UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

JOÃO VICTOR CAMPOS MORAES

Γ-IRT: An Item Response Theory Model for Evaluating Regression Algorithms

Recife
2021

JOÃO VICTOR CAMPOS MORAES

Γ-IRT: An Item Response Theory Model for Evaluating Regression Algorithms

A M.Sc. Dissertation presented to the Center of Informatics of Federal University of Pernambuco in partial fulfillment of the requirements for the degree of Master of Science in Computer Science.

Concentration Area: Artificial Intelligence
Advisor: Ricardo Bastos Cavalcante Prudêncio
Co-Advisor: Telmo de Menezes e Silva Filho

Recife
2021

Catalogação na fonte
Bibliotecária Monick Raquel Silvestre da S. Portes, CRB4-1217

M828i Moraes, João Victor Campos
Γ-IRT: an item response theory model for evaluating regression algorithms / João Victor Campos Moraes. – 2021.
54 f.: il., fig., tab.

Orientador: Ricardo Bastos Cavalcante Prudêncio.
Dissertação (Mestrado) – Universidade Federal de Pernambuco. CIn, Ciência da Computação, Recife, 2021.
Inclui referências.

1. Inteligência artificial. 2. Aprendizagem de máquina. I. Prudêncio, Ricardo Bastos Cavalcante (orientador). II. Título.

006.31 CDD (23. ed.) UFPE - CCEN 2023-89

João Victor Campos Moraes

"Γ-IRT: An Item Response Theory Model for Evaluating Regression Algorithms"

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de Pernambuco, como requisito parcial para a obtenção do título de Mestre em Ciência da Computação.

Aprovado em: 09/03/2021.

_____

Orientador: Prof. Dr. Ricardo Bastos Cavalcante Prudêncio

BANCA EXAMINADORA

_____

Profa. Dra. Renata Maria Cardoso Rodrigues de Souza
Centro de Informática/ UFPE

_____

Profa. Dra. Giselė Lobo Pappa
Departamento de Ciência da Computação / UFMG

_____

Prof. Dr. Telmo de Menezes e Silva Filho
Departamento de Estatística / UFPB

I dedicate this dissertation to my family and friends.

ACKNOWLEDGEMENTS

Throughout the master's program, I received support and encouragement from all people involved in the writing of this work.

I would first like to thank my supervisor, Dr. Ricardo Prudêncio, for all his patience and for helping me to take the most important steps of my professional and academic careers. Through your guidance, insights and constant feedback, I was able to build all the knowledge I acquired in this research area.

I also thank my co-advisor, Dr. Telmo Silva Filho, for all the hard work and hours spent in modeling and analyzing the experiments. Your great collaboration contributed not only to the writing of this work, but to my personal development of analytical and critical thinking.

Finally, I would like to acknowledge my colleagues, Chaina Oliveira and Jessica Reinaldo, for all their help and advice throughout this journey.

To all the people who are part of CIn UFPE, you contribute to make the center a place of professional and human development. This incredible place has become the breeding ground for my dreams and achievements.

ABSTRACT

Item Response Theory (IRT) is used to measure latent abilities of human respondents based on their responses to items with different difficulty levels. Recently, IRT has been applied to algorithm evaluation in Artificial Inteligence (AI), by treating the algorithms as respondents and the AI tasks as items. The most common models in IRT only deal with dichotomous responses (i.e., a response has to be either correct or incorrect). Hence they are not adequate in application contexts where responses are recorded in a continuous scale. In this dissertation we propose the Γ-IRT model, particularly designed for dealing with positive unbounded responses, which we model using a Gamma distribution, parameterised according to respondent ability and item difficulty and discrimination parameters. The proposed parameterisation results in item characteristic curves with more flexible shapes compared to the traditional logistic curves adopted in IRT. We apply the proposed model to assess regression model abilities, where responses are the absolute errors in test instances. This novel application represents an alternative for evaluating regression performance and for identifying regions in a regression dataset that present different levels of difficulty and discrimination.

Keywords: item response theory; regression tasks; machine learning; evaluation.

RESUMO

Teoria da Resposta ao Item (IRT) é usada para medir habilidades latentes de respondentes humanos com base em suas respostas a itens com diferentes níveis de dificuldade. Recentemente, IRT tem sido aplicada à avaliação de algoritmos de Inteligência Artificial (IA), tratando os algoritmos como respondentes e as tarefas de IA como itens. Os modelos mais comuns em IRT lidam apenas com respostas dicotômicas (ou seja, uma resposta deve ser correta ou incorreta). Portanto, não são adequados em contextos de aplicação onde as respostas são registradas em escala contínua. Nesta dissertação propomos o modelo Γ-IRT, especialmente concebido para lidar com respostas positivas ilimitadas, que modelamos usando uma distribuição Gama, parametrizada de acordo com a habilidade do respondente e parâmetros de dificuldade e discriminação do item. A parametrização proposta resulta em curvas características de itens com formatos mais flexíveis em relação às curvas logísticas tradicionais adotadas em IRT. Aplicamos o modelo proposto para avaliar as habilidades do modelo de regressão, onde as respostas são os erros absolutos nas instâncias de teste. Esta nova aplicação representa uma alternativa para avaliar o desempenho da regressão e para identificar regiões em um conjunto de dados de regressão que apresentam diferentes níveis de dificuldade e discriminação.

Palavras-chave: teoria da resposta ao item; tarefas de regressão; aprendizagem de máquina; avaliação.

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

AI	Artificial Inteligence
ANN	Artificial Neural Networks
ASR	Automatic Speech Recognition
CAPES	Coordenação de Aperfeiçoamento de Pessoal de Nível Superior
DT	Decision Tree
ICC	Item Characteristic Curve
IJCNN	International Joint Conference on Neural Networks
IRT	Item Response Theory
KNN	K-Nearest Neighbors
KNR	K-Neighbors Regression
LDA	Linear Discriminant Analysis
LR	Linear Regression
MAE	Mean Absolute Error
ML	Machine Learning
MLP	Multilayer Perceptron
NB	Naive Bayes
NLP	Natural Language Processing
PC	Principal Component
PCM	Partial Credit Model
PCR	Principal Component Regression
PLS	Partial Least Squares
QDA	Quadratic Discriminant Analysis
RF	Random Forest
SVM	Support Vector Machines
SVR	Support Vector Regression

LIST OF SYMBOLS

$\mathscr{B}$	Beta distribution
$\mathbb{E}$	Expectation value
Γ	Gamma distribution
α, β	Beta distribution parameters
$\bar{e}$	Normalised error
δ	Difficulty parameter
ε	Gaussian noise
$\in$	Set membership
λ	Intensity parameter
$\mathscr{N}$	Gaussian distribution
σ	Standard deviation
θ	Respondent ability
a	Discrimination parameter
c	Guessing parameter
e	Absolute error

CONTENTS

1 INTRODUCTION

This introduction chapter shows the context in which the work is inserted. Furthermore, this chapter shows the main contributions obtained with the elaboration of the work.

1.1 PROBLEM AND MOTIVATION

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 (EMBRETSON & REISE, 2013). 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 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 parameters of items 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. For instance, IRT was adopted in Machine Learning (ML) classification (MARTÍNEZ-PLUMED *et al.*, 2016, 2019), in which items correspond to instances in a dataset, respondents are classifiers and the dichotomous responses, also referred to as binary responses, are right or wrong classification outcomes collected in a cross-validation experiment. In another application of IRT for ML classification, CHEN *et al.* (2019) 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 (MARTÍNEZ-PLUMED & HERNÁNDEZ-ORALLO, 2018) and Natural Language Processing (NLP) (LALOR *et al.*, 2016).

Despite useful insights, previous works are limited to the application of IRT for binary (right or wrong predictions) and bounded responses (class probabilities). The IRT models adopted in previous work 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 positive 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.

1.2 DISSERTATION PROPOSAL

We propose a new IRT model which focuses on positive unbounded responses, which have not been adequately treated in the IRT literature. 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 application of IRT to AI evaluation, which is still in its early stage of investigation.

We apply the proposed model in two case studies. First, we use Γ-IRT to fit absolute errors produced in regression tasks. Second, 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 inside the dataset and we propose ability as a complementary measure to evaluate regression models.

Our contributions can be summarised as follows:

1. We propose Γ-IRT, a new IRT model which focuses on positive unbounded responses, which have not been adequately treated in the IRT literature;
2. We use Γ-IRT to evaluate regression algorithms, which is a novel application in literature;
3. We use Γ-IRT to identify difficult and discriminative data instances on regression tasks;

1.3 STRUCTURE

The dissertation is organised in six chapters, as follows.

- Chapter 1 - Introduction;
- Chapter 2 - Literature Review: a brief history and related work on IRT;
- Chapter 3 - The Γ-IRT Model: description of the Γ-IRT model;
- Chapter 4 - Experiments with Regression Models: application of Γ-IRT to analyse regression models and datasets;
- Chapter 5 - Experiments with Noise: analysis on the influence of data noise to the Γ-IRT model;
- Chapter 6 - Conclusion: final remarks and discussion about future works.

2 LITERATURE REVIEW

In psychometrics, IRT consists of a family of mathematical and statistical models used in the design, construction and evaluation of educational and psychological tests (EMBRETSON & REISE, 2013). Psychometrists have advanced this new measurement system to address several shortcomings in common measurement practices at the time.

IRT has emerged as an alternative way of evaluating agents, or respondents, who respond to certain items within a specific context. For example, respondents can be students who answer exam questions, or algorithms which respond to computational tasks. Traditional methods of evaluation generally capture the average or total correct items in a test. Therefore, the respondent's skill is not fully separable from the test characteristics. A major limitation of traditional methods is that they do not take into account the response of different respondents to items with different levels of difficulty. For example, a student gets difficult questions right, but misses easy questions. Is this a good student? In the context of machine learning, should a satisfactory classification algorithm classify more easy or difficult instances? IRT takes into account the difficulties of these tasks. Therefore, respondents are assessed based on a latent ability, which is inferred from responses to more or less difficult items.

The beginning of IRT is often addressed to Frederick M. Lord and Melvin R. Novick, which was a milestone in psychometrics. The book was well connected to leading and emerging scholars in psychometric methods at the time (LORD & NOVICK, 1968). Another line of research of IRT can be attributed to Georg Rasch, who developed a family of IRT models that were applied to perform reading measurements and test development. He was interested in discovering the properties of measurement models and observed that all parameters, from both item and respondent, were completely separable in his models, a property he developed as specific objectivity (RASCH, 1960). Erling B. Andersen, who was Rasch's student, consequently developed effective methods of estimation for the respondent and item parameters in Rasch's models (ANDERSEN, 1973). IRT was one of the dominant subjects among measurement scholars in the 1980's.

This chapter is organized as follows. Section 2.1 presents the traditional IRT models used in psychometrics. Section 2.2 presents IRT models within the context of AI, which is the focus of this dissertation. AI areas such as machine learning, speech recognition and deep learning are shown in the second section.

2.1 TRADITIONAL IRT

Item Response Theory is based on two postulates: the performance of a respondent to an item can be measured and predicted by a latent ability or set of traits; and the relationship between a respondent's performance to an item and the latent characteristics behind the item's performance can be represented by a monotonically increasing function called the Item Characteristic Curve

(HAMBLETON *et al.*, 1991). This curve specifies the growth in the probability of a correct response to an item along the respondent's latent ability. There are several IRT models that differ in mathematical formulation, either in the shape of the curve or in the number of item parameters. The main models are described in the next subsections.

2.1.1 Dichotomous models

The most popular one-dimensional IRT models are the one, two and three item parameters logistic models. These models fit into dichotomous response modeling problems, which means responses can either be correct or incorrect (binary responses). For this subsection, it is assumed that the respondents are examinees who respond to an exam, and the questions that make up the exam are the items..

2.1.1.1 One-parameter logistic model (1PL)

The 1PL model, or Rasch model, assumes that there is only one feature of the item, referred to as difficulty, which influences the respondent's performance. The ICCs produced by the model are given by Equation 2.1.

$$\mathbb{E}[p_{ij}|\theta_i,\delta_j] = \frac{e^{(\theta_i-\delta_j)}}{1+e^{(\theta_i-\delta_j)}} \tag{2.1}$$

where p_{ij} is the probability that the respondent i, with ability θ_i, will get a correct answer in item j, with difficulty δ_j. An equation produces an S-shaped curve with values between 0 and 1 along the θ scale.

The only item parameter in this model (δ_j) represents the ability value in which the probability of obtaining a correct answer is 0.5. The difficulty is also known as the location parameter and indicates the displacement of the curve along the ability scale. The higher the difficulty of an item, the higher the ability required to obtain a correct answer.

Figure 1 illustrates three examples of ICCs with different difficulty values. As difficulty increase, the curves shift to the right, therefore, higher ability values are needed to obtain the same probability of correct response.

