1st International Workshop on Evaluating Progress in Artificial Intelligence - EPAI 2020
In conjunction with the 24th European Conference on Artificial Intelligence - ECAI 2020
Santiago de Compostela, Spain

Towards Efficient and Robust Model Benchmarks with Item Response Theory and Adaptive Testing

Hao Song and **Peter Flach** [1]

Abstract. Recent progress in machine learning, on predictive tasks in particular, is usually claimed on the basis of performance comparisons. Due to the limitation of application scenarios and computational resources, most people cannot evaluate their work on a large variety of tasks and datasets, and it has become tricky for both reviewers and readers to verify individual performance gains for a given approach. In this paper we investigate possible approaches to achieve better efficiency on model benchmarking. For a large collection of datasets, rather than training and testing a given approach on every individual dataset, we seek methods that allow us to pick only a few representative datasets to quantify the model's goodness, from which to extrapolate to performance on other datasets. To this end, we adopt existing approaches from psychometrics: specifically, Item Response Theory and Adaptive Testing. Both are well-founded frameworks designed for educational tests. We propose certain modifications following the requirements of machine learning experiments, and present some initial results and insights in this paper.

1 Introduction

Thanks to the recent popularity of machine learning and artificial intelligence techniques, researchers and practitioners now have very considerable choice of models and learning algorithms when facing a given task. However, as choices come with deliberations, to select an appropriate model is also becoming more and more challenging. From a traditional view, performing a good model selection typically involves two steps. (1) To gather related work and hence seeking for existing comparisons. (2) Prepare a shortlist and run the models within the targeted task for more detailed and local comparisons. However, given the number of research areas and datasets available now, there are few if any research papers that provide a comprehensive benchmark on all related datasets. Furthermore, for publication reasons, people tend to show the datasets where the proposed approaches have improvements, making it even harder to obtain a fair and comprehensive view of different methods [8]. Regarding the second step, given the rapidly rising in computational demands among recent approaches, it is also often impractical to cover a broad set of experiments simultaneously. While there have been platforms like OpenML [18] that aims to collect running results via standard configurations, it still requires relatively large sets of new experiments once a novel task/method is introduced. These additional experiments could take a non-trivial time to run given OpenML's crowdsourcing nature. Although certain research areas and methods can come with formal guarantees, these only cover limited scenarios, and most practices in the field still rely on experiments and empirical evaluations. Therefore, in this paper, we consider the problem of efficiently obtaining fair and reliable benchmarking on a set of models and datasets.

To set up the scenarios of our benchmarking, we first discuss a few standard experimental settings according to their difficulties to be provided with a unified benchmark. The most common setting, which is also the simplest one to benchmark, is the typical predictive machine learning task. For such tasks, we usually can assume there exist some labelled datasets and several model classes that can be trained and tested on any possible combination. Typically, an experiment includes a set of evaluation measures, and we can simply read the measurements to reflect the performance on any given model-dataset pair. Another typical setting is to experimentally verify trade-offs between predictive performance and computational costs, widely seen in approaches requiring computational approximation, such as Bayesian inference [1]. While being similar to the first setting, for such experiments, the predictive performance can no longer be directly taken to compare. That is, we need additional evaluation to incorporate the evaluation measures, such as running speed and memory requirements. While the settings above are mainly quantitative evaluation, the hardest setting to benchmark is the so-called qualitative performance. For such experiments, there is not any fixed or well-accepted evaluation measure, and subjective opinions decide the goodness of results. One significant example of this setting is the images synthesised by Generative Adversarial Networks [6], where people often provide comparisons by case study and annotator voting.

As a starting step, we focus on the first setting and investigate approaches that, while maintaining the overall computational costs, can accurately quantify the performances on a large variety of models and datasets. For this purpose, we refer to the fields of psychometrics and testing in education, and borrow the frameworks of Item Response Theory (IRT) [11, 17] and Computerised Adaptive Testing (CAT) [7, 20]. Both frameworks work under the same scenario, where a participant is assigned several items to answer (response). A typical example would be the case of educational tests, where each student is a participant, and each test question is an item. IRT hence refers to a set of statistical models built on the responses from the participants and items. In IRT, a representative setting is to assume each participant has an ability parameter, and each item gets a difficulty parameter. Both parameters can affect the expectation of the responses. The aim is hence to learn these parameters with some collected responses, which can later provide a ranking or benchmarking. CAT is a framework further built on top of IRT. While in IRT, the availability of many responses from different participant-item combinations is expected. Sometimes a specific combination might not be necessary. For instance, it is less informative to give a harder question to a student who just failed to answer a much simpler one. The

[1] University of Bristol, UK, email: {hao.song, peter.flach}@bristol.ac.uk

purpose of CAT is hence to adaptively select the items according to previous responses so that the total number of items used in the test is kept at a relatively low level. From these points, we can see IRT and CAT indeed fulfils our need for model-dataset benchmarking.

