Assessing the Quality of Mathematics Questions Using Student Confidence Scores

Jessica McBroom*
School of Computer Science
The University of Sydney
Australia
jmcb6755@uni.sydney.edu.au

Benjamin Paassen
School of Computer Science
The University of Sydney
Australia
benjamin.paassen@uni.sydney.edu.au

Abstract

Techniques for automatically assessing the quality of course materials are an important avenue for course development. This paper presents our solution to the *Global Question Quality Assessment task (Task 3)* of the NeurIPS 2020 Education Challenge, which required an automatic ranking of mathematics questions according to their quality. Our solution is simple: We estimate question quality by the average confidence of students when responding to the question. With this simple approach we achieved a tied first place with 80% agreement to expert ratings. We have two explanations for this success: First, the ground truth for question quality were expert rankings, which are highly noisy and thus make overfitting more likely. Second, expert rankings were based on criteria that relate to student confidence, making student confidence a weak but robust estimator. Overall, we believe that this challenge highlights that students' self-assessment can be valuable in estimating quality, and that quality is ill-defined, making it difficult to construct models that do better than our simple baseline.

1 Competition and Task

The NeurIPS 2020 Education Challenge was an international competition held from July 1th, 2020 to October 23rd, 2020. It consisted of four educational data mining tasks on mathematics data, which included the prediction of student responses, the assessment of question quality, and the production of personalized question orderings. Details of the competition can be found at [8] and `https://competitions.codalab.org/competitions/25449#learn_the_details-`.

In this paper, we focus on Task 3, the "Global Question Quality Assessment". This tasks involved the automatic ranking of the quality of 948 multiple-choice mathematics questions without access to labels. The quality rankings were then compared against rankings of domain experts.

1.1 Dataset

The dataset consisted of 1,382,727 answers to 948 multiple-choice mathematics questions from 4,918 students, which was sourced from the online education provider Eedi. For each question, an image of the question was included along with the question topic, such as "Algebra", "Data and Statistics" or "Geometry and Measure". In addition, each answer was annotated with the time, the correctness, and a self-rated confidence. Finally, student demographics data, such as gender and age, were also included

*Corresponding author. Code is available at `https://gitlab.com/quokka_jess/neurips-task-3-solution`

1st NeurIPS Education Challenge (NeurIPS 2020), Vancouver, Canada.

1.2 Evaluation procedure and ground truth

The automatic ranking of questions was validated against ranked pairs provided by experts. In other words, the ground truth was generated by showing a domain expert two mathematical questions and letting them decide which one was - in their opinion - of higher quality. The competition objective was to achieve as much ranking agreement with at least one of the experts as possible.

The expert rankings were distributed into a public data set, against which submissions could be evaluated before the end of the competition, and a private data set (or test set) which yielded the final ranking (refer to `https://competitions.codalab.org/competitions/25449#results`).

Importantly, when generating the ground truth, domain experts had a certain notion of question quality in mind which is described in [1]. In particular, Barton suggests five golden rules to create high-quality questions:

1. They should be clear and unambiguous.
2. They should test a single skill/concept.
3. Students should be able to answer them in less than 10 seconds.
4. You should learn something from each incorrect response without the student needing to explain.
5. It is not possible to answer the question correctly whilst still holding a key misconception.

2 Method

Our method is particularly simple. For each question, we computed the average confidence of students in their answers. We then ranked the question according to the average confidence (higher confidence is better) and submitted this ranking. Importantly, this estimate has no free parameters. We do not fit any model to the data. We merely compute an average and then sort.

The reasoning behind this very simple approach is two-fold.

First, we expect expert rankings to be a highly noisy signal. Attempting to train a model that maximizes the agreement on the public data set would thus likely overfit, i.e. a model that fits spurious or noisy patterns that do not generalize. Indeed, we observe this effect in practice (see next section).