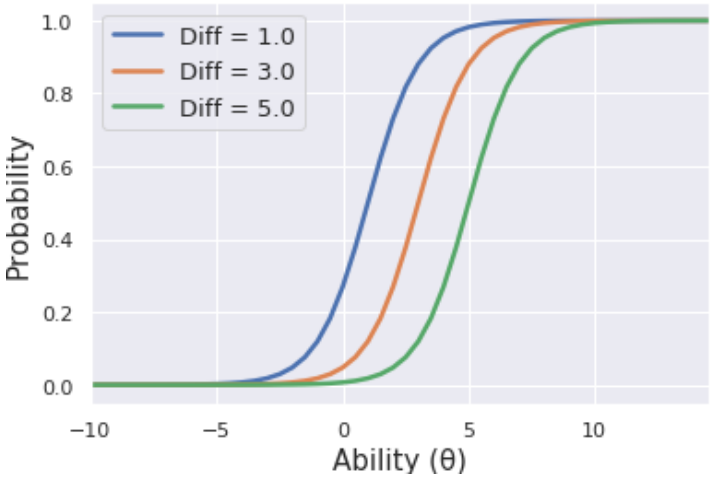
2.1.1.2 Two-parameter logistic model (2PL)

BIRNBAUM (1968) developed the ICC for the 2PL model, which is described next (Equation 2.2).

$$\mathbb{E}[p_{ij}|\theta_i,\delta_j,a_j] = \frac{e^{a_j(\theta_i-\delta_j)}}{1+e^{a_j(\theta_i-\delta_j)}} \tag{2.2}$$

where p_{ij} and δ_j are defined as in Equation 2.1. The two-parameter model is practically the same as the one-parameter model, except for the presence of the new parameter a_j, commonly

Figure 1 – Examples of one-parameter ICCs.

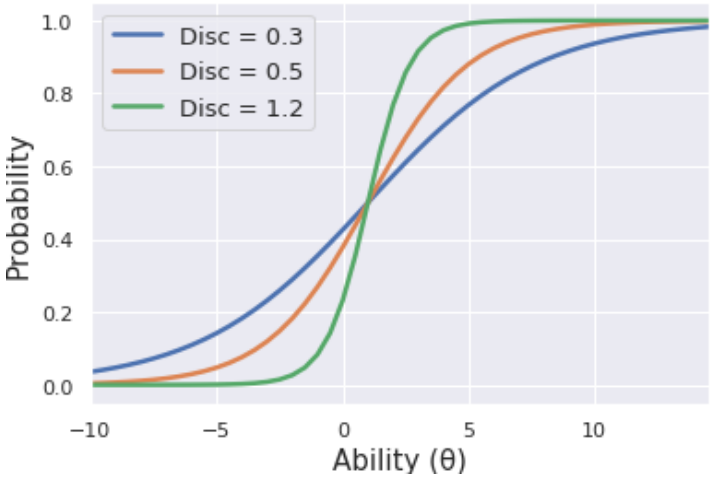


Source: Author.

referred to as item discrimination parameter. The discrimination parameter represents the slope of the ICC, that is, it measures the rate of change of the response along the ability. For ICCs with higher discrimination, the probability of a correct answer is more sensitive to the respondent's ability and grows at a higher rate when compared to a low discrimination ICC.

Figure 2 illustrates three examples of ICCs with different discrimination values and fixed difficulty value of 1. The difference between the curves lies in the slope of the curve. The less discriminating curve grows at a lower rate of change in ability. Similarly, the curve with higher discrimination grows at a higher rate as the ability increases.

Figure 2 – Examples of two-parameter ICCs (fixing the difficulty parameter).



Source: Author.

2.1.1.3 Three-parameter logistic model (3PL)

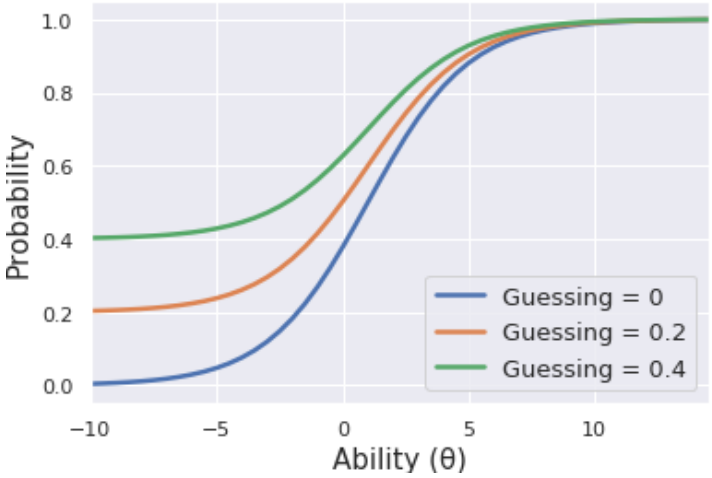
BIRNBAUM (1968) also proposed a model with a third parameter, in addition to the two mentioned in the previous subsections. The new parameter, referred to as the guessing parameter,

take into account non-zero responses for low ability respondents in tasks with multiple-choice items. The ICC is given by the following Equation:

$$\mathbb{E}[p_{ij}|\theta_i, \delta_j, a_j, c_j] = c_j + (1 - c_j)\frac{1}{1 + e^{-a_j(\theta - \delta_j)}} \tag{2.3}$$

where c_j is the guessing parameter of item j. Figure 3 illustrates three ICCs with different guessing parameters. Difficulty and discrimination parameters are the same for the three ICCs, with values of 1 and 0.5, respectively. The curves differ from each other in the lower asymptote value. When $c = 0$ it is assumed that models with low abilities (close to zero) can have responses close to zero.

Figure 3 – Examples of three-parameter ICCs (fixing the difficulty and discrimination parameters).



Source: Author.

2.1.2 Toy example

To better understand IRT in practice, an example of a 2PL model is explained next. In a hypothetical test, there are six questions that a group of ten students need to answer. In the context of IRT, each question represents an item and students are the respondents. The ltm package in R was used in the example to calculate the items' parameters and the abilities of the respondents (RIZOPOULOS, 2006).

The responses of the ten students to the six questions in the test are shown below in Table 1. It is important to note that the response value of 1 indicates correctness and, similarly, response 0 indicates that the item was answered incorrectly. At first glance, it is possible to guess which respondents performed better and worse on both tests. For example, Student 1 got all questions right, hence it is expected that this student's ability stands out better than the other students. On the other hand, Student 10 got all questions wrong, so it is expected to have a low ability. Analogously, respondent 2 almost got all items correct, missing only Question 6, which only three students responded correctly. Despite the incorrect response, it is expected that such a

respondent will obtain a high ability value. For other respondents, it is more difficult to compare performance with each other.

If the students were evaluated using the classic approach, just by counting the number of correct answers, we would have the ranking shown in the last column of Table 1. Students who have the same number of correct answers, in theory, would have the same grade on the test. However, it is observed that the six test questions do not have equivalent "weights" for these students. In Question 1 (Q1), for example, nine out of ten students got it right. In Question 5 (Q5), three out of ten students got it right, the same students who are among the best evaluated. This raises the need for question difficulty to be taken into account in student assessment. A possible solution would be for the evaluator to give different weights to the questions. However, the evaluation would be biased by the evaluator's opinion of what an easy or difficult question is. The students' reality and performance may not reflect the evaluator's expectations. IRT manages to overcome this limitation in the classic approach.

Table 1 – Students' responses to a test composed by 6 questions.

Student	Q1	Q2	Q3	Q4	Q5	Q6	Count
1	1	1	1	1	1	1	6
2	1	1	1	1	1	0	5
3	1	1	1	1	0	1	5
4	1	1	1	0	1	0	4
5	1	1	1	0	0	0	3
6	1	1	0	0	0	1	3
7	1	1	0	0	0	0	2
8	1	0	0	1	0	0	2
9	1	0	0	0	0	0	1
10	0	0	0	0	0	0	0

Source: (RIZOPOULOS, 2006)

Table 2 shows the difficulty and discrimination parameters of the 2PL IRT model of the six test questions. As expected, Q1 has the least difficult question. Q5 and Q6, on the other hand, have the highest difficulty values. Although Q3 has relatively low difficulty, its discrimination is the greatest, as it can best separate a good student from a bad or regular student. Similarly, Q6 has the second lowest discrimination, despite the high difficulty, due to the fact that a regular student got it right. However, this may suggest that the student correctly guessed the question. This will be better explained when we analyse the students' abilities.

Table 3 shows the abilities of the ten students. Ability takes into account the difficulty and discrimination of each question. As expected, Student 1 has the highest ability and Student 10 has the lowest ability. When we compare Students 3 and 4, something interesting happens. Student 3 got 5 questions right and Student 4 got 4 questions, but the ability of Student 3 is lower than Student 4. This happens because it is less coherent to correctly answer the most difficult question in the test (Q6) and to miss a question with almost half its difficulty (Q5). This

Table 2 – Items' parameters (difficulty and discrimination).

Question	Difficulty	Discrimination
Q1	-1.357	2.6941
Q2	-0.658	1.9928
Q3	0.027	3.3679
Q4	0.663	0.1079
Q5	0.672	2.0196
Q6	1.270	0.5920

Source: (RIZOPOULOS, 2006)

response pattern may suggest that Student 3 correctly guessed Q6. This is more evident with the ability of Students 5 and 6. The two students have exactly the same number of correct answers, however, what differentiates them are the different questions. Student 5 answered only the three easiest questions in the test (Q1, Q2 and Q3), while Student 6 answered the two easiest and the most difficult question in the test (Q1, Q2 and Q6). The ability of Student 6 strongly suggests that the response to Q6 was a guess. The application of IRT in this context not only solves the limitations of the classical approach to evaluation, but also provides insight into the performance of respondents.

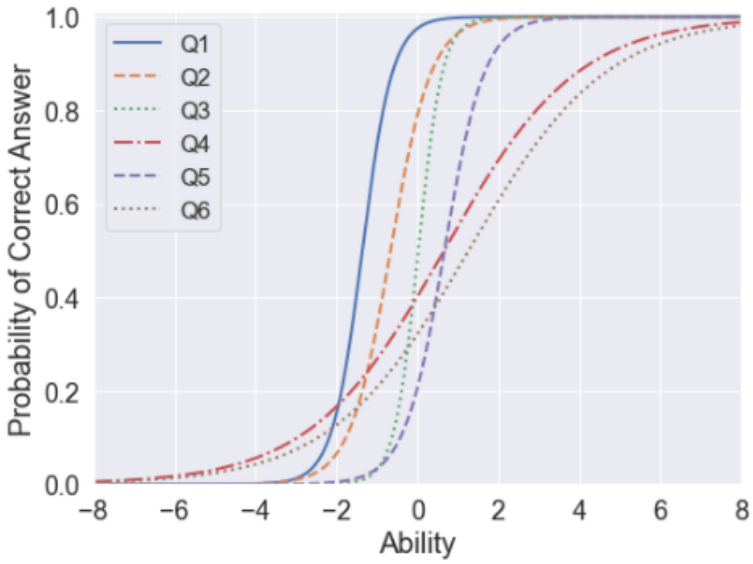
Table 3 – Respondents' abilities.

Student	Ability
1	0.993
2	0.836
3	0.507
4	0.789
5	0.139
6	-0.098
7	-0.369
8	-0.782
9	-0.838
10	-1.468

Source: (RIZOPOULOS, 2006)

Figure 4 shows the ICC of the six test questions. The curves are fitted to the students' responses to each question. It is noted that as the difficulty increases, the curves shift to the right, requiring higher ability values to increase the probability of success (e.g. Q1 and Q6). As for discrimination, as its value increases the slope of the curve also increases. This translates into the question's ability to separate Students with different abilities. The probability of success increases considerably along the ability (e.g. Q3 and Q5).

Figure 4 – ICC of questions from hypothetical test.



Source: Author.

2.1.3 Polytomous models

The models presented in the previous subsection fit only tasks in which the response is binary-valued. However, many other tasks require different response modeling rather than dichotomous models, where the response format does not fit the traditional approach of "true versus false" or "correct versus incorrect". Many measurement instruments used in psychology extract categorical and ordinal measurements, with more than two categories. Polytomous IRT models were developed in the need to model item-response data with multiple categories and to represent the non-linear relationship between the respondent's ability and the probability of an item belonging to a certain category.

SAMEJIMA (1969) developed the Graded-Response Model, which is a generalization of the 2PL model, fits ordered categorical responses (e.g. Likert rating scales). MASTERS (1982) proposed the Partial Credit model (PCM) to model items that have discrete levels of correctness and that are assigned partial credits as the respondents improves their response. Therefore, this model is appropriate to fit item responses where partially correct answers can be achieved (e.g. exam questions).

2.1.4 Continuous response models

Although binary and polytomous models are widely used, loss of information in responses may occur when summarizing the result of a task in well-defined categories. Several other application contexts require that the response is measured on a continuous scale.

In Psychometrics, nonnegative continuous responses have been previously analysed in the context of student reading speed RASCH (1960); MARIS (1993); LINDEN (2006), 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 (RASCH, 1960) modelled t_{ij} with a gamma density given by:

$$\mathbb{E}[t_{ij}|\theta_i,\delta_j] \equiv \frac{(\theta_i/\delta_j)^m}{\Gamma(m)} t_{ij}^{(m-1)} e^{-\theta_i t_{ij}/\delta_j}, \tag{2.4}$$

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.

