Implicit Performance Estimation for Score-Based Classifiers Using Cognitive Diagnosis

A Preprint

Nikita Breskanu
Lomonosov Moscow State University
nbreskanu73@gmail.com

Archil Maisuradze
Lomonosov Moscow State University
artchil@mail.ru

Abstract

Score-based binary classifiers are widely used in machine learning. When it comes to their validation, well-known performance metrics, such as ROC-AUC, F1-score, and Accuracy, are most often used. However, all these metrics have their flaws and represent the performance of the classifier only from a certain angle. This work attempts to aggregate all traditional performance metrics using cognitive diagnosis models. Cognitive diagnosis models are widely researched in smart education and proved to be successful in estimating students' latent knowledge from their exercise solutions. In the context of binary classification, classifiers can be viewed as students and performance metrics as exercises. This reduction represents a novel approach to the validation of binary classifiers and produces latent knowledge attributes, which can be interpreted as new implicit performance attributes.

Keywords Validation · Binary classification · Cognitive diagnosis · Item Response Theory · Machine learning

1 Introduction

The validation is an important step in the machine learning models lifecycle. For this reason, many performance metrics (Accuracy, F1-score, ROC-AUC, etc.) have been developed. However, none of them can entirely characterize the model behavior [6, 5], and there is no clear agreement on which of them to use. It may be impossible to get the true attributes of the model by performing direct aggregation of the answers.

In psychometrics, it is believed that the desired attributes of the subject are only partially manifested in direct measurements. Applying this idea to ML validation, researchers actively try to create a better performance metric by using Item Response Theory (IRT) [17], a classical tool of the psychometrician [10, 12, 3]. However, IRT estimates only one latent attribute, which is not enough to completely describe the model, and the only non-one-dimensional approach approach only estimated the Recall equivalents in multi-class classification [8].

Validating models by their performance metrics can be reduced to the cognitive diagnosis task in smart education, where the goal is to estimate certain predefined attributes of the students by their exercise solutions. For the latter, a wide variety of models have been developed [9, 4, 16, 7]. This work is the first attempt to apply cognitive diagnosis models to create new, potentially better metrics from traditional ones.

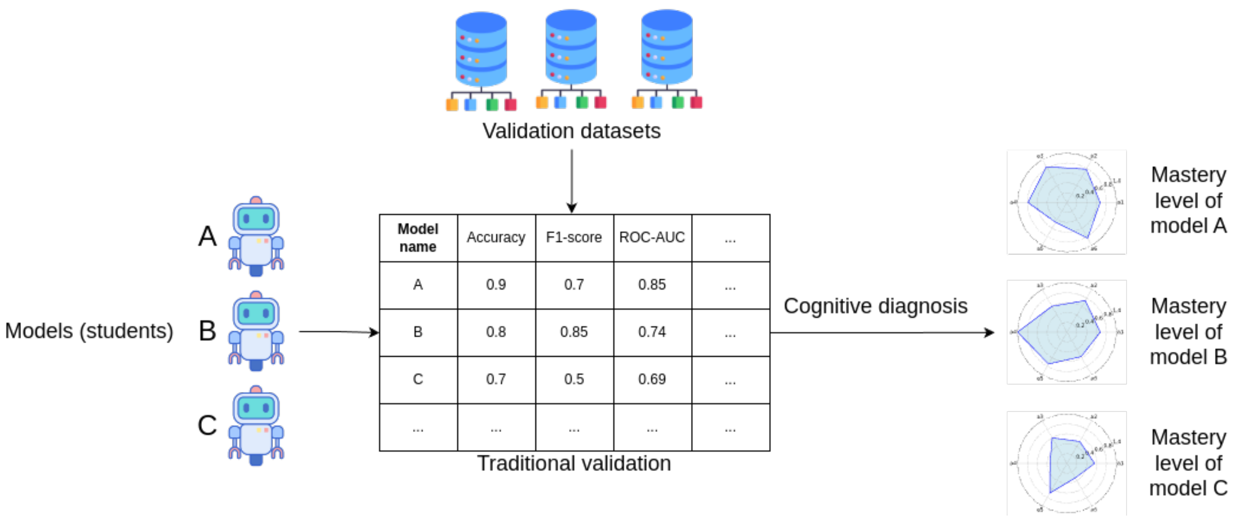

Figure 1: Validation using cognitive diagnosis framework

Our task is to create a framework that would allow estimating mastery of machine learning models for each predefined skill. For that problem, first traditional performance metrics from validation datasets are calculated, and then cognitive assessment is performed, which assigns mastery for each skill to every model. It's important to mention that this approach only works for a pool of models, not for a single one, due to the nature of cognitive diagnosis models.

