2 Part B

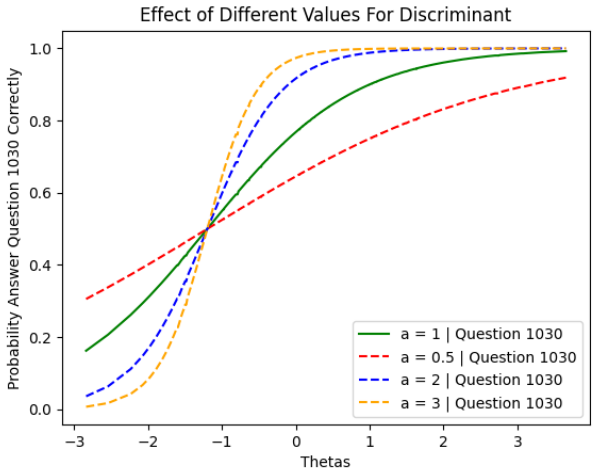
2.1 Algorithm Extension

In question 2 we found that the validation accuracy capped at ∼70% and the rate of increase in the training accuracy slowed significantly (we could get a training accuracy of ∼70% in 25 iterations of gradient descent whereas 100 iterations improved the training accuracy by only ∼4% and saw little change in the validation accuracy). This suggests that the base model is under fitting the training data so our attempt to improve the model accuracy was to increase the model complexity to reduce under fitting. We decided on the 3 parameter logistic item response theory[1] (3PL IRT) model with additional covariate parameters corresponding to the student metadata. The probability that the question j is correctly answered by student i is formulated as:

$$P(c_{ij} = 1|\theta, \beta, \gamma, \zeta_1, ..., \zeta_n, x) = \gamma_j + \frac{(1 - \gamma_j)}{1 + e^{-a_j(\theta_i - \beta_j + \zeta_{1j}x_{i1} + ... + \zeta_{nj}x_{in})}} \tag{1}$$

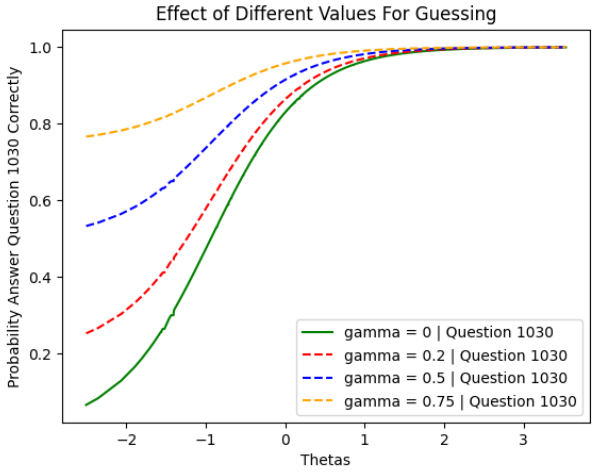
The parameters not present in the original model are: effect of covariate student metadata (ζ), discrimination (α), and guessing (γ). These parameters are estimated for each question in the training set.

1. Discrimination parameter (α): Represents how well a question can differentiate between high-ability and low-ability students. Questions with high discrimination are more informative and can better distinguish between students with different levels of ability. By including the discrimination parameter, the 3PL IRT model can better identify questions that are more informative and can more accurately measure students' ability levels. This can lead to more accurate estimates of students' abilities and better prediction accuracy.



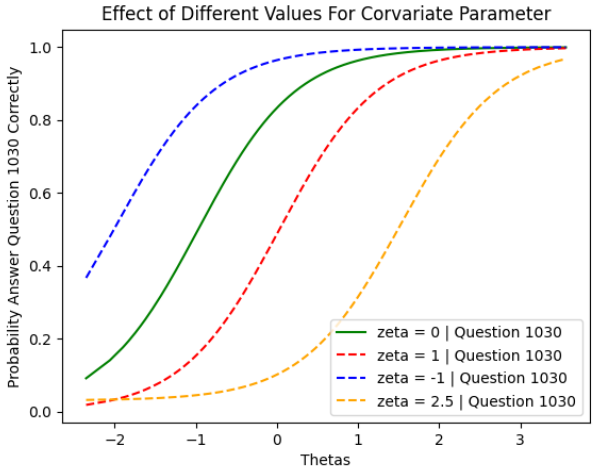
Different values of the discriminant changes the steepness of the probability curve.