Although these models are designed for nonnegative responses, we focus on a different context, therefore their assumptions do not apply here. Previous works that tackled the problem of IRT models for continuous responses mainly focused on responses with bounded support. CHEN *et al.* (2019) introduced the β^3-IRT model, which can generate a rich family of ICCs for responses in the $[0,1]$ range. Equation (2.5) below gives the model definition, where M is the number of respondents, N is the number of items and p_{ij} is the observed response of respondent i to item j, which is drawn from a Beta distribution.

$$\begin{aligned}
&p_{ij} \sim \mathscr{B}(\alpha_{ij},\beta_{ij}),\\
&\alpha_{ij} = \mathscr{F}_\alpha(\boldsymbol{\theta}_i,\boldsymbol{\delta}_j,\mathbf{a}_j) = \left(\frac{\boldsymbol{\theta}_i}{\boldsymbol{\delta}_j}\right)^{\mathbf{a}_j},\\
&\beta_{ij} = \mathscr{F}_\beta(\boldsymbol{\theta}_i,\boldsymbol{\delta}_j,\mathbf{a}_j) = \left(\frac{1-\boldsymbol{\theta}_i}{1-\boldsymbol{\delta}_j}\right)^{\mathbf{a}_j},\\
&\boldsymbol{\theta}_i \sim \mathscr{B}(1,1),\ \boldsymbol{\delta}_j \sim \mathscr{B}(1,1),\ \mathbf{a}_j \sim \mathscr{N}(1,\sigma_0^2)
\end{aligned} \tag{2.5}$$

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 $\mathscr{B}(\alpha_{ij},\beta_{ij})$ and assuming the form given by Equation (2.6).

$$\mathbb{E}[p_{ij}|\boldsymbol{\theta}_i,\boldsymbol{\delta}_j,\mathbf{a}_j] = \frac{\alpha_{ij}}{\alpha_{ij}+\beta_{ij}} = \frac{1}{1+\left(\frac{\boldsymbol{\delta}_j}{1-\boldsymbol{\delta}_j}\right)^{\mathbf{a}_j}\left(\frac{\boldsymbol{\theta}_i}{1-\boldsymbol{\theta}_i}\right)^{-\mathbf{a}_j}} \tag{2.6}$$

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.

2.2 IRT IN ARTIFICIAL INTELLIGENCE

A recent application area of IRT is Artificial Intelligence. According to FLACH (2019), a highly promising opportunity in evaluating machine learning algorithms involve the use of latent-variable models. Instead of human respondents, algorithms are evaluated based on responses to different items or tasks. More specifically, machine learning fits well into the IRT approach: the classification models are the respondents with ability values and each instance of the dataset is an item with a particular difficulty. This way, a joint assessment of the model ability and instance hardness is performed.

2.2.1 Supervised machine learning

A contribution in (PRUDÊNCIO *et al.*, 2015) work was to analyse instance hardness in machine learning tasks using IRT. The experiments consisted of training different Random Forests (RF), varying the number of trees. In the context of IRT, each RF classifier represented a respondent and each instance of the dataset the item. The binary response indicated the right or wrong classification of a RF for a given instance. For the case study presented in the paper, the Heart-Statlog dataset was used and IRT models were generated for each instance and the ability measured for each RF classifier. Results suggested different levels of discrimination among data instances and possible presence of noise in the dataset. Another main point was to use the ability as an alternative way to decide which classifier is better than the other. Item characteristic curves were generated for the classifiers to model the probability of a correct responses given the instance hardness levels. Such curves can be used to select and reuse models for different distributions and levels of instance hardness in a problem.

MARTÍNEZ-PLUMED *et al.* (2016) carried out a series of experiments with different datasets and classification models. To obtain a large population of classifiers, 128 classifiers were generated by varying the parameters of 15 different families: Decision Trees (DT), rule-based methods, Linear Discriminant Analysis (LDA), Bayesian, Artificial Neural Networks (ANN), Support Vector Machines (SVM), boosting, bagging, stacking, Random Forests, K-Nearest Neighbors (KNN), Partial Least Squares (PLS), Principal Component Regression (PCR) and logistic and multinomial regression. The data used in the experiments were the Cassini toy dataset and 8 real datasets extracted from the UCI repository. Each instance of the data sets is represented by 3PL models, whose difficulty, guessing and discrimination parameters are obtained from the responses of all different models, which have different ability values. It was found that difficulty can increase in case of borderline instances, higher number of a different neighbors and outlier presence. On the other hand, discrimination can be very useful to identify noise in the datasets, and also to analyse model overfitting. Ability is an interesting measure

that portrays a different information than accuracy. IRT evaluates classifiers in terms of the other classifiers that are included in the pool of classifiers. This relativeness is a good property, especially if a range of diverse classifiers are in the pool.

CHEN *et al.* (2019) also applied β^3-IRT in the machine learning context. Again, respondents were represented by classifiers and items by instances of datasets. Since β^3-IRT models continuous responses, the response was represented by the probability of correctly classifying an instance to a particular class. Two synthetic binary classification datasets, MOONS and CLUSTERS, available in scikit-learn, were used in the experiments. Two classes from the MNIST dataset (3 vs 5) were also chosen since they are similar and contain difficult instances. Noise was injected in the test set by flipping the label for 20% of randomly chosen data instances. Also, twelve classifiers were built in the experiments: Naive Bayes (NB), Multilayer Perceptron (MLP), AdaBoost, Logistic Regression, K-Nearest Neighbors, Linear Discriminant Analysis, Quadratic Discriminant Analysis (QDA), Decision Tree, Random Forest and three synthetic classifiers. Results showed that item parameters provided useful insights for difficult or noisy instances. Also, latent ability was useful to evaluate classifiers on an instance-wise basis in terms of probability estimation.

To solve the problem of insufficient and over-fitting data, extra training data can be generated artificially through human learning. Knowing that the process of labeling data manually is prone to human errors, LI *et al.* (2016) generated two machine learning algorithms to identify erroneous data instances in linear regression. IRT was used to model the distribution of human errors in labeling, so it was possible to reconstruct a training set with more sparse errors. Simulations showed that the two algorithms are effective in resolving the insufficient training and human labeling error problems.

MARTÍNEZ-PLUMED & HERNÁNDEZ-ORALLO (2017) analysed the behaviour of around 40 learning techniques for one of the most popular general purpose AI benchmarks in the recent years: the Arcade Learning Environment (ALE), based on the Atari 2600 games. Martinez-Plumed used item response theory, and logistic models in particular, to create item characteristic curves to determine which games in the benchmark are more difficult but also more discriminating.

CHEN & AHN (2020) proposed a novel probabilistic framework to improve the accuracy of a weighted majority voting algorithm. In order to assign higher weights to the classifiers which can correctly classify hard-to-classify instances, they built the IRT framework to evaluate the samples' difficulty and classifiers' ability simultaneously. To explain the models, they illustrated how the IRT ensemble model constructs the classifying boundary. They also compared their performance with other widely used methods and show that the model performed well on 19 datasets.

KANDANAARACHCHI, S. & SMITH-MILES (2020) built an IRT based framework for evaluating a portfolio of algorithms and extract characteristics that describe different aspects of algorithm performance. They evaluated 10 classification algorithms: Naïve Bayes, Linear

Discriminant Analysis, Quadratic Discriminant Analysis, Classification and Regression Trees, J48 decision tree, k-Nearest Neighbors, Support Vector Machines with linear, polynomial and radial basis kernels and Random Forests. They also used 235 datasets from UCI and OpenML repositories. Using polytomous IRT models, Kandanaarachchi introduced measures for quantifying the stability, effectiveness and the anomalous nature of algorithms. The framework was used on 5 diverse algorithm portfolios, demonstrating the applicability of this method as an algorithm evaluation tool.

2.2.2 Speech recognition

OLIVEIRA *et al.* (2020) proposed the evaluation of speech synthesizers using IRT models, in which an item is a sentence to be synthesized and a respondent is a speaker. Four speech synthesizers were used in the experiments: Amazon Polly, Google Text to Speech API, IBM Watson Text to Speech and the Microsoft Azure Text to Speech. Each service generated speeches by adopting different speakers, each one associated to a different voice type, language, accents and genre. Each response is the transcription accuracy observed when a given sentence and speaker are adopted for testing the Automatic Speech Recognition (ASR). A total number of 62 speakers were evaluated (respondents) along 12 different sentences (items). Hence, the ability of the synthesis services was estimated for each speaker to produce audio test files that can be well recognized by the ASR system. They found that IRT can identify sentences with different levels of difficulty and discrimination power between good and poor synthetic speakers.

2.2.3 Natural Language Processing (NLP)

Recently, IRT has also been applied to problems with Deep Learning models. LALOR *et al.* (2016) introduced the idea of applying IRT evaluation to NLP tasks. They built a set of scales using IRT and evaluated a single LSTM neural network to demonstrate the effectiveness of the evaluation.

2.3 FINAL CONSIDERATIONS

In this section, we reviewed the literature on IRT and its various applications. In addition to the IRT models in the context of psychometrics and the assessment of human respondents, the recent IRT application in AI evaluation was also reviewed. AI applications includes the use of IRT in speech recognition and NLP, but greater emphasis is given to machine learning, which is within the scope of this work. Since most IRT models are for dichotomous or limited responses, there were no models adaptable to regression tasks, which is also a field of supervised learning. Given this motivation, we present in the next chapter the Γ-IRT model, which focuses on positive unbounded responses. Hence, an appropriate model for evaluating regression tasks can be analysed in depth.

3 THE Γ-IRT MODEL

We now propose Γ-IRT to model unbounded nonnegative responses, such as students' answers to open-ended questions or the absolute values of errors coming out of a regression model. To the best of our knowledge, the task of fitting IRT models to nonnegative continuous responses, such as errors associated to regression models, is still an open problem.

3.1 FORMULATION

The central idea of Γ-IRT is to model continuous errors using a Gamma distribution, 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, drawn from a Gamma distribution:

$$\begin{aligned} & e_{ij} \sim \Gamma(\alpha_{ij}, \beta_{ij}), \\ & \alpha_{ij} = \mathscr{F}_{\alpha}(\boldsymbol{\theta}_i, \boldsymbol{\delta}_j, \mathbf{a}_j, c_j) = c_j \left(\frac{\boldsymbol{\delta}_j}{\boldsymbol{\theta}_i} \right)^{\mathbf{a}_j}, \\ & \beta_{ij} = \mathscr{F}_{\beta}(\boldsymbol{\theta}_i, \boldsymbol{\delta}_j, \mathbf{a}_j) = \left(\frac{1-\boldsymbol{\delta}_j}{1-\boldsymbol{\theta}_i} \right)^{\mathbf{a}_j}, \\ & \boldsymbol{\theta}_i \sim \mathscr{B}(1,1),\ \boldsymbol{\delta}_j \sim \mathscr{B}(1,1),\ \mathbf{a}_j \sim \mathscr{N}(1, \sigma_0^2). \end{aligned} \tag{3.1}$$

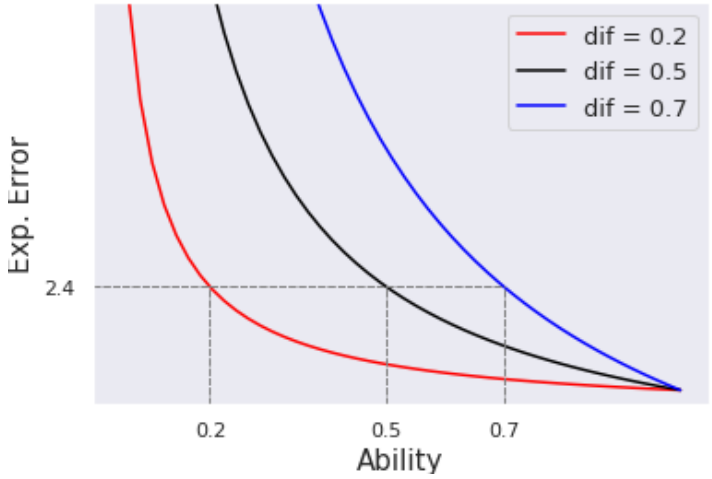
In the model above, $\boldsymbol{\delta}_j \in (0;1)$ is the difficulty parameter of item j, $\mathbf{a}_j$ is the discrimination parameter and $c_j > 0$ is the guessing parameter. For respondents, $\boldsymbol{\theta}_i \in (0;1)$ is the ability of respondent i. In this model, the ICC is the expectation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

$$\mathbb{E}[e_{ij} | \boldsymbol{\theta}_i, \boldsymbol{\delta}_j, \mathbf{a}_j, c_j] = \frac{\alpha_{ij}}{\beta_{ij}} = c_j \left(\frac{\boldsymbol{\delta}_j}{1-\boldsymbol{\delta}_j} \right)^{\mathbf{a}_j} \left(\frac{\boldsymbol{\theta}_i}{1-\boldsymbol{\theta}_i} \right)^{-\mathbf{a}_j} \tag{3.2}$$

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 $\mathbb{E}[e_{ij}] = c_j$.

Figure 5 – Examples of IRTs for different values of difficulty. In all cases, $c_j = 2.4$ and $a_j = 1$.



Source: Author.

3.1.1 Guessing Parameter

c_j can be set as the expected error obtained by a random regression model, i.e. $c_j = \mathbb{E}[e_{Rj}] = |y_j - \mathbb{E}[y]|$. 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 $\boldsymbol{\delta}_j = 0.5$.

3.1.2 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 $\boldsymbol{\theta}_i = 0.5$.

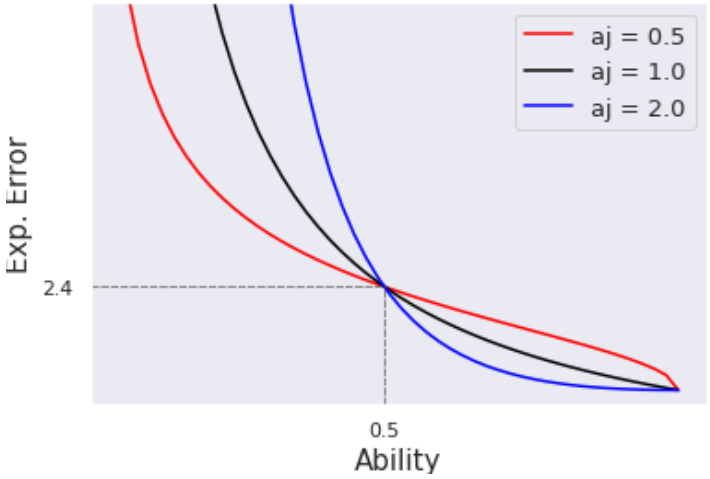
In the first case (easy items), even respondents with low ability will have errors lower than the guessing error. In the second case (difficult items) in turn, only respondents with high ability will outperform the guessing error.