We defined specific skills, and then performed experiments on score-based classifiers to obtain their cognitive mastery levels (1). The new metrics (mastery levels) turned out to be competitive with the traditional ones, describing the model from a slightly different angle and considering other models results, and, most importantly, allow using multiple validation datasets, thus capturing model behavior in different learning contexts.

The proposed validation framework can be used to perform validation and comparison of multiple binary classifiers. We also propose a method for adding newly created model to the pool of existing ones, and obtaining it's mastery levels. This might open the way to creating multi-skill ML model leaderbords, capturing multiple various datasets.

2 Related work

It is known that traditional performance metrics can't fully derscribe the model's performance, and each one of them has their flaws [5]. For example, in binary classification, Accuracy does not see the difference between errors in the positive and negative classes; Precision and Recall do not know the number of correctly identified negative classes (True Negative); ROC-AUC is sensitive to class imbalance [6].

In 2016, the first attempt of applying psychometric tools for ML validation was made [10], where the author tried to apply IRT [17] for estimated a better version of Accuracy for multi-class classifiers. That study created a great interest for other researchers in applying IRT in machine learning. IRT-based ensembles were proposed [3], where the weights are the IRT scores. IRT-based leaderboard for NLP models validation [12]. Researchers also tried to use IRT to reduce the validation dataset [11], or to make manual validation more efficient [14]. An attempt was made to use IRT for clustering examples in multi-dataset NLP benchmarks [13].

One of the advantages of using IRT for evaluation is that it assigns parameters to every item (object in the dataset), which can later be used to enhance interpretation [12]. The framework for estimating this parameters for newly generated questions was proposed [2].

Another widely researched area is the cognitive diagnosis task, where it is required to estimate students' mastery levels on every predefined skill by looking at their exercise solutions, and exercise-skill correspondence matrix, which is also known as Q-matrix. Previously, only classical models like DINA [4] or multidimensional IRT (MIRT) [15] models were used. But in 2022, there was a first attempt of using deep cognitive diagnosis model with trainable interaction function, this deep model was called NeuralCD [16]. Later, a lot of extensions of NeuralCD appeared, which were designed to fix some of its flaws, most apparent of which is the lack of knowledge association [16, 7, 9].

To our knowledge, there has been only one attempt of using cognitive diagnosis models for machine learning models validation. In 2023, Camilla framework was proposed for validating deep computer vision multi-class classifiers [8]. The authors estimated the new equivalents of respective Recalls for each class and argued that they describe the performance better by taking into account difficult and easy samples. However, despite their success, we believe that for binary classifiers estimating 2 Recall equivalents is not enough for the full description of the model performance. Our approach is different from the described above in several ways:

- Binary classification task is considered instead of the multi-class.
- Performance metrics are used as exercises instead of the objects.
- Multiple validation datasets are used instead of one, with retraining classifiers for each dataset. This can potentially test the algorithm performance in different learning contexts.

3 Problem statement

Table 1: Definitions

Quantity	Description
$\mathcal{D} = \{\mathcal{D}_1, \ldots, \mathcal{D}_L\}$	Set of datasets
$\mathcal{S} = \{s_1, \ldots, s_N\}$	Set of models (students)
$\hat{\mathcal{E}} = \{\hat{e}_1, \ldots, \hat{e}_{\hat{M}}\}$	Set of performance metrics (exercises)
$\mathcal{E} = \{e_1, \ldots, e_M\}, M = \hat{M} \times D$	Set of performance metrics, taking datasets into account
$\mathcal{K} = \{\mathcal{K}_1, \ldots, \mathcal{K}_K\}$	Set of concepts
L	Number of datasets
N	Number of models (students)
$\hat{M}$	Number of performance metrics (exercises)
K	Number of concepts
T	Number of response logs
$l \in \{1, \ldots, L\}$	Index of the dataset
$i \in \{1, \ldots, N\}$	Index of the student
$j \in \{1, \ldots, M\}$	Index of the exercise
$k \in \{1, \ldots, K\}$	Index of the concept
$t \in \{1, \ldots, T\}$	Index of the response log
$x^s \in \{0, 1\}^N$	One-hot representation of the model (student)
$x^e \in \{0, 1\}^M$	One-hot representation of the performance metric (exercise)
$Q = \{Q_{jk}\}_{\hat{M} \times K} \in [0, 1]^{\hat{M} \times K}$	Q-matrix
$G \in \{0, 1\}^{K \times K}$	Directed Acyclic Graph (DAG) of concept dependency
$y \in [0, 1]$	Model output
$R = \{(x^s_t, x^e_t, r_t)\}_{t=1}^T$	Response logs
$r \in [0, 1]$	Result of solving the exercise (value of the performace metric)
$\mathcal{M} = \{m_{ik}\}_{N \times K} \in [0, 1]^{N \times K}$	Latent students' mastery levels
$\mathcal{L}$	Loss function

