Item Response Theory for Evaluating Regression Algorithms

João V. C. Moraes[1], Jessica T. S. Reinaldo[2],
Ricardo B. C. Prudencio[3]
Centro de Informática
Universidade Federal de Pernambuco
Recife, Brazil
E-mail: jvcm@cin.ufpe.br[1], jtsr@cin.ufpe.br[2],
rbcp@cin.ufpe.br[3]

Telmo M. Silva Filho
Departamento de Estatística
Universidade Federal da Paraíba
João Pessoa, Brazil
E-mail: telmo@de.ufpb.br

***Abstract*—Item Response Theory (IRT) is a tool developed in psychometrics to measure latent abilities of human respondents based on their responses to items with different levels of difficulty. Recently, IRT has been applied to evaluation in AI, by treating the algorithms as respondents and the AI tasks as items. Particularly in machine learning, IRT has been applied for evaluation of classifiers based on their predictions to each test instance. Based on a matrix of responses (classifiers vs instances), the IRT model estimates the latent difficulty and discrimination of each instance, as well as the ability of each classifier, in such a way that a classifier receives high ability value when it tends to correctly classify the most difficult instances. The IRT models previously adopted for evaluation in classification are not directly applied for regression, since they rely on dichotomous responses (i.e., a response has to be either correct or incorrect). In this paper we propose a new IRT model, particularly designed for dealing with nonnegative unbounded responses, which is adequate for modelling the absolute errors of regression algorithms. In the proposed model, responses follow a gamma distribution, parameterised according to respondents' abilities and items' difficulty and discrimination parameters. The proposed parameterisation results in item characteristic curves with more flexible shapes compared to the logistic curves widely adopted in IRT. The proposed model was evaluated with diverse regression algorithms and two benchmark datasets, one synthetic and one real. Useful insights were derived by inspecting regions in these datasets that present different levels of difficulty and discrimination.**

***Index Terms*—Regression, Item Response Theory, Machine Learning, Evaluation**

I. Introduction

Psychometrics is a research field focused on the objective measurement of cognitive traits, including personality, attitude and intelligence. Item Response Theory (IRT) comprises a set of psychometric models aiming to estimate the latent ability of humans based on their responses to test items with different levels of difficulty [1]. The concept of item depends on the application, and can represent, for instance, test questions, judgements or choices in exams. IRT has been commonly applied to assess the performance of students in exams and in health applications.

In practice, an IRT model produces for each item an Item Characteristic Curve (ICC), which is a function returning the probability of a correct response for the item based on the respondent ability. The ICC is usually a logistic curve determined by two item parameters: difficulty, which is the location parameter of the logistic function; and discrimination, which affects the slope of the ICC. Both latent item traits and the latent abilities of respondents are jointly estimated based on observed responses in a test. Respondents who correctly answer the most difficult items will be assigned high ability values if they also correctly answer easier items, otherwise the model will implicitly assume that said respondents were guessing.

More recently, IRT has been applied for evaluation in AI, where items are tasks and respondents are AI models, although this field is still in an early stage. IRT was adopted in Machine Learning (ML) classification tasks [2, 3], in which items correspond to instances in a dataset, respondents are classifiers, and the binary responses are right or wrong classification outcomes collected in a cross-validation experiment. In another application of IRT for ML classification, [4] proposed the β^3-IRT to model continuous responses in the $[0, 1]$ range, which was then applied to fit class probabilities returned by ML models. IRT has also been used to evaluate AI techniques in other contexts, such as AI games [5] and NLP [6].

Despite useful insights, previous works are limited to the application of IRT for binary and bounded responses. These are not directly applicable, for instance, to evaluate regression models, in which outcomes are continuous unbounded errors. This is also true in many other contexts, in which success is measured in a continuous unbounded scale. In order to overcome this limitation, we propose the Γ-IRT model, which models nonnegative continuous responses by adopting the Gamma distribution. The model offers a wide range of ICCs by defining the Gamma parameters as a proper combination of item difficulty and discrimination and respondent ability.

We apply the proposed model to fit normalised errors produced in regression tasks. In the experiments, noise was gradually injected into the regression datasets, thus inducing changes in the item parameters and model abilities. We demonstrate the use of Γ-IRT to identify regions of high difficulty within the dataset and we propose ability as a complementary

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

measure to evaluate regression models. Our contributions can be summarised as follows:

1) We propose Γ-IRT, a new IRT model which focuses on nonnegative unbounded responses, which have not been adequately treated in the IRT literature;
2) We use Γ-IRT to perform regression evaluation, which is a novel application in literature.

The paper is organised as follows. Initially, we present related work on IRT, followed by the description of the Γ-IRT model. Then we use Γ-IRT to analyse regression models and datasets. Finally, we present our final remarks and discuss future works.

II. Related Work

In Psychometrics, nonnegative continuous responses have been previously analysed in the context of student reading speed [7, 8, 9], where responses correspond to the total time t_{ij} a respondent i takes to finish reading an item j, which is a text consisting of m words. The first of these works [7] modelled t_{ij} with a gamma density given by:

$$p(t_{ij}|\theta_i, \delta_j) \equiv \frac{(\theta_i/\delta_j)^m}{\Gamma(m)} t_{ij}^{(m-1)} e^{-\theta_i t_{ij}/\delta_j}, \tag{1}$$

where, similarly to standard IRT, θ_i is the ability of the i-th student, δ_j is the difficulty of the j-th item and $\Gamma(m) \equiv (m-1)!$ is the gamma function. In this gamma density, the intensity parameter is $\lambda_{ij} \equiv \theta_i/\delta_j$, thus the expected number of words to be read in a given time unit is assumed to be a function of the student's speed and the item's difficulty.