See Figure 5 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.

3.1.3 Discrimination Parameter

a_j characterises the slope of the curve at the difficulty level. Figure 6 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 $a_j = 0.5$,

Figure 6 – Examples of IRTs for different values of discrimination. In all cases, $c_j = 2.4$ and $\boldsymbol{\delta}_j = 0.5$.



Source: Author.

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.

3.2 NORMALISED ERRORS

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

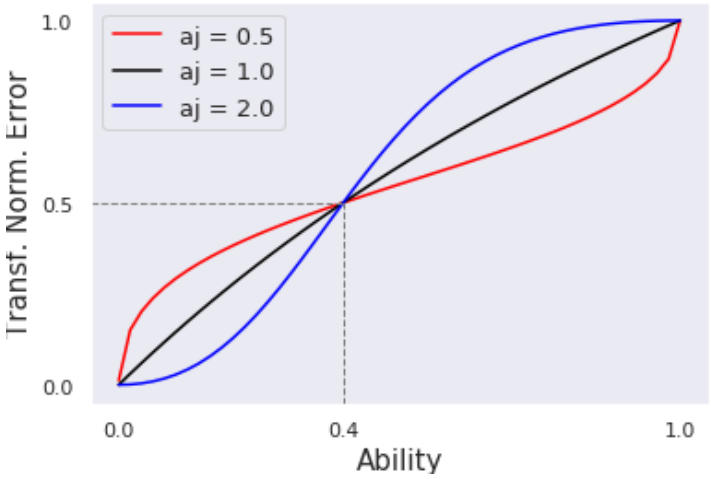
$$\begin{aligned} \bar{e}_{ij} &= \frac{e_{ij}}{c_j} \sim \Gamma(\alpha_{ij}, \beta_{ij} c_j) \\ \alpha_{ij} &= c_j \left(\frac{\boldsymbol{\delta}_j}{\boldsymbol{\theta}_i} \right)^{\mathbf{a}_j}, \; \beta_{ij} c_j = \left(\frac{1-\boldsymbol{\delta}_j}{1-\boldsymbol{\theta}_i} \right)^{\mathbf{a}_j} \end{aligned} \tag{3.3}$$

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:

$$\mathbb{E}[\bar{e}_{ij} | \boldsymbol{\theta}_i, \boldsymbol{\delta}_j, \mathbf{a}_j, c_j] = \frac{\alpha_{ij}}{\beta_{ij} c_j} = \left(\frac{\boldsymbol{\delta}_j}{1-\boldsymbol{\delta}_j} \right)^{\mathbf{a}_j} \left(\frac{\boldsymbol{\theta}_i}{1-\boldsymbol{\theta}_i} \right)^{-\mathbf{a}_j} \tag{3.4}$$

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.

Figure 7 – Examples of β^3-IRT for regression. All curves were produced by setting $\boldsymbol{\delta} = 0.4$.



Source: Author.

3.3 RELATION TO β^3-IRT

The following transformation of normalised errors produces a β^3-IRT curve:

$$\frac{1}{1+\mathbb{E}[\bar{e}_{ij}]} \tag{3.5}$$

Figure 7 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.

NOTE: For estimation we can transform the normalised errors using $\frac{1}{1+\bar{e}_{ij}}$ and produce a β^3-IRT curve. Then, we can transform this ICC back into a Γ-IRT curve using the inverse of this transformation.

3.4 DISCRETE ERROR COUNTS

The model described above is directly applied to continuous nonnegative responses, but its formulation can also be applied to discrete nonnegative responses, such as the number of errors made by students while answering open-ended questions. Let d_{ij} be the error count of the i-th student's answer to the j-th question on a test. These error counts follow a Poisson distribution $d_{ij} \sim Poisson(\lambda_{ij})$. Since the conjugate prior of the rate parameter of a Poisson distribution is the Gamma distribution, we have:

$$\lambda_{ij} \sim \Gamma(\alpha_{ij}, \beta_{ij} c_j) \tag{3.6}$$

$$\mathbb{E}[\lambda_{ij} | \boldsymbol{\theta}_i, \boldsymbol{\delta}_j, \mathbf{a}_j, c_j] = \frac{\alpha_{ij}}{\beta_{ij} c_j} = \left(\frac{\boldsymbol{\delta}_j}{1 - \boldsymbol{\delta}_j} \right)^{\mathbf{a}_j} \left(\frac{\boldsymbol{\theta}_i}{1 - \boldsymbol{\theta}_i} \right)^{-\mathbf{a}_j} \tag{3.7}$$

Therefore the Γ-IRT formulation can be used to estimate the rate parameter λ_{ij}, which also happens to be the expected value of the d_{ij} response, i.e. $\lambda_{ij} = \mathbb{E}[d_{ij}]$. Therefore, we can model discrete nonnegative error counts using the Γ-IRT model.

4 EXPERIMENTS WITH REGRESSION MODELS

In this section, we apply the Γ-IRT model to machine learning regression problems. Each respondent is a different regression model and items are test instances in a dataset. The idea is to provide insights from the absolute regression errors of a pool of models in a dataset, by simultaneously analysing data instance difficulty and discrimination as well as regression model ability.

4.1 METHODOLOGY

We selected 12 datasets with different characteristics and a pool of 13 diverse regression models to evaluate. For each dataset, a hold-out experiment was adopted to collect the absolute error, also refered to as 'error', of each regression model in each test instance (i.e., error e_{ij} of respondent i to item j). The absolute errors observed for a dataset are modelled by our proposed Γ-IRT model, as described in Equation (3.1). Given the matrix of absolute errors in a dataset, the Γ-IRT model is applied to derive the ICC for each test instance as well as the abilities θ_i for all regression models. The Γ-IRT model is built 40 times for each response matrix, therefore we take the average of both items' parameters and abilities.

The next sections provide more details about the datasets and regression models adopted in the experiments.

4.2 DATASETS

In the performed experiments, 12 regression datasets with different characteristics were chosen, presented in Table 4. The first 3 datasets - *Poly 5100*, *Poly 1011* and *Sin 1100* - are univariate regression problems, artificially generated by uniformly sampling the predictor attribute in a specific interval and applying a chosen function (either polynomial and sinusoidal functions) in order to generate the target attribute. All functions are described next, in Equation (4.1.) The other 9 datasets are real benchmark regression problems collected from either UCI or OpenML repositories.

$$\begin{aligned} y_{Poly5100} &= 0.7 + 0.7x + 1.1x^3, x \in [0,1] \\ y_{Poly1011} &= x - x^3, x \in [-4,6] \\ y_{Sin1100} &= \sin(20\frac{x}{\pi} + \frac{\pi}{3}), x \in [0,1] \end{aligned} \tag{4.1}$$

Figure 8 illustrates the datasets used in the experiments. For multiple regression problems, Figure 8 presents the first principal component (PC) and the target attribute. The diversity of

Table 4 – Description of all regression datasets used in the experiments.

ID	Dataset	Size	Attributes	Type	Source repository
1	Poly 5100	500	1	Artificial	-
2	Poly 1011	300	1	Artificial	-
3	Sin 1100	500	1	Artificial	-
4	Auto 93	83	20	Real	OpenML
5	Bike Sharing Day	731	14	Real	UCI
6	Bodyfat	252	14	Real	OpenML
7	Boston Corrected	506	17	Real	OpenML
8	CPU	209	6	Real	OpenML
9	Disclosure Z	662	3	Real	OpenML
10	Human Devel	130	2	Real	OpenML
11	Mileage per Gallon	398	5	Real	UCI
12	Real Estate	414	7	Real	UCI

Source: Author.

regression problems is important to verify how the difficulty and discrimination parameters of the Γ-IRT model behave according to different dataset features. We adopted datasets with little or no noise (e.g. *Poly 5100*), as well as highly noisy datasets (e.g. Disclosure Z), with no apparent relation between predictor and target attributes.

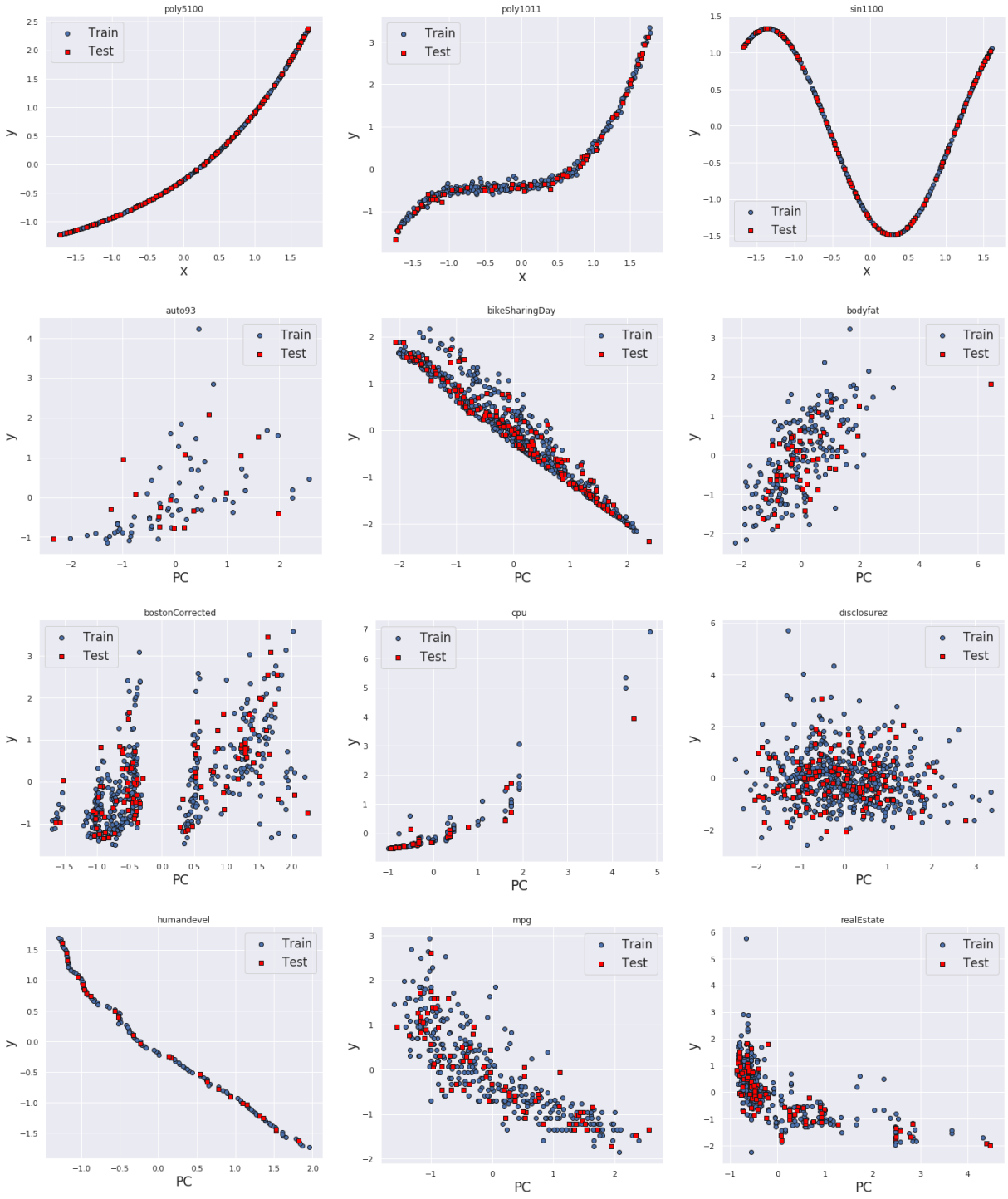
Just to be clear, Principal Component Analysis was used for analysis purposes only. Regression models were trained with all valid attributes from the datasets. To simplify data visualization, since some datasets have more than 10 attributes, the first principal component was obtained in order to locate the instances succinctly.

4.3 REGRESSION MODELS

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, explained next, and all test items for each dataset. From now on, the absolute error will be referred to as 'error' for simplification purposes, since we are not interested in the error's sign.

For each regression dataset, we trained and tested 10 regression models (both 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. All regression models were implemented using the Scikit-learn library. Unless the algorithm's parameters are explicitly specified above, Scikit-learn's default configurations were adopted. In these experiments, a hold-out procedure was used for evaluation, in which 80% of the data instances were randomly chosen for model training and the remaining

Figure 8 – Train and test partitions of all regression datasets. PCA is applied when the dataset has more than 1 attribute (for visualization purposes).



Source: Author.

instances adopted for testing.

In addition to the mentioned regression models, we adopted three baseline models: (i) Optimal - for each instance, it takes the lowest error among all regression models; (ii) Average - it always returns the average of all errors; (iii) Worst - it takes the worst error amongst all regression models. These models are adopted as baselines for comparison.

After collecting all errors, we applied the Γ-IRT model to estimate the difficulty and discrimination values of all instances for each dataset, as well as the ability values for the regression models.

