Spaced Review Prediction for Language Learning

Broderik Craig, Joe Housley, Jared Suchomel, Ben Winsor

April 2023

Abstract

This paper presents an overview of how time series models can be used to predict learner knowledge in language learning, specifically in spaced review. We discuss how these models can be used to detect changes in knowledge levels, identify trends, and make predictions about learners' performance. We also provide a description of our data from the Missionary Training Center. Finally, we outline the advantages of time series models for predicting learner knowledge and the challenges associated with their use. Our findings suggest that time series models can provide valuable insight into learner knowledge and can be used to inform effective spaced review strategies.

1 Problem Statement and Motivation

The Missionary Training Center of the Church of Jesus Christ of Latter-day Saints uses a web-based app called Embark to assist missionaries in the process of learning a foreign language. When a user learns a letter, word, or phrase (grouped collectively as a vocabulary "concept"), the concept is sent to a separate activity called Spaced Review. Spaced Review functions as a means to help users continue to practice words with which they are already somewhat familiar. One problem is that some users who only occasionally take advantage of Spaced Review will rack up dozens of concepts to review, further discouraging them from using Spaced Review. The purpose of this project is to overcome this hurdle by optimizing the order in which Embark presents concepts to users. If we can determine the probability that a user will accurately answer a question, we could prioritize questions to help users focus on their greatest needs first. If we are able to accurately determine the time the concept will be remembered, we can prompt users to review the concept before they forget it. Both of these tasks can be simplified into predicting the learner's knowledge of a given concept.

Similar projects use a technique known as Bayesian Knowledge Tracing (BKT) to track learner mastery of a skill. One of the limitations of BKT is that it assumes user mastery is Bernoulli where the learner has either mastered the skill or not. For the purpose of choosing when to prompt a user to review a concept, it is more useful to treat learner mastery as a continuous variable. Ye, Su, and Cao (2022) attempted a similar approach to spaced review scheduling in which they used Hidden Markov Models to significantly improve predictions compared to previous methods. We attempt to recreate their results by implementing a similar HMM approach. We also attempt to implement an ARIMA model, as the learning process is not necessarily Markov.

2 Data

We use activity attempt data collected from the Spaced Review activity in the Embark application. Our data contains questions answered by native English speakers learning Spanish from July 2020 to March 2023. Each observation contains a unique id for user and concept, the time at which the attempt occurred, the duration of the attempt, and the result of the attempt. The data contains over 8 million individual observations for 2,283 concepts. Since we intend to train separate models for each concept, that is approximately 3,500 observations for each concept.

We noticed that for a single concept, learners will often have multiple attempts in one day. This indicates that if the learner incorrectly answers the question, they will immediately try again. We are interested in the first attempt because it indicates whether the learner had forgotten the concept, so we ignore any attempts after the first attempt for a given day. For each concept we created a training and testing set using scikit-learn's TimeSeriesSplit package.

Once we reduce to one observation per day, we now create a dataset for a given concept. Each entry includes a sequence with the result of the latest attempt by a single user and all the previous attempt results for the same user. We then include the number of days elapsed between the last attempt and the current attempt as an additional feature. So for a given concept, we have a sequence of attempts for a given user. The average length of these attempts is 5, meaning that on average, users will attempt a given concept 5 times over the time period that our data covers. Some users only attempted a concept once or twice. These sequences are not long enough to be considered time series data. To account for this we determine

a sequence length threshold and remove any sequences shorter than the minimum sequence length.

3 Methods and Results

3.1 Gaussian HMM

At first glance, a Gaussian Hidden Markov Model (HMM) fit our needs perfectly. We would be trying to model knowledge (the hidden layer) while being able to measure question accuracy (quantifiable layer). However, when implemented, GHMMs did not fit the data nearly as well as hoped. In a Gaussian Hidden Markov Model parameters are assumed to be distributed normally. Due to the fact that the majority of the questions were answered correctly, the model's main job was to identify when a question would be answered incorrectly. Questions answered wrong follows more of a Poisson distribution. We can also observe that memory tends to degrade with

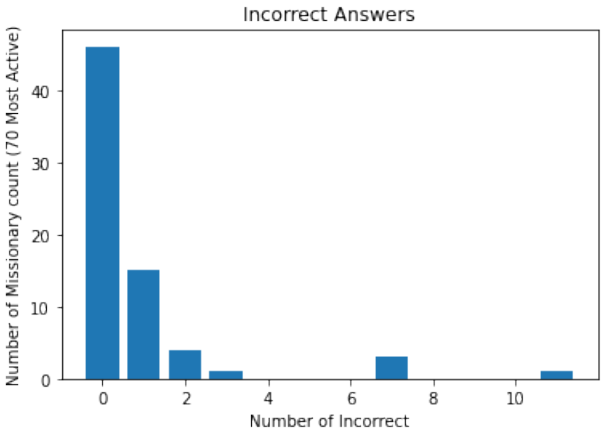

Figure 1: Incorrect Answer Count vs Frequency.

a half-life pattern, so retention is not likely to distribute normally either. Our hypothesis that Gaussian HMM would not fit well to our data set was confirmed when running a Gaussian HMM model produced accuracies consistently lower than random.