Another model for continuous nonnegative responses was proposed by [10] where the goal was to estimate the probability that respondent i would give response $z_{ij} \in (0, 1)$ to item j. This model was later generalised to work with multiple latent respondent traits [11] and its relation to linear Factor Analysis was shown by [12].

Although these models are designed for nonnegative responses, we focus on a different context, therefore their assumptions do not apply here. Our task is not to estimate the probability of a particular response value, given respondent and item latent traits, but to estimate the actual value. In this context, previous works that tackled the problem of IRT models for continuous responses mainly focused on responses with bounded support. Noel and Dauvier [13] proposed an IRT model for responses in the $[0, 1]$ range, adopting the Beta distribution as follows:

$$\begin{aligned} m_{ij} &= \exp\left(\frac{\theta_i - \delta_j}{2}\right), \\ n_{ij} &= \exp\left[-\left(\frac{\theta_i - \delta_j}{2}\right)\right] = \frac{1}{m_{ij}}, \\ p_{ij} &\sim \mathcal{B}(m_{ij}, n_{ij}). \end{aligned} \tag{2}$$

In this model, p_{ij} is the continuous response given by respondent i to item j. The model gives a logistic ICC mapping ability to expected response for item j of the form:

$$\mathbb{E}[p_{ij}|\theta_i, \delta_j] = \frac{m_{ij}}{m_{ij} + n_{ij}} = \frac{1}{1 + \exp\left(-(\theta_i - \delta_j)\right)}. \tag{3}$$

This is very similar to the ICC for dichotomous responses, with a key difference: its output is the expected value of a continuous variable in the $[0, 1]$ range. This model does not have a discrimination parameter, being similar to a standard 1PL IRT model, and it always produces logistic ICCs. To solve these limitations Chen et al. [4] introduced the β^3-IRT model, which can generate a rich family of ICCs for responses in the $[0, 1]$ range. Equation (4) below gives the model definition, where p_{ij} is the observed response of respondent i to item j, which is drawn from a Beta distribution.

$$\begin{aligned} p_{ij} &\sim \mathcal{B}(\alpha_{ij}, \beta_{ij}), \\ \alpha_{ij} &= \mathcal{F}_\alpha(\boldsymbol{\theta}_i, \boldsymbol{\delta}_j, \mathbf{a}_j) = \left(\frac{\boldsymbol{\theta}_i}{\boldsymbol{\delta}_j}\right)^{\mathbf{a}_j}, \\ \beta_{ij} &= \mathcal{F}_\beta(\boldsymbol{\theta}_i, \boldsymbol{\delta}_j, \mathbf{a}_j) = \left(\frac{1 - \boldsymbol{\theta}_i}{1 - \boldsymbol{\delta}_j}\right)^{\mathbf{a}_j}, \\ \boldsymbol{\theta}_i &\sim \mathcal{B}(1, 1),\ \boldsymbol{\delta}_j \sim \mathcal{B}(1, 1),\ \mathbf{a}_j \sim \mathcal{N}(1, \sigma_0^2) \end{aligned} \tag{4}$$

The Beta 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). Both $\boldsymbol{\theta}_i$ and $\boldsymbol{\delta}_j$ are drawn from Beta distributions, i.e. they are measured on a $[0, 1]$ scale, which means that their values are arguably easier to interpret than in other IRT models, in which abilities and difficulties are unbounded. The new parameterisation is able to model non-logistic ICCs defined by the expectation of $\mathcal{B}(\alpha_{ij}, \beta_{ij})$ and assuming the form given by Equation (5).

$$\mathbb{E}[p_{ij}|\boldsymbol{\theta}_i, \boldsymbol{\delta}_j, \mathbf{a}_j] = \frac{\alpha_{ij}}{\alpha_{ij} + \beta_{ij}} = \frac{1}{1 + \left(\frac{\boldsymbol{\delta}_j}{1-\boldsymbol{\delta}_j}\right)^{\mathbf{a}_j} \left(\frac{\boldsymbol{\theta}_i}{1-\boldsymbol{\theta}_i}\right)^{-\mathbf{a}_j}} \tag{5}$$

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. The ICCs can have different shapes depending on $\mathbf{a}_j$, such as sigmoid shapes similar to standard IRT, anti-sigmoidal behaviours and parabolic curves.

Another contribution in Chen et al.'s work was a machine learning application of β^3-IRT, inspired by previous work by Martínez-Plumed et al. [2], where classifiers were evaluated using their ability values, which show interesting properties as a performance measure. Additionally, item discrimination values were used to detect noisy instances.

Aside from IRT, modelling responses in Psychometrics has long been done using Factor Analysis (FA) [14, 15, 16], which assumes that responses are continuous and unbounded. One difference between FA and IRT, which is particularly important if these models are applied in machine learning, is the interpretation of its factors, which are not as clearly defined as IRT's respondent and item parameters.

III. The Γ-IRT Model

We now propose Γ-IRT to model unbounded nonnegative responses, such as students' errors to open-ended questions or the absolute values of errors coming out of regression models.

Particularly in the machine learning domain, to the best of our knowledge, the task of fitting IRT models to regression errors is still an open problem.