In this paper, where we focus on predictive machine learning, every dataset is an item, and each model class is a participant. We aim to investigate the possibilities of using the IRT and CAT frameworks to obtain accurate benchmarks on each model-dataset combination while limiting the total number of experiments. We first give a brief introduction of the existing approaches from both IRT and CAT in section 2, following proposed modifications on them for our benchmarking requirements in section 3. Experiments on some common model classes and datasets will be presented in section 4, and finally additional discussions and insights are provided in section 5.

2 Background

In this section we give a brief review of related methods in both IRT and CAT, together with some notations. We also discuss some existing work on IRT in machine learning.

2.1 Item Response Theory

Item Response Theory refers to a collection of methods that aims to measure individual abilities, question (item) difficulties, and other potential attributes by checking individual responses to a set of questions (items). Statistically, IRT models are latent variable models, where the responses are the observations, and abilities, difficulties and other related parameters are the latent variables to be inferred. IRT models are of particular use when the responses distribute differently according to different items, and simply averaging the responses does not represent a participant's ability. IRT is therefore quite handy while dealing with educational exams, as well as many physiological tests. When it comes to machine learning experiments, where different datasets typically come with different baseline performances, IRT hence provides an opportunity to treat the performance gains among these datasets fairly.

In the following, we introduce two conventional IRT models and discuss their parameter settings and applications. We use the notation θ to denote the parameter of a particular candidate, and use other notations for item parameters according to the type of IRT models. The notation R denotes the random variable of the responses.

Two-parameter logistic model

The two-parameter logistic model is defined as follows:

$$R \mid \Theta = \theta, \Delta = \delta, A = a \sim \textbf{Bernoulli}(\mu_{(\theta,\delta,a)}) \tag{1}$$

$$\mu_{(\theta,\delta,a)} = \frac{1}{1 + \exp\big(-a \cdot (\theta - \delta)\big)} \tag{2}$$

And:

$$\mathrm{E}\Big[R \mid \Theta = \theta, \Delta = \delta, A = a\Big] = \mu_{(\theta,\delta,a)} \tag{3}$$

$$\mathrm{Var}\Big[R \mid \Theta = \theta, \Delta = \delta, A = a\Big] = \mu_{(\theta,\delta,a)} \cdot \Big(1 - \mu_{(\theta,\delta,a)}\Big) \tag{4}$$

Here $R \in \{0,1\}$, $\theta \in \mathbb{R}$ is the ability parameter, and $\delta \in \mathbb{R}$ is the difficulty parameter. The two-parameter logistic model additionally has a discrimination parameter a on the items, which controls how the response distribution changes when candidate ability varies. Therefore, assume we have two participants with different abilities, an item with high discrimination tends to have higher differences between the responses from the two participants respectively. Positive discrimination indicates that higher ability leads to higher expectation on the responses, and vice versa. Besides the two-parameter setting, there also exists a few variants on the Logistic IRT. The three-parameter setting further adds a guessing parameter which lower-bounds the response expectation. A multinomial setting can also be adapted to support categorical responses beyond the binary setting.

Three-parameter Beta model

While the logistic model supports binary responses, a recently proposed IRT model extends the support to continuous response [5]:

$$R \mid \Theta = \theta, \Delta = \delta, A = a \sim \textbf{Beta}(\alpha_{(\theta,\delta,a)}, \beta_{(\theta,\delta,a)}) \tag{5}$$

$$\alpha_{(\theta,\delta,a)} = \Big(\frac{\theta}{\delta}\Big)^a \tag{6}$$

$$\beta_{(\theta,\delta,a)} = \Big(\frac{1-\theta}{1-\delta}\Big)^a \tag{7}$$

$$\tag{8}$$

And:

$$\mathrm{E}\Big[R \mid \Theta = \theta, \Delta = \delta, A = a\Big] = \frac{\alpha_{(\theta,\delta,a)}}{\alpha_{(\theta,\delta,a)} + \beta_{(\theta,\delta,a)}} \tag{9}$$

$$\mathrm{Var}\Big[R \mid \Theta = \theta, \Delta = \delta, A = a\Big] = \frac{\alpha_{(\theta,\delta,a)} \cdot \beta_{(\theta,\delta,a)}}{(\alpha_{(\theta,\delta,a)} + \beta_{(\theta,\delta,a)})^2 \cdot (\alpha_{(\theta,\delta,a)} + \beta_{(\theta,\delta,a)} + 1)} \tag{10}$$

Here $R \in [0,1]$ (a bounded continuous response), $\theta \in [0,1]$, $\delta \in [0,1]$ and $a \in \mathbb{R}$. Similar to the logistic case, here a is still the discrimination parameter, and can control the change rate of responses according to the ratio between ability and discrimination. In addition to supporting continuous responses, one advantage is that the response curve of the three-parameter Beta model has various shapes beyond the usual sigmoid shape (for $a > 1$), including inverse-sigmoid ($0 < a < 1$), parabolic ($a = 1$) and even identity ($a = 1, \delta = 1/2$). .

