Diagnostic Questions - The NeurIPS 2020 Education Challenge

Anonymous Author(s)
Affiliation
Address
email

Abstract

Tasks 1 and 2 of the NeurIPS 2020 Education Challenge are based on student modelling and response prediction. This paper explores classical Item Response Theory (IRT) based approaches to the problem. It can be demostrated that simple 1 parameter IRT model provides reasonable prediction accuracy.

1 Task 1: Answer correctness prediction

Classical Item Response Theory has been widely used for estimating student ability and predicting student performance [2, 7, 4, 1]. The simplest model, also called the one parameter model fits a difficulty parameter for each question and an ability parameter for each student. The probability of student s answering the question i correctly is the logit of the difference between student ability and question difficulty. The following equation represents the One parameter IRT model

$$P(C = 1) = logit(\sigma_s - \beta_i) \text{ (1)}$$

Where σ_s indicates the ability of the student s and β_i indicates the difficulty of the question i. IRT models can be extended to incorporate multidimensional ability and difficulty parameters and called as Multidimensional IRT (MIRT). [6] is a great resource on MIRT models and their applications. The MIRT equations are the same as IRT except that each parameters is a vector. Furthermore, if the skills (or knowledge components) associated with each question is available, equation 1 can be modified as

$$P(C = 1) = logit((\sigma_s - \beta_i) < . > q_i) \text{ (2)}$$

where q_i is an n dimensional binary vector with 1s in the columns of the skills associated with question i, and $< . >$ indicates the dot product. The collection of q_is is called Q Matrix.

The two parameter model adds a second parameter to each question called discimination denoted by α. The equations for probability of correctness for two paramter IRT and MIRT are given by equations 3 and 4 below.

$$P(C = 1) = logit((\sigma_s - \beta_i) * \alpha_i) \text{ (3)}$$
$$P(C = 1) = logit(((\sigma_s - \beta_i) * \alpha_i) < . > q_i) \text{ (4)}$$

The addition of discrimination parameter allows for questions difficulty to be scaled with student ability. As a consequence, students with differnt abilities might find the relative difficulties of 2 questions to be swapped at a threshold value of their ability. For example students with ability <=2.3 might find Q1 easier than Q2 wheras students with higher ability find Q2 easier than Q1.

Traditionally, IRT parameters are estimated using Estimation Maximisation (EM) or Monte Carlo Markov chain (MCMC) techniques. This limits the size of dataset on which IRT could be applied. Recent work by [5] and [3] showed that Stochastic Variational Inference (SVI) can be used to estimate

Submitted to the 1st NeurIPS Education Challenge (NeurIPS Education Challenge 2020). Do not distribute.

IRT parameters. This increases scalability and allows IRT with datasets containing millions of records to be compatible with IRT estimation. To the best of our knowledge IRT has not been applied to a dataset with the size comparable to NeurIPS Education challenge.

1.1 Models

For Task 1 and Task 2, five models were evaluated for response prediction.

1. One paramter IRT (1pIRT)
2. Two paramter IRT (2pIRT)
3. One paramter MIRT with binary Q Matrix (1pMIRT)
4. Two paramter MIRT with binary Q Matrix (2pMIRT)
5. Two paramter MIRT with continuous valued Q Matrix (2pMIRTq)

1pIRT is the simplest model based on equation 1. 1pMIRT is its multidimensional equivalent based on equation 2. 2pIRT and 2pMIRT are based on equations 3 and 4 respectively. For the construction of Q matrix in 2pMIRT model, only the top 2 level skills were considered. This was done to limit the number of dimensions. This resulted in 9 dimensional vectors (since there are only 9 level 0 and level 1 skills). 2pMIRTq is based on equation 4 with the q matrix being estimated or learned rather than provided to the model. Findings in [1] indicate that expert assigned skill / knowledge component tags add little to no value in many datasets. Also, a continuous valued Q matrix can represent temporal information better than binary Q matrix associated with manual tagging. 20 dimensions was arbitrarily chosen for 2pMIRTq. Higher dimensions converged quicker and had higher training set accuracy but lost generalizability and had a much lower test set accuracy.

1.2 Results

Table 1: Task 1 results

Model Name	Training set accuracy (%)	Public test set accuracy (%)
1pIRT	74.28	**73.80**
2pIRT	72.77	72.11
1pMIRT	74.65	73.28
2pMIRT	75.35	73.63
2pMIRTq	**80.72**	73.62

...