2. Guessing parameter (γ): Represents the probability that a student with low ability will respond correctly to a question by chance. Questions with high guessing parameters are more likely to be answered correctly by low-ability student, regardless of whether they actually know the content. The guessing parameter is particularly important for multiple-choice questions where students have to choose one answer from several options. By including the guessing parameter, the 3PL IRT model can better account for the effect of guessing on students' responses, which can improve the accuracy of ability estimates.



Different values of the guessing parameter compresses the probability curve to be between z and 1.

3. Covariate parameters (ζ_k): Accounts for the effect of a covariate (In our case specifically the student metadata: age, gender, and premium pupil) on a question response. Including covariate parameters in a 3PL IRT model can improve the model's accuracy by controlling for the influence of the covariates on the relationship between the student's latent ability and question response. In our equation above, the $\zeta_{kj}x_{ki}$ term is the coefficient for covariate k for question j (ζ_{kj}) times student i's metadata value for covariate k (x_{ki}).



Different values of the covariate parameter correspond to a horizontal translation of the probability curve.

Overall, the 3PL IRT model with additional covariate parameters can provide a more accurate and reliable assessment of students' abilities and therefore should also make better predictions than the model which only considers the latent ability and difficulty parameters.

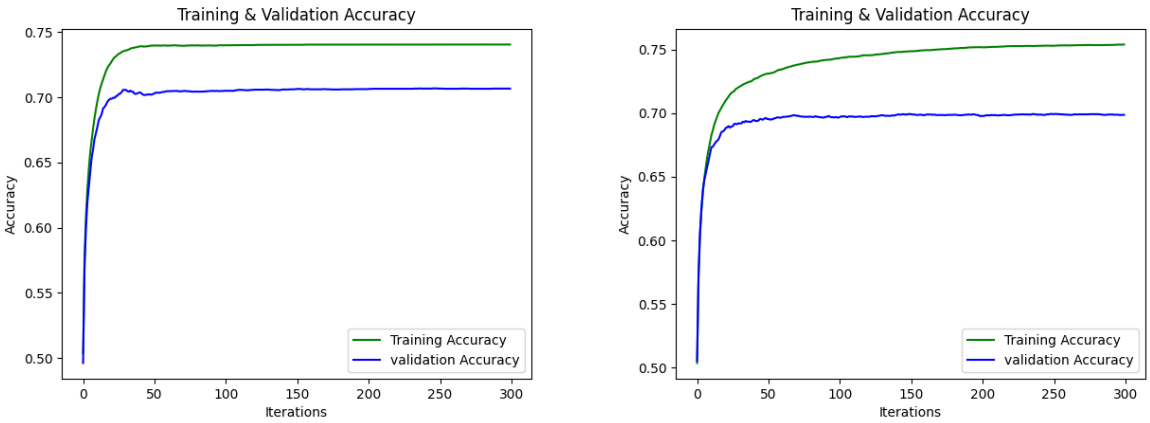
2.2 Student Metadata Covariates

Our EDA of the student metadata found there to be differences in the mean number of questions answered correctly between premium students and non-premium / those with missing premium values so we included it in the extended model as a dummy variable (our only covariate parameter).

2.3 Performance Comparison

We can compare the performance of the 2 models by looking at the graphs of their respective training and validation accuracies over the course of training. We can see that the base model's training accuracy reaches near 75% but then the rate of increase in the training accuracy approaches zero and we see a similar pattern in the validation accuracy. For the extended model, we see that the training accuracy does not see a as dramatic decrease in the rate of change but we do see a similar leveling off in the validation accuracy at around $\sim 70\%$ as was the case with the base model. It appears that the changes to the algorithm have improved the issue of underfitting the training data but it appears that those gains have not been generalized in a way to realize them in the accuracy on the validation set.

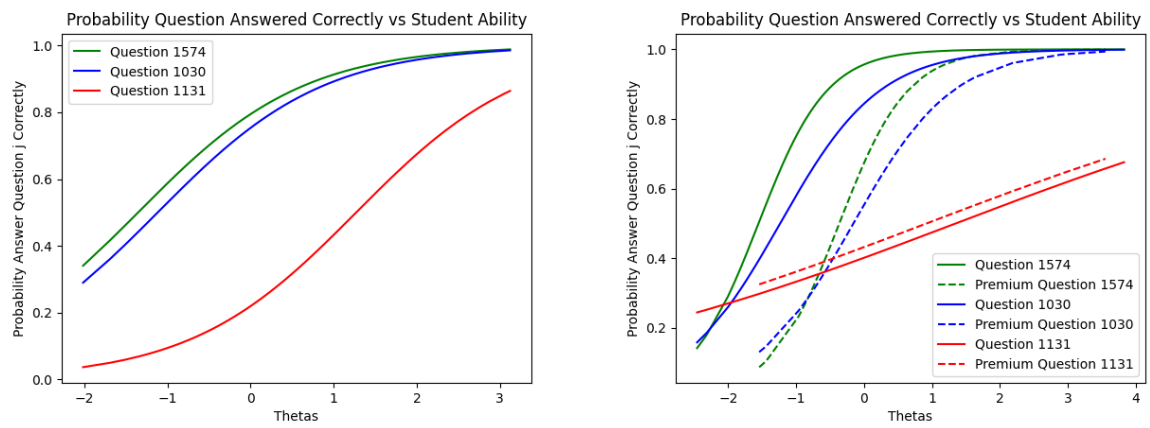
Comparison of Training and Validation Accuracies Throughout Training