Estimation of IRT parameters

The estimation of IRT parameters proceeds as follows. We assume to have a bag of L items, denoted as $\mathbb{D} = \{1, \ldots, L\}$, and a bag of M participants, denoted as $\mathbb{F} = \{1, \ldots, M\}$. With a given experiment protocol, we can collect a set of N item-participant-response tuples, denoted as $\{(d_1, f_1, r_1), \ldots, (d_N, f_N, r_N)\}$. Here $d_i \in \mathbb{D}$, $f_i \in \mathbb{F}$ represents a particular item / participant respectively. Denote $\boldsymbol{\theta} = \{\theta_1, \ldots, \theta_M\}$ as the parameter vector of abilities of all participants, $\boldsymbol{\omega} = \{\omega_1, \ldots, \omega_L\}$ as the vector of item parameters, and $g(r; \theta, \omega)$ as the likelihood function of a selected IRT model. The maximum likelihood estimation can then be given as:

$$(\hat{\boldsymbol{\omega}}, \hat{\boldsymbol{\theta}}) = \mathsf{argmax}_{(\boldsymbol{\omega},\boldsymbol{\theta})} \sum_{i=1}^{N} \ln g(r_i; \omega_{d_i}, \theta_{f_i}) \tag{11}$$

Among specific applications, we can also see a Bayesian treatment [19, 5], where the aim is to calculate the full posterior of the parameters, rather than the maximum likelihood solution. In this work, we primarily use the maximum likelihood fitting in order to keep the computational cost manageable.

2.2 Computerised Adaptive Testing

The fundamental idea of CAT is that, rather than testing a participant with all the questions or a random sequence of questions, the participant is given questions selected in real-time (difficulty) based on the current estimation of ability. We can then update the ability estimation with the response to the selected question, and proceed to select the next question. Therefore, it is quite common to apply CAT based on a pre-trained IRT model, where we have estimated the difficulties (and other parameters) and abilities on a pool of items/participants.

As a result, most CAT approaches include three main components: an IRT model, an item selection method, and an item exposure method. As the name suggests, an item selection method determines, given the current ability estimation, how we select an item with appropriate difficulty to be the next question, so that we can estimate the ability better. Intuitively, we do not want the item to be too complicated or too simple for the actual ability, as for both cases, the responses do not give much information on the one's ability. We introduce two common item selection methods in the following sections.

The item exposure method, on the other hand, controls the marginal frequency/probability that a particular item appears to the participants. The motivation is that we do not want a small number of questions to be exposed to the participants continually. Such high exposure can potentially leak these questions to further participants hence affects later responses. In this work, we focus on the item selection criterion and provide some discussion on item exposure methods at the end of the paper.

Fisher item information

We start with the most commonly adopted approach for item selection, which uses the criterion of Fisher information [10, 2]. Given the current candidate ability θ, a selected IRT model with the likelihood function $g(r;\omega,\theta)$, and a set of L items with parameters $\{\omega_1,\ldots,\omega_L\}$, the Fisher item information on the j_{th} item is then calculated as:

$$\mathsf{I}(\theta;g,\omega_j) = \mathsf{E}_{R\sim\mathbb{P}(\omega_j,\theta)}\Big[\big(\frac{\partial \ln g(R;\omega_j,\theta)}{\partial\theta}\big)^2\Big] \tag{12}$$

$$= \int_r \Big(\frac{\partial \ln g(r;\omega_j,\theta)}{\partial\theta}\Big)^2 g(r;\omega_j,\theta)\mathsf{d}\,r \tag{13}$$

Here $\mathbb{P}(\omega_j,\theta)$ refers to the corresponding probability measure of the IRT model. The Fisher item information calculates the variance of the likelihood gradient, so that we can find the item(s) that can potentially change the likelihood function to a larger extent.

Kullback-Leibler item information

The KL item information [3, 2] is constructed based on the Kullback-Leibler divergence between the IRT likelihood g with current ability θ and the one with a updated ability θ_*. The divergence on the j_{th} item with parameter ω_j is defined as:

$$\mathsf{KL}_{\omega_j}(\theta_* \,||\, \theta) = \mathsf{E}_{R\sim\mathbb{P}(\omega_j,\theta)}\Big[\ln\frac{g(R;\omega_j,\theta)}{g(R;\omega_j,\theta_*)}\Big] \tag{14}$$

However, during application time we do not have access to the updated parameter θ_*. Therefore we cannot calculate the KL-divergence directly. As a solution, we consider the potential information from the j_{th} item to be the integrated divergence around the current ability θ, given the fact that the KL divergence is non-negative:

$$\mathsf{KI}(\theta,g,\omega_j) = \int_{\theta_*=\theta-\varepsilon}^{\theta+\varepsilon} \mathsf{KL}_{\omega_j}(\theta_*||\theta)\mathsf{d}\theta_* \tag{15}$$

This KL item information is hence an aggregated gain around the current ability estimation, hence can be used to select the item with maximal information.

