Machine Learning Engineer Nanodegree

Capstone Proposal

Lucas Camelo Casagrande
January 16th, 2018

Development of a LSTM Network to Predict Students' Answers on Exam Questions

Domain Background

The high dropout rate in universities during the first years of college had challenged the universities to improve their teaching quality in order to keep their students motivated (Brito *et al.*, 2015; Manseira, 2016; Pascoal *et al.*, 2016; Sweeney, 2016).

In this context, professors have fundamental importance in the quality of teaching and in the motivation of their students (Thiede *et al.*, 2015; Blazar, 2016). Effective academic advising is one of the most effective ways for helping students complete their degree in the expected time (Sweeney, 2016). Therefore, it is very important that professors have accurate evaluations of their students in advance so that it is possible to adapt their pedagogy to meet the individualities and difficulties of each student in order to improve the graduation rate (Brito *et al.*, 2015; Thiede *et al.*, 2015; Pascoal *et al.*, 2016; Cakmak, 2017).

Classical evaluation methods such as tests and exams only permit the evaluation of a student after the test has being taken. Therefore, those methods do not allow a complete beforehand evaluation at the beginning of the course of all the expected abilities that a student should already have or learn during the course (Manseira, 2016).

In other hand, predictive models are capable of predicting future information based on historical data. Data collected from a Learning Management System (LMS) can serve as the basis for the training of a model capable of predicting if a student has the sufficient knowledge to answer questions not yet seen and answered, a problem known as Knowledge Tracing (KT).

Therefore, a LSTM network can be trained and used to predict student knowledge based on a set of historical questions answered and classified in their domain of knowledge. Based on this information, a professor can identify their students who are at risk of failing and take the necessary precautions in time to improve the student success on the class, reducing the rate of evasion (Piech *et al.*, 2015; Elbadrawy *et al.*, 2016; Khajah, Lindsey and Mozer, 2016; Schiller, 2016; Xiong *et al.*, 2016).

Problem Statement

Classical evaluation methods such as tests and exams do not allow a complete beforehand evaluation of all the expected abilities that a student should already have or learn while a course is being taken. In other hand, predictive models are capable of predicting future information based on historical data and a model could be trained to predict student's answers on future exam questions.

Therefore, a LSTM network can be trained to find the dependencies between the questions answered in a dataset and use that to predict the probabilities that a student will answer correctly an exam question not yet seen by him. Then, this information can be used by professors to trace the knowledge of their students in advance right at the beginning of the course helping them to identify which students can have more difficulties than others during the course.

Datasets and Inputs

Supervised training of a neural network requires a dataset with examples that already contains the expected answers. With this in mind, this project uses the examples contained in the latest version of the public dataset called "ASSISTments Skill-builder data 2009-2010" and available free on the web (Feng, Heffernan and Koedinger, 2009; ASSISTments, 2010; Piech *et al.*, 2015). In Table 1 is defined some characteristics of this dataset.

Table 1 - Characteristics of the dataset used in this project.

Characteristic	Total
Answered Questions	401.756
Types of questions	26.688
Types of knowledge components	124
Students	4.217

Assistments (2010)

This dataset consists of more than 400,000 answered binary questions in which most of them are already classified in one knowledge component. Of all the attributes available in this dataset, only three were selected to be used in this project: the user identifier (ID), which is the student ID; the ID of the question knowledge component; and a binary variable that indicates whether the question was answered correctly or not by the student. All other attributes have been discarded and the questions that are not classified into a knowledge component were removed from the dataset.

Solution Statement

One possible solution to the problem is to train an Artificial Neural Network (ANN) to predict the answers of future questions based on historical questions. A Feedforward Neural Network could be used to achieve this goal, however it does not take into account the dependency between the answered questions as time passes. The questions are dependent because if a student has answered a math question correctly in the past he is more likely to correctly answer another math question in the future.

In this case, a Recurrent Neural Network (RNN) is better suited to this task, since it takes into account the dependency between the questions. A standard RNN might be sufficient, but as recent research has stated, an LSTM network, which is an RNN composed of LSTM units, has better performance and can store information in its memory for longer time. Therefore, the present project proposes the use of an LSTM network as a solution to the stated problem.

