
- we demonstrate the use of this model in an ML setting, providing a method for assessing the 'ability' of classification algorithms that is based on an instance-wise metric in terms of class probability estimation.

The paper is organised as follows. Section 2 gives a brief introduction of Binary Item Response Theory (IRT). Section 3 presents the β^3-IRT model, followed by related work on IRT in Section 3.3. Section 4 presents the experiments on real students, while Section 5 presents the use of β^3-IRT to evaluate classifiers. Finally, Section 6 concludes the paper.

2 BINARY ITEM-RESPONSE THEORY

In Item Response Theory, the probability of a correct response for an item depends on the latent respondent ability and the item difficulty. Most previous studies on IRT have adopted binary models, in which the responses are either correct or incorrect. Such models assume a binary response x_{ij} of the i-th respondent to the j-th item. In the IRT model with two item parameters (2PL), the probability of a correct response ($x_{ij} = 1$) is defined by a logistic function with location parameter $\boldsymbol{\delta}_j$ and shape parameter $\mathbf{a}_j$. Responses are modelled by the Bernoulli distribution with parameter p_{ij} as follows:

$$x_{ij} = \mathcal{B}ern(p_{ij}),\ p_{ij} = \sigma(-\mathbf{a}_j d_{ij}),\ d_{ij} = \boldsymbol{\theta}_i - \boldsymbol{\delta}_j \quad (1)$$

where $\sigma(\cdot)$ is the logistic function. N is the number of items and M is the number of participants.

The 2PL model gives a logistic Item Characteristic Curve (ICC) mapping ability $\boldsymbol{\theta}_i$ to expected response as follows:

$$\mathbb{E}[x_{ij}|\boldsymbol{\theta}_i, \boldsymbol{\delta}_j, \mathbf{a}_j] = p_{ij} = \frac{1}{1 + e^{-\mathbf{a}_j(\boldsymbol{\theta}_i - \boldsymbol{\delta}_j)}} \quad (2)$$

At $\boldsymbol{\theta}_i = \boldsymbol{\delta}_j$ the expected response is 0.5. The slope of the ICC at $\boldsymbol{\theta}_i = \boldsymbol{\delta}_j$ is $\mathbf{a}_j/4$. If $\mathbf{a}_j = 1, \forall j = 1, \ldots, N$, a simpler model is obtained, known as 1PL, which describes items solely by their difficulties. Generally, discrimination $\mathbf{a}_j$ indicates how probability of correct responses changes as the ability increases. High discriminations lead to steep ICCs at the point where ability equals difficulty, with small changes in ability producing significant changes in the probability of correct response.

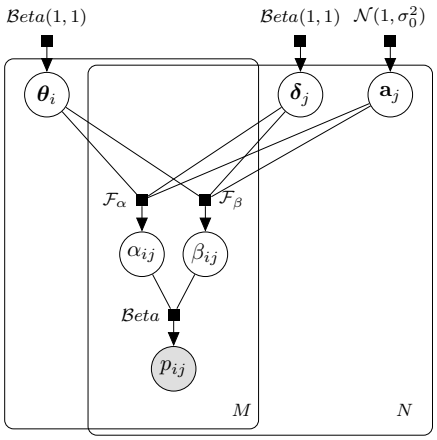


Figure 1: Factor Graph of β^3-IRT Model. The grey circle represents observed data, white circles represent latent variables, small black rectangles represent stochastic factors, and the rectangular plates represent replicates.

The standard IRT model *implicitly* uses a noise model that follows a logistic distribution (in the same way that logistic regression does). This can be replaced with Normally distributed noise, which results in the logistic link function being replaced by a probit link function (Gaussian Cumulative Distribution Function (CDF)), as is the case in the DARE model [Bachrach et al., 2012]. In practice, the probit link function is very similar to the logistic one, so these two models tend to behave very similarly, and the particular choice is often then based on mathematical convenience or computation tractability.

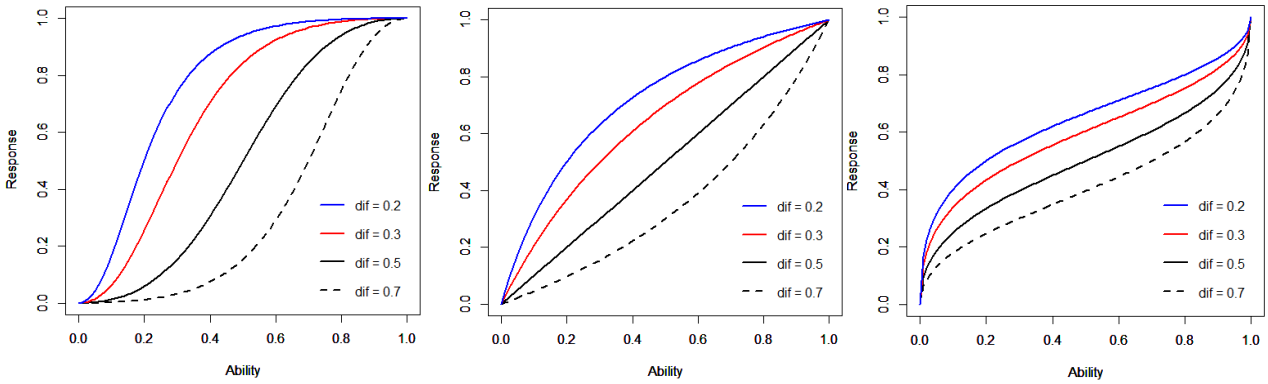
Despite their extensive use in psychometrics, binary IRT models are of limited use when responses are naturally produced in continuous scales. Particularly in ML, binary models are not adequate if the responses to evaluate are class probability estimates.