A. Formulation

The central idea of Γ-IRT is to model continuous errors as random variables following Gamma distributions, parameterised according to item difficulty and discrimination and respondent ability. Let $e_{ij} \in (0, \infty)$ be the observed error of respondent i to item j. For regression in our work, e_{ij} is the absolute error of a regression model in a test instance. Thus, we have:

$$\begin{aligned} & e_{ij} \sim \Gamma(\alpha_{ij}, \beta_{ij}), \\ & \alpha_{ij} = \mathcal{F}_\alpha(\boldsymbol{\theta}_i, \boldsymbol{\delta}_j, \mathbf{a}_j, c_j) = c_j \left(\frac{\boldsymbol{\delta}_j}{\boldsymbol{\theta}_i}\right)^{\mathbf{a}_j}, \\ & \beta_{ij} = \mathcal{F}_\beta(\boldsymbol{\theta}_i, \boldsymbol{\delta}_j, \mathbf{a}_j) = \left(\frac{1-\boldsymbol{\delta}_j}{1-\boldsymbol{\theta}_i}\right)^{\mathbf{a}_j}, \\ & \boldsymbol{\theta}_i \sim \mathcal{B}(1,1),\ \boldsymbol{\delta}_j \sim \mathcal{B}(1,1),\ \mathbf{a}_j \sim \mathcal{N}(1, \sigma_0^2). \end{aligned} \tag{6}$$

In the model above, $\boldsymbol{\delta}_j \in (0; 1)$ is the difficulty parameter of item j, $\mathbf{a}_j$ is the discrimination parameter and $c_j > 0$ is the guessing parameter. For respondents, $\boldsymbol{\theta}_i \in (0; 1)$ is the ability of respondent i. In this model, the ICC is the expectation of $\Gamma(\alpha_{ij}, \beta_{ij})$ along ability, which assumes the following form:

$$\mathbb{E}[e_{ij}|\boldsymbol{\theta}_i, \boldsymbol{\delta}_j, \mathbf{a}_j, c_j] = \frac{\alpha_{ij}}{\beta_{ij}} = c_j \left(\frac{\boldsymbol{\delta}_j}{1-\boldsymbol{\delta}_j}\right)^{\mathbf{a}_j} \left(\frac{\boldsymbol{\theta}_i}{1-\boldsymbol{\theta}_i}\right)^{-\mathbf{a}_j} \tag{7}$$

The following properties can be pointed out from the ICCs for special cases of ability:

- If $\boldsymbol{\theta}_i \to 0$, then $\mathbb{E}[e_{ij}] \to \infty$, i.e., very large errors are expected for respondents with very low ability;
- If $\boldsymbol{\theta}_i \to 1$, then $\mathbb{E}[e_{ij}] \to 0$, i.e., in turn respondents with very high ability tend to produce very low errors;
- If $\boldsymbol{\theta}_i = \boldsymbol{\delta}_j$, then $\mathbb{E}[e_{ij}] = c_j$.

1) Guessing Parameter: The parameter c_j can be set as the expected error obtained by a random respondent. In the domain of regression, an item j is a test instance in a regression dataset. Let y_j be target variable associated to the test instance j. The regression model R is a naive regressor which always returns the average of the target attribute as its predictions. In this case, $c_j = \mathbb{E}[e_{Rj}] = |y_j - \bar{y}|$, in which $\bar{y}$ is the average of the target attribute.

In the ICC, a respondent has random performance when it faces an item for which difficulty equals her ability (if $\boldsymbol{\theta}_i = \boldsymbol{\delta}_j$, then $\mathbb{E}[e_{ij}] = c_j$). In particular, a model with ability $\boldsymbol{\theta}_i = 0.5$ will perform randomly when facing an item with difficulty $\boldsymbol{\delta}_j = 0.5$.

2) Difficulty Parameter: Item difficulty can be analysed regarding a middle point of ability $\boldsymbol{\theta}_i = 0.5$:

- If $\boldsymbol{\delta}_j < 0.5$, then $\mathbb{E}[e_{ij}] < c_j$ for $\boldsymbol{\theta}_i = 0.5$.
- If $\boldsymbol{\delta}_j > 0.5$, then $\mathbb{E}[e_{ij}] > c_j$ for $\boldsymbol{\theta}_i = 0.5$.

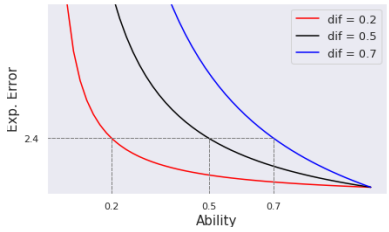


Fig. 1: Examples of ICCs for different values of difficulty. In all cases, $c_j = 2.4$ and $a_j = 1$.

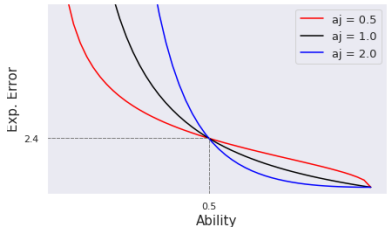


Fig. 2: Examples of ICCs for different values of discrimination. In all cases, $c_j = 2.4$ and $\boldsymbol{\delta}_j = 0.5$.

In the first case (easy items), even respondents with low ability will have errors lower than the guessing error. In the second case (difficult items), only respondents with high ability will outperform the guessing error.

See Figure 1 for examples of IRT curves for different difficulties. When $\boldsymbol{\delta}_j = 0.2$, some respondents with low ability (e.g., $0.2 < \boldsymbol{\theta}_i < 0.5$) are better than random. Only respondents with $\boldsymbol{\theta}_i < 0.2$ are worse. On the other hand, when $\boldsymbol{\delta}_j = 0.7$, there is a range of good respondents ($0.5 < \boldsymbol{\theta}_i < 0.7$) that do worse than random.