4.4 RESULTS: ITEM PARAMETERS

In this section, we analyse the outputs of the Γ-IRT related to the items. In Figure 9, difficulty and discrimination values of the test instances are represented by the color intensity and marker size, respectively. In the group of artificial datasets, which have similar characteristics, such as the absence of noise and a single attribute variable, we observed the presence of clusters of instances with similar difficulties and discrimination values. On the other hand, for the real datasets, the pattern of difficulty may be quite different depending on the problem. Regions of similar difficulty and discrimination are not regular and well defined since real datasets are often noisier. Nevertheless, it is still possible to observe some regularities in the difficulty values within the datasets as follows:

- Artificial datasets (*Poly 5100*, *Poly 1011* and *Sin 1100*): in the *Poly5100* dataset, there are two relatively low difficulty regions (instances either around -1 or around 1). In such regions, all fitted models (both linear and nonlinear) obtained low regression errors. In turn, since this dataset presents a nonlinear pattern, the regions of higher difficulty are determined by the poor fit of the linear models. In general, this dataset has low regression errors, which makes it difficult to distinguish between good and bad regression models. It is also noted that regions of high discrimination coincide with regions of low difficulty and vice versa. For the other two artificial datasets (*Poly 1011* and *Sin1100*), we also observe clusters of similar difficulty and discrimination values along the test set. In *Sin1100*, although discrimination appears to be evenly distributed along the x-axis, the regions of high discrimination coincide with the extreme regions in relation to the y-axis (“hill" and “valley" of the sinusoidal function). It turns out that in these regions the linear models have higher errors, clearly distinguishing high and low ability models.

- *Auto 93* and *CPU*: both datasets have similar curve shapes, although apparently *Auto 93* has a higher variance in the target attribute. Both datasets are less noisy in the initial portion (for lower values of the PC), which reflects in a clear division between the regions of low and high difficulty. When the variance is lower, i.e. when $PC < 0.1$ for *Auto 93* and when $PC < 1$ for *CPU*, difficulty is relatively low. As data becomes more dispersed, due to the possible presence of noise, difficulty values increase. The opposite is true for discrimination, which is relatively low for regions with higher variance.

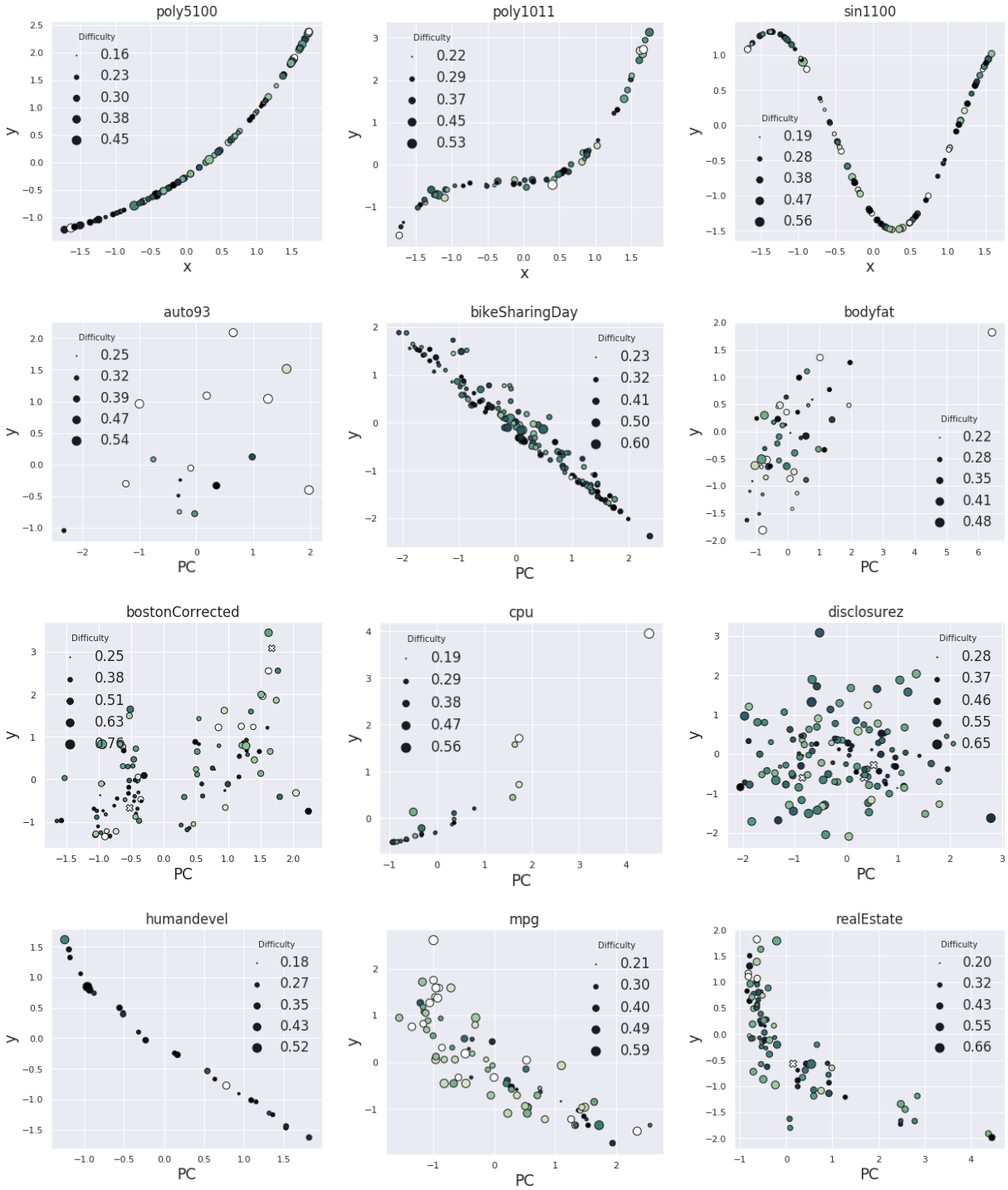
- *Bike Sharing Day*, *Bodyfat* and *Human Devel*: the three datasets have approximate linear patterns, although *Bike Sharing Day* and *Human Devel* are descending curves and *Bodyfat* is a roughly linear ascending curve with noise. They differ in the number of instances and noise (apparently noise is much lower in *Human Devel*). In *Human Devel*, most difficult instances are observed when $PC < -0.5$, where there is an apparent nonlinear pattern compared to the rest of the data. *Bike Sharing Day* looks different as almost all instances have similar difficulty and discrimination values, despite being a larger dataset and having noisy data. Instances closer to the centre of the curve have lower difficulty. The region of highest difficulty, where *PC* is around 0, matches the region with higher noise. We highlight the presence of an outlier in the *Bodyfat* dataset. Higher errors are expected for this particular instance, which reflects in its high difficulty value and low discrimination. In general, the three datasets have low error values, as will be shown in this section, due to their less complex data patterns.

- *Boston Corrected* and *Disclosure Z*: In *Disclosure Z*, the target attribute appears to be randomly distributed along the PC. Instances with target attribute around 0, along the PC, tend to have lower difficulty values. This is reasonable by considering a high level of randomness in the target attribute. Difficulty gradually increases as the points move away from 0 to more extreme regions. The same does not seem to apply to *Boston Corrected*, as this dataset has a slight increasing pattern. However, it is possible to notice that the instances located in the centre of the apparent curve have lower difficulty in comparison to the more extreme instances. The dataset also has a gap between regions where PC is approximately equal to 0. The first region ($PC < 0$) presents less dispersed data and lower difficulties when compared to the second region ($PC > 0$).

- *MPG* and *Real Estate*: both datasets have very similar shapes (nonlinear descending curves). In the *MPG* dataset, instances closer to the centre of the curve usually have lower difficulty values. The farther from the centre, the higher the difficulty values (the target attribute is noisy for these instances). In the *Real Estate* dataset, the region of greatest difficulty coincides with the region of higher noise in the data (where $PC \in [-0.8, -0.2]$). Similarly, less noisy data that more clearly follow a curved pattern have less difficulty.

Figure 10 illustrates the relation between difficulty and discrimination for each instance across all regression models. In most of the studied datasets, difficulty and discrimination are negatively correlated as illustrated in the figure, i.e., the greater the difficulty, the less discriminative the instance. The analysis suggests that instances that do not differentiate well regression models

Figure 9 – Mapping of Difficulty and Discrimination in the test set (Darker colour indicates higher discrimination and bigger markers indicate higher difficulty).

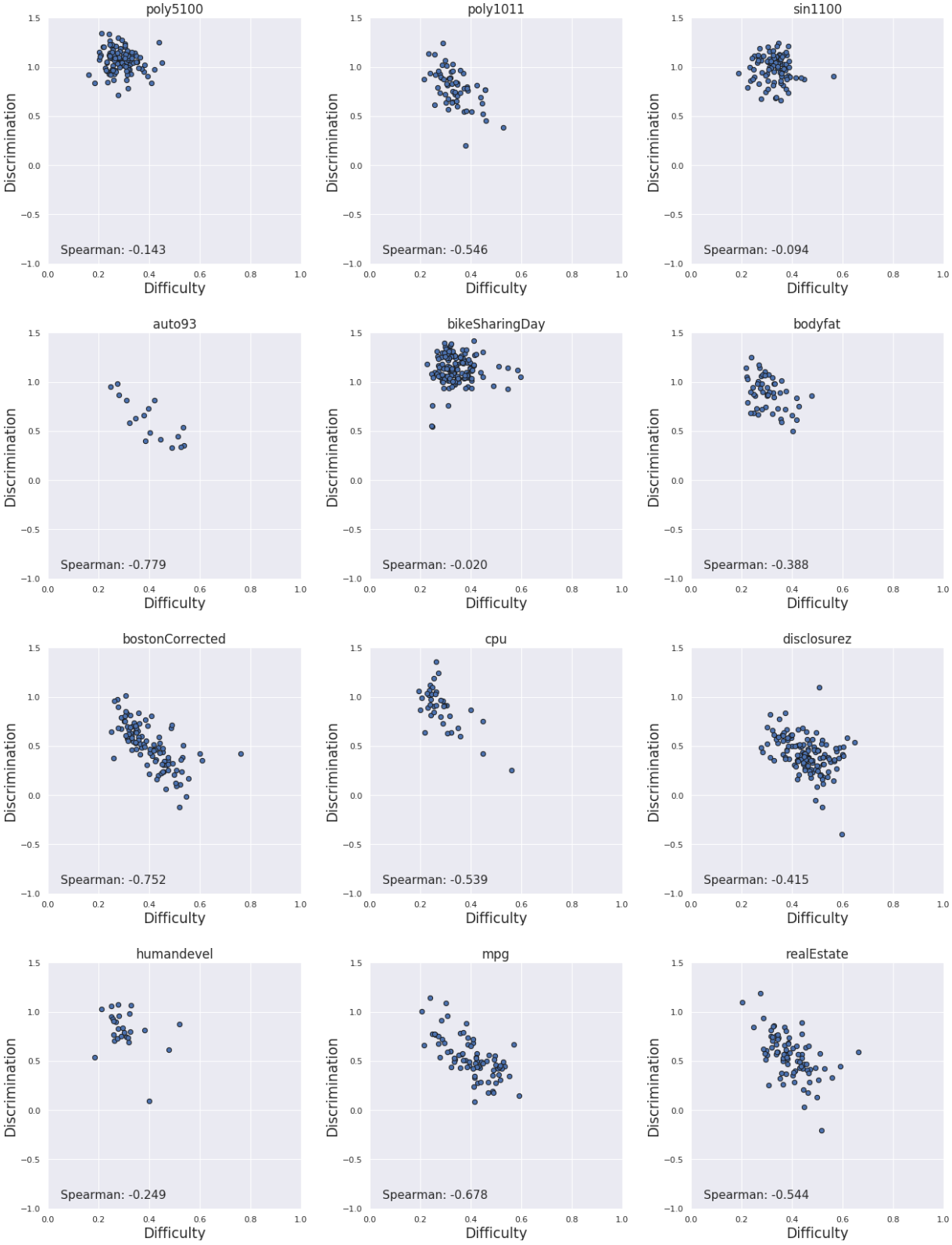


Source: Author.

with different abilities tend to be more difficult, while presenting low discrimination. Similarly, instances that can differentiate models well and have higher errors as the ability increases tend to have high discrimination and low difficulty. This will be exemplified later with some instances extracted from a real dataset.

Figure 11 shows the relationship between difficulty and the average error across all

Figure 10 – Difficulty vs Discrimination.