Item exposure control

The problem of controlling exposure rate is as follows, assume we have pre-trained an IRT model with L items and M participants, and have picked an item selection method. There exists a marginal Bernoulli distribution with mean e_j, for each item $j\in\{1,\ldots,L\}$, indicating how likely an item appears to the participants. Since most item selection method tends to prioritise items that are most helpful to quantify participant abilities, we can expect for some items this number will be close to one, that is, the item tends to be selected for every participant. In contrast, there will also be items that are unlikely to be assigned at all.

To solve the above issue and ensure the robustness of the testing, the Sympson-Hetter method [16] proposes to define a maximal exposure rate $\lambda_j\in[0,1]$ for each item j. We can then define another Bernoulli distribution with parameter $\tau_j\in[0,1]$ so that $\tau_j\cdot e_j\le\lambda_j$. For implementation, since e_j is not known for a specific item selection method, we usually approximated it in an online manner with existing assignment counts for each item.

An alternative item selection-exposure method is the discrimination-stratified multistage (or a-stratified multistage) [4]. It controls the exposure rate by dividing items into several groups according to their discrimination parameters. During the procedure, a participant goes from the highest discrimination group to the lowest discrimination group. Within each group, we select an item purely based on the distance between current ability and item difficulties. As this approach avoids considering the information from the discrimination, the exposure rate of all items tends to be more balanced. Another benefit is that the item selected in this method does not require an information criterion to be defined, as it only requires to compare ability and difficulty.

2.3 Applications in Machine Learning

There has been some recent work adopting the IRT framework for machine learning model analysis [13, 9, 5]. All three apply IRT on a model-instance level, that is, seeing a model as a participant and treating an instance (within a given dataset) as an item. In [13, 9] the authors use the Logistic model and discuss the interpretation of the learnt IRT parameters, including models like the always-correct model (e.g. model predicts the ground truth). The response reflects whether a model correctly predicts the target class. In [5], the authors propose the three-parameter Beta model and learn its parameter in a Bayesian setting (e.g. posterior of the parameters). As the Beta IRT model supports bounded continuous response, in [5], the authors selected the predicted probability of the correct class as the response.

3 Modifications

We now introduce some modifications on top of existing IRT and CAT methods so that we can apply them to the problem of model-dataset evaluation. In general, we consider the following two require-

ments for the IRT and CAT methods. (1) They should support standard machine learning evaluation metrics, that is, to support the modelling of continuous gain/loss measures. (2) The corresponding item information should be obtainable analytically or through efficient approximations. Furthermore, we discuss the preference for non-negative discrimination in the scenario of a model-dataset benchmark.

3.1 Modified logistic IRT

The first modification is on the logistic IRT family. Due to its original application scenario, the logistic IRT family was mainly used to model binary responses. As introduced above, to support CAT with a continuous response, the IRT needs to model a continuous response and provides the corresponding likelihood. The original logistic IRT works on a Bernoulli assumption and the model calculate a mean parameter in the closed interval $[0,1]$. While in the case of Bernoulli distribution, the mean parameter is sufficient to calculate the likelihood, we need to consider other parameterisation for the continuous case. Although the Beta-3 IRT model uses the Beta likelihood and supports continuous response by default, it would also be valuable to keep an IRT model with sigmoid shape for better comparison. To achieve this, we hence replace the Bernoulli assumption with a logit-normal assumption in the IRT model. We use the orginal logistic function to calculate the mean of the response, and add a extra s parameter as the standard deviation:

$$R \mid \Theta = \theta, \Delta = \delta, A = a, S = s \sim \textbf{Logit-normal}(\mu_{(\theta,\delta,a)}, \sigma_{(s)}) \tag{16}$$

$$\mu_{(\theta,\delta,a)} = -a \cdot (\theta - \delta) \tag{17}$$

$$\sigma_s = s \tag{18}$$

The likelihood is then given as:

$$p(r \mid \theta, \delta, a, s) = \frac{1}{\sqrt{2\pi s^2}} \frac{1}{r(1-r)} \exp\Big(- \frac{\big(\ln(\frac{1-r}{r}) + a \cdot (\theta - \delta)\big)^2}{2 * s^2} \Big) \tag{19}$$

While the expectation and variance are not analytically available, they can be obtained by importance sampling:

$$\mathrm{E}[R \mid \theta, \delta, a, s] = \frac{1}{Q} \sum_{i=1}^{Q} \frac{1}{1 + \exp(\tilde{r}_i)} \tag{20}$$

$$\mathrm{Var}[R \mid \theta, \delta, a, s] = \frac{1}{Q} \sum_{i=1}^{Q} \Big(\frac{1}{1 + \exp(\tilde{r}_i)} - \mathrm{E}[R \mid \theta, \delta, a, s] \Big) \tag{21}$$

$$\tilde{r}_i \sim \textbf{Normal}(-a \cdot (\theta - \delta), s) \tag{22}$$

With these modifications, the IRT model and corresponding CAT approaches can work with any bounded continuous response. While other possible extensions support continuous response [15, 14], we experimented particularly with the logistic and Beta-3 models given their close connection.