Base Model (left) compared to Extended Model (right) for 300 training iterations each.

From the probability curves for the student ability, we can clearly see the effect of the additional parameters in the extended model. The range of student ability estimates has increased and the slopes have been changed by the discriminant parameters where for example question 1030 got steeper and 1131 got less steep. As we might expect, the premium parameter corresponds to a horizontal translation to the right as there may be economic factors limiting those students ability to answer the questions correctly. In the case of question 1131, the increase in probability for premium students can be attributed to the smaller sample of premium students or perhaps the nature of the particular question (even students with high estimated ability still having difficulty with it).

Comparison of Probability Answer Correctly vs Estimated Latent Student Ability



Probabilities as function of estimated student ability for Base Model (left) compared to Extended Model (right) on questions 1574, 1030 and 1131.

By looking directly at the parameter values we can see that, for example, premium student with id 99 is estimated to have a higher latent ability than the base model estimates. This is likely due to the extended model discriminating between the influence of different questions on determining a students ability (the discriminant value ~ 1.5 for question 1030) and from the portion of the student's difficulty answering questions attributable to being premium. Both models estimate similar values for question difficulty.

Comparison of Estimated Parameters of Base Model to Extended Model for Student 99 Question 1030

Parameters	Base Model Estimates	Extended Model Estimates
Discriminant	N/A	1.4930398232269373
Ability	0.8802620866496668	1.0582566434396208
Difficulty	-1.1001665570478716	-1.1194754039979278
Guessing	N/A	0.0
Is Premium	N/A	1.13802960125593722

Lastly, we see both models have similar accuracy on the validation and test sets.

Validation and Test Accuracies

Model	Validation Accuracy	Test Accuracy
Base	0.7013830087496472	0.6999717753316399
Extended	0.6989839119390348	0.698278295230031

2.4 Limitations

The following are some possible issues that maybe limiting the extended model's performance:

1. Non-normal distribution of ability: The 3PL IRT model assumes that the latent ability follows a normal distribution[1] If the distribution of ability is highly skewed or has heavy tails, the model may not accurately capture the underlying relationship between ability and question response.
2. Violation of model assumptions: The 3PL IRT model assumes that the questions are independent and the question parameters (difficulty, discrimination, guessing and covariates) are constant across different groups of students.[2] If any of these are violated, this may impact the accuracy of model predictions.

As stated above, the extended model appears to be failing to generalize gains in training accuracy. Some possible changes/extensions to address this problem and more generally issues with low accuracy are:

1. Momentum: a momentum term may improve the training speed to get the increase in training accuracy with fewer iterations.
2. Mini Batch GD: Bootstrapped mini batches introduce noise that help with generalization and may improve training speed as well.
3. More rigorous diagnostic testing and checking model assumptions[2].

2.5 Citations

1. Li Cai, David Thissen. 16 Dec 2014,
   Modern Approaches to Parameter Estimation in Item Response Theory from: Handbook of Item Response Theory Modeling, Applications to Typical Performance Assessment Routledge
   Accessed on: 28 Mar 2023
   https://www.routledgehandbooks.com/doi/10.4324/9781315736013.ch3

2. Wood, J. (2017, November 12).
   Logistic IRT Models. QuantDev Methodology. PennState University
   Retrieved March 28, 2023, from
   https://quantdev.ssri.psu.edu/sites/qdev/files/IRT_tutorial_FA17_2.html

3 Individual Contributions

3.1 Ivan

- kNN code and writeup
- Neural network code and writeup
- Ensemble writeup
- Checking over work

3.2 Sam

- 1PL IRT code and writeup
- Ensemble code and writeup
- 3PL IRT code and writeup
- Checking over work