CS F407: Artificial Intelligence

Creating an AI Model To Assess The Effectiveness and Viability of Intelligent Tutoring Systems (ITS)

Group Members:

Ishaan Kale - 2022A7PS0084P
Sriram Sudheer Hebbale - 2022A7PS0147P
Arunaav Padmapati - 2022AAPS0240P

Abstract:

Intelligent Tutoring Systems (ITS) are computer-based educational programs that provide personalised and adaptive learning experiences. Unlike traditional learning environments, ITS dynamically adjusts instruction based on individual learners' needs, preferences, and performance. They utilise artificial intelligence and machine learning algorithms to analyse learner data, offer real-time feedback, and tailor instruction to optimise learning outcomes. ITS can cover various subjects and educational levels, from elementary mathematics to advanced scientific concepts. These systems aim to enhance learning effectiveness, engagement, and retention by delivering customised instruction that matches the pace and style of each learner. Several studies have attempted to validate the effectiveness of intelligent tutoring systems by comparing their results with those of a teacher in a physical classroom. Our project aims to test how effectively Duolingo - a language teaching ITS, can do its job.

Introduction:

Duolingo is an application used to teach users various languages by showing words of a language the user wants to learn and then testing the user through a plethora of methods, one of which is showing a given word from the language to the user at different intervals of time.

We've attempted to test the intelligent tutoring system (ITS), which tries to predict how probable it is for the user to guess a given word correctly in the future. The results would be used to gauge how effective the ITS was in fulfilling its purpose.

We have done this by calculating the parameter 'p_recall' value, which is the ratio of how often a user gets the word correct to how many times the user has been shown the word.

The data set we've used contains information about a large set of words taken from the app Duolingo. The dataset includes information on how often users have seen the word in the past, how many times users have seen a word in a current session and the probability of a user getting the particular word correct while being tested.

Additionally, this program could be extended to predict how difficult it would be to guess a given word.

Literature Review:

We referred to multiple research papers while conducting our study for this project.

The papers we referred to mainly catered to the following categories:

1. The process and difficulties involved in creating and training an ITS
2. Creating an ITS through programmable and non-programmable methods (Using Cognitive Authoring Tutoring Tools (CTATs))
3. The intricacies of creating an ITS, based on what it is attempting to teach (Such as Electrical/Manufacturing Engineering, Computer Programming, Or engineering simulation software such as SPICE and LASAR)
4. Comparing the efficiency of an ITS with other teaching pedagogies.

Intelligent Tutoring Systems (ITS) aim to guide students through step-by-step problem-solving processes using specific methodologies tailored to the subject being taught. The development of an ITS varies depending on the subject matter. For example, in middle school mathematics, an ITS may employ "Example Tracing," offering hints and feedback as students work through problems. Similarly, an ITS for teaching simulation software like SPICE breaks down learning into stages, from basic tool introduction to advanced coding. Both types of ITS adapt teaching based on student responses, leading to effective learning. Research indicates that ITS can perform as well as, if not better than, other teaching methods, falling slightly behind human-led instruction in small class settings.

Although there are multiple advantages to employing an ITS, we must also consider the following possible drawbacks:

1. **Limited Generalizability:** Creating an ITS may only be possible for some domains. ITSs may only be limited to subjects such as Computer Science and Mathematics.
2. **Limited sample size and difficulty obtaining a dataset:** Creating an efficient and reliable ITS would require a large amount of data, which needs to be specific to the model being developed, and it is a problem we are facing ourselves.
3. **Short-Term Assessment:** ITSs may only give learners short-term feedback and may ignore the long-term learning outcome of the students.

Motivated by the above papers, we attempted to create an AI model that would assess the efficiency and viability of Duolingo - a Language teaching ITS.

Assignment Problem:

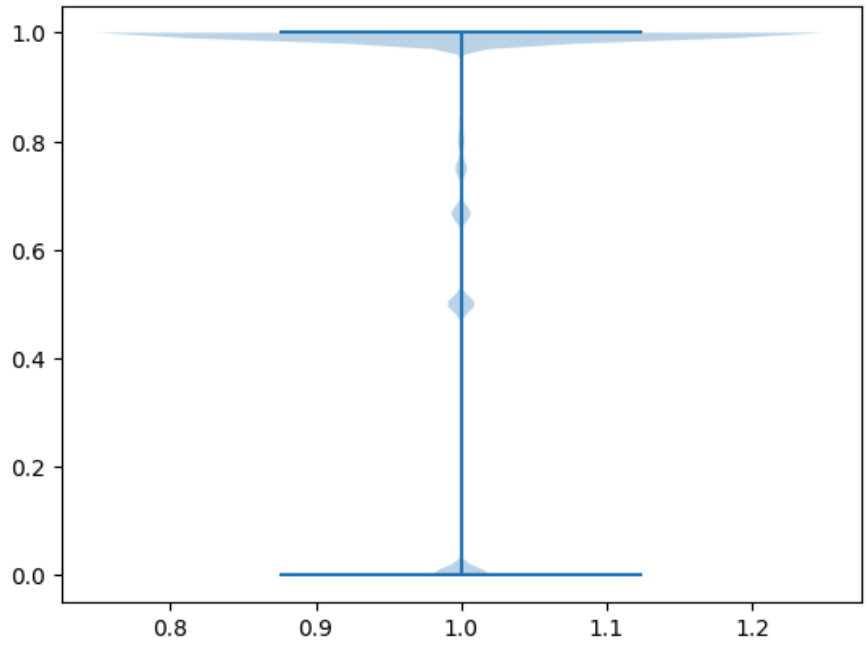
The goal of the assignment was to create a model which would be able to predict how often a learner would be able to guess a word correctly, given that they've been exposed to the given word before and that they have also guessed it correctly in the past. We have done this by feeding the original data from the dataset into the model and obtained results for the spanish language. We also attempted to normalise the data from the dataset and then fed it into the model to check if we would get more accurate results. Following this, we trained the model on another set of data for the Dutch language and then compared the results from our tests with the Spanish data set to verify the scalability of our model.

