Question Selection based on Item Discrimination Estimates

TARANVEER SINGH
MASTERS IN COMPUTATIONAL DATA SCIENCE
CARNEGIE MELLON UNIVERSITY

Abstract

In this challenge the task was to select set of questions from the given question bank that are able to estimate best relative ranking between students. Quality of relative ranking between students depends on information each question provides and how well a question is able to discriminate between students. In this challenge I have used two different approaches, Item Response Theory and Relative Ability estimates & Collaborative Filtering to find question discrimination estimates and done a comparative study between them. These estimates are used for question selection analysis based on their discriminative power. Also, the set of questions selected as a whole should provide good coverage over the range of abilities.

Note: Python Used for Relative Ability Estimate and Collaborative Filtering and R used for rest of the analysis

Table of Contents

1.Introduction and Methodology

1.1 Overview

In this challenge I have used two approaches - (1) Item Response theory and (2) Relative Ability estimate & Collaborative filtering to find item discrimination. This was followed by Question Selection Analysis in order to select the best discriminative questions that also provides a good coverage on possible student ability range.

1.2 Item Response theory

Item Response theory is a standard in psychometric testing for measuring latent ability of the user based on item responses. The idea behind the theory is students with latent ability less than the item difficulty will have low probability of answering the question and students with latent ability more than the item difficulty will have higher probability of answering the question, hence the difference between student ability and item difficulty vs probability of answering follows a sigmoid curve known as “Item Characteristic Curve”. The slope of the sigmoid at the pivot point (item difficulty) gives us item discrimination parameter, steeper the slope, higher it’s discriminative power in other words it provides high information about student ability at that point. In this challenge I have used a 2PL model which has discrimination and difficulty of the question as the two parameters. These are estimated through EM approach

$$P(\theta) = \frac{1}{1+e^{-L}} = \frac{1}{1+e^{-a(\theta-b)}}$$

θ : *Ability Level*

b : *Item discrimination*

a : *Discrimination Parameter*

Drawbacks of estimating parameters via EM especially on sparse data:

1. Slow to converge
2. May not converge at all
3. May get stuck in a local minimum.

Because of these drawbacks, I also worked on alternate methods for finding Item Discrimination apart from this

Code: IRT.R - LTM Package (Standard CRAN package) required to run EM takes 2 hour to find good estimates.

Item discrimination curve for Table 1 question is plotted below

	question_id	Diff	Discrm	index
177	1591	-1.565	8.741	177
281	12716	1.117	0.169	281
289	12734	-0.144	1.343	289