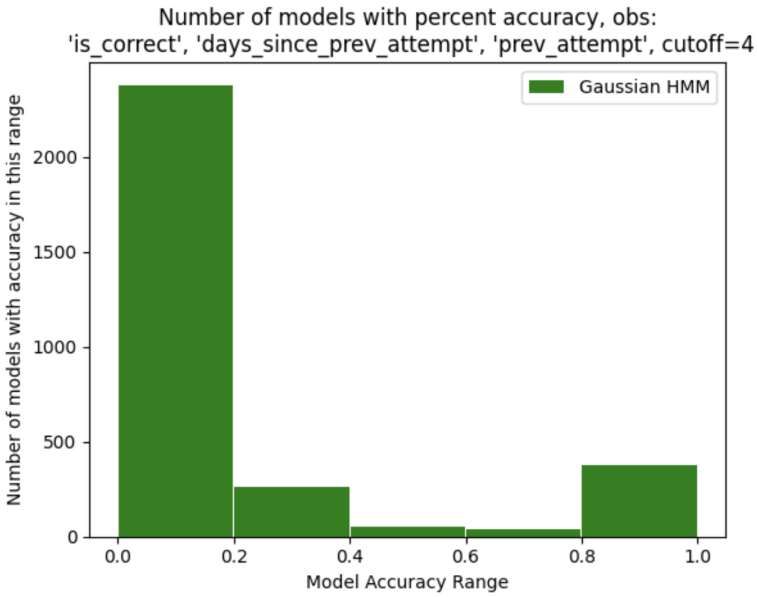

Figure 2: Distribution of model accuracy for Gaussian HMM with minimum sequence length of 4 attempts

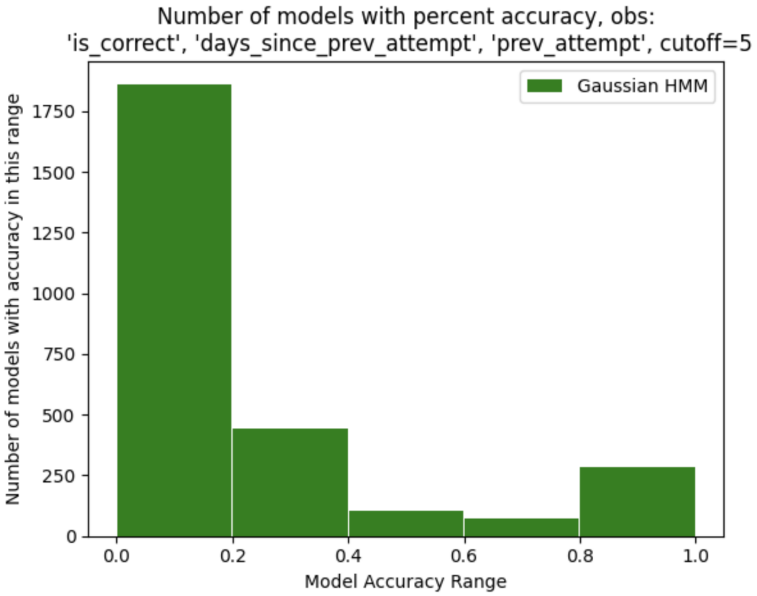

Figure 3: Distribution of model accuracy for Gaussian HMM with minimum sequence length of 5 attempts

3.2 Poisson HMM

We observed much better results when using a Poisson Hidden Markov Model. Poisson distributions describe the likelihood of events occurring in a given period of time, or in our case how many times the user answers correctly within a given number of attempts. Poisson HMM produced much more accurate models than Gaussian HMM, however the accuracy was highly reliant on the minimum length of the attempt sequences. This seems to have little to do with the number of samples produced by limiting the length of the sequences, so sequence length is an important and unique hyperparameter for each concept. The following graphics show the number of trained models that achieved an accuracy threshold (shown on the x-axis).

Extremely poor models can be made useful because our we measured accuracy using a binary prediction assumed to be 0 or 1. So if we can simply identify the poor models, and reverse their prediction, those models can achieve a high accuracy as well.

