A Brief Overview of Metrics for Evaluation of Student Models

Radek Pelánek
Masaryk University Brno
pelanek@fi.muni.cz

ABSTRACT

Many different metrics are used to evaluate and compare performance of student models. The aim of this paper is to provide an overview of commonly used metrics, to discuss properties, advantages, and disadvantages of different metrics, and to summarize current practice in research papers. The paper should serve as a starting point for workshop discussion about the use of metrics in student modeling.

1. INTRODUCTION

A key part of intelligent tutoring systems are models that estimate the knowledge of students. To compare and improve these models we use metrics that measure quality of model predictions. Metrics are also used (sometimes implicitly) for parameter fitting, since many fitting procedures try to optimize parameters with respect to some metric.

At the moment there is no standard metric for model evaluation and thus researchers have to decide which metric to use. The choice of metric is an important step in the research process. Differences in predictions between competing models are often small and the choice of metric can influence the results more than the choice of a parameter fitting procedure. Moreover, fitted model parameters are often used in subsequent steps in educational data mining and thus the choice of metric can indirectly influence many other aspects of the research.

However, despite the fact that the choice of metric is important and that there is no clear consensus on the usage of performance metrics, the topic gets very little attention in most research papers. Most authors do not provide any rationale for their choice of metric. Sometimes it is not even clear what metric is exactly used, so it may be even difficult to use the same metric as previous authors. The main aim of this paper is to give an overview of performance metrics relevant for evaluation of student models and to explicitly discuss points that are in most papers omitted.

2. OVERVIEW OF METRICS

To attain clear focus we discuss only models that predict probability of a correct answer. We assume that we have data about n answers, numbered $i \in \{1, \ldots, n\}$, correctness of answers is given by $c_i \in \{0, 1\}$, a student models provides predictions $p_i \in [0, 1]$. A model performance metric is a function $f(\vec{p}, \vec{c})$. Note that the word "metric" is here used in a sense "any function that is used to make comparisons", not in the mathematical sense of a distance function. Since we are interested in using the metrics for comparison, monotone transformations (square root, logarithm, multiplication by constant) are inconsequential and are used mainly for better interpretability (or sometimes rather for traditional reasons).

2.1 Mean Absolute Error

This basic metric consider the absolute differences between predictions and answers: $MAE = \frac{1}{n} \sum_{i=1}^{n} |c_i - p_i|$. This is not a suitable performance metric, because it prefers models which are biased towards the majority results. As a simple illustration, consider a simulated student which answers correctly with constant probability 0.7. If we compare different constant predictors with respect to this metric, we get that the best model is the one which predicts probability of correct answer to be 1. This is clearly not a desirable result. As this example illustrates, the use of MAE can lead to rather misleading conclusions. Despite this clear disadvantage, MAE is sometimes used for evaluation (although mostly in combination with other metrics, which reduces the risk of misleading conclusions in published papers).

2.2 Root Mean Square Error

A similar metric is obtained by using squared values instead of absolute values: $RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (c_i - p_i)^2}$. Note that from the perspective of model comparison, the important part is only the sum of square errors (SSE). The square root in RMSE is traditionally used to get the result in the same units of as the original "measurements" and thus to improve interpretability of the resulting number. In the particular context of student modeling and evaluation of probabilities, this is not particularly useful, since the resulting numbers are hard to interpret anyway. In order to get better interpretability researchers sometimes use R^2 metric: $R^2 = 1 - \sum_{i=1}^{n} (c_i - p_i)^2 / \sum_{i=1}^{n} (c_i - \bar{c})^2$. With respect to comparison of models, R^2 is equivalent to RMSE since here again the only model dependent part is the sum of square errors. In the context of the standard linear regression (where it is

most commonly used) R^2 has a nice interpretation as "explained variability". In the case of logistic regression (which is more similar to student models) this interpretation does not hold and different "pseudo R^2" metrics are used (e.g., Cox and Snell, McFadden, Nagelkerke). Thus a disadvantage of R^2 is that unless the authors are explicit about which version of R^2 they use (usually they are not), a reader cannot know for sure which metric is reported.

In educational data mining the use of RMSE metric is very common (it was also used as a metric in KDD Cup 2010 focused on student performance evaluation). In other areas, particularly in meteorology, mean square error (RMSE without the square root) is called the Brier score [1]. The Brier score is often decomposed into additive components (e.g., reliability and refinement) which provide further insight into the behaviour of the predictor. Moreover, in an analogy to AUC metric and ROC curve (described below), this metric can be interpreted as area under Brier curves. These methods may provide interesting inspirations for student modeling.