3.2 Approximate item information with importance sampling

The second minor modification also aims to incorporate continuous response. While using binary responses, both Fisher item information and KL item information are available analytically. These analytical results generally are no longer reachable when switching to IRT models with continuous response. However, as the integration in both Fisher item information and KL item information is to calculate an expectation upon a density function, we can approximate them with importance sampling. We hence use this sampling solution for both the logistic and Beta-3 IRT in the following experiments.

3.3 Constraint of non-negative discrimination

An implicit assumption for traditional IRT and CAT applications is that a given item only be tested on a participant once. This assumption is intuitive if we consider the student examination scenario, where the student will tend to remember the question after seen it multiple times, and most likely to result in the same response for any repeated question. The same interpretation applies to the work on IRT with model-instance combinations, where it gains little information when we ask the model to predict the same instance twice. For such settings, it is understandable that specific items might have negative discrimination where stronger participants tend to make the wrong response.

However, in the scenario of model-dataset testing, as seen in many research experiments, repeated experiments can often provide useful statistical information. Using the student examination example again, in machine learning experiments, we include both teaching and testing, but only test on a pre-trained student. Therefore, one key difference between model-dataset testing from existing IRT applications is that we are evaluating a learning process. For each test, instead of having a pre-trained participant to respond to items, we always start from a blank participant and train them before gathering the responses. As long as we assume this, there is a learnable pattern from the dataset, and we can further assume a more robust model should statistically have better performances, hence non-negative discrimination. The worst case would be a dataset containing random noise, and hence gives 0 discrimination. In practice, we can either achieve this via constrained optimisation during the estimation of IRT parameters or directly to estimate the logarithm of discrimination parameters via unconstraint approaches. Alternatively, we can also do it the Bayesian way, which assumes positive discrimination is more likely via the prior distribution. However, as we only consider the maximal likelihood case in this paper, we leave this option as future work.

4 Experiments

In this section, we experimentally investigate the performance of the IRT and CAT methods discussed above. As an initial investigation, we address the following questions. (1) Given a set of standard machine learning evaluation measures, which types of IRT provides better modelling on the responses? (2) For similar settings, which item selection method provides the most benchmarking efficiency?

We first introduce the settings of our experiments. Then, on the IRT side, we compare the inference errors on responses using a standard train-test split setting. Regarding the CAT methods, we compare them on three types of results: (1) The decay of the mean squared error (MSE) on the inferred response, given a validation set. (2) The decay of negative log-likelihood (NLL) on inferred responses, given a validation set. (3) The ranks of the selected datasets.

4.1 Setup

We select six standard evaluation measures: (1) multi-class accuracy, (2) Brier score, (3) Log-loss, (4) weighted averaged binary accuracy, (5) weighted averaged binary AUC, (6) weighted averaged binary

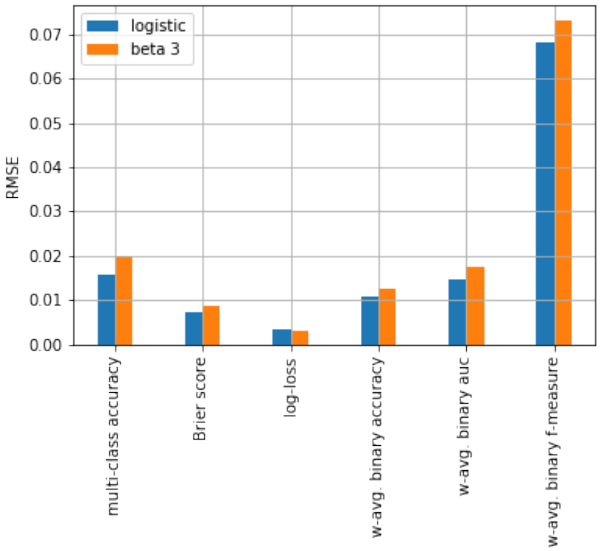


Figure 1: Inference errors of the IRT models on different evaluation measures

F-measure. Here all the losses are bounded with $[0, 1]$ except the log-loss. We hence perform post-processing on the log-loss. We define an upper bound by calculating the logarithm of a tiny positive number, then the entire bound is rescaled to the range of $[0, 1]$. Furthermore, we use the negative value of Brier score and log-loss to fit the IRT models, so that they become measures on gain (e.g. larger values indicate better results), in line with the other evaluation measures.

We select a set of datasets and model classes (described below), and run each model-dataset combination with 50% random train-test split ten times. We use these results to train both Beta-3 and logistic IRT models.

For adaptive testing we selected the gradient boosting classifier as the candidate model, and run it with all the datasets ten times using the same setting above. These results are using as a validation set. During the adaptive testing, each time we update the model ability, we use the trained IRT to infer an expected response/pdf on response. We also calculate the corresponding MSE / NLL with the validation set and compare different IRT and CAT approaches. In principle, a better IRT-CAT combination should have a lower inferred error, as well as a faster convergence speed to the final MSE / NLL.