Table 1 shows the performance of the different models. Although 2pMIRTq had the highest training set accuracy, it was prone to overfitting and 1pIRT had the best test set accuracy overall. The number of parameters seem to have little effect on the accuracy. However, 2 parameter models converged faster. The number of dimensions also seems to have a pronounced effect. Higher dimensions tended to overfit and the disparity between training set accuracy and test set accuracy was large (~12%).

1.3 Challenges and Limitations

The most obvious limitation of using IRT for prediction in this challenge is the characteristic of IRT student ability estimates. The estimates learned this way must conform to the assumption that answering the questions does not affect their ability. While this may be a reasonable assumption for tests like GRE, student ability continuously improves as they progress through Intelligent Tutoring Systems (ITSs). This almost always leads to overestimation of student abilities. Also ITS student responses are usually skewed towards correct responses (since many ITSs require students to practice skills to mastery). To validate the theory of overestimation, we performed a 90-10 train test split based on timestamp with training set data occuring before the test data stratified on students. This improved the discrepancy between training and test set accuracies by a small margin.

The next biggest limitation is the compensatory nature of MIRT. If a student has a high ability in one of the dimensions associated with the question, their response is predicted to be correct despite very low abilities in other dimensions. To mitigate this, we tried to threshold student ability contribution in each individual dimension while predicting correct response. The results were mixed, with improvement in prediction of certain students and degradation in prediction of others. We also tried implementing a possible alternative model called partial compensatory model presented in [6] which requires students to master all skills associated with a question to predict their response as correct. However, we were unsuccessful in estimating parameters with good fit using SVI.

At this stage, we are unable to explain the overfitting of 2pMIRTq model. Further study and discussion with experts is required to better understand the factors affecting this. We believe 2pMIRTq to have tremendous potential as it removes the need to manually tag each question. IRT models also have very high interpretability and transferability. For example, the question difficulties estimated from 1 batch can be retained and only the student ability be estimated for a consecutive batch. Similarly, student abilities from one subject can be used to estimate question difficulty from another subject as long as the skills overlap.

2 Task 2: Predicting student response

Our approach to Task 2 was to borrow the predictions from Task 1 and substitute the correct answers for predictions which indicated correct response. For the predictions involving incorrect responses, we clustered students based on their abilities and selected the most frequent answer for each question from the clusters the students belonged to. We arbitrarily chose 1000 clusters and bin values for each cluster from the range of student abilities. Student ability is normally distributed and hence the number of students in each cluster varies. A possible alternative is to cluster students by having varying bin lengths such that the number of students in each group is approximately equal. Due to time constraints, we were unable to try this. Since 1pIRT model had the best test set accuracy, we used the student ability from that model for prediction. The best accuracy achieved was 63.77%

2.1 Challenges and Limitation

The biggest limitation of the approach is tied to skewed predictions resulting from IRT which are used to derive student resposes. This is explained in detail in section 1.3. The second challenge is computational complexity. Since we need to compute the most frequent answer for each question for each cluster, the time complexity is O(N*I*n). Where N = number of students, I = number of questions and n = number of clusters. This severly limited our abiliy to repeat the experiment for different number of clusters.

References

[1] Theophile Gervet, Ken Koedinger, Jeff Schneider, Tom Mitchell, et al. When is Deep Learning the Best Approach to Knowledge Tracing? *JEDM| Journal of Educational Data Mining*, 12(3):31–54, 2020.

[2] Jeff Johns, Sridhar Mahadevan, and Beverly Woolf. Estimating student proficiency using an item response theory model. In *International conference on intelligent tutoring systems*, pages 473–480. Springer.

[3] John P Lalor, Hao Wu, and Hong Yu. Learning latent parameters without human response patterns: Item response theory with artificial crowds. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, volume 2019, page 4240. NIH Public Access.

[4] Youngjin Lee. Estimating student ability and problem difficulty using item response theory (irt) and trueskill. *Information Discovery and Delivery*, 2019.

[5] Prathiba Natesan, Ratna Nandakumar, Tom Minka, and Jonathan D Rubright. Bayesian prior choice in irt estimation using mcmc and variational bayes. *Frontiers in psychology*, 7:1422, 2016.

[6] Mark D Reckase. Multidimensional item response theory models. In *Multidimensional item response theory*, pages 79–112. Springer, 2009.

[7] Wim J van der Linden and Ronald K Hambleton. *Handbook of modern item response theory*. Springer Science & Business Media, 2013.