3 THE β^3-IRT MODEL

In this Section, we will elaborate on the parametrisation and inference method of β^3-IRT model, highlighting the differences with existing IRT models.

3.1 Model description

The factor graph of β^3-IRT is shown in Fig 1 and Eq (3) below gives the model definition, where M is number of respondents and N is number of items, p_{ij} is the observed response of respondent i to item j, which is drawn from a



(a) $\mathbf{a}_j = 2$. (b) $\mathbf{a}_j = 1$. (c) $\mathbf{a}_j = 0.5$.

Figure 2: Examples of Beta ICCs for Different Values of Difficulty and Discrimination. Higher discriminations lead to steeper ICCs, higher difficulties need higher abilities to achieve higher responses.

Beta distribution:

$$
\begin{aligned}
p_{ij} &\sim \mathcal{B}(\alpha_{ij}, \beta_{ij}), \\
\alpha_{ij} &= \mathcal{F}_\alpha(\boldsymbol{\theta}_i, \boldsymbol{\delta}_j, \mathbf{a}_j) = \left(\frac{\boldsymbol{\theta}_i}{\boldsymbol{\delta}_j}\right)^{\mathbf{a}_j}, \\
\beta_{ij} &= \mathcal{F}_\beta(\boldsymbol{\theta}_i, \boldsymbol{\delta}_j, \mathbf{a}_j) = \left(\frac{1-\boldsymbol{\theta}_i}{1-\boldsymbol{\delta}_j}\right)^{\mathbf{a}_j}, \\
\boldsymbol{\theta}_i &\sim \mathcal{B}(1,1),\ \boldsymbol{\delta}_j \sim \mathcal{B}(1,1),\ \mathbf{a}_j \sim \mathcal{N}(1, \sigma_0^2)
\end{aligned} \tag{3}
$$

The parameters α_{ij}, β_{ij} are computed from $\boldsymbol{\theta}_i$ (the ability of participant i), $\boldsymbol{\delta}_j$ (the difficulty of item j), and $\mathbf{a}_j$ (the discrimination of item j). $\boldsymbol{\theta}_i$ and $\boldsymbol{\delta}_j$ are also drawn from Beta distributions and their priors are set to $\mathcal{B}(1,1)$ in a general setting. The discrimination $\mathbf{a}_j$ is drawn from a Normal distribution with prior mean 1 and variance σ_0^2, where σ_0 is a hyperparameter of the model. These uninformative priors are used as default setting of the model, but could be parametrised by further hyperparameters when there is further prior information available. The default prior mean of $\mathbf{a}_j$ is set to 1 rather than 0 because the discrimination $\mathbf{a}_j$ is a power factor here.

When comparing probabilities we use ratios (e.g. the likelihood ratio). Similarly, here we use the ratio of ability to difficulty since in our model these are measured on a $[0, 1]$ scale as well: a ratio smaller/larger than 1 means that ability is lower/higher than difficulty. These ratios are positive reals and hence map directly to α and β in Eq (3). Importantly, the new parametrisation enables us to obtain non-logistic ICCs. In this model the ICC is defined by the expectation of $\mathcal{B}(\alpha_{ij}, \beta_{ij})$ and then assumes the form:

$$
\mathbb{E}[p_{ij}|\boldsymbol{\theta}_i, \boldsymbol{\delta}_j, \mathbf{a}_j] = \frac{\alpha_{ij}}{\alpha_{ij}+\beta_{ij}} = \frac{1}{1+\left(\frac{\boldsymbol{\delta}_j}{1-\boldsymbol{\delta}_j}\right)^{\mathbf{a}_j}\left(\frac{\boldsymbol{\theta}_i}{1-\boldsymbol{\theta}_i}\right)^{-\mathbf{a}_j}} \tag{4}
$$

As in standard IRT, the difficulty $\boldsymbol{\delta}_j$ is a location parameter. The response is 0.5 when $\boldsymbol{\theta}_i = \boldsymbol{\delta}_j$ and the curve has slope $\mathbf{a}_j/(4\boldsymbol{\delta}_j(1-\boldsymbol{\delta}_j))$ at that point. Fig 2 shows examples of Beta ICCs for different regimes depending on $\mathbf{a}_j$:

$$
\begin{aligned}
\mathbf{a}_j > 1: &\quad \text{a sigmoid shape similar to stardard IRT,} \\
\mathbf{a}_j = 1: &\quad \text{parabolic curves with vertex at 0.5,} \\
0 < \mathbf{a}_j < 1: &\quad \text{anti-sigmoidal behaviour.}
\end{aligned}
$$

Note that the model allows for negative discrimination, which indicates items that are somehow harder for more able respondents. Negative $\mathbf{a}_j$ can be divided similarly:

$$
\begin{aligned}
-1 < \mathbf{a}_j < 0: &\quad \text{decreasing anti-sigmoid,} \\
\mathbf{a}_j < 1: &\quad \text{decreasing sigmoid.}
\end{aligned}
$$

We will see examples in Section 5 that negative discrimination can in fact be useful for identifying 'noisy' items, where higher abilities getting lower responses.

3.2 Model inference

We tested two inference methods on β^3-IRT model, one is conventional Maximum Likelihood (MLE) which we applied to experiments with student answers (Section 4) for response prediction, using the likelihood function shown in Eq (4).