We use the 165 datasets provided by PMLB [12], which is a pre-processed collection of UCI datasets on various classification tasks. For computational efficiency, for all the datasets with more than $10,000$ instances, we sample it down to $10,000$ instances while approximately keeping the marginal distribution of the target variable. We leave testing larger datasets for future work.

We selected 9 model classes from the sklearn package: (1) multi-layer perceptron (MLP), (2) K nearest neighbours (KNN), (3) support vector machine (SVM), (4) pseudo Gaussian process (GP), (5) decision tree (TREE), (6) random forest (RF), (7) Ada boosting (ADA), (8) naive Bayes (NB), (9) logistic regression (LR).

For each model class, we selected eight different parameter settings to form different model instances, resulting in a total number of 72 models. For instance, for the MLP we choose various numbers of hidden units in a two-layer setting. Regarding the GP, here we call it pseudo models as sklearn does not support any sparse modelling. We hence perform a simple random sampling on the training set. We first randomly select one data point for each class, then further sample random data points from the entire training set.

4.2 Comparing IRT approches

The first experiment we performed was to investigate whether the IRT models can accurately model and infer the test results. For this purpose, we perform a random split experiment on the collected responses from the 165 datasets and 72 models. That is, we divided the collected responses into a training set and a test set, and use the training set to train the IRT models, then the test set can be applied to verify the expected responses from each IRT model. Figure 1 shows the results, where we compare them with the root mean squared errors on the inferred performances on the expectations, note here these errors are computed based on the measure after rescaling. As the results show, Logistic IRT achieves a lower RMSE on 5 out of 6 evaluation measures, and Beta-3 IRT only performs better on log-loss. One of the possible reasons is that, after the rescale, log-loss tends to distribute close to the upper bound (e.g. as the lower bound can be hardly achieved by any reasonable model), thus making the Beta distribution more suitable to fit the distribution. Otherwise, the Logistic model tends to be more stable, given its Gaussian assumption.

4.3 Comparing IRT and CAT pairs

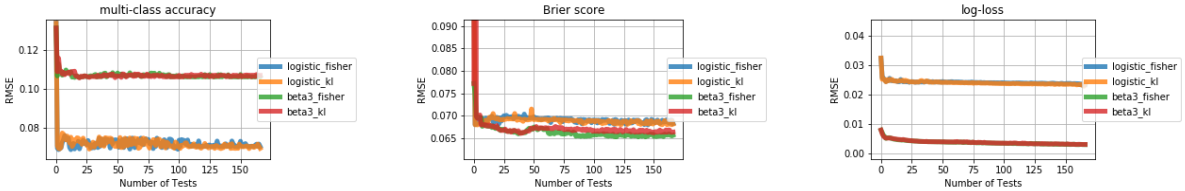
As discussed above, for the second experiment, we use different IRT and item selection approaches to test the candidate gradient boosting classifier. We start the testing by assuming the candidate model to have an averaged ability, then keep testing the model and updating its ability until we have tested all the datasets. At each test step, we record the RMSE and NLL using the validation set. Figure 2 shows the decay of the root mean square error on each IRT model and item selection criterion pairs, and Figure 3 gives the negative log-likelihood.

The results sow that, while the logistic IRT mostly performs better in the previous experiments, here we have more cases where the Beta-3 IRT achieves a lower inference error, as seen e.g. with F-measure. In general, the performance on RMSE and NLL seem to be dominated by the IRT models. These results are mostly different when we switch from the Logistic IRT model to the Beta-3 IRT model or vice versa, while not being affected much by different item selection criterion. In terms of decaying speed, most results show an immediate decay within the first five tests, and the rest tests do not contribute much on the inference error. Exceptions include the F-score for both MSE and NLL, as well as the log-loss in NLL.

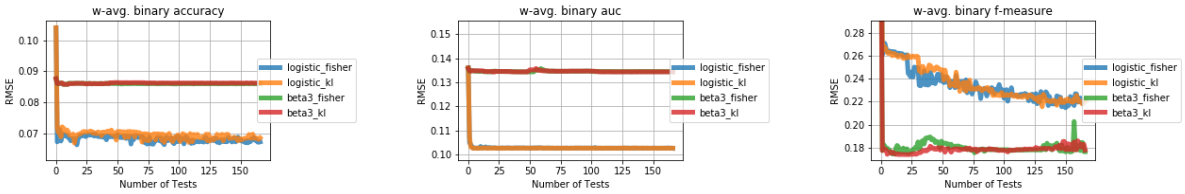
To further understand the selected test sequence from each IRT and item selection approach, here we further calculate the pair-wise rank correlation (Kendall's tau), and the results are in Figure 4. On the top level, it shows that the test sequences tend to be similar between the same IRT model while being significantly uncorrelated between the Beta-3 IRT and Logistic IRT. This observation indicates that the item selection methods play a less important role here, and this is mainly the case for the Logistic IRT. A strong correlation appears between the Fisher item information and KL item information among all the evaluation measures. However, Beta-3 IRT only shows a similar level of correlation in multi-class accuracy, and the correlation is weaker among other evaluation measures.

