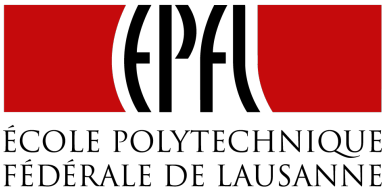

School of computer and communication sciences

Optional Master semester project

Predicting STEM Career Choice from Middle School Interaction with Mathematics Educational Software

---

Author:
Dario Anongba Varela

Supervisor:
Faucon Louis

Professor:
Dillenbourg Pierre

A report submitted in fulfillment of the requirements for the degree of Master in Computer Science

Computer-Human interaction in learning and instruction (CHILI)

January 17, 2018

Abstract

Demand for qualified workers in science, technology, engineering, and mathematics (STEM) fields is ever increasing. Researchers are willing to identify the factors that make a student choose a STEM career over a non-STEM career as early as middle school. In this paper, we discuss a neural network model that has the ability to predict, given click-stream data from middle-school mathematics ASSISTments assignments, if a student will choose a STEM career or not after college. With data from 467 college students, we developed a model that can successfully distinguish 85% of the time if a student will choose a STEM career. With this model, we provide educators with a way of identifying interest or disinterest in STEM from their students.

The model implementation and data processing can be found at: `https://github.com/darioAnongba/ASSIStments-Competition`. It is widely documented.

1 Introduction

Nowadays, due to advanced research in Artificial Intelligence, it is possible to predict with great certitude many complex human decisions. Applying this knowledge to education is the purpose of this project.

The project's model implementation is part of the ASSISTments Data Mining Competition hosted by the Northeast Big Data Innovation Hub. It uses data from a longitudinal study, now over a decade long, led by Professor Ryan Baker and Professor Neil Heffernan. The data is composed of extensive (but carefully deidentified) click-stream data from middle school ASSISTments (an online learning platform) use, as well as carefully curated data on first job out of college.

The final objective of the project was to be able to predict, using this data, whether or not students will choose a career in a STEM field after their studies. Several papers [PBSPG13, SPBBH13, SPOBH14] have shown that behavior in ASSISTments in middle school can predict high school and college outcomes. The question here is whether we can monitor interest for STEM fields at an early age in order to assess if the learning methods are appropriate and if not, try to reignite the students interest for STEM using other methods.

Many STEM jobs require a postsecondary degree or other advanced technical training. However, research shows a gap between the number of students who express interest in STEM degree programs and the number who actually enter them, which is driven by inadequate preparation for higher level STEM skills and other aspects of college readiness [Wan13]. This lack of preparation often begins as early as middle school.

Using a Deep Learning technique called a Recurrent Neural Network (vastly used in Natural Language Processing but innovative in the educational field), we develop a prediction model to distinguish whether or not students will choose a STEM career after finishing college.

2 Methodology

2.1 The ASSISTments system

This study predicts student outcomes from their interactions with the ASSISTments system [RFNJ+05], a free web-based mathematics tutoring system for middle-school mathematics, provided by Worcester Polytechnic Institute (WPI). ASSISTments assesses a student's knowledge while assisting them in learning, providing teachers with formative assessment of students as they acquire specific knowledge components. Within the system, each mathematics problem maps to one or more cognitive skills. When students answer correctly, they proceed to the next problem. When they answer incorrectly, the system scaffolds instruction by dividing the problem into component parts, stepping students through each before returning them to the original problem. Once the original problem is correctly answered, the student advances to the next.

2.2 The Data

The data is composed of action log files from ASSISTments of 467 students, generating a total of 251,488 actions. From these students, 117 have chosen a STEM career and 350 have not (STEM careers are defined by the National Science Foundation).

The data is composed of features of multiple types that need to be processed differently:

- *Floating averages*: Values in the range [0, 1] representing averages. Ex: the average number of correct actions from the student.
- *Categorical*: Values representing categories or indexes. These values need to be treated carefully because a value of "1184832848" for example represents an index and not a numerical value. This means that we need to transform these values in order to use them in a neural network. Ex: The assignment id the student is currently working on.
- *Numerical*: Values representing integers. Ex: The time taken to solve a

problem.

