CSC311 Project Final Report

Group Members:
Zixuan Zeng 1008533419

July 30, 2024

Part A

Q1

1.(a)

```
knn_impute_by_user:
Validation Accuracy: 0.6244707874682472
Validation Accuracy: 0.6780976573525261
Validation Accuracy: 0.6895286480383855
Validation Accuracy: 0.6755574372001129
Validation Accuracy: 0.6692068868190799
Validation Accuracy: 0.6522720858029918
```

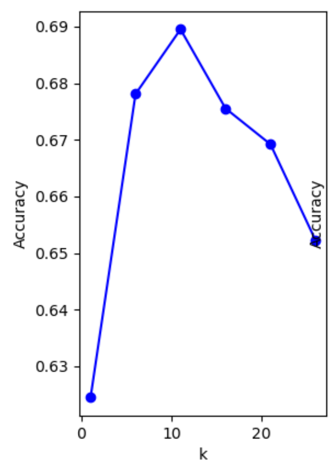

Figure 1: Accuracy vs k for KNN Impute by User

1.(b)

```
Chosen argmax k*: 11 , Test accuracy: 0.6841659610499576
```

Figure 2: Test accuracy with k*

1.(c)

The underlying assumption is that if question A is answered similarly by many students as question B, A's predicted response from specific students matches that of question B.

```
knn_impute_by_item:
Validation Accuracy: 0.607112616426757
Validation Accuracy: 0.6542478125882021
Validation Accuracy: 0.6826136042901496
Validation Accuracy: 0.6860005644933672
Validation Accuracy: 0.6922099915325995
Validation Accuracy: 0.69037538808919
```

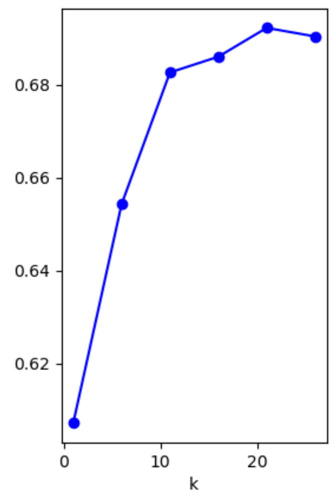

Figure 3: Accuracy vs k for KNN Impute by Item

```
Chosen argmax k*: 21 , Test accuracy: 0.6683601467682755
```

Figure 4: Test Accuracy with k*

1.(d)

The test accuracy for the user-based method (0.6842) is higher than that for the item-based method (0.6684). Therefore, the user-based collaborative filtering method performs better than the item-based collaborative filtering method in this case.

1.(e)

- Computationally expensive. KNN practically has no training process. With large datasets, as the number of students/questions grow, the time required to compute the distances and to identify the nearest neighbors at test time grows significantly.
- Curse of Dimensionality. When the sparse_matrix has too many missing values, it's hard to find good nearest neighbors, since most points will be about the same distances. This affects the prediction accuracy.

Q2

2.(a)

Given the probability that student i correctly answers question j as:

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

The log-likelihood for all students and questions, given the sparse matrix C, is:

$$\log p(C|\theta, \beta) = \sum_{(i,j)\in\text{observed}} [c_{ij}(\theta_i - \beta_j) - \log(1 + \exp(\theta_i - \beta_j))]$$

The derivative of the log-likelihood with respect to the ability parameter θ_i is:

$$\frac{\partial \log p(C|\theta, \beta)}{\partial \theta_i} = \sum_{j\in\text{observed}} [c_{ij} - \sigma(\theta_i - \beta_j)]$$

The derivative of the log-likelihood with respect to the difficulty parameter β_j is:

$$\frac{\partial \log p(C|\theta, \beta)}{\partial \beta_j} = \sum_{i\in\text{observed}} [-c_{ij} + \sigma(\theta_i - \beta_j)]$$

Part B