Table 1

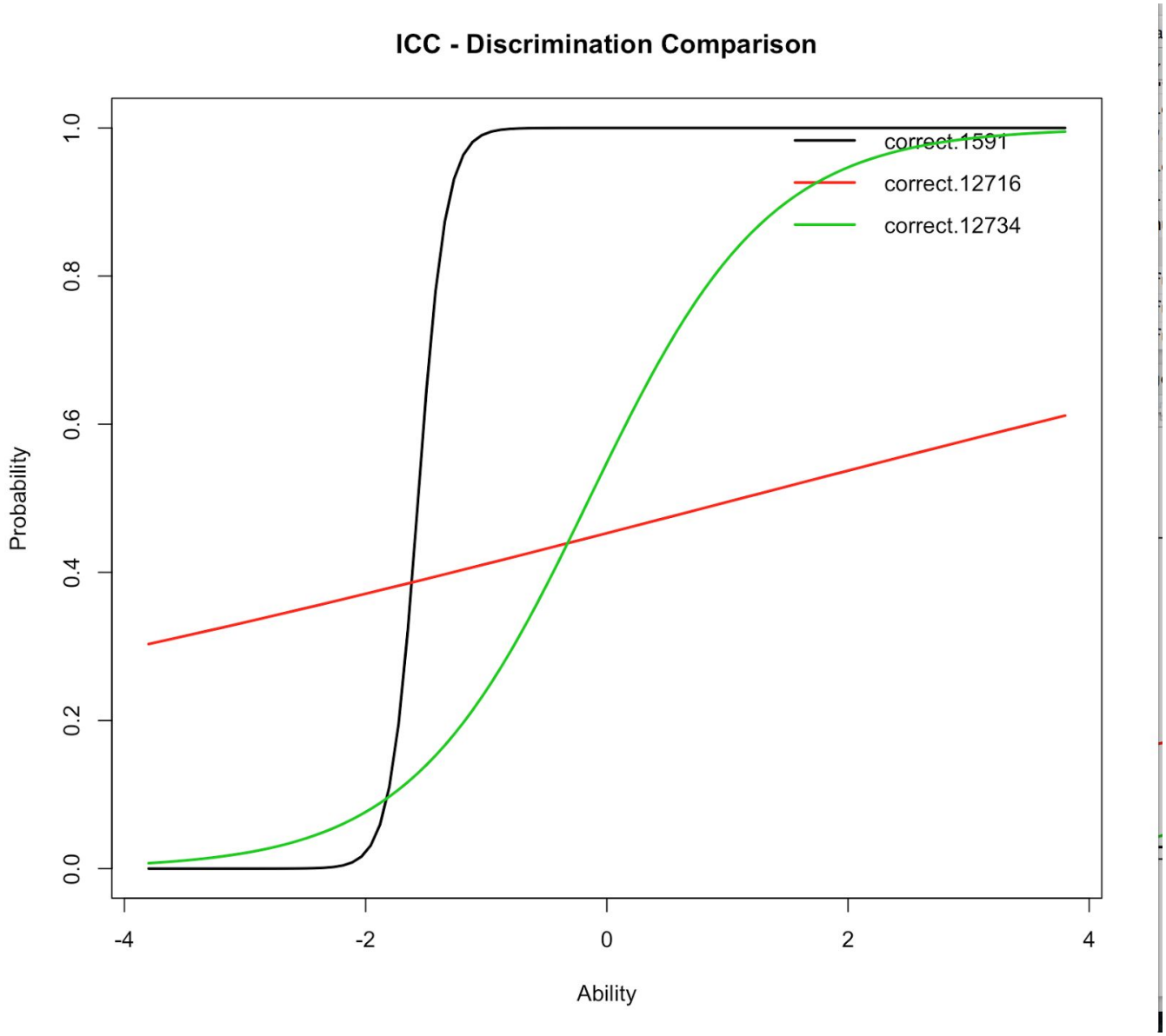

Figure 1: Item Discrimination comparison for low **high** and medium discrimination

1.3 Relative Ability Estimate

Since estimating parameters of Item Characteristics curve is computationally heavy especially with incomplete data I have tried to solve the problem through Relative ability estimate which is computationally inexpensive and is a reasonable approximation. Relative Ability tells us how better is a student from his other peers. Student's relative ability can be estimated based on how well he performed on the questions he was given and how well his peer responded to those questions or in other how difficult were the questions he was given. This estimate for a student j over his observed responses can be calculated by the following equation:

$$\textit{Student Mean (j)} = \frac{no\ of\ correct\ responses}{Obs\ (j)}$$

$$\textit{Question Difficulty (i)} = \frac{no\ of\ correct\ responses}{ObsStudents\ (i)}$$

$$\textit{Relative Ability (j)} = \frac{Student\ Mean\ (j)}{\frac{1}{Obs\ (j)} \sum_{i}^{Obs\ (j)} Question\ Difficulty(i)}$$

Obs (j) : No of questions observed by student j
ObsStudents (i) : No of students who answered Question i

These relative ability estimates for each student can then be used to predict response of student j on question i and hence we can predict missing values in our data

Response Question (j,i) = Relative Ability (j) x Question Difficulty (i)

This estimate was used to perform data imputation which is required to find item discrimination estimate.

Code : RelAbilityEstimate.pynb (Jupyter notebook)

1.4 Collaborative Filtering

Relative ability did some approximation as it was accumulating responses and not capturing the pattern of responses. In reality a student might be weak in certain concepts and strong in the other. This idea led me to find user-user similarity based on their response pattern and predict a student's response based on how highly correlated user responded to that question.

Collaborative filtering was done only to predict question responses that could be done with good confidence i.e if highly correlated user has answered the question which his peer has not answered. On top of this imputed data I performed Relative Ability estimation. This imputed data was the used to find Relative Ability estimates so that the estimates also captures similarity of the user in some sense. As, a result data imputation accuracy using Relative ability estimate improved by adding the Collaborative filtering model.

Note: In order to find user-user similarity I used **Pearson correlation with damping factor** to get high quality correlation.

Accuracy of data Imputation:
Relative Ability Estimate : 75.1%
Collaborative Filtering + Relative Ability Estimate: 78.7%

This shows some amount of information was captured by Collaborative Filtering which boosted the accuracy.

Code : RelAbilityEstimate.pynb (Jupyter notebook)

1.5 Item Discrimination

Item Discrimination as seen in Item response theory estimate how well it separates upper ability group from lower ability group. Hence it can be approximated as below

Item Discrimination ≃ % (Upper Group Correct) - %(Lower Group correct)

To be more stringent we can calculate correlation of students responses on a question with their total score. So correlation will be high when higher ability user answer correct and low ability user answer incorrectly.