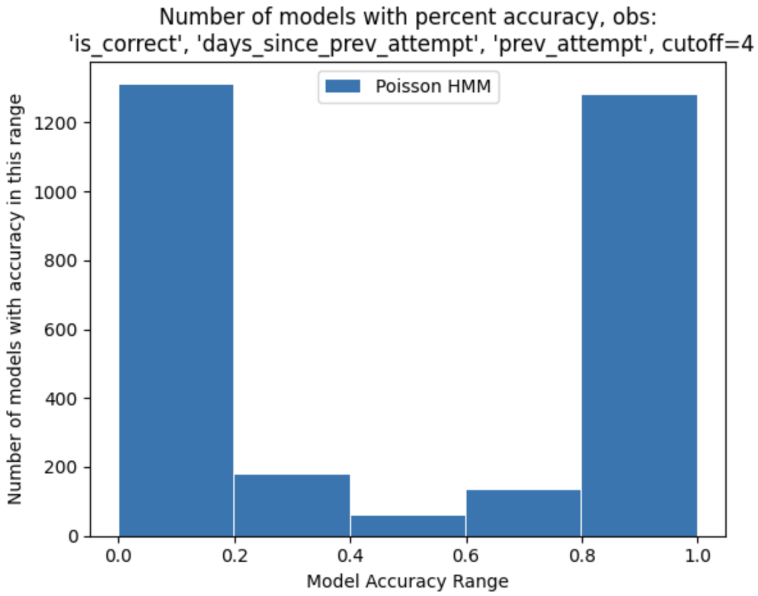

Figure 4: Distribution of model accuracy for Poisson HMM with minimum sequence length of 4 attempts

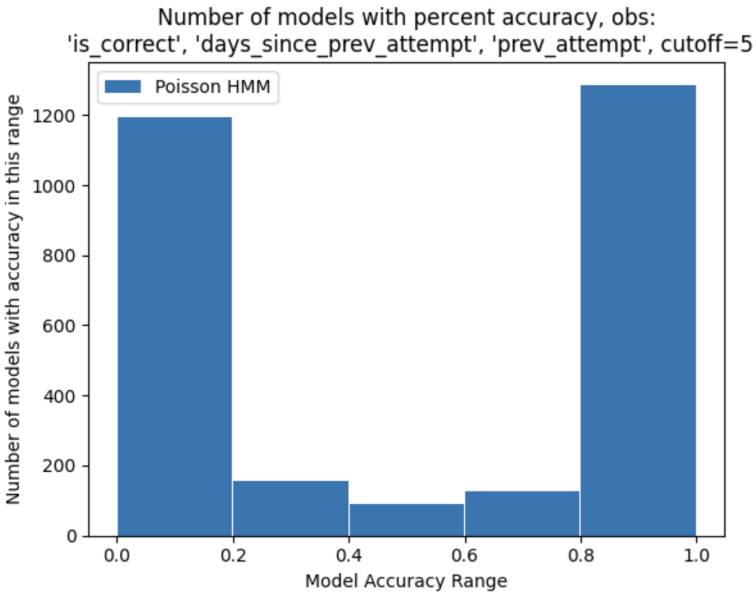

Figure 5: Distribution of model accuracy for Poisson HMM with minimum sequence length of 5 attempts

3.3 ARIMA

We also hypothesized that an ARIMA model would be a reliable tool for modeling missionary comprehension. There are 3 parts of an ARIMA model: 1) the trend, 2) the seasonal component, and 3) random variable built into the model. Due to the fact that memory is an multiplicative model (half-life) rather than a linear one, we would use the natural logarithm to break up the calculations into additive parts to apply this model. The trend would compose of 2 components: 1) a gradual increase in knowledge and 2) Time passed since the topic was last seen. We would also use the random component to account for the random chance that someone knows the answer and gets it wrong, or correctly guesses the right answer without having sufficient knowledge of the material.

The results of the ARIMA model did not yield satisfactory results for our purposes. We attempted to model knowledge, and use a binary classifier (whether they answered the question right or wrong) to check learner knowledge. Our project and our data violated several important assumptions for the ARIMA model to work appropriately: 1) Assume that missionaries learn at the same rate. Not every individual learns at the same rate, and their

retention rates vary strongly. We tried using their average global accuracy to estimate their average accuracy for the topic, but it had negligible impact on their results for a given topic (see graph below). While it is true that some topics had a simple positive trend (Topic 4976), the majority had almost no correlation (topic 1450). Without a reliable method to quantify individual missionary capability, It is impossible to personalize the trend to each missionary. 2) It is very difficult for this model to predict whether