Task definition *Suppose there are L datasets $\mathcal{D} = \{\mathcal{D}_1, \ldots, \mathcal{D}_L\}$, N machine learning models (algorithms) $\mathcal{S}\{s_1, \ldots, s_N\}$, $\hat{M}$ traditional performance metrics $\hat{\mathcal{E}} = \{\hat{e}_1, \ldots, \hat{e}_{\hat{M}}\}$, and K predefined skills (or concepts) $\mathcal{K} = \{\mathcal{K}_1, \ldots, \mathcal{K}_K\}$. Q-matrix $Q \in \mathbb{R}^{\hat{M} \times K}$ is a binary matrix that represents correspondence between performance metrics and concepts: $Q_{jk} = 1 \iff$ knowledge of concept $\mathcal{K}_k$ is required for having a high value of $\hat{e}_j$. Computing performance metrics $\hat{\mathcal{E}}$ for each dataset forms a set of response logs $R = \{(x^s, x^e, r)\}_{t=1}^T$, where $T = N \times M$, $M = \hat{M} \times L$ — a triples consisting of one-hot representation of model (student), performance metric (exercise), taking the index of dataset in the account, and the metric $r \in [0, 1]$, normalized to [0, 1]. The desired models' mastery levels for all concepts can be represented as matrix $\mathcal{M} = \{m_{ik}\} \in [0, 1]^{N \times K}$, where m_{ik} is the mastery level of student s_i for the concept $\mathcal{K}_k$; $m_{ik} = 1$ represents total knowledge of the concept, and $m_{ik} = 0$ — total ignorance. The task is to infer the mastery matrix $\mathcal{M}$ using cognitive diagnosis model by predicting the responses r_t.*

4 Cognitive diagnosis dataset preparation

4.1 Defining datasets

Cognitive models need a large pool of students and exercises, and according to our reduction, students are classifiers and exercises are performance metrics. So, to prepare a dataset for cognitive diagnosis, one needs to evaluate a large number of models on large number of metrics.

In order to test models in different learning scenarios and simultaneously create more exercises, we collected diverse open-source datasets from OpenML. In fact, we used a subset of datasets from paper [1]. The reason for dropping some of the datasets is too large computational complexity when there are more than 1000 features. In total, we have 16 datasets (2).

For each dataset, we transformed it into a pipeline that would be acceptable for every classifier:

- One-hot encoding was performed for all categorical features.
- Standard scaling was performed for all numerical features.

Table 2: Dataset characteristics. All datasets are taken from OpenML. Some of them contain categorical features which will be one-hot encoded. Datasets have diverse class balances: one of them has 99-1.

Dataset name	Samples × features	Numerical × categorical features	Class balance
Banknote-authentication	1372 × 5	5 × 0	55–45%
Blood-transfusion-service-center	748 × 5	5 × 0	76–24%
Breast-w	683 × 10	10 × 0	65–35%
Climate-model-simulation-crashes	540 × 21	21 × 0	99–1%
Cylinder-bands	277 × 40	25 × 15	64–36%
Dresses-sales	99 × 13	2 × 11	59–41%
Diabetes	768 × 9	9 × 0	65–35%
ilpd	583 × 11	10 × 1	71–29%
kc1	2109 × 22	22 × 0	84–16%
kc2	522 × 22	22 × 0	79–21%
pc1	1109 × 22	22 × 0	93–7%
pc3	1563 × 38	38 × 0	89–11%
Phoneme	5404 × 6	6 × 0	70–30%
qsar-biodeg	1055 × 42	42 × 0	66–34%
wdbc	569 × 31	31 × 0	62–38%
wilt	4839 × 6	6 × 0	94–6%