3) Discrimination Parameter: a_j characterises the slope of the curve at the difficulty level. Figure 2 presents examples of IRT curves, fixing difficulties and guessing parameters and varying discrimination. In all curves, the same expected error is obtained at the difficulty level $\boldsymbol{\theta}_i = \boldsymbol{\delta}_j = 0.5$. For $a_j = 0.5$, the expected errors are close to 2.4 (the guessing error) in a wide range of abilities, but when $a_j = 2$ we observe very high errors just before $\boldsymbol{\theta}_i = 0.5$ and very low errors just after this ability point. Thus, this item is more discriminative.

B. Normalised Errors

The guessing parameter can be avoided by taking the normalised errors and then deriving the corresponding ICC:

$$\begin{aligned} & \bar{e}_{ij} = \frac{e_{ij}}{c_j} \sim \Gamma(\alpha_{ij}, \beta_{ij} c_j) \\ & \alpha_{ij} = c_j \left(\frac{\boldsymbol{\delta}_j}{\boldsymbol{\theta}_i}\right)^{\mathbf{a}_j},\ \beta_{ij} c_j = \left(\frac{1-\boldsymbol{\delta}_j}{1-\boldsymbol{\theta}_i}\right)^{\mathbf{a}_j} \end{aligned} \tag{8}$$

Note that if $X \sim \Gamma(\alpha, \beta)$ then $\frac{1}{k}X \sim \Gamma(\alpha, k\beta)$. The normalised errors are drawn from a Gamma distribution, which is simply rescaled according to c_j. The expected normalised error is then:

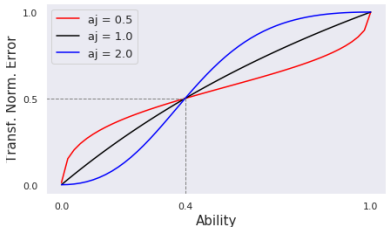


Fig. 3: Examples of β^3-IRT for regression. All curves were produced by setting $\boldsymbol{\delta} = 0.4$.

$$\mathbb{E}[\bar{e}_{ij}|\boldsymbol{\theta}_i, \boldsymbol{\delta}_j, \mathbf{a}_j, c_j] = \frac{\alpha_{ij}}{\beta_{ij} c_j} = \left(\frac{\boldsymbol{\delta}_j}{1-\boldsymbol{\delta}_j}\right)^{\mathbf{a}_j} \left(\frac{\boldsymbol{\theta}_i}{1-\boldsymbol{\theta}_i}\right)^{-\mathbf{a}_j} \tag{9}$$

As a special case, for $\boldsymbol{\theta}_i = \boldsymbol{\delta}_j$ then $\mathbb{E}[\bar{e}_{ij}] = 1$, which then serves as a reference for normalised responses better than random.

C. Relation to β^3-IRT

The following transformation produces a β^3-IRT curve, as given by Equation (5):

$$\begin{aligned}\mathbb{E}[p_{ij}|\boldsymbol{\theta}_i, \boldsymbol{\delta}_j, \mathbf{a}_j] &= \frac{1}{1+\mathbb{E}[\bar{e}_{ij}|\boldsymbol{\theta}_i, \boldsymbol{\delta}_j, \mathbf{a}_j, c_j]} \\ &= \frac{1}{1+\left(\frac{\boldsymbol{\delta}_j}{1-\boldsymbol{\delta}_j}\right)^{\mathbf{a}_j} \left(\frac{\boldsymbol{\theta}_i}{1-\boldsymbol{\theta}_i}\right)^{-\mathbf{a}_j}}\end{aligned} \tag{10}$$

This relationship is convenient for estimation since one can transform the normalised errors and produce the β^3-IRT curves, i.e., estimate the β^3-IRT curves from the responses in the form $\frac{1}{1+\bar{e}_{ij}}$. Then, the β^3-IRT curve can be transformed back into a Γ-IRT curve using the inverse of this transformation.

Figure 3 presents examples of β^3-IRT curves for regression. In the extremes, a transformed response close to 1 means an expected error close to 0. When $\bar{e}_{ij} \to \infty$, the transformed response tends to 0. When ability equals difficulty, the expected error is c_j and consequently the transformed normalised error is 0.5. This level can be used to visually distinguish a success from a failure. Models with ability $\boldsymbol{\theta}_i > 0.4$ in this case will be better than the random regression model.

D. Model inference

The Γ-IRT models were estimated using the transformation to β^3-IRT, as discussed in Section III-C. We then applied the Bayesian Variational Inference method (VI), as proposed by [4], using code available at `https://github.com/yc14600/beta3_IRT`, which is based on the Variational Inference package from the Edward and TensorFlow Python libraries.

IV. Experiments with Regression Models

In this Section, we apply Γ-IRT to machine learning regression problems. Each respondent is a different regression model and items are instances from a dataset. The idea is to extract insights from data and regression models, simultaneously analysing data instance difficulty and discrimination, as well as regression model ability.

Different datasets were simulated by injecting target noise in a benchmark regression task. Finally, we discuss about the items' characteristics and compare the ICCs of instances in specific regions in the dataset in order to seek for specific ICC patterns.

A. Dataset

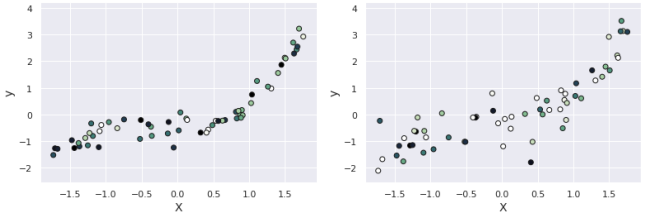
We performed experiments using a synthetic regression task derived from a third-degree polynomial function, as follows:

$$y = -x + x^3 + \epsilon, \ \epsilon \sim N(0, \sigma_y^2) \tag{11}$$

The predictor feature x is uniformly distributed inside the interval $[-4, 6]$ and ϵ is a random target noise.