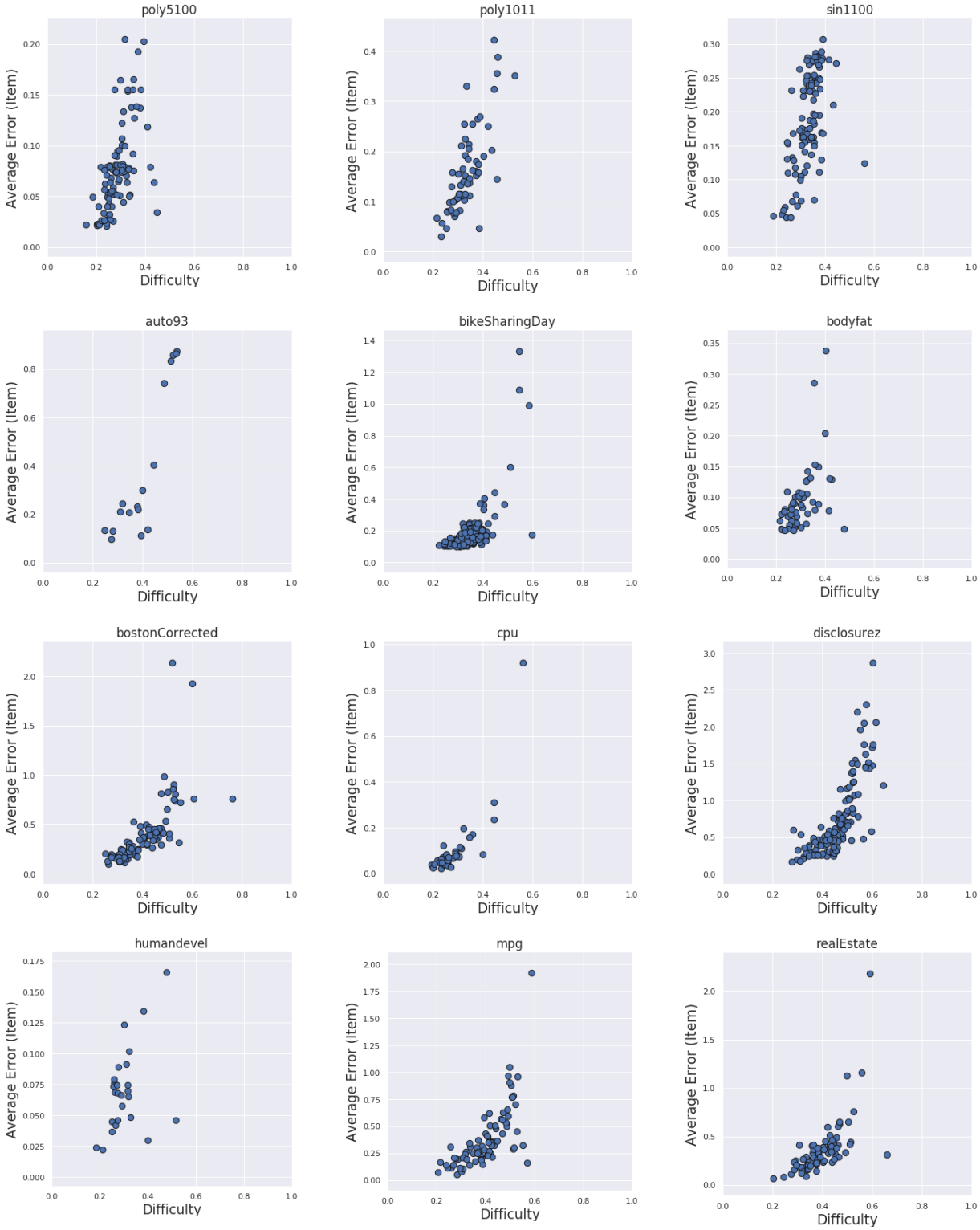
Source: Author

regression models. Items with higher difficulty values usually result from higher errors, as expected. However, the relation between difficulty and average errors is not strictly linear. In

some cases, even instances with low errors may have high difficulty. This can happen when regression models with high ability respond worse than low ability models.

Figure 12 illustrates how the expected error behaves along the respondent's ability in

Figure 11 – Difficulty vs Average Error (Instance).



Source: Author.

different data instances. The red marks indicate the errors and abilities of all regression models for the particular instance. Each instance belongs to a specific region within the test partition of the *Auto93* dataset and is explained next.

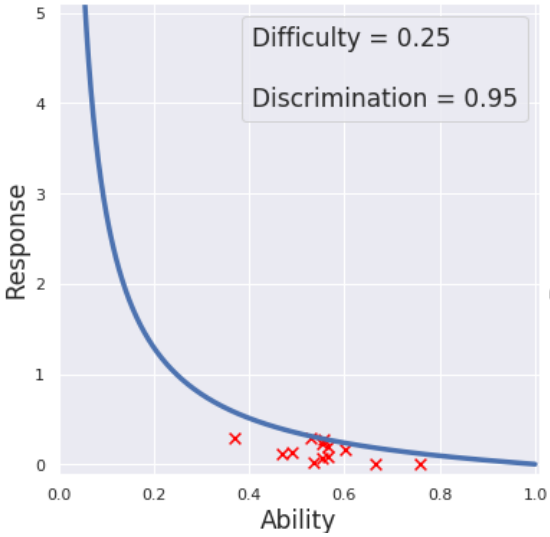
- Instance (a) has the lowest difficulty of the test set. The low overall error explains the low difficulty itself. When we look more deeply at the models that predict better in each instance, we take useful insights on errors and discrimination parameter. Discrimination is higher in instance (a) because regression models respond better to ability. It should be noted that discrimination is the parameter that determines the slope of the expected error curve along the ability.

- Instance (b) has the highest difficulty value as a result of higher overall errors. In instance (b) the relationship between error and discrimination is not so clear, since high-ability models for this set, for example DT and KNR, respond worse than lower-ability models such as MLP100 and MLP50-50. This instance does not discriminate well between good and bad models.

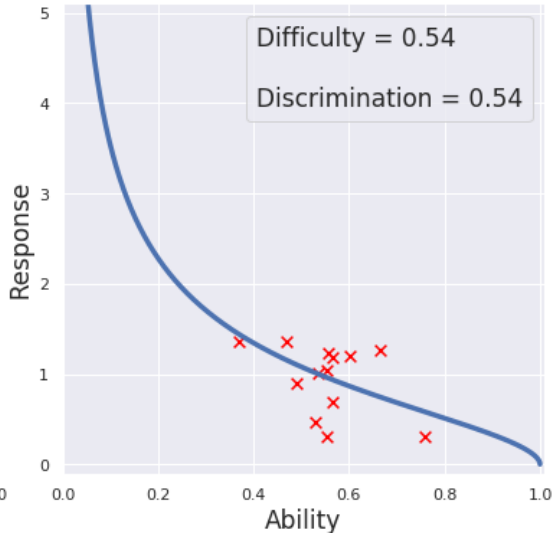
The two instances have well-defined characteristics. Table 5 shows the regression errors given by all regression models for both instances presented above. Looking at Figure 11, for the corresponding dataset, the two instances mentioned above are at the extremes, both in relation to Difficulty and to Average Error per instance.

Figure 12 – Representative Item Characteristic Curves from *Auto 93* dataset.

(a) Instance (a): Low difficulty and high discrimination.



(b) Instance (b): High difficulty and low discrimination.



Source: Author.

Table 5 – Example of error values for two data instances (items) and all regression models (respondents).

	(a)	(b)
LR	0.22322	1.04041
Bayes	0.026483	1.01294
SVR Linear	0.09114	1.17762
SVR Rbf	0.19750	0.68596
KNR	0.16784	1.20252
DT	$1.0e-04$	1.26373
RF	0.10761	1.35852
AdaB	0.28137	1.22883
MLP100	0.28712	0.46188
MLP50-50	0.060775	0.305412
Avg	0.13605	0.90195
Opt	$1.0e-04$	0.305412
Wrs	0.28712	1.35852

Source: Author.

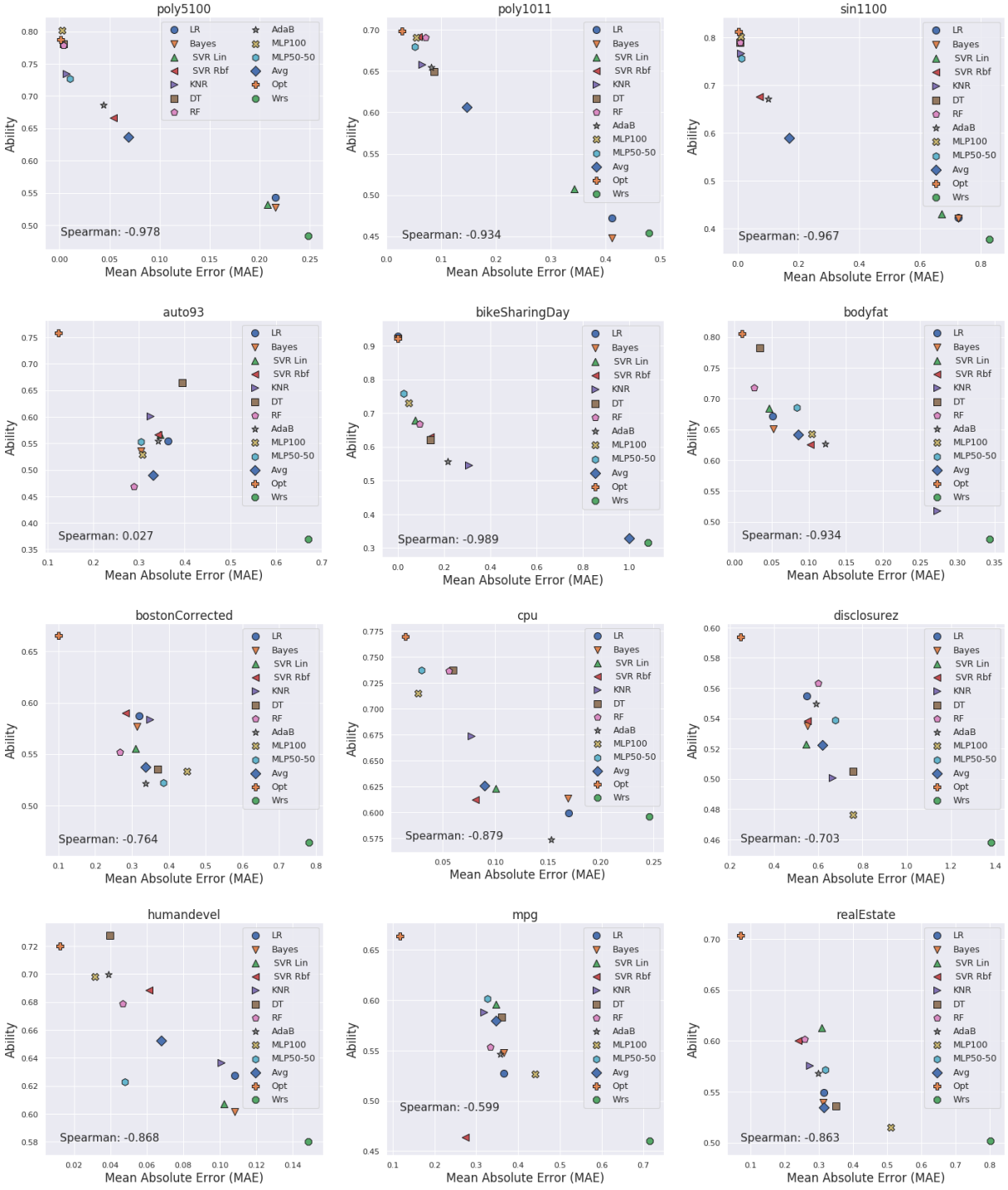
4.5 RESULTS: RESPONDENTS

In this Section, we analyse the performance of all regression models by comparing the ability, obtained after applying the Γ-IRT model, and the Mean Absolute Error (MAE). Figure 13 shows the relation between these two evaluation metrics. In almost all cases, except for *Auto 93* and *MPG* datasets, ability and MAE have 'strong' negative correlation, which is also presented in the figures. This is an expected result since models with the lowest errors most likely have the highest ability values. However, this is not a rule, as the results suggest that ability considers whether models produce higher errors for easy or difficult instances. When we evaluate regression models with conventional metrics, such as MAE, we have only an average result across an entire test set. A major advantage of the ability as an evaluation metric is to take into account the distinct difficulty regions throughout the dataset.

In the *Poly 5100* dataset, we analyse two models to demonstrate the benefit of using ability as a measurement metric. KNR and MLP100 have very similar errors throughout the test set, but their abilities differ as shown in Figure 13. MLP100 has the highest ability as it performs better on average on the 20 most difficult instances. Not only does it perform better in the most difficult instances, it also outperforms KNR on the easiest instances on average.

The *Auto 93* dataset presents a peculiar behaviour in which the regression model with the highest MAE value is also the one with the highest ability (DT model). Similarly, the RF model has the lowest MAE value, yet has the worst ability value among the other models. By directly comparing the two models mentioned, the RF model outperforms the DT model in most easy instances, but when looking at the difficult instances the DT model is slightly better than RF. Something that can influence this result is the low number of items, since this test set has only 17 instances. This can lead to inaccurate ability estimates as well as item estimates.

Figure 13 – Mean Absolute Error (MAE) vs Ability (Spearman's correlation coefficient between both variables is showed in the figure).



Source: Author.

As mentioned above, there are a few exceptions to the relation between ability and MAE. In the *Real Estate* dataset, SVR Linear has higher MAE than SVR Rbf, KNR and RF, however its ability is higher than the three models. This is likely due to SVR Linear performing better in more difficult instances than DT and Bayes. We also note that the closer the performance of the models, the smaller the ability range.

Another advantage of using ability as a performance metric is that the range is always limited between 0 and 1, unlike MAE which can range from 0 to a large number, depending on the scale of the target variable in the dataset.

5 EXPERIMENTS WITH NOISE

In this Section, we analyse the effects of noise on difficulty and discrimination parameters, as well as on model ability. As mentioned in Ferri et al. FERRI *et al.* (2014) it is very common that the training data is under "idealistic" conditions, with features that are carefully measured and preprocessed. So in these experiments, we trained the regression models with simulated training data without noise (the ideal condition) but tested them with test data with different noise levels. Then we analyse if the presence of noise changes the values of instance difficulty and discrimination. Additionally we evaluated the robustness of the ability measure under different noise levels.

5.1 METHODOLOGY

In this experiment, we gradually injected Gaussian noise (ε_y) in the target attribute of the 3 artificial datasets presented in the previous section (*Poly 5100*, *Poly 1011* and *Sin 1100*). Target noise is formally described as follows (Equation 5.1):

$$\begin{aligned} y &\leftarrow y + \varepsilon_y \\ \varepsilon_y &\sim \mathcal{N}(0, \sigma_y) \end{aligned} \tag{5.1}$$

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 *original data set*.