4.2 Defining classifiers

We defined 295 binary classifiers by varying hyperparameters. In order to create more diversity, similarly to [10], several artificial classifiers were implemented:

- Optimal classifier always predicts the correct class (either 0 or 1).
- Pessimal classifier always predicts the incorrect class (either 0 or 1).
- Majority classifier always predicts the majority class (either 0 or 1).
- Minority classifier always predicts the minority class (either 0 or 1).
- Mean target classifier always predicts the mean target value. For example, in 90-10 class balanced it will always output 0.9.
- Uniform random classifier predicts classes randomly with equal probabilities.
- Balanced random classifer predicts classes randomly with probabilities proportional to their class balance. For example, in 90-10 class balance there will be a 90% probability of class 0.

Table 3: Binary classifiers, replicated by varying hyperparameters.

Classifier	Implementation	Varying parameters	Number of models
Logistic regression	sklearn	C, solver	120
Decision tree	sklearn	max_depth, criterion	60
Random forest	sklearn	max_depth, n_estimators	12
Gradient boosting	sklearn	n_estimators, learning_rate	9
Gradient boosting	LGBM	n_estimators, num_leaves	9
SVM	sklearn	C, kernel	30
K nearest neighbors	sklearn	n_neighbors, weights	40
Multilayer perceptron	sklearn	hidden_layer_sizes, activation	15
Optimal classifier	<manual>	<absent>	1
Pessimal classifier	<manual>	<absent>	1
Majority classifier	<manual>	<absent>	1
Minority classifier	<manual>	<absent>	1
Mean target classifier	<manual>	<absent>	1
Uniform Random classifier	<manual>	<absent>	1
Balanced Random classifier	<manual>	<absent>	1

In total, there are 302 score-based binary classifers.

4.3 Defining performance metrics

Performance metrics that were used, are defined in table (4). We used Equal Error Threshold (EER, a point where 2 recalls are the same) to obtain label-based metrics in conjunction with default score-based ones.

There was a possibility of using also default 0.5 threshold, but since not all the classifiers are self-calibrated, we decided not to include it. Moreover, we found that EER threshold metrics were very similar to the respective 0.5 threshold ones.

Table 4: Performance metrics used for score-based classifiers

Performance metric	Definition	Description
ROC-AUC	AUC	Area under the ROC curve
PR-AUC for class 0	PRAUC0	Area under the PR curve, where class 0 is the positive class
PR-AUC for class 0	PRAUC1	Area under the PR curve, where class 0 is the positive class
Gain chart AUC for class 0	GCAUC0	Area under the gain chart, where class 0 is the positive class
Gain chart AUC for class 1	GCAUC1	Area under the gain chart, where class 1 is the positive class
KS statistic	KS	Kolmogorov-Smirnov statistic
Kendall's tau	KTAU	Kendall's correlation between predicted and actual results
Accuracy (EER)	ACC	Accuracy at EER threshold
Precision for class 0 (EER)	PR0	Precision with 0 as positive class at EER threshold
Precision for class 1 (EER)	PR1	Precision with 1 as positive class at EER threshold
Recall (EER)	REC	Recall at EER threshold
Balanced accuracy (EER)	BA	Balanced accuracy at EER threshold
F1-score for class 0 (EER)	FS0	F1-score with 0 as positive class at EER threshold
F1-score for class 1 (EER)	FS1	F1-score with 1 as positive class at EER threshold
Average F1-score (EER)	AVGFS	Average F1-score at EER threshold
Fowlkes–Mallows index for class 0 (EER)	FM0	Fowlkes–Mallows index with 0 at positive class at EER threshold
Fowlkes–Mallows index for class 1 (EER)	FM1	Fowlkes–Mallows index with 1 at positive class at EER threshold
Markedness (EER)	MKNS	Markeness at EER threshold
MCC (EER)	MCC	Matthews correlation coefficient at EER threshold
Jaccard index (EER)	JAC	Jaccard index at EER threshold
Cohen's kappa (EER)	KAPPA	Cohen's kappa at EER threshold

After training all classifiers on all datasets and obtaining performance metrics, there appeared to be 54 classifier duplicates, mostly logistic regressions. After removing them, we obtained the final 248 classifiers $\times$ 353 metrics dataset, ready for cognitive assessment.