2.3 Metrics Based on Likelihood

The likelihood of data (the answers) given a model (predicted probabilities) is $L = \prod_{i=1}^{n} p_i^{c_i} \cdot (1-p_i)^{(1-c_i)}$. Since we are indifferent to monotonic transformations we typically work with the numerically more stable logarithm of the likelihood $LL = \sum_{i=1}^{n} c_i \log(p_i) + (1-c_i)\log(1-p_i)$. This metric can also be interpreted from information theoretic perspective as measure of data compression provided by a model [4]. The log-likelihood metric can be further extended into metrics like Akaike information criterion (AIC) and Bayesian information criterion (BIC). These metrics penalize large number of model parameters and thus aim to avoid overfitting. In the context of student modeling it is typically much better to address the issue of overfitting by cross-validation. Since AIC and BIC provide a faster way to assess models than cross-validation, they may be useful as heuristics in some algorithms (e.g., learning factor analysis), but they are not serious contenders for proper model comparison.

MAE, RMSE and LL have all the form of "sum of penalties for individual errors" and differ only in the function which specifies the penalty. For RMSE and LL values of penalty functions are quite similar, the main difference is in the interval [0.95, 1], i.e., in cases where the predictor is confident and wrong. These cases are penalized very prohibitively by LL, whereas RMSE is relatively benevolent. In fact the LL metric is unbounded, so single wrong prediction (if it is too confident) can ruin the performance of a model. This property is usually undesirable and an artificial bound is used. This corresponds to basically forcing a possibility of a slip and guess behaviour into a model. After this modification the penalties for RMSE and LL are rather similar. Nevertheless, the LL approach "penalize mainly predictions which are confident and wrong" is reasonable thus it is rather surprising that this metric is used only marginally in evaluation of student models (it is used mostly in connection with AIC or BIC).

2.4 Area Under an ROC Curve

Another popular metric is based on the receiver operating characteristics (ROC) curve. If we want to classify predictions into just two discrete classes (correct, incorrect), we need to select a threshold for the classification. For a fixed threshold we can compute standard metrics like precision, recall, and accuracy. If we do not want to use a fixed threshold, we can use the ROC curve, which summarises the behaviour of the prediction model over all possible thresholds. The curve has "false positive rate" on x-axis and "true positive rate" on the y-axis, each point of the curve corresponds to a choice of a threshold. Area under the ROC curve (AUC) provides a summary performance measure across all possible thresholds. It is equal to the probability that a randomly selected correct answer has higher predicted score than a randomly selected incorrect answer. The area under the curve can be approximated using a A' metric, which is equivalent to the well-studied Wilcoxon statistics [2]. This connection provides ways to study statistical significance of results (but requires attention to assumptions of the tests, e.g., independence).

The ROC curve and AUC metric are successfully used in many different research areas, but their use is sometimes also criticised [3], e.g., because the metric summarises performance over all possible thresholds, even over those for which the classifier would never be used in practice. From the perspective of student modeling the main reservation seems to be that this approach focuses on classification and considers predictions only in relative way – note that if all predictions are divided by 2, the AUC metric stays the same.

In the context of student modeling we are usually not interested in classification, we are often interested directly in absolute values of probabilities and we need these values to be properly calibrated. The probabilities are often compared to a fixed constant (typically 0.95) as an indication of a mastered skill and the specific value is meant to carry a certain meaning. Probabilistic estimates can be also used to guide the behaviour of a system to achieve suitable challenge for students, e.g., by choosing question of right difficulty or modifying difficulty by number of options in multiple choice questions.

Nevertheless, despite this disadvantage, AUC is widely used for evaluation of student models, often as the only metric. It seems that in some cases AUC is used as the only metric for final evaluation, but the parameter fitting procedure uses (implicitly) different metric (RMSE or LL). Particularly in cases of brute force fitting this approach seems strange and should be at least explicitly mentioned.

3. REFERENCES

[1] G. W. Brier. Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1):1–3, 1950.

[2] J. Fogarty, R. S. Baker, and S. E. Hudson. Case studies in the use of ROC curve analysis for sensor-based estimates in human computer interaction. In *Proc. of Graphics Interface 2005*, pages 129–136, 2005.

[3] J. M. Lobo, A. Jiménez-Valverde, and R. Real. AUC: a misleading measure of the performance of predictive distribution models. *Global ecology and Biogeography*, 17(2):145–151, 2008.

[4] M. S. Roulston and L. A. Smith. Evaluating probabilistic forecasts using information theory. *Monthly Weather Review*, 130(6), 2002.