The other is Bayesian Variational Inference (VI) [Bishop, 2006] which we applied to experiments with classifiers (Section 5) for full Bayesian inference on latent variables. In VI, the optimisation objective is the lower bound of Kullback–Leibler (KL) divergence between the true posterior and variational posterior of latent variables, which is also referred as Evidence Lower Bound (ELBO). However, the model is highly non-identifiable because of its symmetry [Nishihara et al., 2013], resulting in undesirable combinations of these variables: for instance, when p_{ij} is close to 1, it usually indicates $\alpha_{ij} > 1$ and $\beta_{ij} < 1$, which can arise when $\boldsymbol{\theta}_i > \boldsymbol{\delta}_j$ with positive $\mathbf{a}_j$, or $\boldsymbol{\theta}_i < \boldsymbol{\delta}_j$ with negative $\mathbf{a}_j$. Hence, we update discrimination as a global

Algorithm 1: Variational Inference for β^3-IRT

```
Set number of iterations L_iter, randomly initialise φ, ψ, λ;
for t in range(L_iter) do
    while Θ = {φ, ψ} not converged do
        Compute ∇_Θ L_1 according to Eq (5);
        Update Θ by SGD;
    end
    Compute ∇_λ L_2 according to Eq (6);
    Update λ by SGD;
end
```

variable after ability and difficulty converge at each step (see also Algorithm 1), in addition to setting the prior of discrimination as N(1,1) to reflect the assumption that discrimination is more often positive than negative. We employ the coordinate ascent method of VI [Blei et al., 2017], keeping $\mathbf{a}_j$ fixed while optimising $\boldsymbol{\theta}_i$ and $\boldsymbol{\delta}_j$ and vice versa. Accordingly, the two separate loss functions are defined as below:

$$\begin{aligned}\mathcal{L}_1 = &\sum_{i=1}^{M}\sum_{j=1}^{N}\mathbb{E}_q[\log p(p_{ij}|\boldsymbol{\theta}_i,\boldsymbol{\delta}_j)]\\ &+\sum_{i=1}^{M}\mathbb{E}_q[\log p(\boldsymbol{\theta}_i)-\log q(\boldsymbol{\theta}_i|\boldsymbol{\phi}_i)]\\ &+\sum_{j=1}^{N}\mathbb{E}_q[\log p(\boldsymbol{\delta}_j)-\log q(\boldsymbol{\delta}_j|\boldsymbol{\psi}_j)] \qquad (5)\end{aligned}$$

$$\mathcal{L}_2 = \sum_{j=1}^{N}\mathbb{E}_q[\log p(p_{ij}|\mathbf{a}_j)+\log p(\mathbf{a}_j)-\log q(\mathbf{a}_j|\boldsymbol{\lambda}_j)] \quad (6)$$

Here, $\boldsymbol{\phi},\boldsymbol{\psi},\boldsymbol{\lambda}$ are parameters of variational posteriors of $\boldsymbol{\theta},\boldsymbol{\delta},\mathbf{a}$, respectively. $\mathcal{L}_1$ and $\mathcal{L}_2$ are lower bounds of $\mathbb{KL}(p(\boldsymbol{\theta},\boldsymbol{\delta}|\mathbf{x})||q(\boldsymbol{\theta},\boldsymbol{\delta}|\boldsymbol{\phi},\boldsymbol{\psi}))$ and $\mathbb{KL}(p(\mathbf{a}|\mathbf{x})||q(\mathbf{a}|\boldsymbol{\lambda})$, respectively. Both are optimised using Stochastic Gradient Descent (SGD). In order to apply the reparameterisation trick [Kingma and Welling, 2013], we use Logit-Normal to approximate the Beta distribution in variational posteriors. The steps to perform VI of the model are shown in Algorithm 1, where the variational parameters can be updated by any gradient descent optimisation algorithm (we use the Adam method [Kingma and Ba, 2014] in our experiments). We implemented Algorithm 1 using the probabilistic programming library Edward [Tran et al., 2016].

3.3 Related work

There have been earlier approaches to IRT with continuous approaches. In particular, [Noel and Dauvier, 2007] proposed an IRT model which adopts the Beta distribution with parameters m_{ij} and n_{ij} as follows:

$$m_{ij} = e^{(\boldsymbol{\theta}_i-\boldsymbol{\delta}_j)/2}, \quad n_{ij} = e^{-(\boldsymbol{\theta}_i-\boldsymbol{\delta}_j)/2} = m_{ij}^{-1},$$
$$p_{ij} \sim \mathcal{B}(m_{ij}, n_{ij}).$$

This model gives a logistic ICC mapping ability to expected response for item j of the form:

$$\mathbb{E}[p_{ij}|\boldsymbol{\theta}_i,\boldsymbol{\delta}_j] = \frac{m_{ij}}{m_{ij}+n_{ij}} = \frac{1}{1+e^{-(\boldsymbol{\theta}_i-\boldsymbol{\delta}_j)}} \quad (7)$$

While there is a superficial similarity to the β^3-IRT model, there are two crucial distinctions.

- Similarly to the standard 1PL IRT model, Eq (7) does not have a discrimination parameter. The ICC therefore has a fixed slope of 0.25 at $\boldsymbol{\theta}_i = \boldsymbol{\delta}_j$ and it is assumed that all items have the same discrimination.
- Eq (7) assumes a real-valued scale for abilities and difficulties, whereas β^3-IRT uses a $[0,1]$ scale. Not only does this avoid problems with interpreting extreme values, but more importantly it opens the door to the non-sigmoidal ICCs as depicted in Figure 2.

4 EXPERIMENTS WITH STUDENT ANSWER DATASETS

We begin by applying and evaluating β^3-IRT to model responses of students, which is a common application of IRT. We use datasets that consist of answers given by students of 33 different courses from an online platform (which we cannot disclose, for commercial reasons). The courses have different numbers of questions, but not all students have answered all questions for a given course (in fact, most did not), therefore we did not have p_{ij} values for every student i and question j. On the other hand, it is possible for a student to answer a question multiple times. These datasets may contain noise derived from user behaviour, such as quickly answering questions just to see the next ones and lending accounts to other students.

