1. Evaluation plan

As discussed in early chapters preliminary task of this research is to predict student ability give the correct answer to a question given a sequence of previous questions, correctness of answers and related learning objectives. For given question there are two outputs. They are correctly answered or not. Hence this is a binary classification problem. We develop mainly two models and use one benchmark model (BKT). It is equally important to predict students being able to give answers correctly or wrong. And dataset is approximately balance data set.

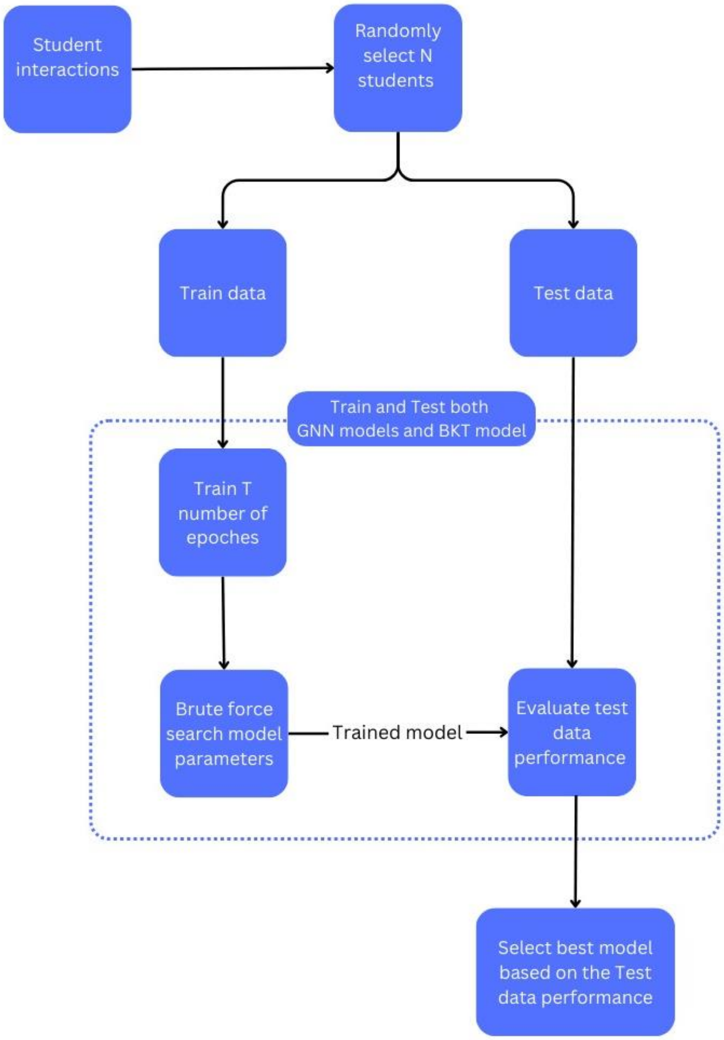
Student interactions
Randomly select N students
Train data
Test data
Train and Test both GNN models and BKT model
Train T number of epoches
Brute force search model parameters
Trained model
Evaluate test data performance
Select best model based on the Test data performance

Figure 1.1 Evaluation process

1.1 Accuracy

Accuracy is a commonly used metric for evaluating the performance of binary classifiers. It is defined as the proportion of correct predictions made by the classifier. In the context of a binary classifier, accuracy is calculated as shown in the Equation 1.1

$$\text{Accuracy} = \frac{\text{Number of correct positive predictions} + \text{Number of correct negative predictions}}{\text{Total number of predictions}}$$

Equation 1.1 Accuracy

Accuracy is a good metric to use when both true positives and true negatives are equally important because it considers both types of correct predictions. Therefore, we use accuracy to evaluate individual model performance. It also frequently used in related literature to compare models.

1.2 Area Under Receiver Operating Characteristic Curve

We also use Area Under Receiver Operating Characteristic Curve (ROC AUC) to compare models and select best model parameters. Receiver Operating Characteristic curve plots true positive rate vs false positive rate. ROC AUC, or Area Under the ROC Curve, is a performance metric for binary classification problems. It measures the ability of a classifier to distinguish between classes and is used as a summary of the ROC curve. An ROC curve is a plot of the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. The TPR is the proportion of positive cases that are correctly identified as positive, while the FPR is the proportion of negative cases that are incorrectly identified as positive.

The AUC is calculated by measuring the area underneath the ROC curve. A perfect classifier would have an AUC of 1, meaning that it can correctly identify all positive cases and correctly reject all negative cases. A random classifier would have an AUC of 0.5, meaning that it is no better than guessing.

AUC is a useful metric because it is not affected by the class imbalance in the data. This means that it can be used to compare the performance of classifiers even when the number of positive and negative cases is not equal. Hence we use ROC AUC as the second performance indicate the compare models.

1.3 Split train and test data

When models are trained, models can be overfitted to the data. Hence, we split data in to test data and train data. Each student's interaction sequence has an order, similar to a time series. Therefore we select the first 80% of the interactions as the training data of each student. Rest of the 20% of the interactions of each student consider as the test data. Each model trained on train on train data and calculate the model performance. Then we predict the student answer correctness using test and calculate the model performance to observe whether model is overfitted.

During the training process we consider the running average loss of each epoch to select best model parameters. We use Binary Cross Entropy function as the loss function, since this is binary classification task. Considering the lengthy time each model takes to train we can not use grid search to find the optimal model parameters, but we use brute force method to find better model parameters. As an example we adjust model learning rate , hidden layer size , number of layers etc.

Since we are using a new data set we can not compare our model performance directly with the previous studies. Hence we use BKT and train and test the same data set. All most all the other studies have used BKT as a benchmark model. This can be used to compare out model performance and data set with other studies.