114-1 Data Mining

Final Project: Knowledge Tracing Prediction

Task Introduction

For each student's interaction with a question, predict whether they will answer correctly.

- You are free to use any model or approach for this task.
- Requirement
  - Upload your submission to Kaggle
  - Submit a report and your source code to E3

Dataset

- train.csv
  - Student interactions and responses for training
- test.csv
  - Student interactions for prediction

Training Data

train.csv - The data is ordered chronologically for each student. each row contains student interaction data with the following columns:

- interaction_id - Integer. A unique identifier for each row/interaction in the entire dataset..
- question_id - Integer. Identifier for the question being answered.
- Problem- String. The full text content of the math question presented to the student.
- category - String. The general subject category of the question.
- student_id - Integer. A unique identifier for the student performing the interaction.
- student_process- String. The step-by-step work or reasoning provided by the student when solving the problem.
- concepts_of_the_problem- String. The specific mathematical concept(s) tagged for this particular question.
- concepts_lacking_in_student - String. An assessment indicating specific concept(s) the student seemed to struggle with during this specific interaction.
- response- Binary. Indicates whether the student's final answer to the question was correct (1) or incorrect (0). This is the primary target variable you need to predict. .

Testing Data

test.csv - The data is ordered chronologically for each student. each row contains student interaction data with the following columns:

- interaction_id - Integer. A unique identifier for each row/interaction in the entire dataset..
- question_id - Integer. Identifier for the question being answered.
- Problem- String. The full text content of the math question presented to the student.
- category - String. The general subject category of the question.
- student_id - Integer. A unique identifier for the student performing the interaction.
- concepts_of_the_problem- String. The specific mathematical concept(s) tagged for this particular question.

Note: test.csv does not include the response' column as this is what needs to be predicted.

Kaggle

- Competition Link
- Create a team with your **group ID**, we use this information for grading. Use this link to find your group ID, your team name should be group_<groupID> .
- If you failed to do so under any circumstances, there will be **a penalty of 5 points** to your score, so be sure to use the correct team name.

Kaggle (cont.)

- Public leaderboard is calculated with 50% of the test set, private leaderboard is calculated with the other 50%, the final standings may be different.
- DO NOT **overtune** your model to fit the public leaderboard, or you will suffer from overfitting
- The scoring metric is **AUROC** .

Kaggle Submission

- Report user with their index (column “interaction_id” in test.csv) and the corresponding prediction to each answer.
- There should be 404 x 2 entries in your csv file, with columns “interaction_id” and “response”.
- Please ensure your model outputs continuous probability values between 0 and 1 (not binary). This approach better reflects the model's confidence in each prediction.
- The order of the ids does not matter. Refer to sample_submission.csv for the correct format.

Report Submission

Answer the 4 questions. Please answer the questions in detail to receive full points for each question.

1. (10%) What preprocessing and feature engineering methods did you apply to the knowledge tracing data? Additionally, please discuss how different feature representations (e.g., question embeddings, concept embeddings, temporal features) contributed to model performance based on your experimental results.

2. (10%) Analyze how your model utilized both the textual content and the associated labels. Discuss different strategies you explored for representing or integrating these sources of information, and evaluate their impact on performance using experimental results.

Report Submission (cont.)

3. (10%) Try different model architecture. Describe your model architecture choices for knowledge tracing. Why did you select this particular architecture (e.g., LSTM, Transformer, or hybrid models)? How did architectural decisions such as hidden dimension size, number of layers, and attention mechanisms affect performance? Support your discussion with experimental evidence.

4. (10%) Conduct an error analysis of your knowledge tracing model's predictions. Identify patterns in misclassified student responses and discuss potential reasons for these errors. Are there specific concepts, question types, or student behavior patterns where your model performs poorly? What insights can be gained from these errors?

Grading policy

- Kaggle (60%)
  - 30% based on the public leaderboard score and 70% based on the private leaderboard score
  - Leaderboard score consists of basic score and ranking score
    - Basic score :
      Over strong baseline : 50
      Over simple bassline : 35
      Under simple baseline : 20
    - Ranking score:
      10-(10/N)*(ranking-1), N=numbers of teams in the interval
- Report (40%)
  - 10% for each question

E3 Submission

Submit your source code and report to E3 before 12/17 (Wed.) 23:59.

No late submission !

One group, one submission

Follow the submission format or there will be a deduction of **5 points** for final project!

- Format
  - source code : final_<Group_ID>.py or other files (.py) you made
  - report : final_<Group_ID>.pdf

If you have any question about final project, please feel free to contact with TA : MinKuang Tsu

through email vctmk.cs13@nycu.edu.tw

Have Fun !

Linear Algebra,
Calculus,
too difficult

Take Advanced Deep Learning, Machine Learning first

@AI迷因