- *Binary*: Values representing binary concepts (like yes or no). Ex: Whether the problem was solved correctly or not.

In addition, average information about the learning process of a student is appended to each of his actions (the average correctness of all actions for example). This information is highly redundant as it is the same at every action and could potentially be extracted in order to reduce the size of the dataset.

We will examine in more depth how to process these different types of features in order to use them in our model efficiently.

2.3 Deep Learning Model: Recurrent Neural Network

As we have seen, the data is composed of a sequence of actions per student. Each student have done a different number of actions spanning from as low as 2 actions to as much as 3057 actions. Each action is chronologically related to the previous one because they represent a series of events in the ASSISTments system (the student being driven to different problems given the answers given).

Classical machine learning methods like Logistic Regression or SVM and standard Deep Learning models like traditional neural networks are not optimal for this type of data for the following reasons:

- *Data aggregation*: In order to use methods like Logistic Regression, Support Vector Machines or even Ridge Regression, we need to create a direct dependency between the data and the corresponding label. In order to achieve this with our data, we would need to somehow aggregate the data, by taking the mean or the standard deviation for example. It is obvious that this method would cause a significant loss of information given that we are discarding the chronological relations in the data. We are also discarding categorical features that cannot be aggregated (like the problem type or the skill being tested).
- *Notion of Persistence*: Traditional neural networks cannot use their reasoning about previous events to inform later ones as they do not have a notion of persistence. In our data, the events are chronologically sorted and past

events can influence current events. This means that using a traditional neural network would once again result in a significant loss of information. Without correlation between student actions, a neural network would have worse results than classical methods like logistic regression.

Recurrent neural networks address this issue by introducing loops in the network, allowing information to persist. For a quick description of what a recurrent neural network is, please see this wonderful article from Christopher Olah, a Google researcher [Chr15].

In fact, a Recurrent Neural Network (or RNN) is not so different than a normal neural network. A RNN can be thought of as multiple copies of the same network, each passing a message to a successor, as seen in Figure 1.

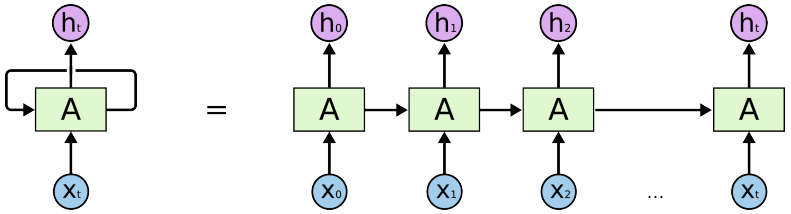

Figure 1: An unrolled recurrent neural network.

As stated by Christopher Olah, *This chain-like nature reveals that recurrent neural networks are intimately related to sequences and lists. They're the natural architecture of neural network to use for such data.*. This corresponds exactly to the type of data that we have.

Because standard RNNs have a problem of long-term dependency (the most ancient information is gradually lost as the memory is decaying over time), we have focused our research on Long Short Term Memory networks, or LSTMs, that solve this particular problem. [Chr15].

2.3.1 Data pre-processing

Obviously, our data cannot be inserted into a RNN as it is because of the different nature of the features and their range. Neural networks in general expect their data to be normalized and scaled in order to reach a global minima optimally.

Additionally, categorical features need to be processed into vector of numbers, this is called embedding. Here is a description of the data preprocessing done for every type of feature.

- *Floating averages*: No special transformation is needed for these features as they are already scaled (in the range [0, 1]) and normalized.
- *Categorical*: These features can either be strings or integers representing categories or indexes, and can potentially be huge. In order to be able to use these features, we need to process them in two steps. The first one consists in encoding the values into integers from 0 to n-1, where n is the number of classes of that category. The second is creating a vector of fixed size representing this value.
- *Numerical*: These values need to be normalized and scaled.
- *Binary*: No special transformation is needed for these features as they are either 1 or 0.

