Student Model Quality Evaluation
What is a student model?
A student model is a representation within the architecture of an intelligent tutoring system (ITS) or intelligent learning environment (ILE) of a student's understanding of material being taught.

Brusilovsky, 1994
Why evaluate a student model?
Reasons to evaluate

☑ Evaluate the student
☑ Improve the ITS / ILE
☑ Guide the behaviour of the system

Choosing the metrics to model is critical!
Types of student models

Skill Modeling
- Most-used
- Comparing predicted to actual performance
- Used to guide adaptive behavior of ES

Models of Affect and Motivation
- Affective States
- Behaviour
- Personalized feedback
Choosing how to evaluate
### Probabilistic Understanding of Errors

<table>
<thead>
<tr>
<th>Metric</th>
<th>Formula</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean Absolute Error (MAE)</td>
<td>[ \frac{1}{n} \sum_{i=1}^{n}</td>
</tr>
<tr>
<td>Root Mean Square Error (RMSE)</td>
<td>[ \sqrt{\frac{1}{n} \sum_{i=1}^{n} (o_i - p_i)^2} ]</td>
</tr>
<tr>
<td>Log-likelihood (LL)</td>
<td>[ \sum_{i=1}^{n} o_i \log(p_i) + (1 - o_i) \log(1 - p_i) ]</td>
</tr>
</tbody>
</table>
Mean Absolute Error (MAE) considers absolute differences between predictions and answers. This is not a suitable performance metric, because it prefers models which are biased towards the majority result, linearly. Despite this clear disadvantage, MAE is sometimes used for evaluation of student models.

\[ \text{MAE} = \frac{1}{n} \sum_{i=1}^{n} |o_i - p_i| \]
Root Mean Square Error

In educational data mining the use of RMSE metric is very common, particularly for evaluation of skill models.

From the perspective of model comparison, the important part is only the sum of square errors. The square root in RMSE is traditionally used to get the result in the same units as the original “measurements” and thus to improve interpretability of the resulting number.
Log-likelihood values are negative, it is not averaged connected with extensions that penalize large number of model parameters and thus aim to avoid overfitting

\[ \text{LL} = \sum_{i=1}^{n} o_i \log(p_i) + (1 - o_i) \log(1 - p_i) \]
QUALITATIVE UNDERSTANDING OF ERRORS

<table>
<thead>
<tr>
<th>Predicted</th>
<th>Positive</th>
<th>Negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pos.</td>
<td>true positive $(TP)$</td>
<td>false positive $(FP)$</td>
</tr>
<tr>
<td>Neg.</td>
<td>false negative $(FN)$</td>
<td>true negative $(TN)$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Accuracy</th>
<th>$(TP + TN)/n$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision</td>
<td>$TP/(TP + FP)$</td>
</tr>
</tbody>
</table>
Metrics and usefulness of models
Impact on student practice

Minimize over and under-practice
For that, the settings of thresholds is important
The study is to analyze their relation to model performance
Impact on parameter fitting

Model parameters can provide us with valuable insight into the learning process. Is typically done by optimizing performance of a model with respect to a chosen metric. Unreliable, good results doesn't mean stability.
Beyond single number
BRIER SCORE DECOMPOSITION

\[ BS = \frac{1}{N} \sum_k n_k (q_k - f_k)^2 - \frac{1}{N} \sum_k n_k (f_k - f)^2 + f(1-f) \]

\[ BS = REL - RES + UNC \]

More insight into the behaviour of the predictor

Reliability
Resolution
Uncertainty
RELIABILITY DIAGRAMS

Histogram of number of cases

reliability - lines
resolution - histogram
Does the choice matter?
Correlations

LL

-0.69

-0.96

MAE

-0.69

0.83

RMSE

-0.96

0.83
Which metric should I use?
Reference:

Thanks!

Any questions?