In order to produce datasets with different levels of difficulty, distinct levels of Gaussian noise were injected in the target attribute. In addition, we standardised feature and target attributes to facilitate the process of noise injection and subsequent data analysis. The standard deviation of the target noise σ_y varied from 0 to 0.5, with increments of 0.025, with $\sigma_y = 0$ referred to as *baseline dataset*. Figure 4 (a) and (b) present two examples of generated datasets.

For each noise-level configuration, 40 datasets were produced with 300 instances each. In order to train and test the regression models, we randomly split the data into training and test sets, with 80% for training. Noise is only injected in the test set, therefore the training set is preserved from the original dataset. As mentioned in Ferri et al. [17] it is very common that the training data is under "idealistic" conditions, with features that are carefully measured and preprocessed.



(a) Target noise injected ($\sigma_y = 0.25$). (b) Target noise injected ($\sigma_y = 0.5$).

Fig. 4: Examples of the *polynomial* test set for each scenario and specific noise levels. (Darker colour indicates higher noise.)

B. Regression models

For every scenario described in the previous Subsection, we trained and tested 10 regression models (linear and nonlinear): (i) Linear Regression; (ii) Bayesian Ridge; (iii) Support Vector

Regression - linear kernel; (iv) Support Vector Regression - radial basis function (RBF) kernel and penalty parameter $C = 5.0$; (v) k-Nearest Neighbours Regression - $K = 5$; (vi) Decision Tree Regression; (vii) Random Forest Regression; (viii) AdaBoost Regression; (ix) Multilayer Perceptron - one hidden layer with 100 neurons; (x) Multilayer Perceptron - two hidden layers with 50 neurons each and logistic activation function. The regression models were implemented using the scikit-learn library. Unless the algorithm's parameters are not explicitly specified above, all models used scikit-learn's default configurations.

In addition to the mentioned regression models, we have artificially created 3 synthetic models: (i) Optimal - for each instance, it takes the best response among all regression models; (ii) Average - always predicts the mean value of the test set; (iii) Worst - takes the worst response amongst all regression models. These models are adopted as baselines for comparison.

The response e_{ij} is the absolute error obtained by the regression model i for instance j in the test set. Hence we produced an item-response matrix with 13 models and 60 test items for each simulated dataset. Finally, the average item parameters and regression model abilities are measured across the 40 datasets for each noise level.

C. Discussion

Figure 6 presents the difficulties of instances from test sets with three different noise levels. In the original test set, high-difficulty items are concentrated inside the interval $x \in [0.4, 1.0]$, while low-difficulty correspond to $x < -1.25$ and $x > 1.25$. As noise increases, difficulty increases in the central region of the curve. The target variable in this region is approximately constant, therefore injecting noise results in larger regression errors. However, significant changes in difficulty are not observed in the extreme regions of the curve since they do not suffer relevant distortions when noise is injected. Note that the difficulty histograms gradually shift to the right, reflecting higher difficulties in the presence of noise. The figure also presents the discrimination observed for the same three test sets. The picture is not as clear as in difficulty, but easier instances tend to show higher discrimination. The discrimination histograms gradually shift to the left, thus when noise is applied in the test set, instances tend to lose their power to discriminate between good and bad regression models. Figure 7 also shows the effects of noise injection in the parameters at instance level (items are represented by the dots). In the last case where maximum noise is applied, the first item with negative discrimination appears. In all cases, difficulty and discrimination have negative correlation.

Table I shows the responses given by all regression models, including the baselines (Average, Optimal and Worst), to instances (a) and (b). Instance (a) presents very low responses when compared to instance (b), thus it has higher relative errors and, consequently, higher difficulty.

Figure 5 illustrates the expected error along respondent ability of the same two data instances showed previously.

	Instance (a)	Instance (b)
LR	0.0099	0.6544
Bayes	0.0099	0.6541
SVR(Lin)	0.0180	0.5935
SVR(Rbf)	0.0944	0.9683
KNR	0.1593	0.9666
DT	0.3399	0.9912
RF	0.2931	0.9857
AdaB	0.1523	0.9331
MLP100	0.0952	0.9891
MLP50-50	0.0900	0.9833
Avg	0.5000	0.5000
Opt	0.5143	0.9915
Wrs	0.0099	0.5000

TABLE I: Example of response values for two data instances (items) and all regression models (respondents).

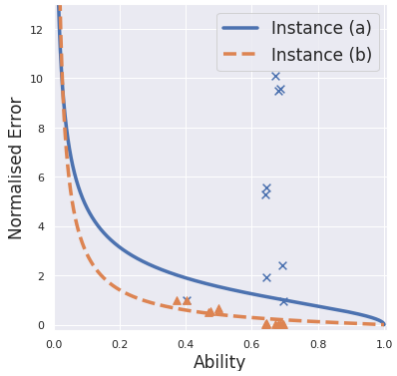


Fig. 5: Representative Item Characteristic Curves from *polynomial* dataset. Marks are the regression models' responses fit by the ICCs.

The marks indicate the normalised errors and abilities of all regression models. Each instance belongs to a specific region within the original test set, although its target value may vary as random noise is injected. Instance (b) represents an instance from a low-difficulty region, while instance (a) is an instance from a high-difficulty region. Notice that the curve shapes change according to each region. Unlike instance (a), which presents relatively high difficulty and low discrimination, instance (b) has a low associated error and appears to respond better to increased ability (high discrimination). Instance (a) and (b) have difficulty values of 0.70 and 0.27, with discrimination values of 52 and 89, respectively.