In this experiment, we adopted the same regression models and training procedure described in Section 4.3: datasets are randomly split into two subsets, with 80/20% for training and testing. However, noise is injected in the test set and for each noise-level configuration, 40 different noisy sets are generated. We obtain the absolute errors, also refered to as 'error' in this section, from the regression models to all the instances of the test set and then apply the Γ-IRT model to derive the item parameters and regression model abilities.

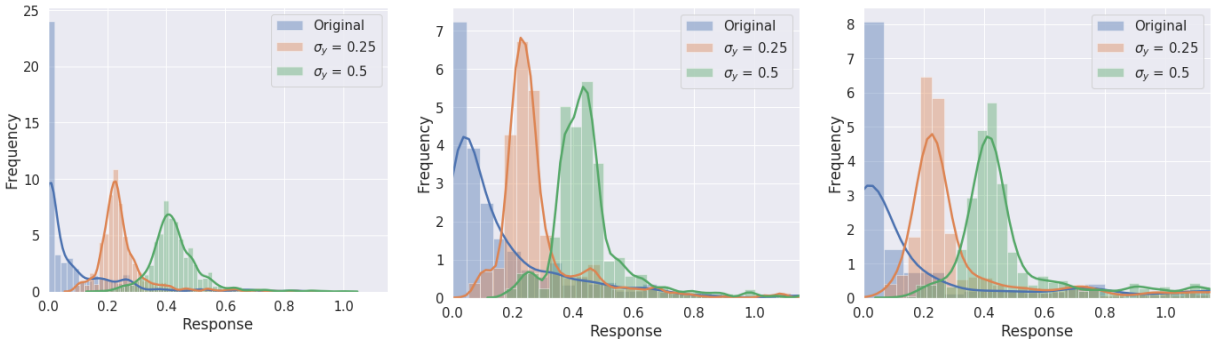
5.2 RESULTS: ITEM PARAMETERS

Before going into detail with the Γ-IRT models, it is important to mention how regression models respond to items in general as noise is inserted. Higher noise in the target attribute increases regression errors, so the error values of all regression models over instances gradually increase as well. Figure 14 shows the shift in error distribution across three specific steps of noise injection.

Figure 15 illustrates the boxplot of difficulty and discrimination along all steps of noise injection. It shows how difficulty and discrimination distributions, respectively, change as noise is injected into the target attribute. In the 3 datasets, difficulty gradually increases as noise

Figure 14 – Effects of noise injection in the error distributions.

(a) Poly 5100 (b) Poly 1011 (c) Sin 1100



Source: Author.

is injected. As already shown in the previous section, difficulty and discrimination correlate negatively. Thus, as difficulty increases, discrimination decreases with noise injection. The behaviour of each dataset through noise injection is explained next:

- Original *Poly 5100* has only two relatively low difficulty regions due to the poor fit of linear models (Figure 15a). In the intermediate noise injection step ($\sigma_y = 0.25$), the low difficulty regions quickly equalise to the rest of the set. This is because all regression models start to produce higher errors in these regions. Discrimination drops significantly due to the gradual loss of the difference among all regression models. Both linear and nonlinear models have high regression errors with high noise levels affecting not only difficulty but also discrimination.

- In the original *Poly 1011* test set, high-difficulty items are concentrated inside the intervals $x \in [-1.4, -1.0]$, $x \in [0.25, 0.75]$ and $x > 1.3$, while low and average difficulty regions correspond to the complementary intervals. The target variable in the central region is approximately constant, therefore, it is expected that injecting noise results in larger regression errors inside this interval. Significant changes in difficulty are not observed in the high-difficulty regions of the curve since they do not suffer relevant distortions when noise is injected. Notice that the difficulty boxplots gradually shift upwards, reflecting higher difficulties in the presence of noise (Figure 15b). The picture is not as clear as in difficulty, but easier instances tend to show higher discrimination. As already observed, 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. If the injected noise is large enough, data tends to become random. Thus the variance of difficulty tends to decrease. This can be seen in Figure 15, which illustrates difficulty and discrimination values along the x-axis.

- In the *Sin 1100* original dataset, according to Figure 9c, there are three regions of

relatively low difficulty: where x is around the values -1.5, -0.6 and 0.9. Similar to the datasets described above, these low difficulty regions exist due to the poor fit of linear models (note that the regions can be linked together by a straight line). In the same regions discrimination is relatively high because the error of the worst models, i.e. linear models, is quite high when compared to models that learn the curve pattern well, i.e. MLP100. Thus there is a clear distinction among models in these regions. Between the original and the middle noise injection stage there is a general increase in difficulty, but the emphasis is on the sharp drop in discrimination across all regions, as shown in Figure 15c. In the final noise injection step, the parameter values keep changing, but at a lower rate.

The figures make it clear that noise makes instances more difficult to predict by regression models. Discrimination decreases, also indicating that noise makes data difficult to differentiate regression models. Looking at the datasets in the last noise step, it can be seen that the data have almost completely lost the nonlinear pattern of the curves. It is more evident from the *Poly 5100* and *Poly 1101* datasets that curves have given rise to noisy data following a more linear pattern, which influences the behaviour of abilities as we will see in the next section. It also suggests for all datasets that the difference between maximum and minimum values of difficulty decreases as noise is inserted.

5.3 RESULTS: RESPONDENT ABILITIES

Figures 16 and 17 show the performance of all regression models as target noise is injected in the test sets. There are 3 main groups of models that present similar behaviour among themselves: linear, nonlinear and baseline models.

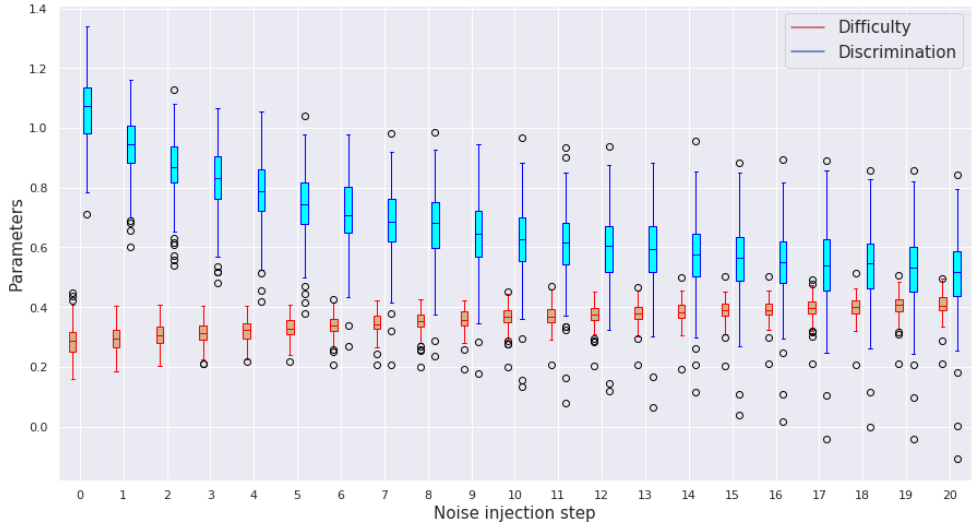
The group of linear models (formed by Linear Regression, Bayes and SVR Linear) does not present significant changes in its ability. Initially, they have the highest MAE values, although they increase slower than nonlinear models as data gets noisier. As already explained in the previous section, as larger noises are injected the datasets gradually lose their curved patterns and give way to a noisy linear pattern. This is reflected in the slight increase of ability throughout the process in linear models only.

The models with the best performance belong to the group of nonlinear models. Looking at their abilities, RF and SVR Rbf stand out as the best regression models among all. The ability of nonlinear models, however, declines significantly as noise is inserted into the target attribute. This may also be due to the loss of patterns in datasets. Nonlinear models that fit well-defined curves now produce high errors not only because of noise, but because of the lower generalisation of data.

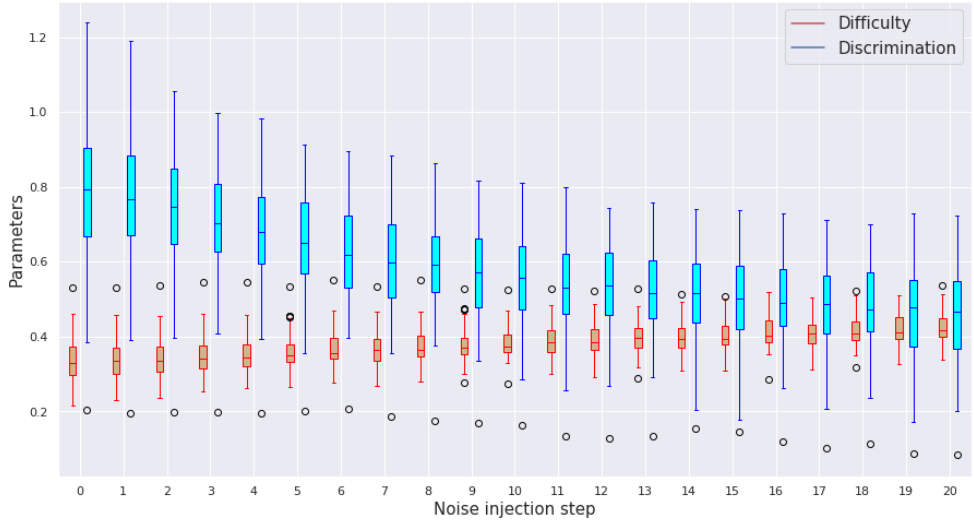
Since models are fitted using non-noisy data, distortions in the test set caused by noise injection result in greater errors, which can cause as a higher uncertainty about the relative performance of different models. Thus we checked whether ability could be a more robust

Figure 15 – Boxplot of parameters along noise injection.

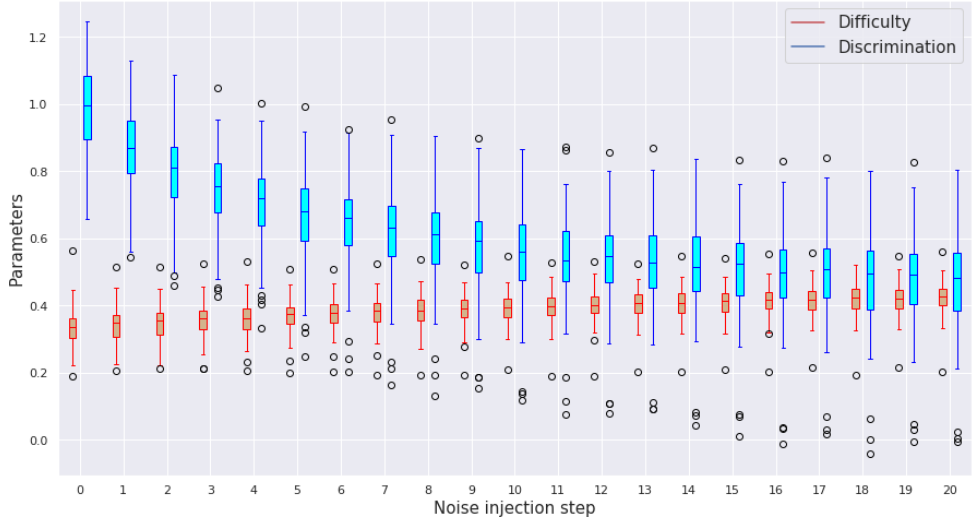
(a) Poly 5100



(b) Poly 1011



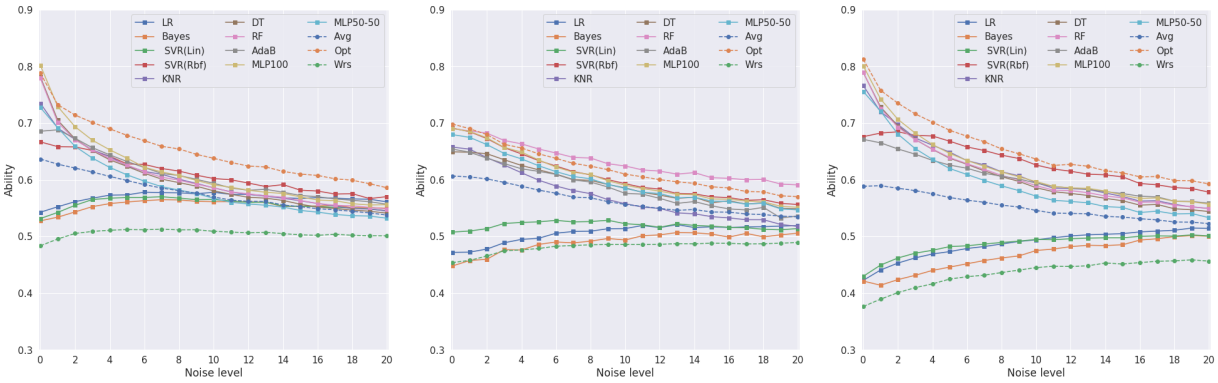
(c) Sin 1100



Source: Author.

Figure 16 – Evolution in the ability of all regression models along noise injection.