Furthermore, we separate dynamic features and static features, avoiding the redundance of the static features at every action.

The final data structure is a Map from student ID, to an array of size 3, as illustrated in Figure 2, containing:

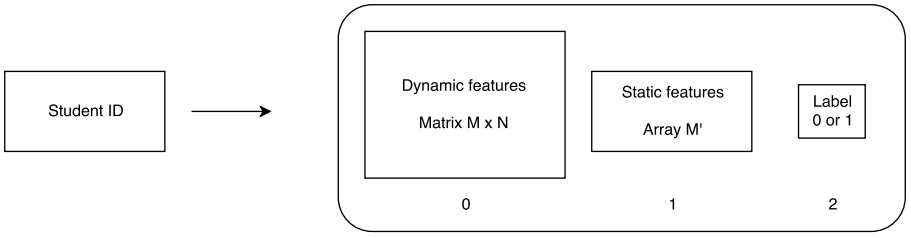

Figure 2: Map containing all processed data

1. *Dynamic features*: A matrix of size M x N of features that can potentially be different at every action, where M is the number of dynamic features and N is the number of actions.
2. *Static features*: A vector of size M' of features that are fixed for a given student, where M' is the number of static features.

3. *Label*: A boolean value, 1 if the student has chosen a STEM career, 0 otherwise.

2.3.2 Model

The complete model implementation can be seen in Figure 3.

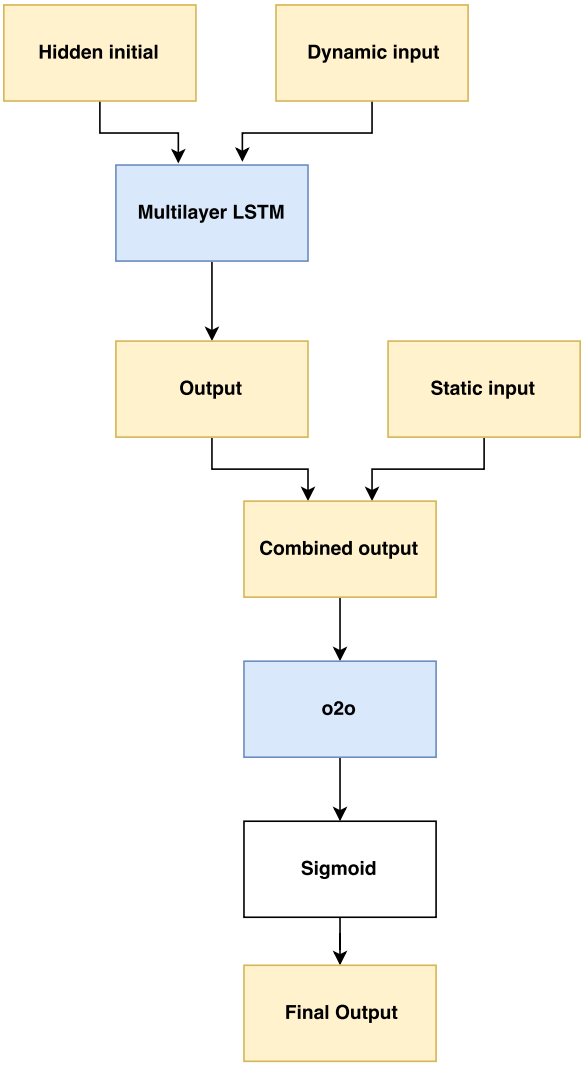

Figure 3: Complete model implementation

It is composed of a **multilayer LSTM (or GRU) RNN** taking as input the dynamic features matrix and an inital hidden state. The inside of the multilayer LSTM can be seen as a black box as we used an already implemented pyTorch

version, the number of layers being a parameter of the system. The RNN will output a vector of size S, where S is the desired hidden dimension. This vector is then concatenated with our static input vector into a new vector of size S + M'. Finally, this new vector is used as parameter to a "decoder" layer (linear layer called o2o as "output to output" in the diagram) that ouputs a value passed into the sigmoid function in order to reduce its range between 0 and 1. The final output is the probability that a student will choose a career in STEM after his studies.