1) Model ability: Figure 8 shows the performance of all regression models as target noise is injected. There are 3 main groups of models that present similar behaviour among themselves: linear, nonlinear and base models.

The group of linear models (formed by Linear Regression, Bayes and Linear SVR) has the highest normalised errors among the regression models, although they slightly increase as data gets noisier. They also present a slight increase in ability, suggesting that models that are more robust to noise have their abilities increased when compared to noise-sensitive models. Initially, the models with the best performance belong to the group of nonlinear models. Looking at the ability,

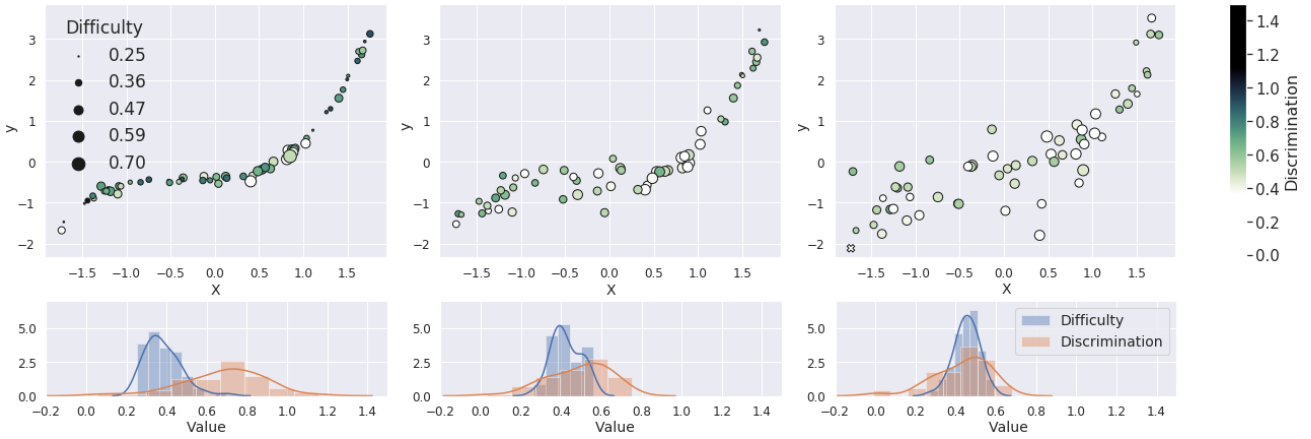


Fig. 6: Difficulty and discrimination (Bigger marker indicates higher difficulty and darker colour indicates higher discrimination.)

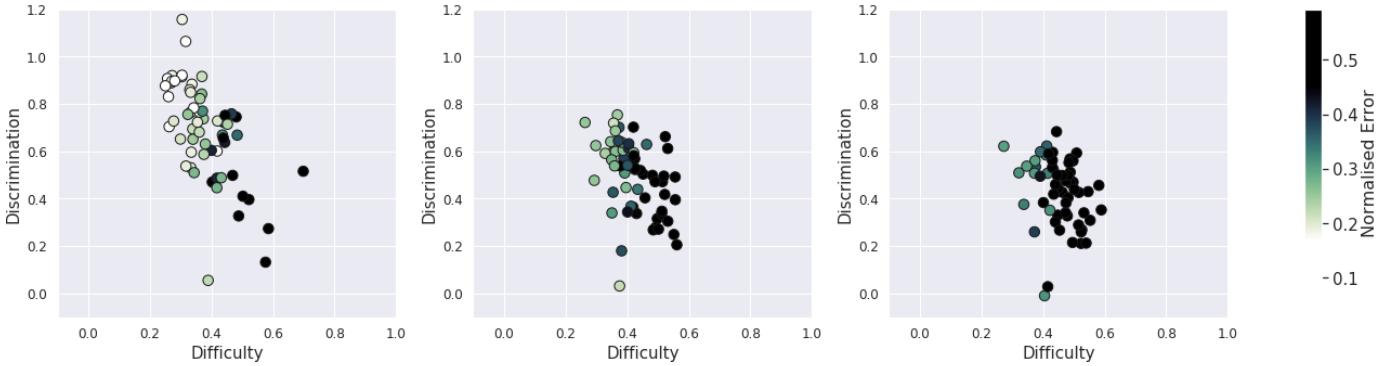


Fig. 7: Difficulty vs Discrimination (Darker colour indicates higher error.)

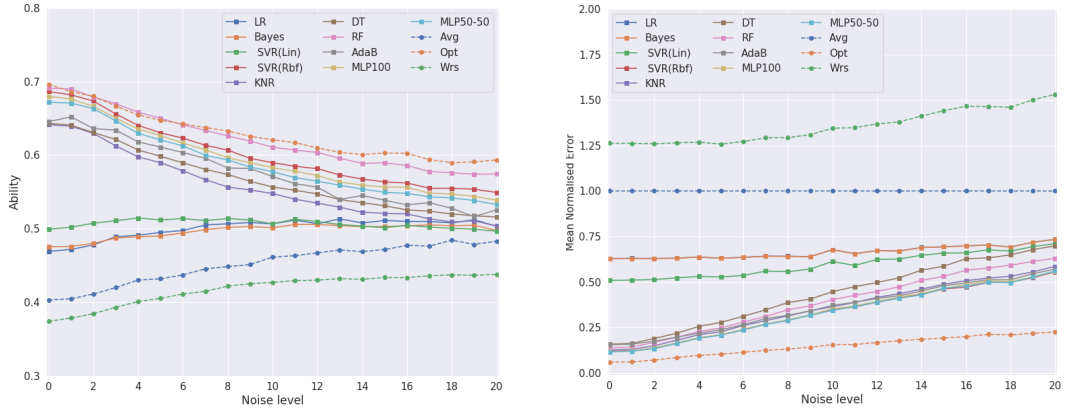


Fig. 8: Evolution in the performance of all regression models along noise injection.