We compare the β^3-IRT model with Noel and Dauvier's continuous IRT model. Since without a discrimination parameter their model is rather weak, we strengthen it by introducing a discrimination parameter as follows:

$$\mathbb{E}[p_{ij}|\boldsymbol{\theta}_i,\boldsymbol{\delta}_j,\mathbf{a}_j] = \frac{m_{ij}}{m_{ij}+n_{ij}} = \frac{1}{1+e^{-\mathbf{a}_j(\boldsymbol{\theta}_i-\boldsymbol{\delta}_j)}} \quad (8)$$

We refer to this model as 2PL-ND.

Our experiments follow two scenarios. In the first scenario, to generate the continuous responses, we take the i-th student's average performance for the j-th question, $p_{ij} = \mathbb{E}[x_{ij}]$. In the second scenario, only the first attempts of each student for each question are considered, leading to

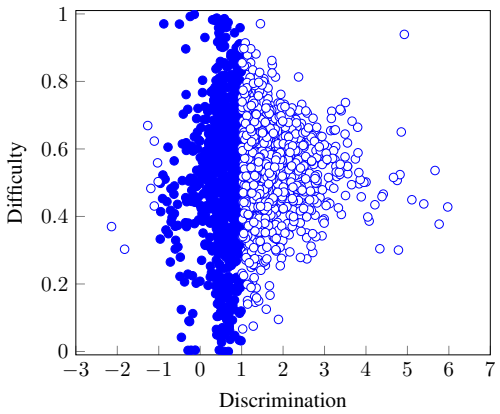


Figure 3: Discrimination versus Difficulty (Inferred in one of the 30 hold-out schemes across all 33 student answer datasets. Solid dots represent discriminations in the $(-1, 1)$ range, which produce non-logistic curves in our approach. These represent approximately 46% of all estimated discriminations.)

a binary problem. The models were trained by minimising the log-loss of predicted probabilities using SGD, implemented in Python, with the Theano library. We trained the models for 2500 iterations, using batches of 2000 answers and an adaptive learning rate calculated as $0.5/\sqrt{t}$, where t is the current iteration. β^3-IRT predicts the probability that the i-th student will answer the j-th question correctly following Eq (4), while 2PL-ND follows Eq (8).

For 2PL-ND, abilities and difficulties are unbounded and drawn from $\mathcal{N}(0, 1)$. For β^3-IRT, abilities are bounded in the range $(0, 1)$, so they were drawn from $\mathcal{B}(m_i^+, m_i^-)$, were m_i^+ (m_i^-) represent the number of correct (incorrect) answers given by the i-th student. Likewise, difficulties were drawn from $\mathcal{B}(n_j^+, n_j^-)$, where n_j^+ (n_j^-) represent the number of correct (incorrect) answers given to the j-th question. For both models, discriminations were drawn from $\mathcal{N}(1, 1)$.

In the experiments, we ran 30 hold-out schemes with 90% of the dataset kept for training and 10% for testing. For a training set to be valid, every student and question must occur in it at least once. Therefore, we sample the training and test sets with stratification based on the students and then we check which questions are absent, sampling 90% of their answers for training and 10% for testing.

Table 1 shows the log-loss for all 33 courses (full details of these courses can be found in the supplementary material). Best results were marked in bold, in case they were statistically significant according to a paired Wilcoxon signed-rank test, with significance level 0.001. The results show that β^3-IRT outperformed 2PL-ND in all cases. We posit that this can be explained by the versatility of the model, which is able to find non-logistic ICCs when discriminations are taken from $(-1, 1)$. Fig 3 shows that roughly 46% of all questions from the 33 datasets are estimated to have discriminations in this interval, supporting this conclusion. As shown in Fig 2c, non-logistic ICCs are better at discriminating lower and higher ability values than logistic ICCs, which focus more on ability values around the corresponding question difficulty, and β^3-IRT is able to model both cases. Additionally, approximately 10% of all questions had negative discriminations. One possible explanation is the presence of noise in these datasets, as mentioned earlier in this section. We will return to the relationship between negative discrimination and noise in Section 5.

Table 1: Student Answer Datasets (Log-loss) for Continuous (Student's Average) and First Attempt Performance.