5 Discussion

From the results above, here we discuss several messages. (1) IRT models have a critical role in the benchmark. A suitable IRT model

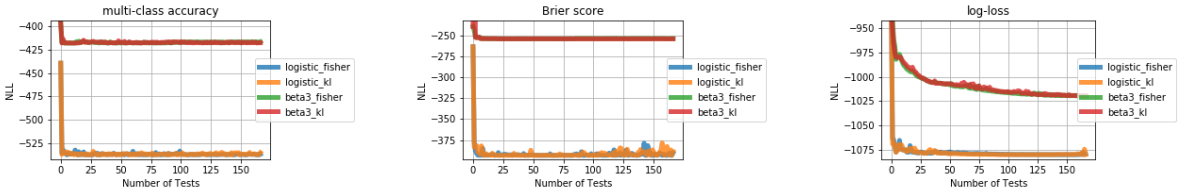


(a) Multi-class accuracy (b) Brier score (c) Log-loss

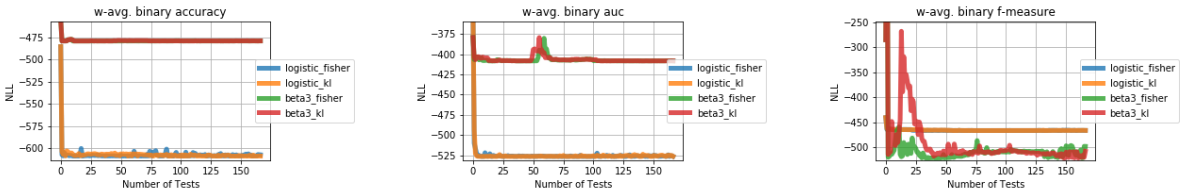


(d) Weighted averaged binary accuracy (e) Weighted averaged binary AUC (f) Weighted averaged binary F-measure

Figure 2: Root mean squared error of the adaptive testing sequence of the gradient boosting classifier



(a) Multi-class accuracy (b) Brier score (c) Log-loss



(d) Weighted averaged binary accuracy (e) Weighted averaged binary AUC (f) Weighted averaged binary F-measure

Figure 3: Negative log-likelihood of the adaptive testing sequence of the gradient boosting classifier

can indeed lead to better inference on the test results, without spending much effort on further testing. However, neither IRT model experimented with in this paper appear to dominate on all evaluation measures, suggesting considerable room for further investigation. (2) Adaptive testing can effectively reduce the total number of experiments. For most evaluation measures, we can indeed observe a significant decay on the inference error with a small number of tests. (3) Item exposure control is worth further consideration in the benchmarking process. The trained IRT model shows a reliable power on dominating the test sequences, meaning the same IRT will prefer to suggest the same combination of datasets at the beginning of the benchmark. While we didn't explore it in this paper, such behaviour can potentially lead to over-fitting, as it encourages the developers to focus on the performance on the first few datasets.

REFERENCES

[1] David M Blei, Alp Kucukelbir, and Jon D McAuliffe, 'Variational inference: A review for statisticians', *Journal of the American statistical Association*, **112**(518), 859–877, (2017).

[2] Hua-Hua Chang, 'Psychometrics behind computerized adaptive testing', *Psychometrika*, **80**(1), 1–20, (2015).

[3] Hua-Hua Chang and Zhiliang Ying, 'A global information approach to computerized adaptive testing', *Applied Psychological Measurement*, **20**(3), 213–229, (1996).

[4] Hua-Hua Chang and Zhiliang Ying, 'A-stratified multistage computerized adaptive testing', *Applied Psychological Measurement*, **23**(3), 211–222, (1999).