Random Forest and RBF SVR stand out as the best regression models among all. Nonlinear models, however, decline significantly as noise is inserted into the target attribute. This is precisely because nonlinear models have a higher error growth rate as noise is injected.

The group of baseline models (formed by Average, Optimal and Worst) present a different behaviour when compared to the others. The Optimal model tracks the performance of the best model as expected, and since the responses of the best models on average tend to decline, its ability declines as well. Notice that all estimated errors and responses are relative to the Average model. Hence it presents a constant and unitary relative error, and since most regression models produce higher errors as noise is injected, the ability of the Average model increases. Opposite to the Optimal model, the Worst model tracks the performance of the worst model, which can often be the Average model or a linear regression model.

Because models are fitted to non-noisy data, distortions

in the test set caused by noise injection result in greater errors, thus we checked if ability could be a more robust performance measure. For this, we calculated the percentage variation in ability and in Mean Absolute Error (MAE) of each regression model that occurred in a given noise injection step relative to the noise-free test set. Figure 9 illustrates the heat map of the difference between the variation in ability and in Mean Absolute Error (MAE) of each regression model, in percentage values, that occurred in a given noise injection step relative to the initial step (noise free experiment). Negative values (darker green cells) indicate that the variation of ability values is smaller than the error variation. The first two linear models (Linear Regression and Bayesian Ridge), show greater variation in ability than in error in almost all noise injection steps. Nonlinear models (except AdaBoost) show on average negative variation as noise increases (from the 10th step forward). Results show that ability varies significantly less than MAE as noise increases, according to a paired t-test (p-value 0.00004). Thus it is a robust performance metric to use in regression tasks.

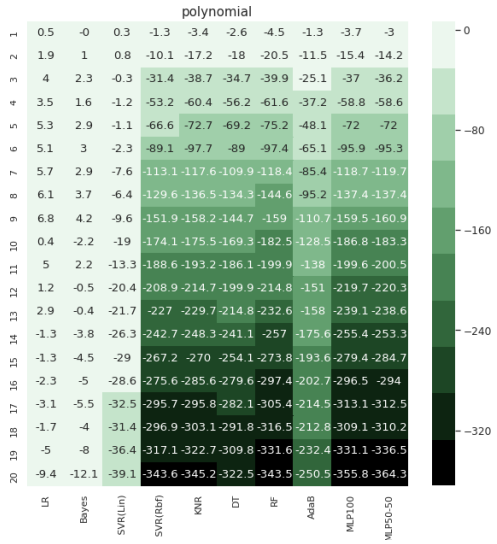


Fig. 9: Heat map of differences between the percentage variation in ability and MAE along noise injection.

V. Real Dataset Analysis

We now analyse the real case dataset *Real Estate*, taken from the UCI public repository [18], in order to check how the Γ-IRT model behaves with real datasets. It has 414 instances and seven attributes, with one being the target variable. The same 13 regression models from Section IV were used as respondents. We trained the models using 80% of all instances and all attributes of the dataset, but for visualisation purposes we use principal component analysis to reduce dimensionality to just one principal component. The remaining instances were used for test purposes.

Figure 10 shows the difficulty and discrimination values of the test set, represented by color intensity and point size, respectively. The data has a descending nonlinear pattern, similar to a hyperbolic curve. The region of greatest difficulty coincides with the region of greatest noise in the data, inside the interval $PC \in [-0.8, -0.2]$. Similarly, less noisy data that more clearly follows the curved pattern of the data has less difficulty.

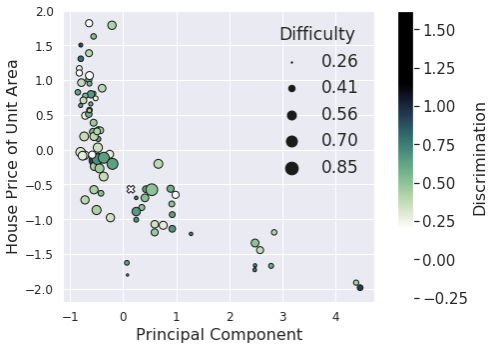


Fig. 10: Difficulty and discrimination in the *Real Estate* dataset (Bigger points indicate higher difficulty and darker colour indicates higher discrimination.)

In Figure 11, instances (c) and (d) have difficulty values of 0.49 and 0.76, and discrimination values of -0.27 and 0.51, respectively. Instance (c) has higher normalised errors than instance (d). Thus, we would expect its difficulty to also be greater. However, discrimination is crucial here: instance (c) is negatively discriminated while instance (d) has positive discrimination, which happened because high-ability regression models had larger errors than weaker ones. In general, instances that do not respond well to high-ability models tend to be difficult with negative discrimination. Such cases were treated as noisy items in [4].

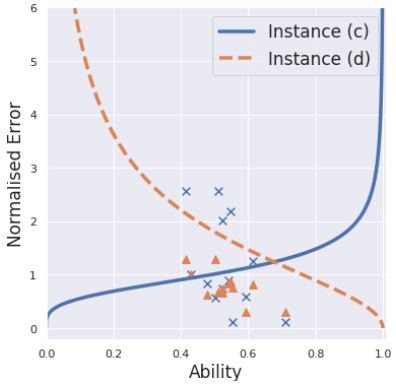


Fig. 11: Representative Item Characteristic Curves from *Real Estate* dataset. Marks are the regression models' responses fit by the ICCs.