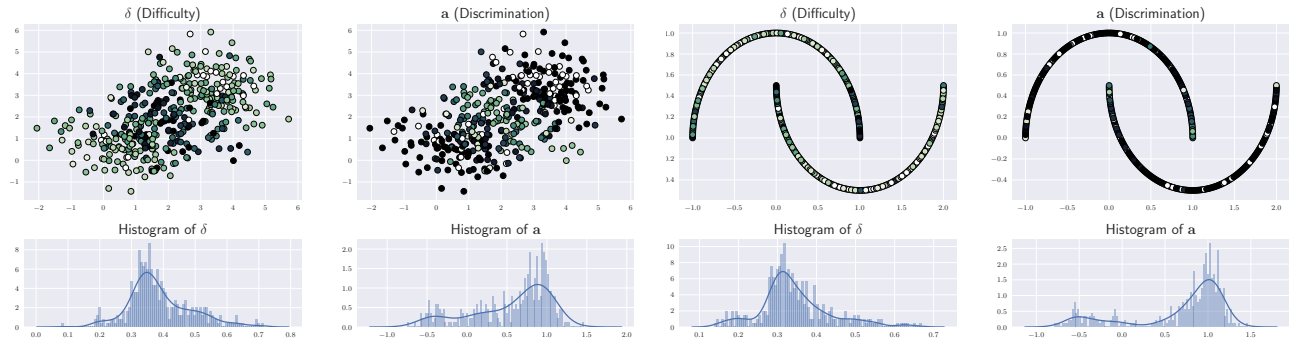
	continuous		first attempts	
course	β^3-IRT	2PL-ND	β^3-IRT	2PL-ND
1	**0.631 ± 0.003**	0.713 ± 0.004	**0.623 ± 0.004**	0.699 ± 0.005
2	**0.630 ± 0.022**	0.972 ± 0.081	**0.623 ± 0.023**	0.953 ± 0.060
3	**0.617 ± 0.004**	0.695 ± 0.004	**0.628 ± 0.024**	0.760 ± 0.086
4	**0.671 ± 0.004**	0.742 ± 0.009	**0.669 ± 0.004**	0.731 ± 0.007
5	**0.594 ± 0.004**	0.692 ± 0.008	**0.597 ± 0.004**	0.696 ± 0.013
6	**0.661 ± 0.009**	0.899 ± 0.039	**0.651 ± 0.009**	0.892 ± 0.030
7	**0.630 ± 0.007**	0.795 ± 0.020	**0.632 ± 0.007**	0.791 ± 0.015
8	**0.648 ± 0.014**	0.941 ± 0.044	**0.641 ± 0.023**	0.967 ± 0.059
9	**0.657 ± 0.011**	0.941 ± 0.030	**0.660 ± 0.011**	0.931 ± 0.032
10	**0.649 ± 0.007**	0.847 ± 0.030	**0.655 ± 0.009**	0.841 ± 0.032
11	**0.633 ± 0.016**	0.889 ± 0.051	**0.630 ± 0.012**	0.891 ± 0.067
12	**0.650 ± 0.013**	0.938 ± 0.063	**0.662 ± 0.016**	0.883 ± 0.051
13	**0.697 ± 0.066**	1.002 ± 0.218	**0.659 ± 0.086**	1.023 ± 0.423
14	**0.642 ± 0.028**	0.936 ± 0.074	**0.623 ± 0.028**	0.909 ± 0.057
15	**0.588 ± 0.002**	0.650 ± 0.003	**0.584 ± 0.003**	0.642 ± 0.003
16	**0.605 ± 0.002**	0.674 ± 0.003	**0.603 ± 0.002**	0.663 ± 0.002
17	**0.603 ± 0.002**	0.665 ± 0.003	**0.596 ± 0.003**	0.657 ± 0.003
18	**0.598 ± 0.006**	0.725 ± 0.008	**0.608 ± 0.005**	0.729 ± 0.011
19	**0.651 ± 0.015**	0.923 ± 0.064	**0.644 ± 0.020**	0.934 ± 0.074
20	**0.640 ± 0.021**	0.959 ± 0.060	**0.636 ± 0.018**	0.933 ± 0.040
21	**0.639 ± 0.016**	0.949 ± 0.072	**0.650 ± 0.014**	0.968 ± 0.094
22	**0.629 ± 0.020**	0.935 ± 0.050	**0.622 ± 0.016**	0.931 ± 0.060
23	**0.602 ± 0.004**	0.692 ± 0.011	**0.609 ± 0.004**	0.682 ± 0.005
24	**0.657 ± 0.014**	0.950 ± 0.046	**0.652 ± 0.011**	0.950 ± 0.044
25	**0.642 ± 0.015**	0.917 ± 0.034	**0.627 ± 0.010**	0.871 ± 0.038
26	**0.572 ± 0.011**	0.836 ± 0.045	**0.593 ± 0.014**	0.874 ± 0.038
27	**0.662 ± 0.022**	0.998 ± 0.093	**0.647 ± 0.028**	0.971 ± 0.073
28	**0.603 ± 0.001**	0.647 ± 0.002	**0.603 ± 0.001**	0.645 ± 0.002
29	**0.553 ± 0.021**	0.916 ± 0.056	**0.558 ± 0.017**	0.856 ± 0.051
30	**0.646 ± 0.019**	1.001 ± 0.075	**0.647 ± 0.019**	0.979 ± 0.048
31	**0.647 ± 0.014**	0.911 ± 0.060	**0.634 ± 0.014**	0.918 ± 0.053
32	**0.578 ± 0.001**	0.627 ± 0.001	**0.578 ± 0.002**	0.626 ± 0.002
33	**0.663 ± 0.021**	0.929 ± 0.060	**0.674 ± 0.025**	0.993 ± 0.074

5 β^3-IRT FOR CLASSIFIERS

We now apply β^3-IRT to supervised machine learning. Here, each 'respondent' is a different classifier and items are instances from a classification dataset[1]. The responses are the probabilities of the *correct* class assigned by the classifiers to each instance. Specifically, the observed response is given by $p_{ij} = \sum_k \mathbb{I}(y_j = k) p_{ijk}$, where $\mathbb{I}(\cdot)$ is the indicator function, y_j is the label of item j, k is the index of each class, and p_{ijk} is the predicted probability of item j from class k given by classifier i. The application of β^3-IRT in classification tasks aims to answer the following:

1. Can item parameters be used to characterise instances in terms of: (i) how difficult it is to estimate class probabilities for each instance; and (ii) how useful

[1]Note that an entire dataset could be considered as an item too, with associated difficulty.



(a) Dataset CLUSTERS (b) Dataset MOONS.

Figure 4: Inferred Latent Variables of Items of Two Synthetic Datasets. Darker colour indicates higher value. Items closer to the class boundary get higher difficulty and lower discrimination.

each instance is to discriminate between good and bad probability estimators?

