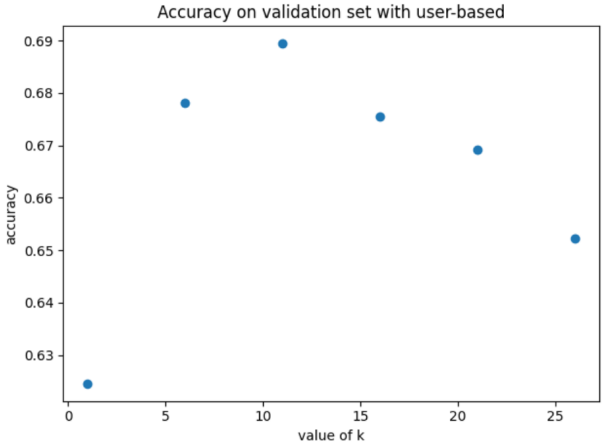
CSC311 Final Project

Kehui Li, Wuyue Lu

Part A

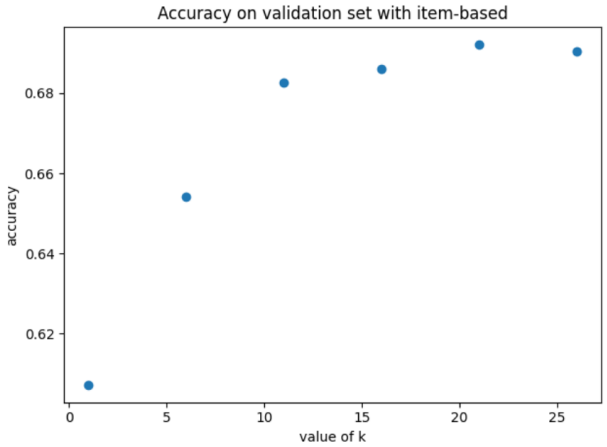
1. k-Nearest Neighbor

a) The accuracy of the validation data is shown below:



b) The test accuracy: 0.6841659610499576 with chosen k* = 11

c) The accuracy of the validation data is shown below:



The test accuracy: 0.6816257408975445 with chosen k* = 21

d) User-based has a higher test accuracy of 0.684 than the item-based of 0.681. User-based will be a better filtering.

e)

1. As the Knn needs to store the entire dataset in memory, and computation must be done for each query at test time, this is really expensive by the standard of a learning algorithm.
2. In high dimensions, “most” points are approximately the same distance. As for each student, it has thousands of questions for comparing similarities, same problem for item-based approach. Thus, the nearest distance might not be filtering well.

2. Item Response Theory

a) Suppose there are M students and N questions.

$$p(c_{ij}=1|\theta_i,\beta_j)=\frac{exp(\theta_i-\beta_j)}{1+exp(\theta_i-\beta_j)}$$

$$p(c_{ij}|\theta_i,\beta_j)=(\frac{exp(\theta_i-\beta_j)}{1+exp(\theta_i-\beta_j)})^{c_{ij}}(\frac{1}{1+exp(\theta_i-\beta_j)})^{1-c_{ij}}$$

$$p(C|\theta,\beta)=\prod_{i=1}^{M}\prod_{j=1}^{N}p(c_{ij}|\theta_i,\beta_j)$$

$$\log p(C|\theta,\beta)=\sum_{i=1}^{M}\sum_{j=1}^{N}c_{ij}\log(\frac{exp(\theta_i-\beta_j)}{1+exp(\theta_i-\beta_j)})+(1-c_{ij})\log\frac{1}{1+exp(\theta_i-\beta_j)}$$

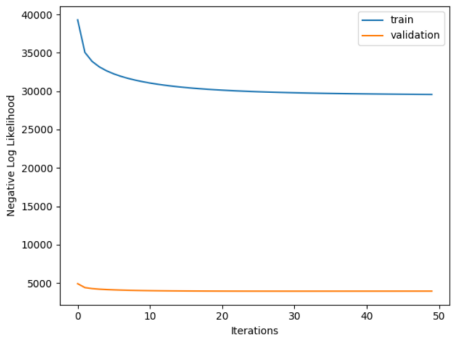
$$=\sum_{i=1}^{M}\sum_{j=1}^{N}c_{ij}(\theta_i-\beta_j)-c_{ij}\log(1+exp(\theta_i-\beta_j))-(1-c_{ij})\log(1+exp(\theta_i-\beta_j))$$

$$=\sum_{i=1}^{M}\sum_{j=1}^{N}c_{ij}(\theta_i-\beta_j)-\log(1+exp(\theta_i-\beta_j))$$

$$\frac{\delta\log p(C|\theta,\beta)}{\delta\theta_i}=\sum_{j=1}^{N}c_{ij}-\frac{exp(\theta_i-\beta_j)}{1+exp(\theta_i-\beta_j)}$$