Note: I used **biserial correlation** to calculate correlation between categorical variable and continuous variable.

1.6 Question Selection Analysis

Coefficients generated by IRT model was used for this analysis. However Relative Ability estimates have correlated Item Discrimination values. Similar Analysis can be done on the same.
Code : QuestionAnalyis.pynb (Jupyter notebook)

1.6.1 Outlier Analysis

Since the matrix was very sparse, Questions with low frequencies will not have statistically significant parameter estimates of item information curve. Outliers in Discrimination and Difficulty had a significant overlap of about 67% in Questions which had less that 40 values observed which proves my hypothesis.

Based on the above result I removed the questions that had frequencies less than (mean - 1.5*sd)

1.6.2 Difficulty Group Estimation

To estimate good cluster means I first removed outlier using IQR Analysis and then used kmeans with 5 centers to determines question corresponding to 5 different difficulty levels.

Reason: Since 5 questions are selected at random from the question bank and in general we want to estimate ranking for students with all kind of abilities. Having equal representation of all level of questions in the question bank will maximize the probability of getting questions at each level. I could have seen the ability range of students in current data to determine questions no of question levels but since the current question bank is not reliable we can't say for sure what the actual ability of the learner is.

1.6.3 Discrimination Quality Estimation

Frank B. **Baker** (1985) grouped Discrimination values into six groups

Discrimination Values	Discrimination Quality
< 0.7	Very low
0.7 - 0.34	low
0.35 - 0.64	Moderate
0.65 - 1.34	High
1.35 - 1.7	Very high
Inf	Perfect

Table 2

Note: Negative Discrimination is suspicious question which needs to be reviewed.

1.6.4 Method 1: Top Discrimination Quality Selection

The following table was obtained based on the current data

Discrimination Quality vs Difficulty Level

	1	2	3	4	5
0_very_high	25	23	21	4	0
1_high	24	41	30	19	6
2_moderate	11	38	39	44	35
3_low	2	0	0	0	4
4_very_low	0	1	1	0	4
5_Suspicious	0	0	1	0	0

Table 3

Based on these I selected “very high” and “high” discriminative questions from every category except level 5 where “moderate” was also considered because of not enough questions in the “very high” and “high category”. In the following method not all difficulty levels have equal representation. Figure 2 shows comparison between Question Kept and Question Eliminated.

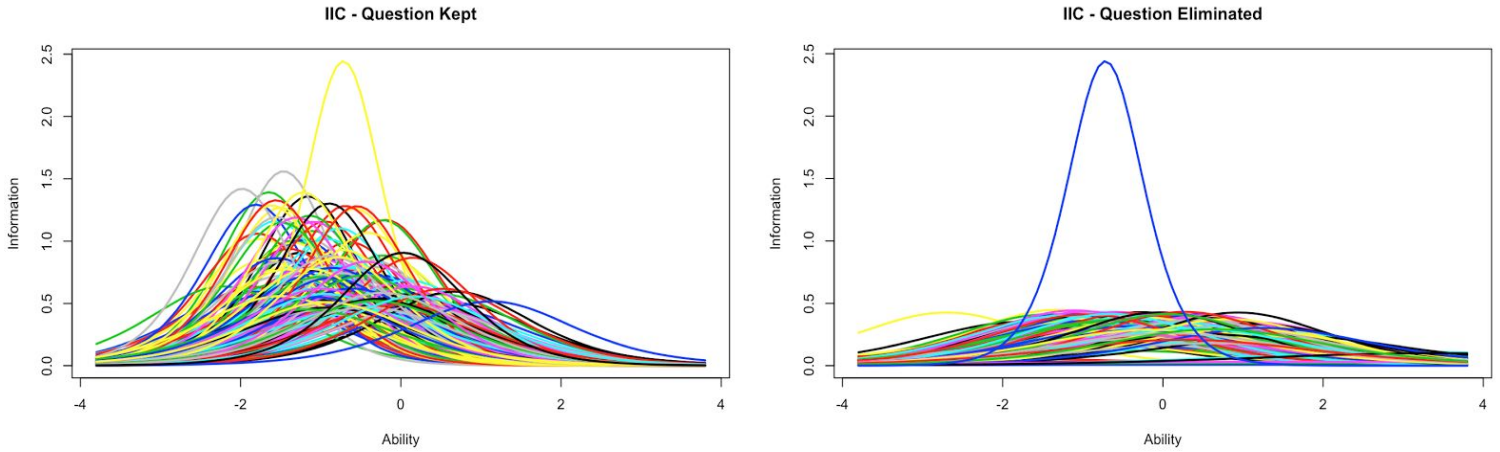

Figure 2: Item Information Curve - Method 1

Figure 2 clearly shows highly discriminative questions (high information) were selected as compared to questions that were eliminated. Blue colored High Information curve is suspicious question with negative discrimination. (question id: 1232) - Generated by **Plot_IIC.R**