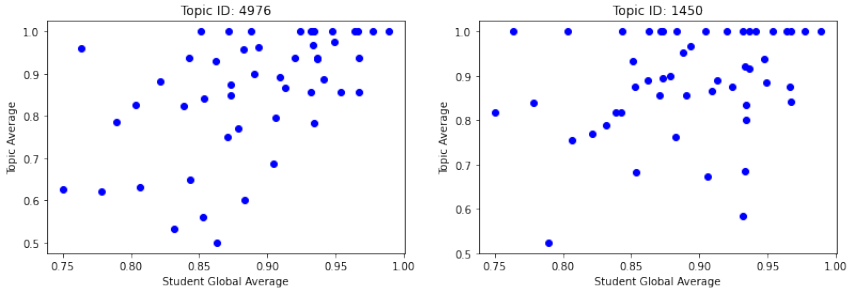

Figure 6: Student Topic Accuracy by vs Student Global Accuracy

or not a given individual would get a correct answer after a given period. As knowledge of the concepts that we trained it on would be very close to 1 after a period in the mission field, some questions would be answered incorrectly. It was also increasingly difficult for the model to correctly identify how much memory would decay, as many missionaries would not see a concept for over 3 months and get it right, only to miss it the very next day. In short, there was too many outliers and no consistent base for the model to linearly separate the data. 3) Missionaries had no consistent basis of which they were tested on. Some would view the materiel once, then wait a few weeks to view it again. Others would see it every day for a week. The goal was to model learning, which is impossible if the learning data is inconsistent. 4) The binary aspect of question response in and of itself was a hamper to gauging the knowledge of the missionary. There was a roughly an eighty five percent accuracy overall, and it is very difficult to scale a binary classifier to a continuous range [0,1], especially with the other problems with the data already addressed.

4 Analysis

One of the issues with the current approach is that not all concepts are equal, meaning that they vary in difficulty and exposure. Some concepts

are commonly seen and so they have a healthy amount of data points, while others are rarely seen and so any model we attempt to train on the limited data available is susceptible to overfitting. In a future iteration of this project, we would cluster the concepts by some feature such as difficulty and train each model on a group of concepts. This would increase the number of samples available to each model and likely improve the accuracy of our predictions.

Poisson HMM modeling proved fruitful, but was not without caveats. Not all concepts had sufficient data to create meaningful test-train splits, so concepts with insufficient data were discarded. Remaining concepts were analyzed by taking each user, extracting concept data, and concatenating the user data together (we assume each missionary learns a given concept similarly). Results were mixed. The "null" prediction rate (predicting a correct answer every time), is about 85%.

Another important hyper-parameter was the cutoff, or minimum sequence length. Many users were only attempted a given concept a few times. So, for example, if a user only attempted a concept two times, and we used a cutoff value of three, that user's data would be dropped from the dataset. Initial tests suggest the cutoff has little effect on the final accuracy of a given model, but that it does significantly affect the number of samples.

The goal of this project was to model individual concept learning. As previously stated, the Poisson HMM was the only model that could be called successful. However, different concepts are not learned independently when learning Spanish. Future study could emphasize different emphasis with the same data-set. Some possible situations would be: Modeling how well a missionary learns the language as a whole, not necessarily each topic, if better accuracy while in the MTC leads to better accuracy while serving in their missions, or if retention of information is independent of their accuracy.

5 Ethical Considerations

Each user is assigned a random user-id number, so anonymity is preserved. The only harm from the misuse or misinterpretation of our results could be to make the Embark language learning app a little less efficient. One of the assumptions we make in this project is that each user learns the same way. Thus our model is best fit for the average user and may hinder the learning of more advanced users, or less advanced users. In a perfect world, we would have enough data, memory, and computational power to train a model for each individual user.

There is some potential for a feedback loop to be created because the predictions of the model will influence which concepts a user is shown for review, which in turn will affect the distribution of new data that is being gathered. When the model is re-evaluated and trained, we need to be mindful of how the results of the model may have affected future data.

6 Conclusion

The use of time series models such as Hidden Markov Models and ARIMA has great potential in predicting learner knowledge for spaced review prediction. The results obtained from the study demonstrate that a Poisson HMM can accurately predict the level of knowledge retention of learners over time. We also want to stress the importance of collecting accurate and comprehensive data on learner performance, which can be used to develop personalized learning models that can cater to the individual needs of each learner. By incorporating time series models into the development of spaced review schedules, we can enhance the learning experience of users and promote deeper and more meaningful language retention. Further research in this area is needed to refine and improve these models, but the potential benefits of these approaches are clear.