There's also a dropout layer inside of the multilayer RNN, which randomly zeros parts of its input with a given probability (another parameter) and is usually used to fuzz inputs to prevent overfitting. Here we're using it towards the end of the network to purposely add some chaos and increase sampling variety.

It is also worth noting that the neural network is bidirectional. Meaning that it not only traverses the data in one direction, but in both directions simultaneously, outputting a number of hidden dimensions 2 times higher.

3 Results

In order to evaluate the network, we used the **Area Under the Receiver Operating Characteristic Curve (ROC AUC)** metric, and not the accuracy. It would be very easy to obtain a good accuracy considering that more than 3/4 of the students choose a non-STEM career. A model that always outputs 0 as an answer would obtain an accuracy of more than 75%, which is reasonably good considering the complexity of the question.

Results were obtained using a cross-validated set of 30 random students. The size of the validation set is significantly small because of the poor amount of data at our disposal for training.

Model fine-tuning was performed by grid search on the most significant parameters that could potentially influence the results, those parameters are the **hidden dimension**, the **number of layers** and the **dropout** of the multilayer RNN (GRU or LSTM). We believe that even better results can be achieved by doing a grid-search on all parameters, but this would take a considerable amount of time and computing power.

We assessed that the model was indeed training by plotting the accuracy and AUC ROC as seen in Figure 4. The system has been trained for 30 epochs with dropout of 10%.

We can see that increasing the hidden dimension allows for faster learning. The greater, the more inclined the system is to overfitting.

Previous studies [SPOBH14] have shown that it was possible to obtain an AUC ROC of 0.66 using logistic regression. Our results greatly surpass those and show that we are able to obtain an AUC ROC score of over 0.875 with an accuracy of 90% on the valiation set, as it can be seen on Figure 5.

Increasing the hidden dimension and the number of layers have shown to cause overfitting and worse results. 32 hidden dimensions and 3 layers are the optimal parameters for our network.

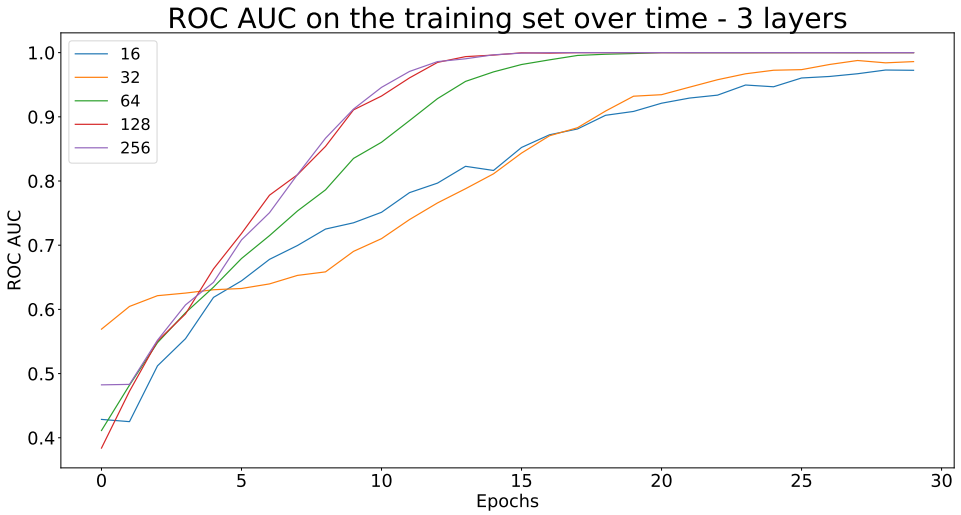

Figure 4: ROC AUC on the training set over time with 3 layers for different hidden dimensions