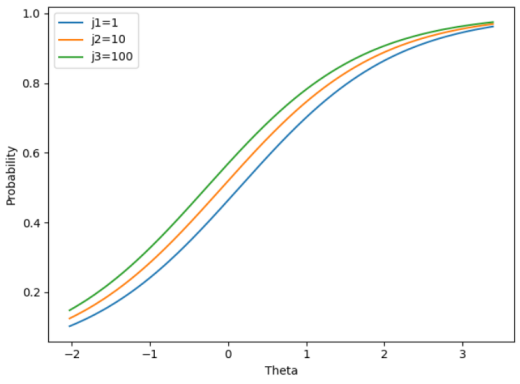
$$\frac{\delta\log p(C|\theta,\beta)}{\delta\beta_j}=\sum_{i=1}^{M}-c_{ij}+\frac{exp(\theta_i-\beta_j)}{1+exp(\theta_i-\beta_j)}$$

b) The hyperparameters are: lr = 0.01, iterations = 50.



c) Validation accuracy is 0.7060400790290714, test accuracy is 0.7067456957380751.

d) Following graph shows the probability of the correct response as a function of students' ability given question 1, 10 and 100. From the graph we can observe that the probability is monotonically increasing with respect to ability. This means that as the ability of the student increases, the probability of correctness of all questions increases. And among the three questions, students have the best performance in answering q100.



3. Neural Networks.

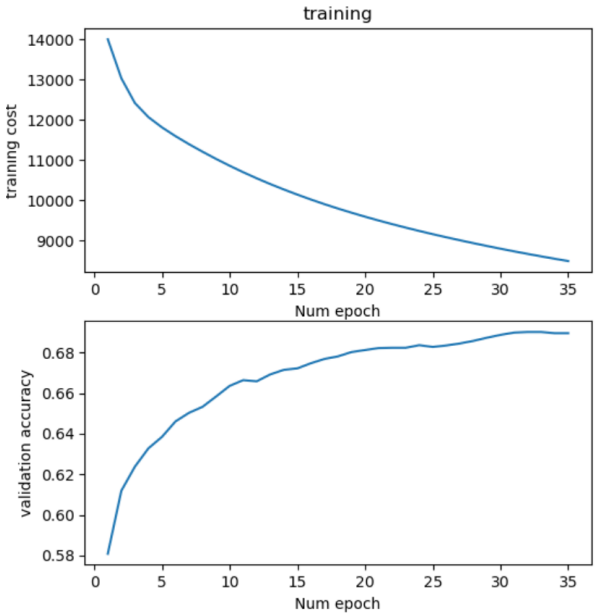
a) 1. For ALS, the objective is non-convex in U and Z jointly, we need to fix Z and optimize U, followed by fix U and optimize Z, and so on until convergence. The Neural network can work with the linear function. And for the objective that is non-convex, the neural network works with gradient descent.

2. ALS is optimized by minimizing loss with respect to both U and Z, while for autoencoders, we only need to minimize loss with respect to the weight matrix.

3. As for each U and Z in ALS for each u_i and z_i same as a linear regression problem, we can derive an optimal solution. However, for neural networks, gradient descent with gradients computed via backprop is used to train the overwhelming majority of neural nets today. We approximate the optimal solution.

c) Select k* = 10, with a learning rate of 0.03 and the number of iterations 35, has the highest validation accuracy of 0.6895286480383855.

d)Test accuracy: 0.6855771944679651



e)Same hyperparameters in (d), that is, k* = 10 , iterations = 35, learning rate = 0.03

lambda	validation accuracy
0.001	0.6874117979113745
0.01	0.6714648602878917
0.1	0.6243296641264465
1	0.6244707874682472

With λ = 0.001, the validation accuracy slightly decreases, test accuracy slightly increases. The validation accuracy is 0.6874117979113745, the test accuracy is 0.6886819079875811

4. Ensemble

Ensemble process:

- Resample three datasets from the training set using bootstrap method.
- Train models independently on three resampled datasets.
- Average the probability $p \in [0, 1]$ produced by the three models.
- If the average probability is greater than 0.5, we predict the student is correct on this question

With learning rate = 0.1 and number of iterations = 50, the final validation accuracy is 0.707874682472481 and test accuracy is 0.7090036692068868.

The ensemble model performs slightly better on both validation and test set than the IRT model. Since we are averaging over independent samples, the amount of variability in the predictions is reduced. Hence, the overfitting of models is eliminated.

