Question 2

(a)

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

Notice that

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

Then

$$p(c_{ij}=0|\theta_i,\beta_j)=1-p(c_{ij}=1|\theta_i,\beta_j)=\frac{1}{1+\exp(\theta_i-\beta_j)}$$

Thus

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

Therefore,

$$p(C|\theta,\beta) = \prod_{i=1}^{N}\prod_{j=1}^{M}\left(\frac{\exp(\theta_i-\beta_j)}{1+\exp(\theta_i-\beta_j)}\right)^{C_{ij}}\left(\frac{1}{1+\exp(\theta_i-\beta_j)}\right)^{1-C_{ij}}$$

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

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

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

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

$$\frac{\partial l}{\partial\theta_i} = \sum_{j=1}^{M} C_{ij} - \frac{\exp(\theta_i-\beta_j)}{1+\exp(\theta_i-\beta_j)}$$

$$\frac{\partial l}{\partial\beta_j} = \sum_{i=1}^{N} -C_{ij} + \frac{\exp(\theta_i-\beta_j)}{1+\exp(\theta_i-\beta_j)}$$

(b)

In this part, I user the following hyperparameters:

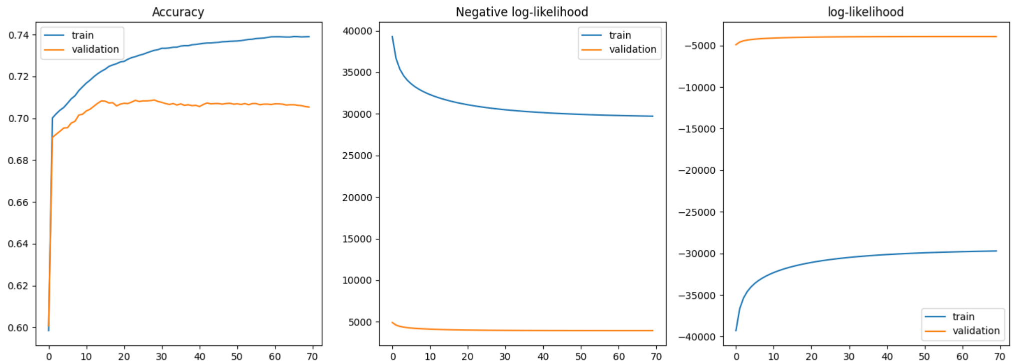
Number of Iterations: 70

Learning Rate: 0.05

Initialize theta: All 0s

Initialize beta: All 0s

Training process is shown as below:



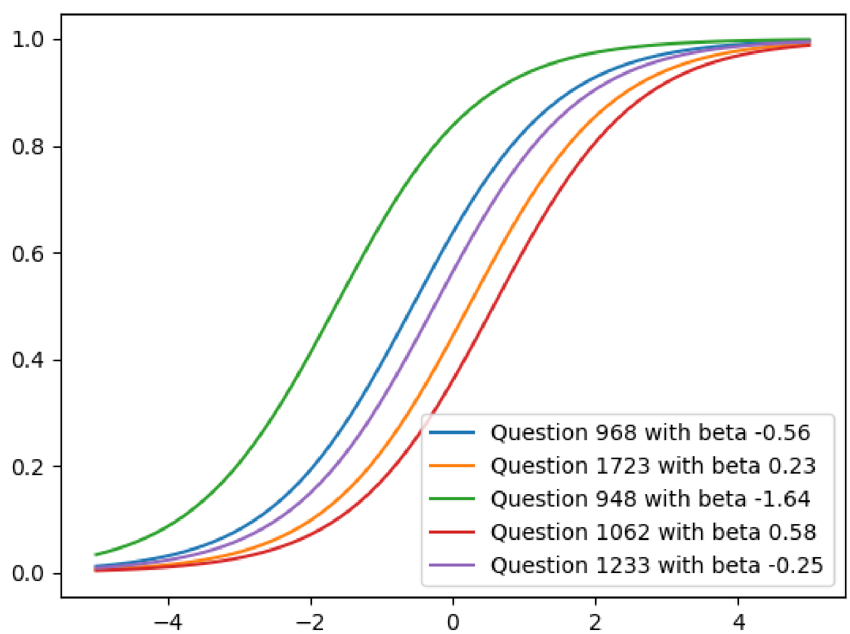
(c)

Final Train Accuracy is: 0.7390100197572679

Final Validation Accuracy is: 0.7053344623200677

Final Test Accuracy is: 0.7047699689528648

(d)



In this part, I use different as $p(c_{ij})$ as a function of θ given 5 different questions. We can see the shape of those 5 curves are in a shape of sigmoid function, which means it is a transformation of sigmoid function. The reason why it looks like sigmoid function is $p(c_{ij}) = sigmoid(\theta_i - \beta_j)$.

We can consider θ as the ability of the students and the β as the difficulty of the question. Then suppose we have the same value of θ, that is we have the same value on x-axis, which means the students have same ability, the higher the value of y-axis is, the easier the question is. This is because the higher value of y-axis represents a higher probability that the students can answer the question correctly. From the graph above, we also can notice the curve of the question with lower β(easier) is at the top of this graph while the ones with larger β(harder) is at the bottom.