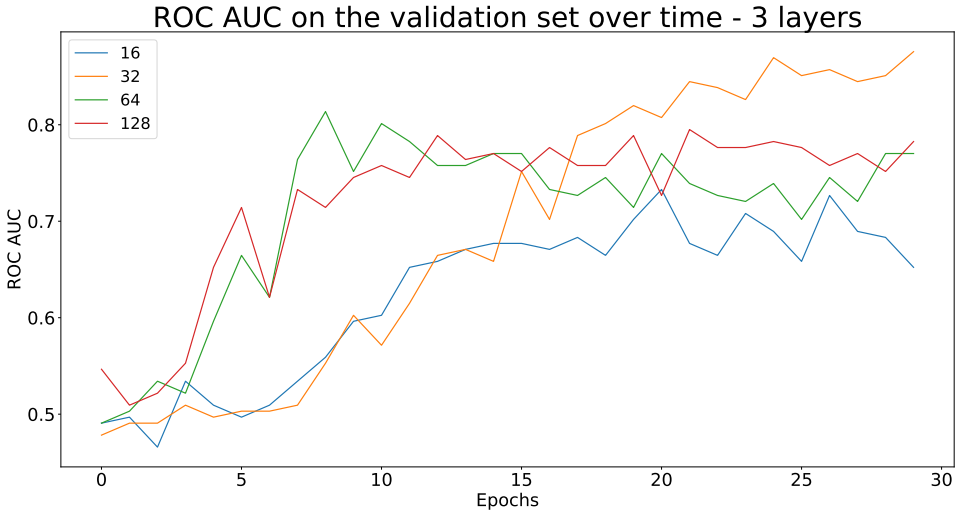

Figure 5: ROC AUC on the validation set over time with 3 layers for different hidden dimension values

4 Discussion and conclusion

This project presents a deep learning model using a recurrent neural network that is able to predict 87% of the time whether or not a student is going to choose a STEM career after his studies. Success within middle school mathematics (indicated by correct answers and high probability of knowledge in ASSISTments) is positively associated with STEM career choice.

One possible use of these findings is to give educators and career counselors a new lens on early indicators of disinterest or disengagement from STEM content and instruction, allowing them to develop counseling strategies that will sustain student interest in pursuing STEM degrees and careers. [SPOBH14].

One fairly important aspect of the model is its incapacity to be interpreted. Indeed, the network can be seen as a black box that outputs whether or not a student is prone to be interested in STEM or not, and we cannot interpret the result by giving more or less importance to certain features for example. This creates a problem as we are incapable of knowing what is the factor that makes a student lose or gain interest in STEM.

It is well-known that neural networks need a considerable amount of data to be efficient. We believe that our model can be greatly improved by increasing the amount of data available for its training and validation.

References

[Chr15] Olah. Christopher. Understanding lstm networks. 2015.

[PBSPG13] Z.A. Pardos, R.S.J.d. Baker, M.O.C.Z. San Pedro, and S.M. Gowda. Affective states and state tests: Investigating how affect and engagement during the school year predict end of year learning outcomes. *Journal of Learning Analytics*, pages 107–128, 2013.

[RFNJ+05] L. M. Razzaq, M. Feng, G. Nuzzo-Jones, N. T. Heffernan, K. R. Koedinger, B. Junker, and K. P. Rasmussen. Blending assessment and instructional assisting. *AIED*, pages 555–562, 2005.

[SPBBH13] M. San Pedro, R. Baker, A. Bowers, and N. Heffernan. Predicting college enrollment from student interaction with an intelligent tutoring system in middle school. *Proceedings of the 6th International Conference on Educational Data Mining*, 2013.

[SPOBH14] M. San Pedro, J. Ocumpaugh, R. Baker, and N. Heffernan. Predicting stem and non-stem college major enrollment from middle school interaction with mathematics educational software. *Proceedings of the 7th International Conference on Educational Data Mining*, pages 276–279, 2014.

[Wan13] X. Wang. Why students choose stem majors motivation, high school learning, and postsecondary context of support. *American Educational Research Journal*, pages 1081–1121, 2013.