5 Requirements for mastery levels

We propose several interpretability requirements for the mastery levels (new metrics generated by the cognitive model):

- Mastery levels (new metrics) are expected fully cover the old metrics, i.e. traditional metrics are expected to be derived from mastery levels by using some formula.
- If one classifier has higher mastery level on some concept $\mathcal{K}_k$ than the other classifier, it is expected to have higher corresponding performance metrics $e_j : Q_{jk} = 1$ than the other.
- If one classifier has higher performance metric e_j than the other classifier, it is expected to have higher corresponding mastery levels $\mathcal{K}_k : Q_{jk} = 1$ than the other.

References

[1] Vitor Cirilo Araujo Santos, Lucas Cardoso, and Ronnie Alves. The quest for the reliability of machine learning models in binary classification on tabular data. *Scientific Reports*, 13(1):18464, 2023.

[2] Luca Benedetto, Andrea Cappelli, Roberto Turrin, and Paolo Cremonesi. R2de: a nlp approach to estimating irt parameters of newly generated questions. In *Proceedings of the tenth international conference on learning analytics & knowledge*, pages 412–421, 2020.

[3] Ziheng Chen and Hongshik Ahn. Item response theory based ensemble in machine learning. *International Journal of Automation and Computing*, 17:621–636, 2020.

[4] Jimmy De La Torre. Dina model and parameter estimation: A didactic. *Journal of educational and behavioral statistics*, 34(1):115–130, 2009.

[5] Peter Flach. Performance evaluation in machine learning: the good, the bad, the ugly, and the way forward. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 9808–9814, 2019.

[6] Nathalie Japkowicz. Why question machine learning evaluation methods. In *AAAI workshop on evaluation methods for machine learning*, volume 6. University of Ottawa, 2006.

[7] Jiatong Li, Fei Wang, Qi Liu, Mengxiao Zhu, Wei Huang, Zhenya Huang, Enhong Chen, Yu Su, and Shijin Wang. Hiercdf: A bayesian network-based hierarchical cognitive diagnosis framework. In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, pages 904–913, 2022.

[8] Qi Liu, Zheng Gong, Zhenya Huang, Chuanren Liu, Hengshu Zhu, Zhi Li, Enhong Chen, and Hui Xiong. Multi-dimensional ability diagnosis for machine learning algorithms. *arXiv preprint arXiv:2307.07134*, 2023.

[9] Shuo Liu, Hong Qian, Mingjia Li, and Aimin Zhou. Qccdm: A q-augmented causal cognitive diagnosis model for student learning. In *ECAI 2023*, pages 1536–1543. IOS Press, 2023.

[10] Fernando Martínez-Plumed, Ricardo BC Prudêncio, Adolfo Martínez-Usó, and José Hernández-Orallo. Making sense of item response theory in machine learning. In *ECAI 2016*, pages 1140–1148. IOS Press, 2016.

[11] Felipe Maia Polo, Lucas Weber, Leshem Choshen, Yuekai Sun, Gongjun Xu, and Mikhail Yurochkin. tinybenchmarks: evaluating llms with fewer examples. *arXiv preprint arXiv:2402.14992*, 2024.

[12] Pedro Rodriguez, Joe Barrow, Alexander Miserlis Hoyle, John P Lalor, Robin Jia, and Jordan Boyd-Graber. Evaluation examples are not equally informative: How should that change nlp leaderboards? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4486–4503, 2021.

[13] Pedro Rodriguez, Phu Mon Htut, John P Lalor, and João Sedoc. Clustering examples in multi-dataset benchmarks with item response theory. In *Proceedings of the Third Workshop on Insights from Negative Results in NLP*, pages 100–112, 2022.

[14] João Sedoc and Lyle Ungar. Item response theory for efficient human evaluation of chatbots. In *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*, pages 21–33, 2020.

[15] Yanyan Sheng and Christopher K Wikle. Comparing multiunidimensional and unidimensional item response theory models. *Educational and Psychological Measurement*, 67(6):899–919, 2007.

[16] Fei Wang, Qi Liu, Enhong Chen, Zhenya Huang, Yu Yin, Shijin Wang, and Yu Su. Neuralcd: a general framework for cognitive diagnosis. *IEEE Transactions on Knowledge and Data Engineering*, 35(8):8312–8327, 2022.

[17] Frances M Yang and Solon T Kao. Item response theory for measurement validity. *Shanghai archives of Psychiatry*, 26(3), 2014.