Figure 12 shows the relation between Ability and Mean Normalised Errors. These two metrics have "strong" negative correlation, which is also presented in the figures. This is an expected result since models with the lowest regression errors most likely have the highest ability values. However, this is not a rule, as the results suggest that ability takes into account whether models produce higher errors for easy or difficult

instances. For example, MLP100 has higher mean normalised error than DT and Bayes, however its ability is higher than the two models. This is likely due to MLP100 performing better in more difficult instances than DT and Bayes.

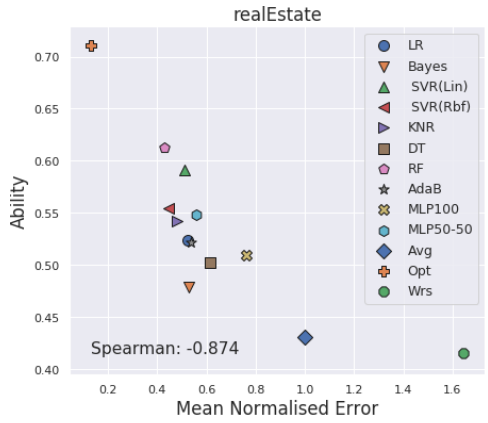


Fig. 12: Mean Normalised Error vs Ability (Spearman's correlation coefficient between both variables is showed in the figure).

VI. Conclusion

In this paper we proposed a new IRT model, called Γ-IRT, developed to fit nonnegative unbounded responses. We applied Γ-IRT in two regression scenarios to analyse the performance of regression models, and also the levels of difficulty and discrimination of data instances located in specific regions in the datasets, with results indicating regions of high and low difficulty. Additionally, noisy data seem to present higher difficulty and lower discrimination when compared to noise-free data. Furthermore, model ability may be used as a robust performance metric as it tracks the normalised error values, but is less affected by noise.

Although initially designed for regression evaluation, the proposed approach can be easily extended to other AI contexts in which models produce continuous responses. Thus our work increases the scope of IRT application to AI evaluation, which is still in its early stage of investigation. Future works may include expanding our experiments to include more regression models and datasets, as well as the analysis of noise injection in the feature attributes and its effects over the item parameters and the ability of regression models. Another possible application of Γ-IRT is feature selection based on difficulty and discrimination of each attribute over the machine learning models.

References

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

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

[3] F. MartÍnez-Plumed, R. B. Prudêncio, A. Martínez-Usó, and J. Hernández-Orallo, "Item response theory in ai: Analysing machine learning classifiers at the instance level," *Artificial Intelligence*, vol. 271, pp. 18 – 42, 2019.

[4] Y. Chen, T. M. Silva Filho, R. B. Prudencio, T. Diethe, and P. Flach, "β^3-irt: A new item response model and its applications," in *Proceedings of Machine Learning Research*, ser. Proceedings of Machine Learning Research, vol. 89. PMLR, 2019, pp. 1013–1021.

[5] F. Martínez-Plumed and J. Hernández-Orallo, "Analysing results from AI benchmarks: Key indicators and how to obtain them," *CoRR*, vol. abs/1811.08186, 2018.

[6] J. P. Lalor, H. Wu, and H. Yu, "Building an evaluation scale using item response theory," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2016, pp. 648–657.

[7] G. Rasch, *Probabilistic Models for Some Intelligence and Attainment Tests*, ser. Studies in mathematical psychology. Danmarks Paedagogiske Institut, 1960.

[8] E. Maris, "Additive and multiplicative models for gamma distributed random variables, and their application as psychometric models for response times," *Psychometrika*, vol. 58, no. 3, pp. 445–469, Sep 1993.

[9] W. J. van der Linden, "A lognormal model for response times on test items," *Journal of Educational and Behavioral Statistics*, vol. 31, no. 2, pp. 181–204, 2006.

[10] F. Samejima, "Homogeneous case of the continuous response model," *Psychometrika*, vol. 38, no. 2, pp. 203–219, 1973.

[11] ——, "Normal ogive model on the continuous response level in the multidimensional latent space," *Psychometrika*, vol. 39, no. 1, pp. 111–121, 1974. [Online]. Available: https://doi.org/10.1007/BF02291580

[12] P. J. Ferrando, "Theoretical and empirical comparisons between two models for continuous item response," *Multivariate Behavioral Research*, vol. 37, no. 4, pp. 521–542, 2002, pMID: 26816326. [Online]. Available: https://doi.org/10.1207/S15327906MBR3704_05

[13] Y. Noel and B. Dauvier, "A beta item response model for continuous bounded responses," *Applied Psychological Measurement*, vol. 31, no. 1, pp. 47–73, 2007.

[14] R. B. Cattell, *Personality and motivation structure and measurement.* Oxford, England: World Book Co., 1957.

[15] C. Spearman, ""general intelligence," objectively determined and measured," *The American Journal of Psychology*, vol. 15, no. 2, pp. 201–292, 1904.

[16] W. Stephenson, "Technique of factor analysis," *Nature*, vol. 136, no. 3434, pp. 297–297, 1935.

[17] C. Ferri, J. Hernández-Orallo, A. Martinez-Usó, and M. J. Ramirez-Quintana, "Identifying dominant models when the noise context is known," 2014.

[18] H. T. K. Yeh, I. C., "Uci machine learning repository," *Applied Soft Computing*, pp. 260–271, aug 2018. [Online]. Available: http://archive.ics.uci.edu/ml