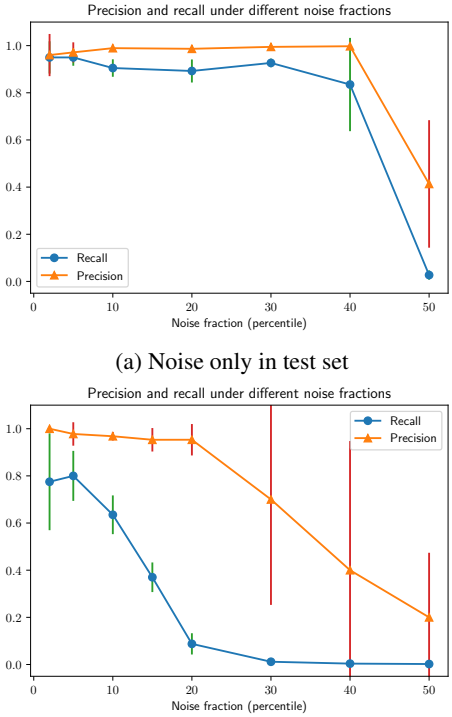
2. Does respondent ability serve as a new measure of performance that can be used to complement classifier evaluation?

The following steps are adopted to obtain the responses from M classifiers in a dataset: 1. Train all M classifiers on a training set; 2. Use all trained classifiers to predict the probability of each class for each data instance of a test set, which gives p_{ijk}; 3. Compute p_{ij} from p_{ijk}. Inference is then performed in the β^3-IRT model using the responses of the M classifiers to N test instances.

5.1 Experimental setup

We first applied the β^3-IRT model on two synthetic binary classification datasets, MOONS and CLUSTERS, chosen because they are convenient for visualisation. Both datasets are available in scikit-learn [Pedregosa et al., 2011]. Each dataset is divided into training and test sets, each with 400 instances. We also tested the model on classes 3 vs 5 from the MNIST dataset [LeCun et al., 2010], chosen as they are similar and contain difficult instances. For the MNIST dataset, the training and test sets have 1000 instances. The classes are balanced in each dataset. We inject noise in the test set by flipping the label y_j for 20% of randomly chosen data instances. The hyperparameter of discrimination σ_0 is set to 1 across all tests unless specified explicitly.

We tested 12 classifiers in this experiment: (i) Naive Bayes; (ii) Multi-layer Perceptron (MLP) (two hidden layers with 256 and 64 units); (iii) AdaBoost; (iv) Logistic Regression (LR); (v) k-Nearest Neighbours (KNN) (K=3); (vi) Linear Discriminant Analysis (LDA); (vii) Quadratic Discriminant Analysis (QDA); (viii) Decision Tree; (ix) Random Forest; (x) Calibrated Constant Classifier (assign probability $p = 0.5$ to all instances); (xi) Positive classifier (always assign positive class to instances); (xii) Negative classifier (always assign negative class to instances). All except the last three are taken from scikit-learn [Pedregosa et al., 2011] using default configuration unless specified explicitly.



(a) Noise only in test set

(b) Noise in training and test set

Figure 5: Denoising Performance of Negative Discrimination in Different Settings. Tested on MNIST dataset, means and standard deviations over 5 runs of all combinations of any two classes.

5.2 Exploring item parameters

Figs 4a and 4b show the posterior distributions of difficulty and discrimination in the CLUSTERS and MOONS datasets. Instances near the true decision boundary have higher difficulties and lower discriminations for both datasets, whereas instances far away from the decision boundary have lower difficulties and higher discriminations. Fig 6 illustrates

ICCs for different combinations of difficulty, discrimination and ability. Our model provides more flexibility than logistic-shape ICC.

There are some items inferred to have negative discrimination: these are mostly items with incorrect labels, as shown in Fig 7a. The negative discrimination fits the case when a low-valued response (correctness) is given by a classifier with high ability to an item with low difficulty. Figs 7a and 7b shows that negative discrimination flips high difficulty to low difficulty in comparison with the results where the discrimination is fixed. Figs 7c and 7d show that when there are no noisy items, no negative discriminations are inferred by the model.

However, unlike the common setting of label noise detection [Frénay and Verleysen, 2014], using negative discrimination to identify noisy labels requires that the training set can only include very few noisy examples. The reason is that noise in the training set introduces noise to all classifiers' abilities, and hence the noise in test set is hard to be identified. The experiment results of such cases are compared in Fig 5. This is a common issue in ensemble-based approaches for noise detection, which has been addressed for instance in [Sluban and Lavrač, 2015]. Our model can be trained on a small noise-free training set and then updated incrementally with identified non-noisy items, which is still practical in real applications. In contrast, in students answer experiments, there is no separate training set to build students' abilities before getting their answers for questions because the students are assumed to be trained by formal education already.

5.3 Assessing the ability of classifiers

Fig 9 shows a linear-like relation between ability and average response except the top right and bottom left of the figure. However, most classifiers are in the non-linear part, with ability between 0.7 and 0.8 and avg. response around 0.72, and the highest ability does not correspond to the highest avg response. This is caused by the element-wise difference which we will discuss below.

Table 2 shows the comparison between abilities and several popular classifier evaluation metrics on the MNIST dataset, while Table 3 gives the Spearman's rank correlation between these metrics. The experiment results of CLUSTERS are provided in the supplementary materials. We can see that ability behaves differently from the other metrics since it is not only estimated using aggregates of predicted probabilities, but also by the difficulties and discriminations of corresponding items. This can be seen from the equation below:

$$\left(\frac{1}{\bar{p}_{ij}} - 1\right)^{\mathbf{a}_j^{-1}} \left(\frac{1}{\boldsymbol{\delta}_j} - 1\right) = \frac{1}{\boldsymbol{\theta}_i} - 1,\ \bar{p}_{ij} = \frac{\alpha_{ij}}{\alpha_{ij} + \beta_{ij}} \tag{9}$$

Where $\bar{p}_{ij}$ is the expected response of item j given by classifier i, α_{ij} and β_{ij} are defined in Eq (3). For example, a low $\bar{p}_{ij}$ for a difficult instance will not give high penalty to the ability $\boldsymbol{\theta}_i$ because the difficulty $\boldsymbol{\delta}_j$ is high and the discrimination $\mathbf{a}_j$ is often close to zero for difficult items as we observed in Fig 4, whereas log-loss will generate high penalty as long as the correctness is low and Brier score will only consider the correctness as well. The ability learned by our model provides a new scaled measurement for classifiers, which evaluates the performance of probability estimation in a sense of weighted instance-wise basis.