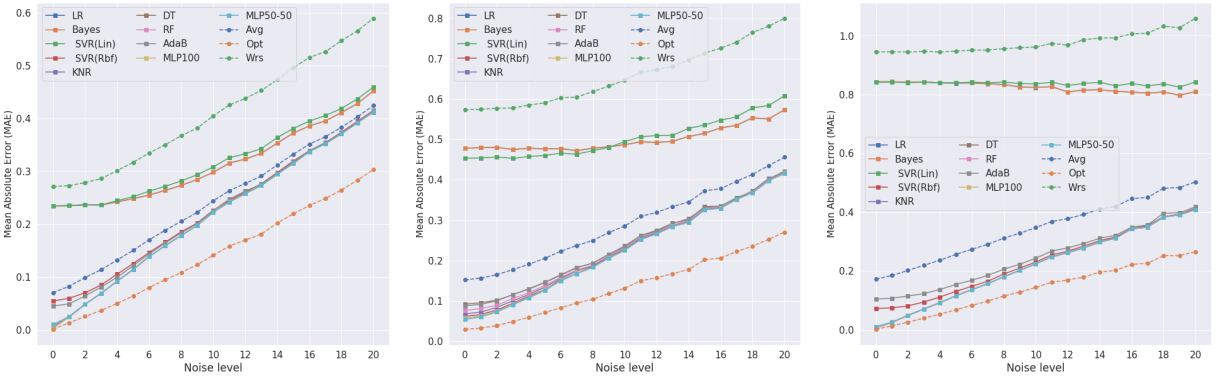
(a) Poly 5100 (b) Poly 1011 (c) Sin 1100



Source: Author.

Figure 17 – Evolution in the Mean Absolute Error (MAE) of all regression models along noise injection.

(a) Poly 5100 (b) Poly 1011 (c) Sin 1100



Source: Author.

performance measure, as it estimates a latent model behaviour. For this, we calculated the percentage variation in ability and in MAE of each regression model that occurred in a given noise injection step relative to the original test set. We expect that a better performance measure would be less sensitive to the presence of noise. The percentage variations of ability and MAE, in noise injection step $k = 1, ..., 20$, are calculated as follows:

$$\frac{|\theta_{ik} - \theta_{i0}|}{\theta_{i0}} \times 100\%$$

$$\frac{|MAE_{ik} - MAE_{i0}|}{MAE_{i0}} \times 100\% \tag{5.2}$$

Where θ_{i0} and MAE_{i0} are the ability and the MAE of the regression model i for the original dataset under analysis. We evaluated which of the two measures varies less for each

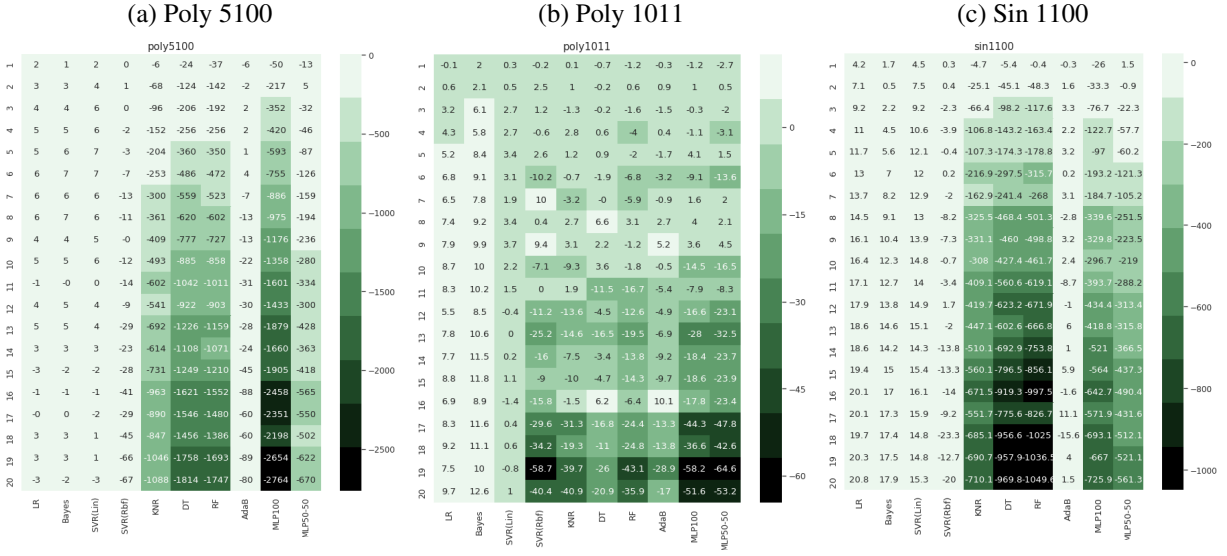
noise injection step k. For example, if in a given step k the ability varies less than the MAE, the subtraction between these two values must be negative. This difference is measured for each regression model in all noise injection steps.

Figure 18 illustrates the heat map of the difference between the percentage variation in ability and in MAE of each regression model that occurred in each noise level. Negative values (darker green cells) indicate that the variation of ability values is smaller than the MAE variation, which favours ability. In general, ability is more robust than MAE, especially when more noise is inserted in the test sets. Results show that ability varies significantly less than MAE as noise increases, for the three artificial test sets. We tested whether the sample of percentage variation in ability came from a distribution with a mean lower than the percentage variation in MAE, which would suggest a lower sensitivity to noise as explained previously. According to a paired t-test, the p-values for *Poly 5100*, *Poly 1011* and *Sin 1100* are equal to 0.000008, 0.00004 and 0.00001, respectively.

The first three linear models (Linear Regression, Bayesian Ridge and SVR Linear) have relatively lower absolute values in the heatmap than the rest of the models in almost all noise injection steps. This is because linear models already produce high error values for the original test set. It is also possible to observe that the slope of the error curve of the linear models presented in Figure 17 is smaller than all other models. Thus the variation in MAE is not as significant as in the other models.

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 errors of the best models on average tend to increase, its ability declines. Notice that the Average model tracks the average performance of all regression models.

Figure 18 – Heat map of differences between the percentage variation in ability and MAE along noise injection.



Source: Author.

Since most regression models produce higher errors as noise is injected and their abilities decrease, the ability of the Average model decreases as well. Opposite to the Optimal model, the Worst model tracks the performance of the worst model, which can often be a linear model.

6 CONCLUSION

In this section the final considerations of the work are raised, as well as the possibilities of future works. Finally, the main academic contribution of this dissertation is shown.

6.1 FINAL CONSIDERATIONS

In this paper we proposed a new IRT model, called Γ-IRT, developed to fit positive unbounded responses. We applied Γ-IRT in a regression scenario to analyse the performance of regression models and also the levels of difficulty and discrimination of data instances located in specific regions in the dataset. Experiments were carried out with 3 artificial datasets and 9 real case datasets extracted from the UCI and OpenML repositories. For each dataset, 10 regression models were built from the open-source library Scikit-learn, and 3 baseline models were generated by inserting synthetic values directly into the response matrix. The models were trained with 80% of the data and test results were obtained with the remaining 20%. Experiments were also carried out to analyse the effects of noise injection on the Γ-IRT model. Noise, extracted from a Gaussian distribution, was injected into the target attribute in several stages, gradually increasing the standard deviation of the noise sample to higher values.

The results of the experiments provided interesting insights for regression tasks. The results suggest that there are regions of high and low difficulties, caused either by more complex data patterns to be learned or by the presence of noise. Noisy data seem to present higher difficulty and lower discrimination when compared to noise free data. For example, in noise free artificial data sets, linear models often have lower abilities because they cannot learn the curves present in the data patterns. In the real datasets, more susceptible to noise and with more feature attributes, the ability values are less discriminated by the type of regression model (linear or nonlinear). In the experiments carried out with noise injection, linear models proved to be more robust than nonlinear models. This is due to the fact that nonlinear models are more sensitive to the presence of noise and can "suffer" more from overfitting. This is suggested when the ability of linear models increases and nonlinear models decrease, as noise is injected into the data. Furthermore, ability vary less than MAE throughout the noise injection, hence, model ability may be used as a robust performance metric as it tracks the error values and is less affected by noise.

The application of IRT to regression evaluation is new in literature. 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.

6.2 FUTURE WORK

From the experiments, it was possible to analyse the behavior of Γ-IRT for different datasets and regression models. However, there are plenty of factors that have not been explored in depth due to the introductory nature of this work. Therefore, future work may include:

- Feature noise injection

  In cases where the collection or treatment of feature data is subject to errors, resulting in the addition of noise, it will be important to analyse the effects over the items' parameters. For example: assuming that sensors collect data in a given operation are degrading, the reliability of the collected data decreases. What insights could be obtained with the use of Γ-IRT model in this process of degradation of features? Furthermore, through the ability of regression models, would it be possible to identify models that are less sensitive to the degradation of feature attributes?

- Feature selection

  It can be challenging to choose an appropriate feature selection method and interpret the results for a specific dataset, despite the number of existing approaches. Another possible application of Γ-IRT is feature selection based on difficulty and discrimination of each attribute over the ML models. For a set of different regression models, it would be interesting to analyse whether any features, or sets of features, are more discriminative than others. Likewise, to check if the ability of the same regression model can vary when including or excluding specific features.

- Time series forecasting

  Since the responses of time series models are positive and unbounded, it would be interesting to analyse the behavior of Γ-IRT parameters throughout the series. Would it be possible to predict a concept drift in the time series using the item parameters? What insights could the ability provide for the different tested models?

6.3 ACADEMIC CONTRIBUTION

Finally, in terms of contributions, the results achieved within this dissertation culminated in an international conference paper presentation with a good rating in the CAPES qualification. The paper is shown below:

- Item Response Theory for Evaluating Regression Algorithms, MORAES *et al.* (2020) - Presented in July 2020 at the International Joint Conference on Neural Networks (IJCNN) - Qualis A2. Its content is directly related to this dissertation.

REFERENCES

ANDERSEN, E. B. (1973). A goodness of fit test for the rasch model. *Psychometrika*, 38(1):123–140.

BIRNBAUM, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In *Statistical Theories of Mental Test Scores*, 395–479.

CHEN, Y., SILVA FILHO, T. M., PRUDENCIO, R. B., DIETHE, T., & FLACH, P. (2019). β^3-irt: A new item response model and its applications. In *Proceedings of Machine Learning Research*, 89:1013–1021.

CHEN, Z. & AHN, H. (2020). Item response theory based ensemble in machine learning. *International Journal of Automation and Computing*, 17(5):621–636.

EMBRETSON, S. & REISE, S. (2013). *Item Response Theory for Psychologists*. Taylor & Francis.

FERRI, C., HERNÁNDEZ-ORALLO, J., MARTINEZ-USÓ, A., & RAMIREZ-QUINTANA, M. J. (2014). Identifying dominant models when the noise context is known.

FLACH, P. (2019). Performance evaluation in machine learning: The good, the bad, the ugly, and the way forward. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33:9808–9814.

HAMBLETON, R. K., SWAMINATHAN, H., & ROGERS, H. J. (1991). *Fundamentals of item response theory*. Sage.

KANDANAARACHCHI, S. & SMITH-MILES, K. (2020). Comprehensive algorithm portfolio evaluation using item response theory.

LALOR, J., WU, H., & YU, H. (2016). Building an evaluation scale using item response theory. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*.

LI, X., CHEN, Y., & ZENG, K. (2016). Integration of machine learning and human learning for training optimization in robust linear regression. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

LINDEN, W. J. V. D. (2006). A lognormal model for response times on test items. *Journal of Educational and Behavioral Statistics*, 31(2):181–204.

LORD, F. & NOVICK, M. R. (1968). Statistical theories of mental test scores. addison. *Reading*.

MARIS, E. (1993). Additive and multiplicative models for gamma distributed random variables, and their application as psychometric models for response times. *Psychometrika*, 58(3):445–469.

MARTÍNEZ-PLUMED, F. & HERNÁNDEZ-ORALLO, J. (2017). Ai results for the atari 2600 games : difficulty and discrimination using irt.

MARTÍNEZ-PLUMED, F. & HERNÁNDEZ-ORALLO, J. (2018). Analysing results from AI benchmarks: Key indicators and how to obtain them. *CoRR*, abs/1811.08186.

MARTÍNEZ-PLUMED, F., PRUDÊNCIO, R. B., MARTÍNEZ-USÓ, A., & HERNÁNDEZ-ORALLO, J. (2016). Making sense of item response theory in machine learning. In *European Conference on Artificial Intelligence, ECAI*, 1140–1148.

MARTÍNEZ-PLUMED, F., PRUDÊNCIO, R. B., MARTÍNEZ-USÓ, A., & HERNÁNDEZ-ORALLO, J. (2019). Item response theory in ai: Analysing machine learning classifiers at the instance level. *Artificial Intelligence*, 271:18 – 42.

MASTERS, G. N. (1982). A rasch model for partial credit scoring. *Psychometrika*, 47(2):149–174.

MORAES, J. V. C., REINALDO, J. T. S., PRUDENCIO, R. B. C., & FILHO, T. M. S. (2020). Item response theory for evaluating regression algorithms. In *2020 International Joint Conference on Neural Networks (IJCNN)*.

OLIVEIRA, C. S., TENORIO, C. C., & PRUDÊNCIO, R. B. (2020). Item response theory to estimate the latent ability of speech synthesizers. In *European Conference on Artificial Intelligence, ECAI*.

PRUDÊNCIO, R. B., HERNÁNDEZ-ORALLO, J., & MARTÍNEZ-USÓ, A. (2015). Analysis of instance hardness in machine learning using item response theory. In *Second International Workshop on Learning over Multiple Contexts in ECML 2015. Porto, Portugal, 11 September 2015*, 1(3).

RASCH, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests*. Studies in mathematical psychology. Danmarks Paedagogiske Institut.

RIZOPOULOS, D. (2006). ltm: An r package for latent variable modelling and item response theory analyses. *Journal of Statistical Software*, 17(5):1–25.

SAMEJIMA, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika*, 34(S1):1–97.