Part B

1.Introduction

In the following parts, we would try to modify the Item Response Theory model to obtain a better prediction. In part A, we have tried to improve the IRT model using the bagging method, which is able to reduce overfitting. However, from what we obtained, the accuracy on validation and test set did not improve much, indicating the overfitting of the model is not severe. Instead, the low training accuracy (0.7067) indicates that the model is too simple to capture the relationship between input and output variables precisely. To address the problem, we came up with two extensions to increase model complexity.

2.Extension 1: Two-Parameter Model

2.1.Description

In the previous part of the IRT model, we assume that for all questions, the curves showing their probability of correctness with respect to ability have the same shape. However, different questions do not necessarily have the same characteristic curve in practice. To improve this, based on the IRT model in part A, we will introduce a new parameter λ called a discrimination parameter, which measures the differential capability of the question. The resulting model is called a Two-Parameter IRT model (which we refer to as IRT2 model in this section). The probability that the question j is correctly answered by student i is formulated as:

$$p(c_{ij} = 1|\theta_i, \beta_j, \lambda_j) = \frac{exp(\lambda_j(\theta_i - \beta_j))}{1 + exp(\lambda_j(\theta_i - \beta_j))}$$

where λ_j is the discrimination parameter for question j, β_j represents the difficulty of question j, and θ_i is the i-th student's ability.

The log-likelihood and its derivative with respect to the three parameters are shown below:

$$\log\, p(C|\theta, \beta, \lambda) = \sum_{i=1}^{M}\sum_{j=1}^{N} c_{ij}(\lambda_j(\theta_i - \beta_j)) - \log(1 + exp(\lambda_j(\theta_i - \beta_j)))$$

$$\frac{\delta \log\, p(C|\theta, \beta, \lambda)}{\delta \theta_i} = \sum_{j=1}^{N} c_{ij}\lambda_j - \lambda_j \frac{exp(\lambda_j(\theta_i - \beta_j))}{1 + exp(\lambda_j(\theta_i - \beta_j))}$$

$$\frac{\delta \log\, p(C|\theta, \beta, \lambda)}{\delta \beta_j} = \sum_{i=1}^{M} -c_{ij}\lambda_j + \lambda_j \frac{exp(\lambda_j(\theta_i - \beta_j))}{1 + exp(\lambda_j(\theta_i - \beta_j))}$$

$$\frac{\delta \log\, p(C|\theta, \beta, \lambda)}{\delta \lambda_j} = \sum_{i=1}^{M} c_{ji}(\theta_i - \beta_j) - (\theta_i - \beta_j)\frac{exp(\lambda_j(\theta_i - \beta_j))}{1 + exp(\lambda_j(\theta_i - \beta_j))}$$

Same as what we did in part A, use alternating gradient descent on θ , β, and λ to minimize the negative log-likelihood.

2.2.Results and Comparison

Accuracy Comparison:

We have tried the IRT2 model with six groups of hyperparameters (lr = 0.01, iterations ∈ [10,20,30,40,50,60]) and observed that the accuracy would slightly improve with the increment of iterations until 50. Therefore, we selected the model trained with lr = 0.01 and number of iterations = 50. Following is the comparison of accuracies:

Model	*Iterations*	*Learning Rate*	*Validation Accuracy*	*Test Accuracy*
KNN	N/A	N/A	68.95%	68.41%
Neural Network	35	0.03	68.95%	68.56%
IRT	50	0.01	70.60%	70.67%
Ensemble	50	0.01	70.70%	70.70%
IRT2	50	0.01	70.72%	70.72%

The new model performs much better compared to the knn and neural network, and slightly better than the base model and IRT using the bagging method.

Parameter Comparison:

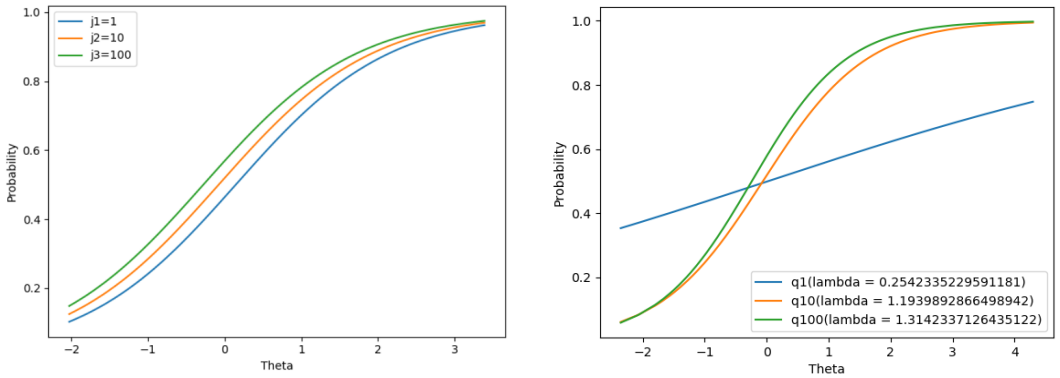


Figure 1: Item Characteristic Curves for IRT and IRT2

Unlike Part A, all questions now have different shapes of characteristic curves. As the theta value changes from 0 to 1, the probability of a correct response changes from 0.6 to 0.9 for q10 and q100, which is much larger than q1. For that reason, questions 10 and 100 can differentiate students whose ability value is around 0.5 more efficiently than question 1 can.

2.3.Limitations and Improvements

While IRT2 considers the difference in differential capability of questions, students might also be different. It is possible that for different groups of students, the model fit for them is quite different, that they might have different best fit values for hyperparameters. One of the possible improvements and our second extension is using multiple-group IRT models, which analyze questions in independent groups to study differential item functioning or invariance.

3.Extension 2: Multigroup Analysis

3.1.Description

It is natural to think about whether students with same features such as ages and genders would have similar reactions towards questions. In the following section, we are going to focus on responses in different groups of students.

We first divided students into three groups by their birth dates: students born in year ~2003, 2004~2007 and 2008~2021. Train three models with these subpopulations and calculate the final accuracy. Then we divided students into three groups by their gender: female, male and unspecified and did the same steps as above.

3.2.Result and Comparison

Group By age:

group_2004: students with birthdays before 2004, including 2004 with data size 133.

group_2007: students with birthdays between 2004 and 2007, including 2007, with data size 168.

group_2021: students with birthdays after 2007, with data size 65.

Students whose birthday is unspecified are not included in any group.

Model	*Iterations*	*Learning Rate*	*Validation Accuracy*	*Test Accuracy*
IRT2	50	0.01	70.72%	70.72%
IRT2(group_2004)	60	0.01	72.36%	72.00%
IRT2(group_2007)	50	0.01	69.08%	69.84%
IRT2(group_2021)	50	0.01	66.78%	60.38%

The validation accuracy and test accuracy increase pretty much in group_2004 compared to the single group model. However, the model for group_2007 and group_2021 predicts worse on the validation set and test set than IRT2.

The better performance in group_2004, even with relatively small data size, compared to the other two groups might be because of the factor of random guess. Students make guesses when encountering a hard question. Older students may have better logical guesses than younger ones.

Group by Gender:
Female: female students with data size 202
Male: male students with data size 277
Unspecified: Group of students whose gender is Unspecified with data size 133

Model	*Iterations*	*Learning Rate*	*Validation Accuracy*	*Test Accuracy*
IRT2	50	0.01	70.72%	70.72%
IRT2(female)	30	0.01	70.11%	70.19%
IRT2(male)	40	0.01	69.62%	71.18%
IRT2(unspecified)	50	0.01	71.01%	66.91%

Validation accuracy only slightly increases in the unspecified group, and test accuracy only increases in the male group. The result shows that it is not a good way of splitting groups, that the relation between gender and their behavior towards questions is weak.

3.3.Limitations and Improvements

The multigroup analysis requires dividing the dataset into subsets, and trains models separately. Size of training data for each model is much smaller than the size of the original dataset. The reduced dataset may lack some crucial features for the models to recognize. Thus, this method works better when the original dataset is big enough. One possible improvement could be resample the subset by ensemble, which helps to reduce model overfitting.

4.Conclusion

In conclusion, we tried two approaches to optimize the model. Through the two-parameter IRT, we are able to make a more complex model on characteristics of questions but there is only little improvement. Then, we applied the two-parameter IRT on different groups of students and observed that its performance on predicting older students improved even with a relatively small data size, but the other groups seemed to have little to no effect. This model can still be improved using three- and four- parameter models, which consider students would have random guessing and ceiling parameters in the response curves.

5.Reference

An, X., & Yung, Y.-F. (n.d.). *Item Response Theory: What It Is and How You Can Use the IRT Procedure to Apply It*. Support.Sas.Com. Retrieved December 1, 2021, from https://support.sas.com/resources/papers/proceedings14/SAS364-2014.pdf

6.Contribution

Part A
Q1: Kehui Li
Q2: Wuyue Lu
Q3: Kehui Li
Q4: Kehui Li & Wuyue Lu

Part B
Both Kehui Li and Wuyue Lu participated in discussion of extended models. Kehui Li contributed to model research and implementation and carried out experiments. Wuyue Lu helps to structure the report and analyze the results.