Second, although the rankings are subjective and noisy, the golden rules of [1] are strongly related to student confidence. In particular, high student confidence suggests that a question appeared unambiguous to the students and that they can answer quickly [5], which covers the relation to the first three rules. Regarding the fourth and fifth rule, we argue that a wrong answer with high confidence is indicative of a key misconception, similar to the Dunning-Kruger effect [2, 3, 4, 6, 7]. In other words, if a question has a significant rate of wrong answers with high confidence, the question can tell us about key misconceptions held by the students. Indeed, the training data for the challenge contains 42621 answers were students rated their confidence as 100% but were still wrong. In summary, we believe that student confidence can be motivated from past literature as a proxy for the kind of question quality suggested by [1], especially when there are many wrong answers with high confidence.

3 Results and Discussion

On the public data set our method (submitted under the label "the Quokka Appreciation Team") achieved a score of 76%, far below the highest score of 96%. However, on the private data set, our method achieved a score of 80%, tying on first place with three other teams. Indeed, our approach was the only top one with increased score from public to private. By contrast, the average score on the public data was 80% but only 70% on the private data. The second-highest score on the public data, 92%, corresponded to only 44% on the private data. Overall, we interpret these results as indication that the expert rankings are highly noisy, where patterns on the public data do not necessarily translate to the private data. Because our approach does not have free parameters which could be fitted to such spurious and noisy patterns, it exhibits higher robustness when applied to the private data. Note that

this issue of overfitting demonstrates the importance of having a separate test set that is only used once when assessing the quality of educational models.

In addition to doing better on the public data than the private data, our model is also much simpler than the other top-performing models, which involved logistic models in item-response theory, convolutional neural networks for text localization, and metrics based on multiple features, including entropy (see descriptions available at `https://eedi.com/projects/neurips-education-challenge`). This suggests that asking students how confident they were in answering a question can provide just as much information about question quality as highly sophisticated models. In addition to this, our model is efficient, easy to interpret, and has minimal requirements (e.g. it does not require images of the questions or for the questions to be multiple choice).

4 Conclusion

This paper has presented our solution to Task 3 of the NeurIPS 2020 Education Challenge, which received a tied first ranking of 80%. In particular, our solution is to rank questions simply according to the average student confidence. We provided two reasons for the success of our model: First, that teacher rankings are noisy, making overfitting likely; but since our model is not fitted to the data, it does not overfit. Second, that confidence is related to the quality definition underlying the teacher rankings and thus provides a weak but robust estimator. Overall, we believe that the result of this competition can provide two helpful lessons: Student self-assessment can yield insight into question quality, even if students are wrong, and simple methods from first principles can be competitive with more involved machine learning if the underlying task is ill-defined - as is the case with quality ratings.

References

[1] Craig Barton. What makes a good diagnostic question? *medium*, 2017.

[2] Alexandra R. Brandriet and Stacey Lowery Bretz. Measuring meta-ignorance through the lens of confidence: examining students' redox misconceptions about oxidation numbers, charge, and electron transfer. *Chemistry Education Research and Practice*, 15:729–746, 2014.

[3] Donald A Curtis, Samuel L Lind, Christy K Boscardin, and Mark Dellinges. Does student confidence on multiple-choice question assessments provide useful information? *Medical education*, 47(6):578—584, June 2013.

[4] David Dunning. Chapter five - the dunning–kruger effect: On being ignorant of one's own ignorance. volume 44 of *Advances in Experimental Social Psychology*, pages 247 – 296. Academic Press, 2011.

[5] Wayne Douglas Powel. *Influence of the amount and relevance of information on the speed and confidence of the response*. PhD thesis, 1989.

[6] Stephen M. Swartz. Acceptance and accuracy of multiple choice, confidence-level, and essay question formats for graduate students. *Journal of Education for Business*, 81(4):215–220, 2006.

[7] Douglas M. Walker and John S. Thompson. A note on multiple choice exams, with respect to students' risk preference and confidence. *Assessment & Evaluation in Higher Education*, 26(3):261–267, 2001.

[8] Zichao Wang, Angus Lamb, Evgeny Saveliev, Pashmina Cameron, Yordan Zaykov, José Miguel Hernández-Lobato, Richard E. Turner, Richard G. Baraniuk, Craig Barton, Simon Peyton Jones, Simon Woodhead, and Cheng Zhang. Diagnostic questions:the neurips 2020 education challenge, 2020.