Implementation:

We took the following steps to implement our model:

1. Importing the dataset. We imported our dataset from the website Kaggle. The same is linked below: https://www.kaggle.com/datasets/aravinii/duolingo-spaced-repetition-data

2. We then checked if the dataset is clean and dropped unnecessary or irrelevant columns from the dataset. In our case, we dropped the columns ‘session_seen’, ‘session_correct’, ‘timestamp’ and ‘user_id’. This was done to reduce noise in the dataset since it prevents the model from training on irrelevant or redundant data.
3. We then restricted the dataset to users who only used the app with the English UI. This was done to slightly ease our study and implementation.
4. We proceeded to obtain results from our model. We first picked out users who were learning Spanish, and then trained the model as a linear regressor, which fits a straight line with all of our columns on the x-axis and the p_recall parameter on our y-axis. We then tested the model on the same dataset. Using this, we obtained an error rate of 0.07.
5. We then tested this model with another dataset which only had the p_recall value as ‘1’, and that had an error rate of 0.08, which was very close to what we obtained with the original dataset. This indicated to us that the data set was very skewed towards the value 1. This was evident from the following plot that we had obtained:



6. To try combatting this, we attempted to create a normal distribution of data by first undersampling the dataset (i.e. removing certain values from the dataset to simulate a normal distribution) and then see how the linear

regression model performs with this data. Thus, the model was trained using this data.

7. We then tested this model with the original data set and found that the error rate with the undersampled data (a normal distribution) was 7 times lesser (0.01) when compared to the error obtained when it was trained with the original dataset(skewed). This could possibly be due to the fact that the model may struggle to accurately predict values away from that central point, leading to a higher error rate for a skewed dataset due to a lack of data points at the extremes and away from the central point. Conversely, a dataset with a normal distribution provides a more balanced representation of the data since it has a symmetrical bell shape, Thus allowing the model to make more accurate predictions across a wider range of values, resulting in a lower error rate.
8. We then tested the model trained with normalized data with unnormalized data, and that resulted in an error rate of 0.23, which was significantly higher. This can be explained because a model trained on a normal dataset cannot be scaled to predict p_recall values from a biased dataset.
9. We then tested the model for people who were trying to learn Dutch in a similar way. (first train with the original dataset, then undersampled dataset). We found that the results were very similar to the ones from the Spanish learners. The error rates were very similar
10. We then tested the Spanish model(trained on normalized data) on the Dutch dataset(normalized data) to see if language was a determining factor for the results obtained. We got an error rate of 0.01(mean squared error). Which was the same as the error rate we obtained on the Spanish normalized dataset. This led us to the conclusion that linear regression models trained on normalized data can be scaled up to other normalized datasets with a similar variance and still work up to a similar accuracy. In conclusion, models trained on near-normal data in an ITS such as this have a much better opportunity to be scaled to different tasks.
11. Since the data is scalable, Transfer Learning would also be possible, i.e., pre-trained/obtained knowledge from a past task may be utilized and leveraged to adapt the model for a different learning language.

Future Work:

The COVID-19 pandemic showed that learning remotely is possible, which significantly increased the use of ITSs. Thus, it has become essential to check the viability and effectiveness of these programs. The model we created is used to check how effective a language teaching ITS.

Using the obtained p_recall value, we could decide which word could be shown to the user next. For example, a word with a lower value of p_recall would be more difficult to guess than one with a higher p_recall value. Thus, the program would attempt to display a question that is neither too easy nor too difficult for a user to guess. Our next step would be determining which range about a p_recall value would be most effective for learning (e.g. 0.5 +/- 0.1 or 0.6 +/- 0.1). This would be done by experimenting with students using both versions and seeing which group had better results.

Using a similar idea of calculating a parameter such as p_recall and with a relevant data set, we could also create a program that would check the viability of ITSs that deal with different subjects such as mathematics and engineering.

References:

1. A Web-based Intelligent Tutoring System for Computer Programming - A Web-based Intelligent Tutoring System for ... - CiteSeerXCiteSeerXhttps://citeseerx.ist.psu.edu › document

2. Application Of An Intelligent Tutoring System In Electrical Engineering Education - https://ieeexplore.ieee.org/document/570302
3. Building Intelligent Tutorial Systems for Teaching Simulation in Engineering Education - https://ieeexplore.ieee.org/document/123417
4. Intelligent Tutoring System Authoring Tool for Manufacturing Engineering Education - https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=355de8cd25d1a5cf6e9bf8239033d8243afc7ad2
5. Work in Progress: Intelligent Tutoring Systems in Computer Science and Software Engineering Education - https://rex.libraries.wsu.edu/esploro/outputs/99900601055601842/filesAndLinks?index=0