Benchmark Model

Piech et al. (2015) was the first to propose the use of LSTM networks to solve the problem of Knowledge Tracking, which was then called Deep Knowledge Tracing (DKT). In his paper it's possible to see a comparison between his approach and another one known as Bayesian Knowledge Tracing (BKT). In the same context, Khajah, Lindsey and Mozer (2016) replicated the work of Piech et al. (2015) to explore his approach more deeply. In their paper, it is also possible to see another comparison between the DKT approach and the BKT, including some of its extensions. In the same dataset used in this project, the classical BKT has achieved an AUC of 0,73.

Both related works uses the same dataset for training and evaluation, therefore the results obtained using the BKT approach that can be seen in their papers will be used to evaluate the performance achieved on the proposed model.

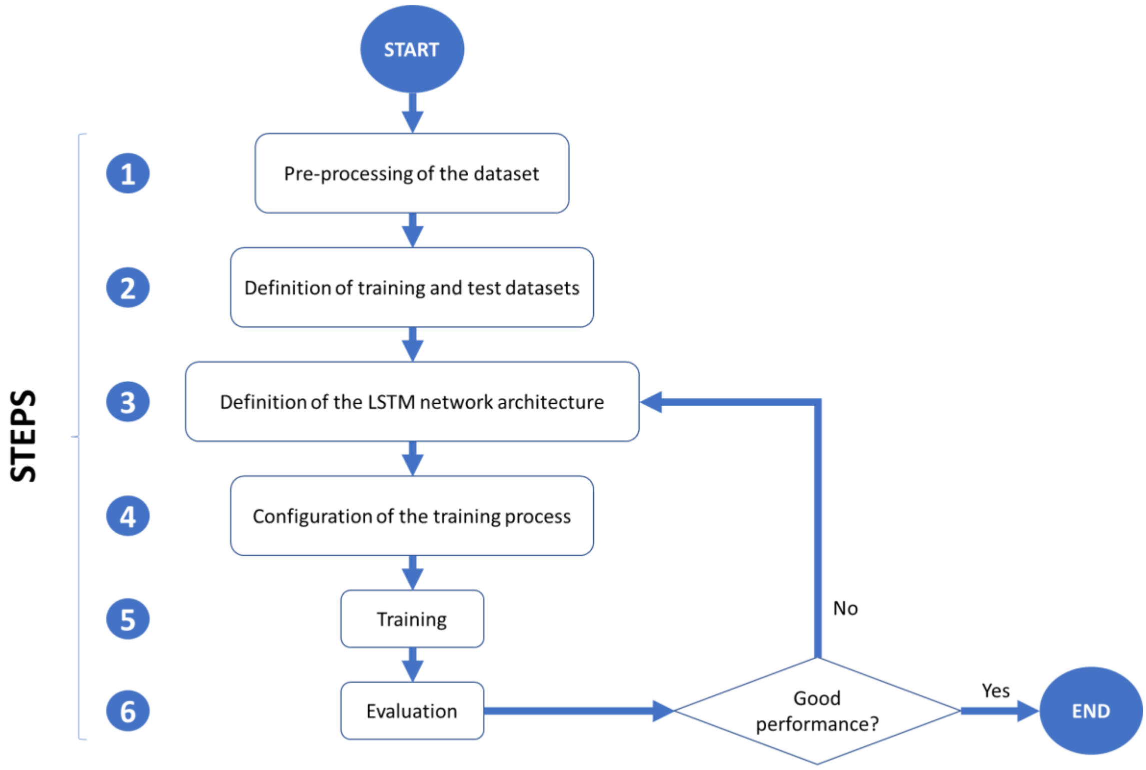
Evaluation Metrics

As related works computes the Area Under the Receiver Operating Characteristic Curve (AUC) to evaluate their models, this project will use the same metric to allow the comparison of its performance. This metric reflects the ability of the model to discriminate correct from incorrect responses in which a score of 1.0 reflects a perfect discrimination.

Project Design

The theoretical workflow defined in this project for approaching the proposed solution for the given problem can be seen in the figure 1. This workflow is composed of six main steps that should be done and result in a good performance before ending. The architecture of the LSTM network and the parameters defined for training should be changed in a way that results in a better performance in each try.