Table 2: Comparison between Ability and other Classifier Performance Metrics (MNIST)

	Avg. Resp.	Ability	Accuracy	F1 score	Brier score	Llog loss	AUC
DT	0.7398	0.7438	0.7425	0.7297	0.2337	1.1537	0.7776
NB	0.6439	0.7423	0.6425	0.6951	0.3533	10.6097	0.6799
MLP	0.7826	0.8384	0.7825	0.774	0.2086	2.457	0.7951
Ada.	0.5621	0.4887	0.775	0.7656	0.2036	0.5959	0.79
RF	0.7215	0.7395	0.7725	0.7573	0.1926	3.4979	0.8119
LDA	0.7185	0.8052	0.72	0.7098	0.2714	5.1328	0.7276
QDA	0.5948	0.5892	0.595	0.6611	0.405	13.9884	0.5854
LR	0.7699	0.8001	0.7775	0.7688	0.2059	1.4246	0.7939
KNN	0.7645	0.8228	0.7675	0.7572	0.2111	6.3816	0.8011

Table 3: Spearman's Rank Correlation between Ability and other Classifier Performance Metrics (MNIST)

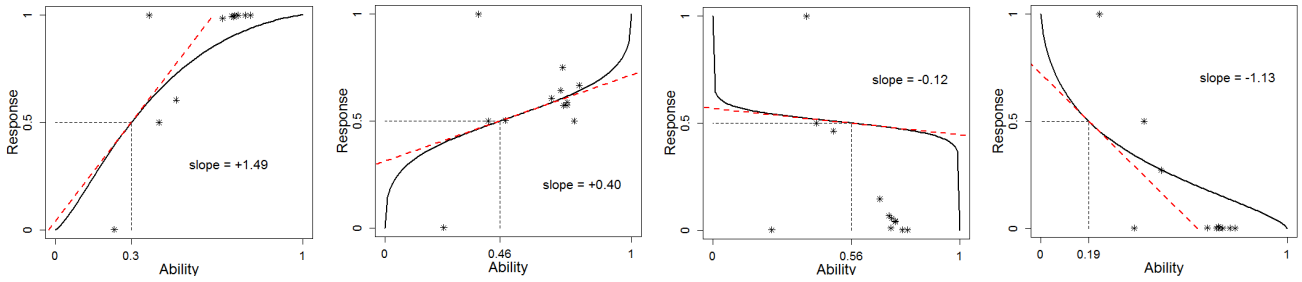
	Avg. Resp.	Ability	Accuracy	F1	Brier	Log loss	AUC
Avg. Resp.	1.0	0.8333	0.6	0.6	0.2833	0.2	0.6
Ability	0.8333	1.0	0.3333	0.3333	-0.05	-0.05	0.35
Accuracy	0.6	0.3333	1.0	1.0	0.8333	0.7	0.75
F1	0.6	0.3333	1.0	1.0	0.8333	0.7	0.75
Brier	0.2833	-0.05	0.8333	0.8333	1.0	0.6833	0.8333
Log loss	0.2	-0.05	0.7	0.7	0.6833	1.0	0.3667
AUC	0.6	0.35	0.75	0.75	0.8333	0.3667	1.0

Another advantage of ability as a measurement of classifiers is that it is robust to noisy test data. Fig 8 demonstrates that the inferred abilities of the classifiers stay nearly constant as the noise fraction in the test set is increased until half of the test points are incorrectly labelled.

6 CONCLUSIONS

This paper proposed β^3-IRT, a new IRT model solves the limitations of a previous continuous IRT approach [Noel and Dauvier, 2007], by adopting a new formulation which allows the model to produce a new family of Item Characteristic Curves including sigmoidal and anti-sigmoidal curves. Therefore, our β^3-IRT model is more versatile than previous IRT models, being able to model a significantly more expressive class of response patterns. Additionally, our new formulation assumes bounded support for abilities and difficulties in the $[0, 1]$ range which produces more natural and interpretable results than previous IRT models.

We evaluated β^3-IRT in two experimental scenarios. First, β^3-IRT was applied to the psychometric task of student



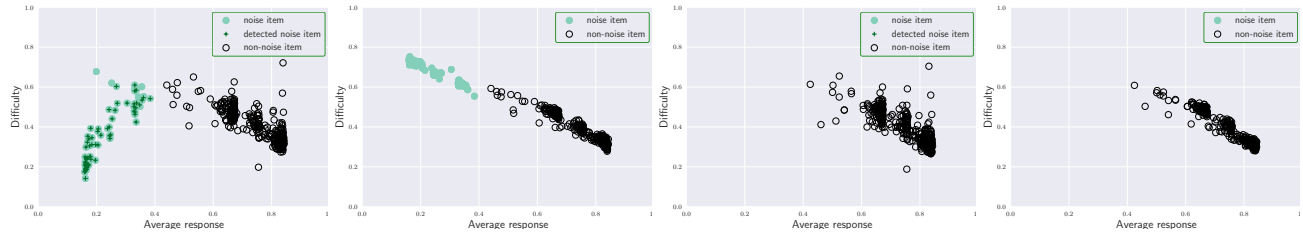
(a) Difficulty low, positive discrimination high. Instance is non-noisy item far from decision boundary.