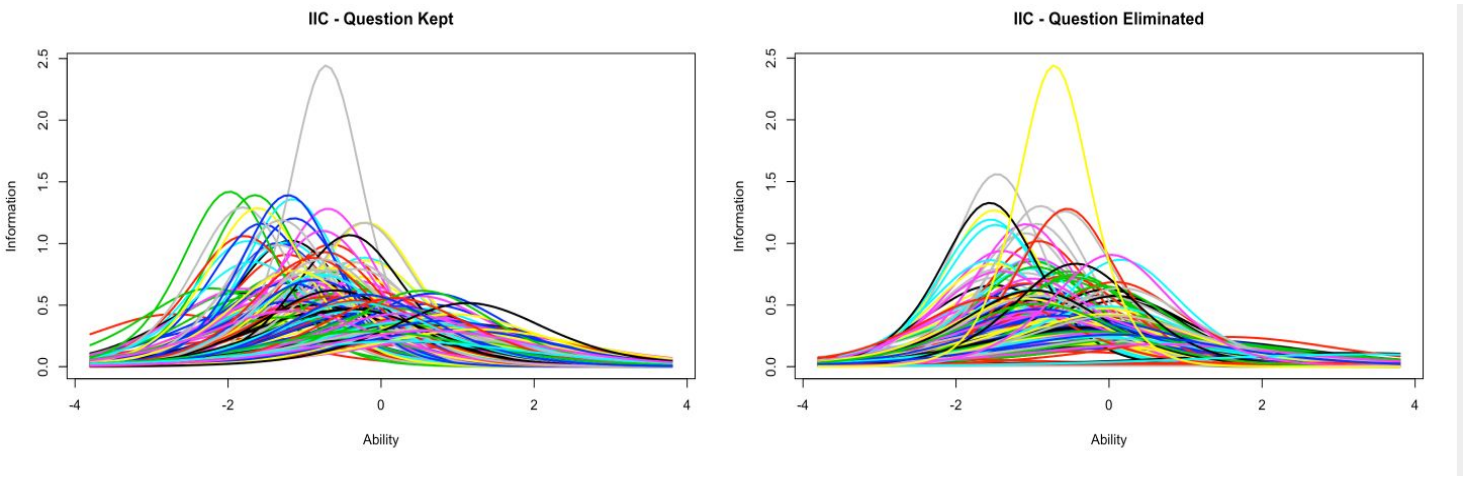

Figure 3: Item Information Curve - Method 2

Figure 3 shows high coverage across abilities as compared to questions that were eliminated

1.6.4 Method 2: Maximize Ability Coverage (Equal proportion) with good Discrimination

The main idea behind this is to

1. Maximize coverage of all kind of abilities
2. Get equal proportion of question across all levels
3. Get reasonable discriminative values keeping in mind the above two constraints

Step 1: Since Level 5 (“Hard”) Questions have very few good discriminative questions as shown in Table 3. So Level 5 is the bottleneck. Total no of questions in the hard category have reasonable discriminative power (moderate and beyond) is the no of questions k (here 41) that will be selected from the remaining levels

Step 2: In this method standard deviation for each difficulty level standard deviation for every discrimination quality is considered, if standard deviation within that group is low then the lower discrimination group is also added to the pool of available options. However nothing below moderate is taken whatsoever

Step 3: Select k questions from the pool so that they are farthest apart from each other.

Figure 3 shows the Item Information curve for Method 2

.

1.6.5 Result Comparison (IRT vs Relative Ability Estimate)

Performance - i7 processor

1. IRT took 120 mins to to a reasonable estimate but it still didn't converge.
2. Relative Ability estimate took 5 mins to calculate all the estimates and item discrimination

Overlap and Correlation

1. % Overlap in low discriminative power group : 55%
2. Correlation in Discriminative estimates: 0.38 with p values < 10^(-10) which means high confidence in correlation.

Suspicious Question Identification

Both identified questions with negative discrimination correctly.

2. Conclusion and Future Work

Both the approaches have decent similarity in estimating discriminative power. This can be further developed. Also I haven't worked on tweaking the parameters for estimating threshold of highly correlated users and damping factor which can be worked upon.Question Selection Process (Method 1 and Method 2) has an exploratory part to it which can be automated by generating scores based on various features like Discrimination, proportion and confidence of estimates

3. Flow Chart - Link between modules