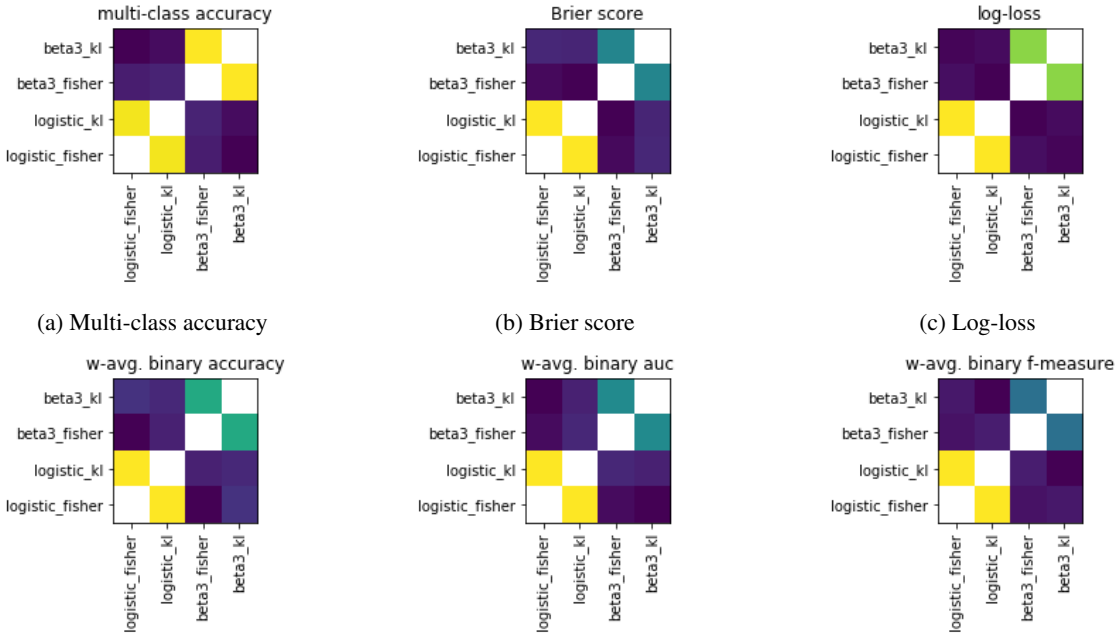
[5] Yu Chen, Telmo Silva Filho, Ricardo B. Prudencio, Tom Diethe, and Peter Flach, 'β^3-IRT: A new item response model and its applications', in *AISTATS 2019*, eds., Kamalika Chaudhuri and Masashi Sugiyama, volume 89 of *Proceedings of Machine Learning Research*, pp. 1013–1021, (2019).

[6] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, 'Generative adversarial nets', in *Advances in neural information processing systems*, pp. 2672–2680, (2014).

[7] Bert F Green, R Darrell Bock, Lloyd G Humphreys, Robert L Linn, and Mark D Reckase, 'Technical guidelines for assessing computerized adaptive tests', *Journal of Educational Measurement*, **21**(4), 347–360, (1984).

[8] Matthew Hutson. Artificial intelligence faces reproducibility crisis, 2018.

[9] Fernando Martínez-Plumed, Ricardo BC Prudêncio, Adolfo Martínez-Usó, and José Hernández-Orallo, 'Making sense of item response theory in machine learning', in *Proceedings of the Twenty-second European Conference on Artificial Intelligence*, pp. 1140–1148. IOS Press, (2016).



(a) Multi-class accuracy (b) Brier score (c) Log-loss

(d) Weighted averaged binary accuracy (e) Weighted averaged binary AUC (f) Weighted averaged binary F-measure

Figure 4: Kendall's Tau between the adaptive testing sequences of the gradient boosting classifier, a brighter yellow colour indicates a stonger correlation and a deeper blue colour corresponds to a weaker correlation.

[10] Rob R Meijer and Michael L Nering. Computerized adaptive testing: Overview and introduction, 1999.

[11] Gary A Morris, Lee Branum-Martin, Nathan Harshman, Stephen D Baker, Eric Mazur, Suvendra Dutta, Taha Mzoughi, and Veronica McCauley, 'Testing the test: Item response curves and test quality', *American Journal of Physics*, **74**(5), 449–453, (2006).

[12] Randal S Olson, William La Cava, Patryk Orzechowski, Ryan J Urbanowicz, and Jason H Moore, 'Pmlb: a large benchmark suite for machine learning evaluation and comparison', *BioData mining*, **10**(1), 36, (2017).

[13] Ricardo BC Prudêncio, José Hernández-Orallo, and Adolfo Martınez-Usó, '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*, (2015).

[14] Fumiko Samejima, 'Graded response model', in *Handbook of modern item response theory*, 85–100, Springer, (1997).

[15] Kojiro SHOJIMA, 'A noniterative item parameter solution in each em cycle of the continuous response model', *Educational technology research*, **28**(1-2), 11–22, (2005).

[16] JB Sympson and RD Hetter, 'Controlling itemexposure rates in computerized adaptive testing, as described in wainer, et al.,(1990)', (1985).

[17] Wim J van der Linden and Ronald K Hambleton, *Handbook of modern item response theory*, Springer Science & Business Media, 2013.

[18] Joaquin Vanschoren, Jan N Van Rijn, Bernd Bischl, and Luis Torgo, 'OpenML: networked science in machine learning', *ACM SIGKDD Explorations Newsletter*, **15**(2), 49–60, (2014).

[19] Bernard P Veldkamp and Mariagiulia Matteucci, 'Bayesian computerized adaptive testing', *Ensaio: Avaliação e Políticas Públicas em Educação*, **21**(78), 57–82, (2013).

[20] David J Weiss and G Gage Kingsbury, 'Application of computerized adaptive testing to educational problems', *Journal of Educational Measurement*, **21**(4), 361–375, (1984).