(b) Difficulty high, positive discrimination low. Instance is non-noisy item close to decision boundary.

(c) Difficulty high, negative discrimination low. Instance is noisy item close to decision boundary.

(d) Difficulty low, negative discrimination high. Instance is noisy item far from decision boundary.

Figure 6: Examples of ICC in the CLUSTERS Dataset. Stars are the actual classifier responses fit by the ICCs.



(a) $\mathbf{a}_j$ is learned.

(b) $\mathbf{a}_j$ is fixed to 1.

(c) $\mathbf{a}_j$ is learned.

(d) $\mathbf{a}_j$ is fixed to 1.

Figure 7: Correlation between Average Response and Difficulty Changes. (Under different settings of discrimination, shown for classes 3 vs 5 of MNIST dataset. (a), (b) are from test data with 20% injected noise; (c), (d) no noise.)

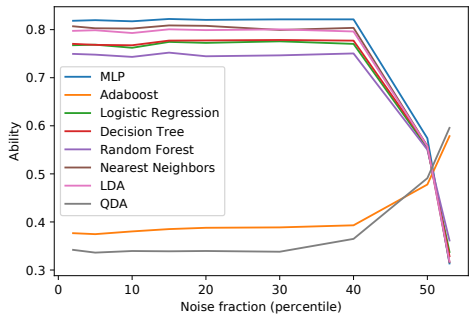


Figure 8: Ability of classifiers with different noise fractions of test data, which shows the ability is robust to noisy test data.

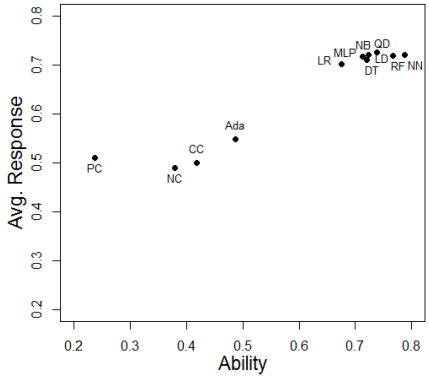


Figure 9: Ability vs Average Response in the CLUSTERS dataset. The classifiers in the top right getting similar avg, response around 0.7, but their abilities are diverse from 0.65 to 0.8.

performance estimation, β^3-IRT outperformed 2PL-ND in all 33 datasets, showing the importance of the versatility of the ICCs that are produced by our approach. We then applied β^3-IRT in a binary classification scenario. The results showed that item parameters inferred by the β^3-IRT model can provide useful insights for difficult or noisy instances, and the inferred latent ability variable serves to evaluate classifiers on an instance-wise basis in terms of probability estimation. To the best of our knowledge, this was the first time that an IRT model was used for these tasks.

Future work includes using the model as a tool for model selection or ensembling. Another use is to design datasets for benchmarking, based on estimated difficulties and discriminations. Another extension would be to replace the Beta with a Dirichlet distribution to cope with multi-class scenarios.

Acknowledgements

Part of this work was supported by The Alan Turing Institute under EPSRC grant EP/N510129/1. Ricardo Prudêncio was financially supported by CNPq (Brazilian Agency).

References

[Bachrach et al., 2012] Bachrach, Y., Minka, T., Guiver, J., and Graepel, T. (2012). How to grade a test without knowing the answers: a Bayesian graphical model for adaptive crowdsourcing and aptitude testing. In *Proc. of the 29th Int. Conf. on Machine Learning*, pages 819–826. Omnipress.

[Bishop, 2006] Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.

[Blei et al., 2017] Blei, D. M., Kucukelbir, A., and McAuliffe, J. (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877.

[Embretson and Reise, 2013] Embretson, S. and Reise, S. (2013). *Item Response Theory for Psychologists*. Taylor & Francis.

[Frénay and Verleysen, 2014] Frénay, B. and Verleysen, M. (2014). Classification in the presence of label noise: a survey. *IEEE transactions on neural networks and learning systems*, 25(5):845–869.

[Kingma and Ba, 2014] Kingma, D. and Ba, J. (2014). Adam: A method for stochastic optimization. In *Proc. of the 3rd Int. Conf. on Learning Representations (ICLR)*.

[Kingma and Welling, 2013] Kingma, D. P. and Welling, M. (2013). Auto-encoding variational Bayes. In *Proc. of the 2nd Int. Conf. on Learning Representations (ICLR)*.

[LeCun et al., 2010] LeCun, Y., Cortes, C., and Burges, C. J. (2010). MNIST handwritten digit database. *AT&T Labs [Online]. Available: http://yann. lecun. com/exdb/mnist*, 2.

[Martínez-Plumed et al., 2016] Martínez-Plumed, F., Prudêncio, R. B., Martínez-Usó, A., and Hernández-Orallo, J. (2016). Making sense of item response theory in machine learning. In *European Conference on Artificial Intelligence, ECAI*, pages 1140–1148.

[Nishihara et al., 2013] Nishihara, R., Minka, T., and Tarlow, D. (2013). Detecting parameter symmetries in probabilistic models. *arXiv preprint arXiv:1312.5386*.

[Noel and Dauvier, 2007] Noel, Y. and Dauvier, B. (2007). A beta item response model for continuous bounded responses. *Applied Psychological Measurement*, 31(1):47–73.

[Pedregosa et al., 2011] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

[Sluban and Lavrač, 2015] Sluban, B. and Lavrač, N. (2015). Relating ensemble diversity and performance: A study in class noise detection. *Neurocomputing*, 160:120–131.

[Tran et al., 2016] Tran, D., Kucukelbir, A., Dieng, A. B., Rudolph, M., Liang, D., and Blei, D. M. (2016). Edward: A library for probabilistic modeling, inference, and criticism. *arXiv preprint arXiv:1610.09787*.