Measuring Predictive Performance of User Models: The Details Matter

Radek Pelánek

Adaptive Learning

Research group, Masaryk university Brno

EvalUMAP 2017

Evaluation over Historical Data

common evaluation approach

- model building
- data collection
- cross-evaluation methodology (train/test division)
- model fitting
- quantifying model quality by measuring predictive accuracy
- comparison of models, interpretation of results

Measuring Predictive Accuracy

prediction	0.7	0.6	0.9	0.95	0.8	0.85	...
outcome	0	1	1	0	1	1	...

$$\Downarrow$$

quality of predictions (RMSE, AUC, ...)

this step:

- gets little attention
- can significantly influence results
- can be nontrivial to do properly

Motivation

- Deep knowledge tracing, Piech et al. NIPS 2015
  - claims of large improvement in model performance as measured by AUC
- How deep is knowledge tracing?, Khajan et al., EDM 2016
  - the "improvement" caused to large degree by methodological differences in computation of AUC

RMSE and AUC Metrics

- Root Mean Square Error (RMSE)
  - $\sqrt{\frac{1}{n}\sum_{i=1}^{n}(o_i - p_i)^2}$
  - closely related to "Brier score"
- Area Under the ROC Curve (AUC)
  - Receiver Operating Characteristics (ROC) curve
  - relative ranking of predictions
  - widely used in many domains, but also widely criticized

Averaging

	skill1	skill2	skill3	...
student1	1	0	-	
student2	0	1	1	
student3	1	-	0	
...				

metric computation:

- global
- averaging across skills
- averaging across students

Analysis of Student Modeling Literature

- little attention to the choice of metric
- details of computation typically not specified
- AUC often used as a single metric

Illustration of Metric Properties

- scenarios with simulated data
- simple "learning curve" model of student behaviour
- illustration of metric properties

Absolute Values of Metrics

Absolute values of metric do not express quality of models, but rather properties of data.

- RMSE: baseline rate of events
- AUC: heterogeneity of data

Do not try to interpret the values.
Do not compare values across data sets.

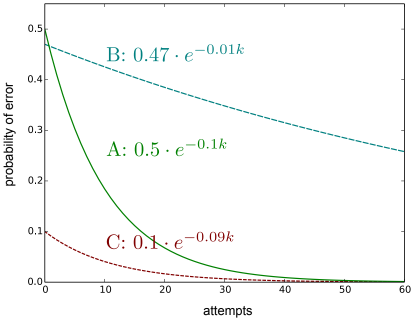
Relative Values of Metrics

relative values = differences in metric values

Very different models can have nearly the same metric value, particularly for AUC.

Do not rely on AUC as a single metric to measure model performance.

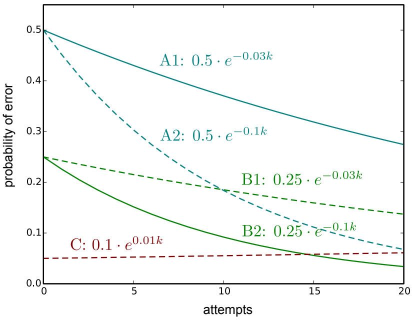
Averaging Across Students



model	RMSE global	RMSE per student
B	0.40	0.46
C	0.35	0.48

curve A
70% of students: 5 attempts
30% of students: 60 attempt

Averaging Across Skills



model	AUC global	AUC per skill
A1, B2 (correct)	0.73	0.63
A2, B1 (speed mismatch)	0.60	0.63
A1, C (negative learning)	0.68	0.45

Summary

- choice of metric matters
- details of metric computation matter
- should we adopt standards?
    - "universal metric" – no
    - "good practice" – yes