Figure 1 – Defined workflow for approaching a solution for the problem stated.



The Author

In the first step, examples that do not contain a defined knowledge component are removed from the dataset. Then, the samples are grouped by student identification and padded with a constant value of -1 to normalize the sequences size. After that, a one-hot encoding is applied on the knowledge component ID of each question in the sequence to transform its integer value into vectors of the same size.

In the second step, the dataset is divided into two sets where 70% of the data is reserved for training and the remainder (30%) for testing. In the third step the architecture and parameters of the LSTM network are defined, some of them are: the number of LSTM layers and units; the weights initialization method; the activation function; the loss function; the optimizer; and the regularization techniques used to avoid overfitting.

In the fourth step, the number of epochs and minibatches are defined for training. In the fifth step the model is trained and in the last step it is evaluated and compared with related works. If the model does not have a good performance, all the steps from step 3 are repeated until a good performance is achieved.

REFERENCES

ASSISTments (2010) *Skill-builder data 2009-2010*. Available at: https://sites.google.com/site/assistmentsdata/home/assistment-2009-2010-data/skill-builder-data-2009-2010 (Accessed: 12 January 2018).

Blazar, D. (2016) *Teacher and Teaching Effects on Students' Academic Performance, Attitudes, and Behaviors: Extensions of the Literature*. Graduate School of Education of Harvard University. doi: 10.1111/0022-1082.00115.

Brito, D. M. De *et al.* (2015) 'Identificação de estudantes do primeiro semestre com risco de evasão através de técnicas de Data Mining', in *Nuevas Ideas en Informática Educativa TISE 2015*. Santiago: Tise, pp. 459–463. Available at: http://www.tise.cl/volumen11/TISE2015/459-463.pdf.

Cakmak, A. (2017) 'Predicting Student Success in Courses via Collaborative Filtering', *International Journal of Intelligent Systems and Applications in Engineering*, 5(1), pp. 10–17. doi: 10.18201/ijisae.2017526690.

Elbadrawy, A. *et al.* (2016) 'Predicting Student Performance Using Personalized Analytics', *Institute of Electrical and Electronics Engineers*. IEEE, 49, pp. 61–69. doi: 10.1109/mc.2016.119.

Feng, M., Heffernan, N. and Koedinger, K. (2009) 'Addressing the assessment challenge with an online system that tutors as it assesses', *User Modeling and User-Adapted Interaction*, 19(3), pp. 243–266. doi: 10.1007/s11257-009-9063-7.

Khajah, M., Lindsey, R. V and Mozer, M. C. (2016) 'How deep is knowledge tracing?', pp. 1–8. Available at: http://arxiv.org/abs/1604.02416.

Manseira, P. R. P. (2016) *Um método para execução de avaliações educacionais diagnósticas*. Unisociesc.

Pascoal, T. *et al.* (2016) 'Evasão de estudantes universitários: diagnóstico a partir de dados acadêmicos e socieconômicos', in *XXVII Simpósio Brasileiro de Informática na Educação*. Uberlândia: SBIE, pp. 926–936. doi: 10.5753/cbie.sbie.2016.926.

Piech, C. *et al.* (2015) 'Deep Knowledge Tracing', in *Neural Information Processing Systems 28*. NIPS, pp. 1–9. Available at: http://arxiv.org/abs/1506.05908.

Schiller, S. (2016) 'Predicting and Improving Academic Success and Student Retention: An Action Research', in *Pre-ICIS SIGDSA/IFIP WG8.3 Symposium: Innovations in Data Analytics*. Dublin, pp. 1–5. Available at: https://aisel.aisnet.org/sigdsa2016/15.

Sweeney, M. (2016) *Intelligent Degree Planning Systems: Advancements in Personalized Achievability Rating*. George Mason University. Available at: http://hdl.handle.net/1920/10759.

Thiede, K. W. *et al.* (2015) 'Can teachers accurately predict student performance?', *Teaching and Teacher Education*. Elsevier Ltd, 49, pp. 36–44. doi: 10.1016/j.tate.2015.01.012.

Xiong, X. *et al.* (2016) 'Going Deeper with Deep Knowledge Tracing', in *9th International Conference on Educational Data Mining*. Raleigh: EDM, pp. 545–550.